machines

Special Issue Reprint

Recent Developments in Machine Design, Automation and Robotics

Edited by
Raul D. S. G. Campilho

mdpi.com/journal/machines

MDPI

Recent Developments in Machine Design, Automation and Robotics

Recent Developments in Machine Design, Automation and Robotics

Guest Editor

Raul D. S. G. Campilho

MDPI

Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Guest Editor
Raul D. S. G. Campilho
Department of Mechanical
Engineering
Polytechnic of Porto
Porto
Portugal

Editorial Office
MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Special Issue, published open access by the journal *Machines* (ISSN 2075-1702), freely accessible at: https://www.mdpi.com/journal/machines/special_issues/SBK53621ME.

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, A.A.; Lastname, B.B. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-7258-6085-2 (Hbk)
ISBN 978-3-7258-6086-9 (PDF)
https://doi.org/10.3390/books978-3-7258-6086-9

© 2025 by the authors. Articles in this book are Open Access and distributed under the Creative Commons Attribution (CC BY) license. The book as a whole is distributed by MDPI under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) license (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

About the Editor . . . . . **ix**

Preface . . . . . **xi**

Raul D. S. G. Campilho
Recent Developments in Machine Design, Automation and Robotics
Reprinted from: *Machines* **2025**, *13*, 683, https://doi.org/10.3390/machines13080683 . . . . . **1**

Davide Fabiocchi, Nicola Giulietti, Hermes Giberti and Marco Carnevale
AI-Driven Virtual Sensors for Real-Time Dynamic Analysis of Mechanisms: A Feasibility Study
Reprinted from: *Machines* **2024**, *12*, 257, https://doi.org/10.3390/machines12040257 . . . . . **10**

Md. Ehsanul Haque, Mahe Zabin and Jia Uddin
EnsembleXAI-Motor: A Lightweight Framework for Fault Classification in Electric Vehicle Drive Motors Using Feature Selection, Ensemble Learning, and Explainable AI
Reprinted from: *Machines* **2025**, *13*, 314, https://doi.org/10.3390/machines13040314 . . . . . **25**

Ghaith Al-refai, Dina Karasneh, Hisham Elmoaqet, Mutaz Ryalat and Natheer Almtireen
Surface Classification from Robot Internal Measurement Unit Time-Series Data Using Cascaded and Parallel Deep Learning Fusion Models
Reprinted from: *Machines* **2025**, *13*, 251, https://doi.org/10.3390/machines13030251 . . . . . **59**

Mouaad Bougoffa, Samir Benmoussa, Mohand Djeziri and Olivier Palais
Hybrid Deep Learning for Fault Diagnosis in Photovoltaic Systems
Reprinted from: *Machines* **2025**, *13*, 378, https://doi.org/10.3390/machines13050378 . . . . . **73**

Mohammed Yousri Silaa, Oscar Barambones, Aissa Bencherif and Ilyes Rougab
An Adaptive Control Strategy with Switching Gain and Forgetting Factor for a Robotic Arm Manipulator
Reprinted from: *Machines* **2025**, *13*, 424, https://doi.org/10.3390/machines13050424 . . . . . **91**

Bartas Lisauskas and Rytis Maskeliunas
Efficient Transformer-Based Road Scene Segmentation Approach with Attention-Guided Decoding for Memory-Constrained Systems
Reprinted from: *Machines* **2025**, *13*, 466, https://doi.org/10.3390/machines13060466 . . . . . **110**

Andrei Raul Osan and Raul Florentin Drenţa
Application of Artificial Neural Networks in Predicting Surface Quality and Machining Time
Reprinted from: *Machines* **2025**, *13*, 561, https://doi.org/10.3390/machines13070561 . . . . . **131**

Patricia Ochoa, Cinthia Peraza, Patricia Melin, Oscar Castillo and Zong Woo Geem
Type-3 Fuzzy Dynamic Adaptation of the Crossover Parameter in Differential Evolution for the Optimal Design of Type-3 Fuzzy Controllers for the Inverted Pendulum
Reprinted from: *Machines* **2025**, *13*, 450, https://doi.org/10.3390/machines13060450 . . . . . **154**

Karthik Subramanian, Liya Thomas, Melis Sahin and Ferat Sahin
Supporting Human–Robot Interaction in Manufacturing with Augmented Reality and Effective Human–Computer Interaction: A Review and Framework
Reprinted from: *Machines* **2024**, *12*, 706, https://doi.org/10.3390/machines12100706 . . . . . **175**

Hai Van Pham, Huy Quoc Do, Minh Nguyen Quang, Farzin Asadi and Philip Moore
Proposed Multi-ST Model for Collaborating Multiple Robots in Dynamic Environments
Reprinted from: *Machines* **2024**, *12*, 797, https://doi.org/10.3390/machines12110797 . . . . . **208**

Georgios Voutsakelis, Ioannis Dimkaros, Nikolaos Tzimos, George Kokkonis and Sotirios Kontogiannis
Development and Evaluation of a Tool for Blind Users Utilizing AI Object Detection and Haptic Feedback
Reprinted from: *Machines* **2025**, *13*, 398, https://doi.org/10.3390/machines13050398 . . . . . **240**

Jie Pan, Yongfeng Cheng, Chunhua Hu, Ming Jiang, Yong Ma, Fanhao Meng and Qiang Shi
The Innovative Design and Performance Testing of a Mobile Robot for the Automated Installation of Spacers on Six-Split Transmission Lines †
Reprinted from: *Machines* **2025**, *13*, 432, https://doi.org/10.3390/machines13050432 . . . . . **261**

Adriana Florescu
Digital Twin for Flexible Manufacturing Systems and Optimization Through Simulation: A Case Study
Reprinted from: *Machines* **2024**, *12*, 785, https://doi.org/10.3390/machines12110785 . . . . . **280**

Loïc Tissot-Daguette, Célestin Vallat, Marijn Nijenhuis, Florent Cosandier and Simon Henein
Near-Zero Parasitic Shift Rectilinear Flexure Stages Based on Coupled n-RRR Planar Parallel Mechanisms
Reprinted from: *Machines* **2025**, *13*, 530, https://doi.org/10.3390/machines13060530 . . . . . **300**

Filippo Foiani, Giulia Morettini, Massimiliano Palmieri, Stefano Carletta, Filippo Cianetti and Marco Dionigi
Multibody Simulation of 1U CubeSat Passive Attitude Stabilisation Using a Robotic Arm
Reprinted from: *Machines* **2025**, *13*, 509, https://doi.org/10.3390/machines13060509 . . . . . **323**

Lei Jia and Yan Pei
Recent Advances in Multi-Agent Reinforcement Learning for Intelligent Automation and Control of Water Environment Systems
Reprinted from: *Machines* **2025**, *13*, 503, https://doi.org/10.3390/machines13060503 . . . . . **345**

Evangelos Tzimas, George Papazetis, Panorios Benardos and George-Christopher Vosniakos
Design and Development of a Flexible Manufacturing Cell Controller Using an Open-Source Communication Protocol for Interoperability
Reprinted from: *Machines* **2024**, *12*, 519, https://doi.org/10.3390/machines12080519 . . . . . **389**

Pedro M. P. Curralo, Raul D. S. G. Campilho, Joaquim A. P. Pereira and Francisco J. G. Silva
Design of Connector Assembly Equipment for the Automotive Industry
Reprinted from: *Machines* **2024**, *12*, 731, https://doi.org/10.3390/machines12100731 . . . . . **410**

Amr Rashed, Yousry Abdulazeem, Tamer Ahmed Farrag, Amna Bamaqa, Malik Almaliki, Mahmoud Badawy and Mostafa A. Elhosseini
Toward Inclusive Smart Cities: Sound-Based Vehicle Diagnostics, Emergency Signal Recognition, and Beyond
Reprinted from: *Machines* **2025**, *13*, 258, https://doi.org/10.3390/machines13040258 . . . . . **430**

Ciprian Paduraru, Marina Cernat and Adelina-Nicoleta Staicu
A Unified Framework for Automated Testing of Robotic Process Automation Workflows Using Symbolic and Concolic Analysis †
Reprinted from: *Machines* **2025**, *13*, 504, https://doi.org/10.3390/machines13060504 . . . . . **466**

Daniele Puri, Leonardo Vita, Davide Gattamelata and Valerio Tulliani
Roll/Tip-Over Risk Analysis of Agricultural Self-Propelled Machines Using Airborne LiDAR Data: GIS-Based Approach
Reprinted from: *Machines* **2025**, *13*, 377, https://doi.org/10.3390/machines13050377 . . . . . **494**

Felipe Klein Fiorentin, Rita Dantas, Jorge Wolfs Gil, Andrea Piga Carboni, Thiago Antonio Fiorentin and Abílio Manuel Pinho de Jesus
On the Specimen Design, Physical Properties and Geometry Effect on Heat Generation and Thermal Gradient in Ultrasonic Fatigue
Reprinted from: *Machines* **2025**, *13*, 380, https://doi.org/10.3390/machines13050380 . . . . . **513**

Min Zhang, Zhijing Zhang, Jian Xiong and Xiao Chen
Accuracy Analysis of Complex Transmission System with Distributed Tooth Profile Errors
Reprinted from: *Machines* **2024**, *12*, 459, https://doi.org/10.3390/machines12070459 . . . . . **534**

Zhuangya Zhang, Baorun Yang, Mingde Duan, Ruijie Gu, Shijie Liang and Yang Chen
Design and Experimental Research of a Non-Destructive Detection Device for High-Precision Cylindrical Roller Dynamic Unbalance
Reprinted from: *Machines* **2024**, *12*, 684, https://doi.org/10.3390/machines12100684 . . . . . **554**

Sheeraz Ahmed, Qasim Ali, Ghulam Jawad Sirewal, Kapeel Kumar and Gilsu Choi
Hybrid Brushless Wound-Rotor Synchronous Machine with Dual-Mode Operation for Washing Machine Applications
Reprinted from: *Machines* **2025**, *13*, 342, https://doi.org/10.3390/machines13050342 . . . . . **570**

Vasileios I. Vlachou and Theoklitos S. Karakatsanis
Development of a Fault-Tolerant Permanent Magnet Synchronous Motor Using a Machine-Learning Algorithm for a Predictive Maintenance Elevator
Reprinted from: *Machines* **2025**, *13*, 427, https://doi.org/10.3390/machines13050427 . . . . . **586**

Tae-Seong Kim, Jun-Won Yang, Kyung-Hun Shin, Gang-Hyeon Jang, Cheol Han and Jang-Young Choi
Analysis of Electromagnetic Characteristics of Outer Rotor Type BLDC Motor Based on Analytical Method and Optimal Design Using NSGA-II
Reprinted from: *Machines* **2025**, *13*, 440, https://doi.org/10.3390/machines13060440 . . . . . **614**

Athanasios Kokkinis, Theodore Frantzis, Konstantinos Skordis, George Nikolakopoulos and Panagiotis Koustoumpardis
Review of Automated Operations in Drilling and Mining
Reprinted from: *Machines* **2024**, *12*, 845, https://doi.org/10.3390/machines12120845 . . . . . **632**

Berat Baris Buldum, Kamil Leksycki and Suleyman Cinar Cagan
Comparative Analysis of Dry, Minimum Quantity Lubrication, and Nano-Reinforced Minimum Quantity Lubrication Environments on the Machining Performance of AZ91D Magnesium Alloy
Reprinted from: *Machines* **2025**, *13*, 430, https://doi.org/10.3390/machines13050430 . . . . . **651**

Sunil Kumar Maurya, Gianni Campatelli, Massimo Veracini, Massimo Arcioni and Dario Clori
Balancing Productivity and Sustainability in EDM: A Comprehensive Analysis of Energy Consumption and Electrode Degradation
Reprinted from: *Machines* **2025**, *13*, 469, https://doi.org/10.3390/machines13060469 . . . . . **667**

Diego Fernando Ramírez-Jiménez and Cristian Alejandro Torres Valencia
Experimental Design and Simulation of a Fly-Cutting Plant for Academic Environment Practices
Reprinted from: *Machines* **2025**, *13*, 15, https://doi.org/10.3390/machines13010015 . . . . . **683**

Vignesh Venkat Raman, Rakshith Badarinath and Vittaldas V. Prabhu
TOPSIS-Based Methodology for Selecting Fused Filament Fabrication Machines
Reprinted from: *Machines* **2025**, *13*, 574, https://doi.org/10.3390/machines13070574 . . . . . **707**

About the Editor

Raul D. S. G. Campilho

Raul D. S. G. Campilho is a researcher at CIDEM, ISEP—School of Engineering, Polytechnic of Porto, and a collaborator at INEGI—Institute of Science and Innovation in Mechanical and Industrial Engineering, Portugal. His main research interests include advanced joining technologies, composite materials, manufacturing systems, and structural design for aerospace and mechanical applications. He has coordinated and participated in numerous national and international research projects focusing on structural performance, hybrid bonding techniques, and digital manufacturing. Raul has published extensively in peer-reviewed journals and is involved in the scientific community as editor in chief, associate editor, and reviewer for scientific journals. He is also a supervisor of MSc and PhD students and regularly collaborates with industry on applied research and innovation initiatives.

Preface

The fields of machine design, automation, and robotics have experienced rapid development in recent years, driven by the need for efficiency, adaptability, and sustainability in industrial processes. Modern production systems increasingly rely on intelligent machines, advanced control strategies, and collaborative robotic solutions to meet the challenges of product quality, flexibility, and cost reduction in competitive markets.

This Reprint, entitled "Recent Developments in Machine Design, Automation and Robotics", compiles 33 research contributions originally published in the journal *Machines*. The works presented here involve different lines of research in the field, including artificial intelligence for predictive analysis and control, deep learning methods for classification and diagnosis, adaptive and robust control approaches for robotic systems, and frameworks for collaborative and multi-agent robotics.

Several papers focus on the integration of digital tools and automation strategies into industrial contexts, such as digital twins for flexible manufacturing, open-source communication protocols for interoperability, and robotic process automation testing. Sustainability and performance trade-offs are also addressed, for instance in machining processes, lubrication strategies, and energy-efficient manufacturing. Complementary to these themes, novel solutions in machine and mechanism design are presented, covering areas such as flexure mechanisms, connector assembly equipment, testing devices, and space robotics applications.

Together, these contributions provide a concise overview of the most recent scientific and technological advances in machine design, automation, and robotics, highlighting both theoretical approaches and applied solutions. The collection aims to serve as a reference point for researchers and industrialists seeking to understand and further develop the state of the art in this quickly evolving domain.

Raul D. S. G. Campilho
Guest Editor

machines

MDPI

Editorial

Recent Developments in Machine Design, Automation and Robotics

Raul D. S. G. Campilho [1,2]

[1] CIDEM, ISEP—School of Engineering, Polytechnic of Porto, 431, 4200-072 Porto, Portugal; raulcampilho@gmail.com
[2] INEGI—Institute of Science and Innovation in Mechanical and Industrial Engineering, Pólo FEUP, 400, 4200-465 Porto, Portugal

This editorial introduces a Special Issue of *Machines* entitled "Recent Developments in Machine Design, Automation and Robotics". This Special Issue arrives at an important time in industrial evolution, when digital transformation, sustainability, and productivity optimization are redefining how machines and systems are conceived, built, and operated [1–3]. Modern industrial competitiveness depends not only on production output or cost efficiency, but also on the ability to rapidly adapt to changing demands and ensure high levels of reliability and traceability while incorporating environmentally responsible practices to help businesses achieve their goals [4]. These objectives can only be achieved by integrating automation, robotics, and intelligent system design into the industrial processes [5,6]. As the fourth industrial revolution (Industry 4.0) continues to progress, the design and deployment of machines and manufacturing systems is facing significant changes [7]. Concepts such as cyber–physical systems, real-time data processing, artificial intelligence (AI), and the Internet of Things (IoT) are giving rise to smart factories—environments in which automation and robotics are not isolated functions but act together to promote autonomous operation and continuous improvement [8,9]. This transformation provides new opportunities and presents novel challenges: the implemented production systems must be increasingly flexible, decentralized, safe for human interaction, and resilient to disturbances [10]. On the other hand, global pressures regarding sustainability and energy efficiency also affect machines' design and operation during their life cycle, influencing material selection, maintenance, and recycling [11]. This Special Issue aims to address these needs by collecting innovative scientific and technological contributions that describe the evolution of machine design, automation, and robotics. The 32 accepted papers cover a wide range of innovative topics in the form of review and research papers, which are organized into five main areas. Each area represents a distinct dimension of the current research and developments in intelligent industrial systems.

1. Artificial Intelligence and Machine Learning in Automation

The increasing user experience and accessibility of AI tools have revolutionized how industrial automation is applied in real settings [12,13]. Machine learning (ML) techniques currently play a major role in improving machine performance, adaptability, and fault tolerance by applying techniques/tools such as advanced diagnostic systems, predictive control, and autonomous decision-making [14]. These tools allow systems to read and act in dynamic environments, enabling self-optimization and proactive maintenance strategies [15]. AI is becoming increasingly prevalent in hardware and edge devices, creating intelligent machines capable of learning from their own operation and improving over time [16,17]. Fabiocchi et al. [18] introduced a vision-based virtual sensor that estimates

ground reaction forces in mechanical systems without the need for physical sensors. Using AI models trained on dynamic behavior data, the system analyzed video input to infer force trends in real time. Validated under realistic conditions, the method was considered both accurate and robust, and it offered a non-invasive and cost-effective solution for force monitoring in industrial machinery. Haque et al. [19] proposed a lightweight and interpretable framework for fault diagnosis in electric vehicle drive motors. By combining optimized ML models with feature selection and a soft voting ensemble, the approach achieved high diagnostic accuracy with low computational demands. The framework was validated using both simulated and real-world datasets, demonstrating strong generalizability and transparency. Al-refai et al. [20] developed two deep learning models for surface classification using internal measurement unit (IMU) time-series data from ground robots as a cost-effective alternative to vision-based systems. A cascaded convolutional neural network-long short-term memory (CNN-LSTM) attention model and a parallel fusion variant were proposed and evaluated. While the cascaded model achieved higher accuracy, the parallel model offered faster processing, making it better suited for real-time applications. Both models benefited from the inclusion of multi-head attention mechanisms. Bougoffa et al. [21] introduced a hybrid deep learning framework combining stacked sparse auto-encoders with an optimized multilayer perceptron to detect faults in photovoltaic systems. Trained and validated on real-world data, the model achieved near-perfect accuracy for various fault types and environmental conditions. It outperformed conventional methods, thus providing a scalable and efficient solution to enhance system reliability and reduce maintenance costs. Silaa et al. [22] proposed an adaptive sliding-mode controller with a forgetting factor to improve trajectory tracking in a two-degrees-of-freedom (DOF) robotic arm. The adaptive mechanism enhanced robustness and reduced chattering, while the forgetting factor helped to prevent gain drift. Stability was proven via Lyapunov theory, and simulations confirmed superior performance over conventional sliding-mode controls (SMCs) in terms of accuracy and noise resilience. Lisauskas and Maskeliunas [23] presented a Transformer-based model for road scene segmentation which was designed to improve object detection and environmental understanding in autonomous systems. A custom decoder with attention-based fusion was developed to preserve spatial detail and enhance contextual awareness. The model achieved strong performance on the Cityscapes dataset and, with its lightweight architecture, was considered suitable for memory-constrained devices. Osan and Drenţa [24] applied artificial neural networks (ANNs) to optimize metal machining processes by predicting surface quality and estimating machining time. One model forecasted surface roughness based on machining parameters, while another estimated operation duration for cost planning. The results showed accurate predictions, supporting improved efficiency and process control in manufacturing. Ochoa et al. [25] introduced a type-3 fuzzy system (T3FS) to dynamically adapt the crossover parameter in differential evolution algorithms. Applied to an inverted pendulum control task, the approach improved optimization accuracy and consistency compared to static parameter settings. The results demonstrated lower error and variance, highlighting the potential of T3FS in tuning controllers for uncertain nonlinear systems.

2. Robotics and Human–Machine Interaction

Modern robotics is shifting to higher levels of collaboration, mobility, and contextual awareness [26,27]. Human–robot interaction (HRI) is no longer limited to industrial safety or physical assistance. It now involves intuitive interfaces, augmented reality, shared decision-making, and accessibility [28]. This area includes developments in collaborative robots (cobots), multi-robot coordination, mobile robotics for specialized tasks, and sensor-driven interaction [29,30]. This research is essential for automating tasks that are

traditionally occupied by human labor, while maintaining safety, flexibility, and ergonomic considerations [31]. As robotics becomes universal, designing systems that can understand and respond to human intentions becomes central to successful implementation and coexistence in manufacturing [32]. Subramanian et al. [33] reviewed the role of augmented reality (AR) in HRI, particularly in manufacturing contexts. It identified the main challenges, such as situational awareness and communication, and proposed a framework to guide the design and evaluation of AR systems. The framework, supported by case studies, emphasized safety, collaboration, and user-centered principles from human–computer interaction (HCI). Pham et al. [34] proposed a path-planning model (named the Multi-ST model) for multi-robot coverage path planning in environments with static and dynamic obstacles. Based on a spiral-spanning tree algorithm with intelligent reasoning, the method improved coordination, obstacle avoidance, and coverage efficiency. The experimental results showed superior performance compared to existing approaches in terms of robustness, scalability, and adaptability to dynamic conditions. Voutsakelis et al. [35] developed and evaluated a smartphone app for blind users that combined real-time object detection with haptic and audio feedback. Using the EfficientDet-lite2 model and the common objects in context (COCO) dataset, the app enabled users to interact with detected objects through touch and sound. Usability tests confirmed that the haptic feedback significantly improved user experience, accessibility, and independence. Pan et al. [36] presented the design and testing of a robot to install spacers on six-split transmission lines, aiming to reduce manual labor and safety risks. The system included a wire-following mechanism, spacer storage, and an assembly arm. The experimental results showed efficient operation with low error rates, demonstrating its potential for use in power line maintenance under complex conditions. Florescu [37] developed a digital twin for a flexible manufacturing system integrating intelligent machine tools and robots, supporting the goals of Industry 4.0. The digital replica was used for system design, simulation, and virtual commissioning of a real cylindrical-part production process. Simulations enabled error detection, collision avoidance, and process optimization through collaborative robotics. Tissot-Daguette et al. [38] introduced flexure-based rectilinear stages using coupled serial revolute joint (RRR) planar parallel mechanisms to minimize parasitic shifts without sacrificing stiffness or range. Finite element method (FEM) simulations and experimental tests confirmed sub-micron precision and robust support stiffness. A silicon prototype with thermally preloaded buckling beams further reduced translational stiffness, demonstrating the design's suitability for high-precision micro- and nano-positioning applications. Foiani et al. [39] evaluated a robotic C-arm manipulator as a passive attitude control device for a 1U CubeSat. Simulations showed that while gravity gradient stabilization and magnetic control were feasible, pointing accuracy was limited and affected by instabilities due to the lack of damping. The approach proved suitable for missions with relaxed attitude requirements typical of CubeSat platforms.

3. Smart Manufacturing and Control Systems

Control strategies play a central role in the architecture of automated operations. In the context of smart manufacturing, they evolve into adaptive, real-time, and often AI-assisted mechanisms that ensure high precision, energy efficiency, and system resilience [40]. This area explores advanced manufacturing cells, digital twins for simulation and optimization, and automated quality assurance techniques [41,42]. It also embraces the design of software architectures and communication protocols that enable interoperability, decentralization, and modularity, which are the main characteristics of Industry 4.0-compliant environments [43]. Overall, smart control systems enhance productivity and are fundamental for integrating sustainability and flexibility into modern production [44]. Jia and Pei [45]

reviewed recent advances in multi-agent reinforcement learning (MARL) applied to the intelligent control of water environment systems. The authors examined core algorithm types, cooperative strategies for tasks like water scheduling and monitoring, and challenges such as communication instability and system heterogeneity. The study also highlighted emerging solutions for generalization and robustness, which provided insights for future applications in water infrastructure. Tzimas et al. [46] presented the design and implementation of a controller for a flexible manufacturing cell with heterogeneous equipment and communication interfaces. Focusing on integration, extensibility, and interoperability, the system was applied to microfluidic device production. Open-source technologies enabled effective control and communication, supporting reusability and reducing dependence on proprietary protocols. Curralo et al. [47] designed semi-automatic equipment to replace manual connector assembly in the automotive sector, aiming to reduce cycle time and labor costs. The system automated component insertion and used AI-based vision for quality control. Structural verification, cost analysis, and validation confirmed the design's effectiveness and adaptability to other industries such as electronics and aerospace. Rashed et al. [48] addressed the lack of sound-based diagnostics in intelligent transportation systems by developing a specialized audio dataset and proposing a multi-model classification framework. Using advanced feature extraction and a Bayesian-optimized ensemble method, the system accurately identified vehicle faults and emergency sounds. The results showed strong potential for real-time diagnostics and accessibility enhancements in smart city environments. Paduraru et al. [49] proposed a hybrid testing framework for robotic process automation (RPA) workflows, which combined symbolic execution and concolic testing to improve validation. The open-source solution enhanced test coverage, reduced manual effort, and supported multiple RPA platforms. Industry-based evaluations confirmed its effectiveness in improving reliability and aligning with best testing practices. Puri et al. [50] addressed agricultural safety by analyzing terrain slope data with terrestrial Light Detection and Ranging (LiDAR) and correlating it with machinery tilt specifications to assess tip/roll-over risks. Conducted in alpine vineyards in Trento, the research supported safer decision-making for machinery use on steep terrain. The findings aimed to improve occupational health and safety in agriculture by informing both operators and manufacturers. Klein Fiorentin et al. [51] addressed challenges in ultrasonic fatigue testing, focusing on specimen design and heat generation during high-frequency loading. A methodology was proposed to quantify heat severity based on material properties and to guide geometric design for resonance at 20 kHz. The work provided practical insights into managing temperature effects and thermal gradients in fatigue testing.

4. Electric Machines, Drives, and Mechatronics

Despite the growing role of software and AI, mechanical and electromechanical components remain the core of industrial systems [52]. Innovations in electric machines, powertrains, sensors, and actuation mechanisms enable more compact, efficient, and robust machines adapted to specific applications [53]. The field of mechatronics, combining mechanical engineering, electronics, and computing, continues to produce highly integrated and responsive systems [54,55]. In this area, attention is given to the design, modeling, and optimization of electric motors, dynamic analysis of mechanical transmissions, and advanced diagnostics for condition monitoring and system reliability [56]. Zhang et al. [57] proposed a modeling approach for gear tooth profile errors using both systematic and random distribution components to better reflect real machining conditions. A mathematical model was developed to assess their impact on contact stress and transmission accuracy. The results showed a 13.8% reduction in output accuracy due to these errors, which showed the importance of precise error prediction during the design stage. Zhang et al. [58] de-

veloped a non-destructive device to detect dynamic unbalance in small cylindrical rollers using air flotation support and resonance amplification. A dynamic model was created and validated through modal and harmonic response analyses. Experimental tests confirmed the system's ability to measure unbalance accurately without damaging the roller surface. Ahmed et al. [59] presented a hybrid brushless wound-rotor synchronous machine with an outer rotor and integrated permanent magnets for enhanced performance. A centrifugal switch enabled dual-mode operation, allowing the machine to function as a permanent magnet synchronous motor (PMSM) at high speeds. Designed for dual-speed applications like washing machines, the concept was validated through two-dimensional (2D) simulations using ANSYS®. Vlachou and Karakatsanis [60] developed a fault-tolerant PMSM for elevator systems and validated its performance through simulation and real-world testing. Operational data were analyzed using a Random Forest classifier, achieving 94% accuracy in detecting motor health states. The integration of robust design with ML enabled effective predictive maintenance and improved system reliability. Kim et al. [61] proposed a 2D equivalent analysis method for outer rotor brushless DC (BLDC) motors, addressing challenges posed by the permanent magnet (PM) overhang structure. The approach incorporated spatial harmonics and Carter coefficients to accurately evaluate electromagnetic performance. Optimization using the Non-dominated Sorting Genetic Algorithm-II (NSGA-II) enhanced efficiency and power density, and the results were validated against FEM simulations for accuracy.

5. Green Design, Machining, and Sustainability

Sustainable engineering is no longer a side consideration but a guiding principle in the design of industrial systems [62,63]. This area addresses green manufacturing processes, energy-efficient machining, environmentally conscious machine design, and multi-criteria decision-making, facilitating the selection of technologies with minimal environmental impact [64]. Concepts such as minimum quantity lubrication (MQL), recycling of materials, and energy-aware optimization are central in this area [65]. As industries face increasing regulatory pressure and societal expectations to implement sustainable practices, the development of machines that support circular economy principles is becoming a strategic imperative [66]. Kokkinis et al. [67] reviewed current trends and emerging technologies in mining exploration and drilling, highlighting the shift toward fully autonomous, infrastructure-independent machines. Advances in AI, IoT, robotics, and computational modeling were shown to be transforming mining equipment design. Their study emphasized the need for new approaches in deep and near-to-face exploration, especially as interest in re-opening abandoned mines grows. Buldum et al. [68] evaluated the turning performance of AZ91D magnesium alloy under dry, MQL, and nano-reinforced MQL (NanoMQL) conditions using carbon nanotube additives. The NanoMQL setup significantly improved surface quality and reduced tool wear compared to other methods. Enhanced lubrication and thermal conductivity led to lower built-up edge formation, making NanoMQL ideal for demanding machining conditions. Maurya et al. [69] analyzed energy consumption and electrode wear in die-sinking electrical discharge machining (EDM) using real-time monitoring under varied machining parameters. High discharge currents significantly reduced machining time and specific energy consumption, improving both efficiency and sustainability. The findings offered practical insights to optimize EDM processes in advanced manufacturing. Ramírez-Jiménez and Torres Valencia [70] presented the design and simulation of a low-cost fly-cutting plant prototype for academic training in automation and control. A three-dimensional (3D) model was created in SketchUp and simulated using Matlab and Simulink, incorporating a conveyor and continuous cutting system. Performance and cutting tests validated the setup, making it suitable for hands-on process control education.

Raman et al. [71] proposed a technique for order of preference by similarity to ideal solution (TOPSIS)-based methodology to support the selection of Fused Filament Fabrication (FFF) machines based on multiple weighted criteria. Sensitivity analysis using Monte Carlo simulations evaluated the effects of decision variability and scoring uncertainty. The approach was tested on three case studies, showing robust and consistent machine rankings for specific part requirements.

This Special Issue presents an innovative and forward-looking compilation of research showcasing the coverage and interdisciplinarity of machine design, automation, and robotics that defines the next generation of industrial technologies. By presenting rigorous theoretical analyses, experimental investigations, numerical simulations, and practical implementations, the contributing authors have collectively advanced the understanding of automation, robotics, and machine design in meaningful and application-oriented ways. As the Guest Editor of this Special Issue, I wish to express my sincere appreciation to all authors for their high-quality contributions, and to the reviewers whose critical evaluations and expert insights were essential to ensuring the high scientific standard of this issue. I also gratefully acknowledge the editorial staff of *Machines* for their support throughout the publication process. It is my strong belief that this Special Issue, facilitated by the open access publication system offered by MDPI, will serve as a valuable reference for academic researchers and as a source of technical guidance for industry professionals dealing with the design, control, and integration of intelligent manufacturing systems.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Li, L. Digital transformation and sustainable performance: The moderating role of market turbulence. *Ind. Mark. Manag.* **2022**, *104*, 28–37. [CrossRef]
2. Han, Y.; Li, Z.; Feng, T.; Qiu, S.; Hu, J.; Yadav, K.K.; Obaidullah, A.J. Unraveling the impact of digital transformation on green innovation through microdata and machine learning. *J. Environ. Manag.* **2024**, *354*, 120271. [CrossRef] [PubMed]
3. Turan, E.; Konuşkan, Y.; Yıldırım, N.; Tunçalp, D.; İnan, M.; Yasin, O.; Turan, B.; Kerimoğlu, V. Digital twin modelling for optimizing the material consumption: A case study on sustainability improvement of thermoforming process. *Sustain. Comput. Inform. Syst.* **2022**, *35*, 100655. [CrossRef]
4. Leng, J.; Liu, Q.; Ye, S.; Jing, J.; Wang, Y.; Zhang, C.; Zhang, D.; Chen, X. Digital twin-driven rapid reconfiguration of the automated manufacturing system via an open architecture model. *Robot. Comput. Integr. Manuf.* **2020**, *63*, 101895. [CrossRef]
5. Duro, R.; Kondratenko, Y. *Advances in Intelligent Robotics and Collaborative Automation*; CRC Press: Boca Raton, FL, USA, 2022.
6. Mineo, C.; Vasilev, M.; Cowan, B.; MacLeod, C.; Pierce, S.G.; Wong, C.; Yang, E.; Fuentes, R.; Cross, E. Enabling robotic adaptive behaviour capabilities for new industry 4.0 automated quality inspection paradigms. *Insight Non Destr. Test. Cond. Monit.* **2020**, *62*, 338–344. [CrossRef]
7. George, A.S.; George, A.H. Industrial revolution 5.0: The transformation of the modern manufacturing process to enable man and machine to work hand in hand. *J. Seybold Rep. ISSN NO* **2020**, *1533*, 1525.68128.
8. Ren, L.; Dong, J.; Liu, S.; Zhang, L.; Wang, L. Embodied intelligence toward future smart manufacturing in the era of AI foundation model. *IEEE/ASME Trans. Mechatron.* **2024**, 1–11, *early access*. [CrossRef]
9. Czimmermann, T.; Chiurazzi, M.; Milazzo, M.; Roccella, S.; Barbieri, M.; Dario, P.; Oddo, C.M.; Ciuti, G. An autonomous robotic platform for manipulation and inspection of metallic surfaces in industry 4.0. *IEEE Trans. Autom. Sci.* **2021**, *19*, 1691–1706. [CrossRef]
10. Leng, J.; Ruan, G.; Xu, C.; Zhou, X.; Xu, K.; Qiao, Y.; Liu, Z.; Liu, Q. Deep reinforcement learning of graph convolutional neural network for resilient production control of mass individualized prototyping toward industry 5.0. *IEEE Trans. Syst. Man Cybern. Syst.* **2024**, *54*, 7092–7105. [CrossRef]
11. Zhai, J.; Chen, X.; Sun, X.; Xie, H. Economically and thermodynamically efficient pressure-swing distillation with heat integration and heat pump techniques. *Appl. Therm. Eng.* **2023**, *218*, 119389. [CrossRef]
12. Lazarenko, B. Manufacturing evolution: The intelligent automation revolution. *Quality* **2025**, *64*, 43.
13. Ghulaxe, V. Robotic process automation with ML and artificial intelligence: Revolutionizing business processes. *Int. J. Eng. Technol. Manag. Sci.* **2024**, *4*, 107–112.

14. Li, L.; Liu, J.; Wei, S.; Chen, G.; Blasch, E.; Pham, K. Smart robot-enabled remaining useful life prediction and maintenance optimization for complex structures using artificial intelligence and machine learning. In Proceedings of the Sensors and Systems for Space Applications XIV, Online, 12–16 April 2021; pp. 100–108.
15. Zhang, Y.; Cai, B.; Gao, C.; Zhao, Y.; Shao, X.; Yang, C. A system-centred predictive maintenance re-optimization method based on multi-agent deep reinforcement learning. *Expert Syst. Appl.* **2025**, *274*, 127034. [CrossRef]
16. Tsochev, G.; Sharabov, M. Artificial intelligence methods used in industry 4.0 in particular industrial control systems. *AIP Conf. Proc.* **2021**, *2333*, 070017. [CrossRef]
17. Yang, T.; Yi, X.; Lu, S.; Johansson, K.H.; Chai, T. Intelligent manufacturing for the process industry driven by industrial artificial intelligence. *Engineering* **2021**, *7*, 1224–1230. [CrossRef]
18. Fabiocchi, D.; Giulietti, N.; Carnevale, M.; Giberti, H. AI-driven virtual sensors for real-time dynamic analysis of mechanisms: A feasibility study. *Machines* **2024**, *12*, 257. [CrossRef]
19. Haque, M.E.; Zabin, M.; Uddin, J. EnsembleXAI-Motor: A lightweight framework for fault classification in electric vehicle drive motors using feature selection, ensemble learning, and explainable AI. *Machines* **2025**, *13*, 314. [CrossRef]
20. Al-refai, G.; Karasneh, D.; Elmoaqet, H.; Ryalat, M.; Almtireen, N. Surface classification from robot internal measurement unit time-series data using cascaded and parallel deep learning fusion models. *Machines* **2025**, *13*, 251. [CrossRef]
21. Bougoffa, M.; Benmoussa, S.; Djeziri, M.; Palais, O. Hybrid deep learning for fault diagnosis in photovoltaic systems. *Machines* **2025**, *13*, 378. [CrossRef]
22. Silaa, M.Y.; Barambones, O.; Bencherif, A.; Rougab, I. An adaptive control strategy with switching gain and forgetting factor for a robotic arm manipulator. *Machines* **2025**, *13*, 424. [CrossRef]
23. Lisauskas, B.; Maskeliunas, R. Efficient transformer-based road scene segmentation approach with attention-guided decoding for memory-constrained systems. *Machines* **2025**, *13*, 466. [CrossRef]
24. Osan, A.R.; Drenţa, R.F. Application of artificial neural networks in predicting surface quality and machining time. *Machines* **2025**, *13*, 561. [CrossRef]
25. Ochoa, P.; Peraza, C.; Melin, P.; Castillo, O.; Geem, Z.W. Type-3 fuzzy dynamic adaptation of the crossover parameter in differential evolution for the optimal design of type-3 fuzzy controllers for the inverted pendulum. *Machines* **2025**, *13*, 450. [CrossRef]
26. Olszewski, M. Modern industrial robotics. *Pomiary Autom. Robot.* **2020**, *24*, 5–20. [CrossRef]
27. Elwin, M.L.; Strong, B.; Freeman, R.A.; Lynch, K.M. Human-multirobot collaborative mobile manipulation: The omnid mocobots. *IEEE Robot. Autom. Lett.* **2022**, *8*, 376–383. [CrossRef]
28. Liang, H.; Zheng, H.; Lu, Y. Context-aware cognitive assistive assembly system based on online human action recognition. In Proceedings of the 2024 IEEE 20th International Conference on Automation Science and Engineering (CASE), Bari, Italy, 28 August–1 September 2024; pp. 381–387.
29. Jin, L.; Qi, Y.; Luo, X.; Li, S.; Shang, M. Distributed competition of multi-robot coordination under variable and switching topologies. *IEEE Trans. Autom. Sci.* **2021**, *19*, 3575–3586. [CrossRef]
30. Hu, K.; Zou, L.; Chen, Z.; Jiang, J.; Jiang, F.; Tao, X. Wireless multi-robot collaboration: Communications, perception, control and planning. *IEEE Netw.* **2024**, *39*, 302–311. [CrossRef]
31. Mitterberger, D.; Dörfler, K.; Sandy, T.; Salveridou, F.; Hutter, M.; Gramazio, F.; Kohler, M. Augmented bricklaying: Human–machine interaction for in situ assembly of complex brickwork using object-aware augmented reality. *Constr. Robot.* **2020**, *4*, 151–161. [CrossRef]
32. Jokinen, K.; Homma, K.; Matsumoto, Y.; Fukuda, K. Integration and interaction of trustworthy ai in a virtual coach–an overview of EU-Japan collaboration on eldercare. In Proceedings of the Annual Conference of the Japanese Society for Artificial Intelligence, Online, 8–11 June 2021; pp. 190–200.
33. Subramanian, K.; Thomas, L.; Sahin, M.; Sahin, F. Supporting human–robot interaction in manufacturing with augmented reality and effective human–computer interaction: A review and framework. *Machines* **2024**, *12*, 706. [CrossRef]
34. Pham, H.V.; Do, H.Q.; Nguyen Quang, M.; Asadi, F.; Moore, P. Proposed multi-ST model for collaborating multiple robots in dynamic environments. *Machines* **2024**, *12*, 797. [CrossRef]
35. Voutsakelis, G.; Dimkaros, I.; Tzimos, N.; Kokkonis, G.; Kontogiannis, S. Development and evaluation of a tool for blind users utilizing AI object detection and haptic feedback. *Machines* **2025**, *13*, 398. [CrossRef]
36. Pan, J.; Cheng, Y.; Hu, C.; Jiang, M.; Ma, Y.; Meng, F.; Shi, Q. The innovative design and performance testing of a mobile robot for the automated installation of spacers on six-split transmission lines. *Machines* **2025**, *13*, 432. [CrossRef]
37. Florescu, A. Digital twin for flexible manufacturing systems and optimization through simulation: A case study. *Machines* **2024**, *12*, 785. [CrossRef]
38. Tissot-Daguette, L.; Vallat, C.; Nijenhuis, M.; Cosandier, F.; Henein, S. Near-zero parasitic shift rectilinear flexure stages based on coupled n-RRR planar parallel mechanisms. *Machines* **2025**, *13*, 530. [CrossRef]

39. Foiani, F.; Morettini, G.; Palmieri, M.; Carletta, S.; Cianetti, F.; Dionigi, M. Multibody simulation of 1U CubeSat passive attitude stabilisation using a robotic arm. *Machines* **2025**, *13*, 509. [CrossRef]
40. Simmons, N. Design of robust manufacturing systems: A systemic approach. *Am. J. Ind. Prod. Eng.* **2025**, *6*, 10–14.
41. Banerjee, S.; Jesubalan, N.G.; Kulkarni, A.; Agarwal, A.; Rathore, A.S. Developing cyber-physical system and digital twin for smart manufacturing: Methodology and case study of continuous clarification. *J. Ind. Inf. Integr.* **2024**, *38*, 100577. [CrossRef]
42. Rossi, A.; Moretti, M.; Fravolini, M.; Senin, N. Autonomous learning of digital twins for intelligent extrusion optimisation in MEX. *J. Manuf. Process.* **2025**, *143*, 321–334. [CrossRef]
43. Lu, Y.; Asghar, M.R. Semantic communications between distributed cyber-physical systems towards collaborative automation for smart manufacturing. *J. Manuf. Syst.* **2020**, *55*, 348–359. [CrossRef]
44. Musaeva, K.; Vyachina, I.; Aliyeva, M. Smart factories and their impact on modern manufacturing enterprises: Prospects and challenges in the era of the digital economy. In Proceedings of the E3S Web of Conferences, Les Mureaux, France, 21–22 November 2024; p. 07010.
45. Jia, L.; Pei, Y. Recent advances in multi-agent reinforcement learning for intelligent automation and control of water environment systems. *Machines* **2025**, *13*, 503. [CrossRef]
46. Tzimas, E.; Papazetis, G.; Benardos, P.; Vosniakos, G.-C. Design and development of a flexible manufacturing cell controller using an open-source communication protocol for interoperability. *Machines* **2024**, *12*, 519. [CrossRef]
47. Curralo, P.M.P.; Campilho, R.D.S.G.; Pereira, J.A.P.; Silva, F.J.G. Design of connector assembly equipment for the automotive industry. *Machines* **2024**, *12*, 731. [CrossRef]
48. Rashed, A.; Abdulazeem, Y.; Farrag, T.A.; Bamaqa, A.; Almaliki, M.; Badawy, M.; Elhosseini, M.A. Toward inclusive smart cities: Sound-based vehicle diagnostics, emergency signal recognition, and beyond. *Machines* **2025**, *13*, 258. [CrossRef]
49. Paduraru, C.; Cernat, M.; Staicu, A.-N. A unified framework for automated testing of robotic process automation workflows using symbolic and concolic analysis. *Machines* **2025**, *13*, 504. [CrossRef]
50. Puri, D.; Vita, L.; Gattamelata, D.; Tulliani, V. Roll/tip-over risk analysis of agricultural self-propelled machines using airborne LiDAR data: GIS-based approach. *Machines* **2025**, *13*, 377. [CrossRef]
51. Klein Fiorentin, F.; Dantas, R.; Wolfs Gil, J.; Piga Carboni, A.; Fiorentin, T.A.; de Jesus, A.M.P. On the specimen design, physical properties and geometry effect on heat generation and thermal gradient in ultrasonic fatigue. *Machines* **2025**, *13*, 380. [CrossRef]
52. Berezin, L.; Oliinyk, O.; Rubanka, M. Innovative trends in industrial machinery engineering and education. *Actual Probl. Mod. Sci.* **2021**.
53. Luo, Y.; Wu, K.; Spielberg, A.; Foshey, M.; Rus, D.; Palacios, T.; Matusik, W. Digital fabrication of pneumatic actuators with integrated sensing by machine knitting. In Proceedings of the Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 30 April–5 May 2022; pp. 1–13.
54. Chen, X. Application of intelligent control in mechatronics system. *Eng. Adv.* **2023**, *3*, 182–185. [CrossRef]
55. Oomen, T.; Steinbuch, M. Model-based control for high-tech mechatronic systems. In *Proceedings of the Mechatronics and Robotics*; CRC Press: Boca Raton, FL, USA, 2020; pp. 51–80.
56. Hammood, A.S. Towards condition monitoring: Fabrication and finite element analysis of a helical gear transmission rig for fault simulation. *Jordan J. Mech. Ind. Eng.* **2024**, *18*, 219–226. [CrossRef]
57. Zhang, M.; Zhang, Z.; Xiong, J.; Chen, X. Accuracy analysis of complex transmission system with distributed tooth profile errors. *Machines* **2024**, *12*, 459. [CrossRef]
58. Zhang, Z.; Yang, B.; Duan, M.; Gu, R.; Liang, S.; Chen, Y. Design and experimental research of a non-destructive detection device for high-precision cylindrical roller dynamic unbalance. *Machines* **2024**, *12*, 684. [CrossRef]
59. Ahmed, S.; Ali, Q.; Sirewal, G.J.; Kumar, K.; Choi, G. Hybrid brushless wound-rotor synchronous machine with dual-mode operation for washing machine applications. *Machines* **2025**, *13*, 342. [CrossRef]
60. Vlachou, V.I.; Karakatsanis, T.S. Development of a fault-tolerant permanent magnet synchronous motor using a machine-learning algorithm for a predictive maintenance elevator. *Machines* **2025**, *13*, 427. [CrossRef]
61. Kim, T.-S.; Yang, J.-W.; Shin, K.-H.; Jang, G.-H.; Han, C.; Choi, J.-Y. Analysis of electromagnetic characteristics of outer rotor type BLDC motor based on analytical method and optimal design using NSGA-II. *Machines* **2025**, *13*, 440. [CrossRef]
62. Yakovenko, I.; Permyakov, A.; Dobrotvorskiy, S.; Basova, Y.; Kotliar, A.; Zinchenko, A. Prospects for the development of process equipment in aggregate-modular design for sustainable mechanical engineering. *Int. J. Mechatron. Appl. Mech.* **2023**, *13*, 145–156.
63. Goodfellow, H.D.; Wang, Y. *Industrial Ventilation Design Guidebook: Volume 2: Engineering Design and Applications*; Academic press: Cambridge, MA, USA, 2021.
64. Khanna, N.; Shah, P.; Sarikaya, M.; Pusavec, F. Energy consumption and ecological analysis of sustainable and conventional cutting fluid strategies in machining 15–5 PHSS. *Sustain. Mater. Technol.* **2022**, *32*, e00416. [CrossRef]
65. Babu, M.N.; Anandan, V.; Babu, M.D. Performance of ionic liquid as a lubricant in turning inconel 825 via minimum quantity lubrication method. *J. Manuf. Process.* **2021**, *64*, 793–804. [CrossRef]

66. Hagedorn, W.; Jäger, S.; Wieczorek, L.; Kronenberg, P.; Greiff, K.; Weber, S.; Röttger, A. More than recycling–The potential of the circular economy shown by a case study of the metal working industry. *J. Clean. Prod.* **2022**, *377*, 134439. [CrossRef]
67. Kokkinis, A.; Frantzis, T.; Skordis, K.; Nikolakopoulos, G.; Koustoumpardis, P. Review of automated operations in drilling and mining. *Machines* **2024**, *12*, 845. [CrossRef]
68. Buldum, B.B.; Leksycki, K.; Cagan, S.C. Comparative analysis of dry, minimum quantity lubrication, and nano-reinforced minimum quantity lubrication environments on the machining performance of AZ91D magnesium alloy. *Machines* **2025**, *13*, 430. [CrossRef]
69. Maurya, S.K.; Campatelli, G.; Veracini, M.; Arcioni, M.; Clori, D. Balancing productivity and sustainability in EDM: A comprehensive analysis of energy consumption and electrode degradation. *Machines* **2025**, *13*, 469. [CrossRef]
70. Ramírez-Jiménez, D.F.; Torres Valencia, C.A. Experimental design and simulation of a fly-cutting plant for academic environment practices. *Machines* **2025**, *13*, 15. [CrossRef]
71. Raman, V.V.; Badarinath, R.; Prabhu, V.V. TOPSIS-based methodology for selecting fused filament fabrication machines. *Machines* **2025**, *13*, 574. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

AI-Driven Virtual Sensors for Real-Time Dynamic Analysis of Mechanisms: A Feasibility Study

Davide Fabiocchi, Nicola Giulietti *, Marco Carnevale and Hermes Giberti

Dipartimento di Ingegneria Industriale e dell'Informazione, Università di Pavia, Via Adolfo Ferrata 5, 27100 Pavia, Italy; davide.fabiocchi01@universitadipavia.it (D.F.); marco.carnevale@unipv.it (M.C.); hermes.giberti@unipv.it (H.G.)

* Correspondence: nicola.giulietti@unipv.it

Abstract: The measurement of the ground forces on a real structure or mechanism in operation can be time-consuming and expensive, particularly when production cannot be halted to install sensors. In cases in which disassembling the parts of the system to accommodate sensor installation is neither feasible nor desirable, observing the structure or mechanism in operation and quickly deducing its force trends would facilitate monitoring activities in industrial processes. This opportunity is gradually becoming a reality thanks to the coupling of artificial intelligence (AI) with design techniques such as the finite element and multi-body methods. Properly trained inferential models could make it possible to study the dynamic behavior of real systems and mechanisms in operation simply by observing them in real time through a camera, and they could become valuable tools for investigation during the operation of machinery and devices without the use of additional sensors, which are difficult to use and install. In this paper, the idea presented is developed and applied to a simple mechanism for which the reaction forces during operating conditions are to be determined. This paper explores the implementation of an innovative vision-based virtual sensor that, through data-driven training, is able to emulate traditional sensing solutions for the estimation of reaction forces. The virtual sensor and relative inferential model is validated in a scenario as close to the real world as possible, taking into account interfering inputs that add to the measurement uncertainty, as in a real-world measurement scenario. The results indicate that the proposed model has great robustness and accuracy, as evidenced by the low RMSE values in predicting the reaction forces. This demonstrates the model's effectiveness in reproducing real-world scenarios, highlighting its potential in the real-time estimation of ground reaction forces in industrial settings. The success of this vision-based virtual sensor model opens new avenues for more robust, accurate, and cost-effective solutions for force estimation, addressing the challenges of uncertainty and the limitations of physical sensor deployment.

Keywords: virtual sensors; measurement uncertainties; artificial intelligence; multi-body; dynamic analysis

1. Introduction

In recent years, the Internet of Things (IoT) has paved the way for new methods of obtaining and processing large amounts of data, playing a crucial role in a variety of applications, such as mobility, industrial automation, and smart manufacturing. The necessity for a large number of sensors to monitor crucial production variables in order to ensure consistent product quality and optimize energy consumption [1] has caused the emergence of virtual sensors, software-based models that emulate the functioning of physical sensors. The main advantage of virtual sensors is related to their cost-effectiveness, gaining an advantage especially in those cases where the installation and maintenance of physical sensors is difficult and expensive [2,3]. The flexibility of virtual sensors, which are easily adaptable to different process conditions and environments, makes them easy to integrate into industrial settings with different aims, ranging from predictive maintenance [4] to

process optimization [5]. Generally, soft sensors are divided into two categories: those based on mathematical models and those based on data-driven methods. The foundational concept of virtual sensors originates from the utilization of mathematical models to estimate critical quantities, without direct measurement. Hu et al. extrapolated the force exerted at the catheter tip within the human body with a mathematical model, exploiting a sensor positioned on the exterior of the catheter tube [6]. In recent years, with the advent of new technologies that provide enhanced data storage capabilities, data-driven methods have gained significant attention thanks to the development of new machine learning algorithms [7,8]. This innovative data-driven approach offers a new, effective means of obtaining accurate readings along the entire processing chain, without installing expensive hardware, providing an effective tool for the online estimation of quantities that are important but difficult to measure with traditional sensors. A clear example of a data-driven virtual sensor is presented by Sabanovic et al. in their work, where different architectures of artificial neural networks are exploited to estimate vertical acceleration in vehicle suspension [9]. Leveraging technological advancements in computational power and increasingly efficient algorithms, recent developments have paved the way for the use of vision-based virtual sensors (i.e., based on image analysis). Indeed, the use of cameras, even across various spectral ranges, allows for real-time information gathering in a non-invasive manner, enabling installation and data acquisition even on pre-existing structures that were not designed for the use of physical sensors. Moreover, the extreme versatility of these systems allows for their use remotely or in dangerous or hard-to-reach situations (i.e., through the use of robots or unmanned aerial vehicles (UAVs) [10]). Byun et al. introduce a methodology for the use of image-based virtual vibration measurement sensors to monitor vibration in structures, offering an alternative to traditional accelerometers [11]. Wang et al. introduce a model-based approach for the design of virtual binocular vision systems using a single camera and mirrors to improve the 3D haptic perception in robotics [12]. Ögren et al. develop vision-based virtual sensors with the aim of estimating the equivalence ratio and concentration of major species in biomass gasification reactors, using image processing techniques and regression models on real-time light reaction zone image data, demonstrating the applicability and effectiveness of vision-based monitoring for process control in a complex industrial setting [13]. Alarcon et al. discuss the integration of Industry 4.0 technologies into fermentation processes by implementing complex culture conditions that traditional technologies cannot achieve. In this context, computer vision techniques are exploited to develop a virtual sensor to detect the end of the growth phase and a supervisory system to monitor and control the process remotely [14]. There are many cases found in the literature related to the use of virtual sensors to measure forces. Physics-informed neural networks (PINNs) are used to estimate equivalent forces and calculate full-field structural responses, demonstrating high accuracy under various loading conditions and offering a promising tool for structural health monitoring [15]. Marban et al. introduce a model based on a recurrent and convolutional neural network for the sensor-less force measurement of the interaction forces in a robotic surgical application. Using video sequences and surgical instrument data, the model estimates the forces applied during surgical tasks, improving the haptic feedback in minimally invasive robot-assisted surgery [16]. Ko et al. developed a vision-based system to estimate the interaction forces between the robot grip and objects by combining RGB-D images, robot positions, and motor currents. By incorporating proprioceptive feedback with visual data, the proposed model achieves high accuracy in force estimation [17]. In the context of smart manufacturing, Chen et al. develop a real-time milling force monitoring system, using sensory data to accurately estimate the forces involved in the process, thus enabling real-time adjustment to optimize the cutting operation [18]. Bakhshandeh et al. propose a digital-twin-assisted system for machining process monitoring and control. Virtual models, integrated with real-time sensor data, are used to measure the cutting forces, enabling adaptive control, anomaly detection, and precision applications without physical sensors [19].

1.1. Novel Work and Motivation

In this research, we intend to demonstrate that properly trained inferential models have the potential to revolutionize the study of the dynamic behavior of real systems and mechanisms. By exploiting virtual sensing techniques, we intend to observe these systems in operation in real time through a camera, eliminating the need for additional, often bulky and difficult-to-install sensors. Our goal is to validate this approach using a simple mechanism as a test case, illustrating how these models can become valuable tools in studying the operation of machinery and devices with ease, efficiency, and robustness. This study is intended to explore the feasibility and effectiveness of inferential models in capturing and analyzing the complex dynamics of mechanical systems, paving the way for their widespread application in industry. Even in the case in which accurate multi-body models or digital twins of the mechanism are available, delivering real-time input to these models can be challenging. Simulating the actual scenario for a multi-body model requires knowledge of the actual motion of law and demands significant computational effort, especially in the case of very complex models. The measurement of the motion law is easy in the case of synchronous motors since their speed is correlated to the frequency of the AC power supply and the number of poles of the motor itself. However, in an industrial context, asynchronous motors are frequently used, and the measurement of their speed is not straightforward. External sensors such as encoders become necessary, demanding both excessive time and high costs for their installation. Our research investigates the implementation of an innovative vision-based virtual sensor that, through data-driven training, is able to emulate traditional sensing solutions for the estimation of reaction forces. The implemented virtual sensor and related multi-layer perceptron (MLP) architecture are trained and tested using a simple mechanism, exploiting a multi-body model. The first model is trained with the ideal inputs and outputs, while the second is instead trained on a dataset that takes into account the uncertainty in the measurement of the input quantities (i.e., closely replicating a real-world scenario). The developed models are tested with new and unobserved trajectories to further assess their effectiveness and robustness.

The motivation behind our study lies in the challenge of studying the dynamics of operational machines that lack installed sensors. In many cases, halting production for the installation of sensors is impractical and time-consuming, despite the potential benefits in terms of analysis, control, and overall system reliability. The interest in exploring non-invasive alternatives, to obtain accurate estimations of machine behavior without disrupting ongoing operations, drove us to implement a specific virtual sensor for the estimation of ground reaction forces in industrial machines.

The concept revolves around exploiting verified multi-body models of machines to train AI-based virtual sensors. These sensors can then be employed to gather real-time information by processing data from cameras, eliminating the need for physical force sensors. Our aim is to develop a robust method that enables the generation of virtual sensors capable of providing valuable insights into the dynamics of existing operational machines. Our approach is well suited for the real-time estimation of forces and other relevant values and it accounts for the variability introduced by various factors, such as the camera positioning, significantly enhancing the reliability and robustness of the sensor. In summary, our study paves the way for the establishment of a robust framework for the creation of virtual sensors that can monitor and gauge the condition of moving machines solely by visually observing them.

1.2. Contribution

The contribution of our work to the existing body of knowledge lies in the development and application of a vision-based virtual sensor, in the context of industrial processes and mechanical systems. The main key points of our study are the following.

- This paper introduces a non-invasive virtual sensor as an alternative to traditional sensing systems, using cameras and leveraging data-driven inferential models to measure the forces involved in a mechanism.

- The exploitation of the proposed vision-based virtual sensor is an efficient solution in reducing the need for external sensors like force transducers or encoders, whose installation is often time-consuming and expensive.
- The virtual sensor has been developed also considering datasets with uncertainties, thus assessing the robustness of the overall measurement system to real-world disturbances.
- This study demonstrates the adaptability of the proposed solution in capturing and analyzing the dynamics of mechanical systems for real-time solutions.

1.3. Organization of the Paper

The article is structured as follows. Section 2 details the design of the multi-body model, the camera model used to simulate frame acquisition, the methodology used for data collection, and the implementation of an AI-based virtual sensor to estimate ground reaction forces. The results of the training of two different models are presented in Section 3. The models thus trained are finally compared with the results that would have been obtained through the use of the multi-body model.

2. Materials and Methods

To first assess the feasibility of the method, the implementation of a vision-based virtual sensor to estimate the reaction forces in industrial mechanisms is applied to a simple but fundamental mechanical system, the four-bar linkage mechanism, with a verified multi-body model (Figure 1). This system is chosen for its basic but significant role in the study of mechanical systems, providing a clear framework to test our hypotheses. The approach focuses on the development of an AI-based virtual sensor designed to estimate the reaction forces (constraint reactions) within the mechanism based solely on visual observation. This method is similar to extracting hypothetical data from a camera monitoring the moving mechanism, mimicking real-world scenarios where direct force measurement is not feasible. The main objective is to demonstrate the practicality of using AI-based algorithms to extract reaction forces in real time. The implemented AI-based virtual sensor is responsible for estimating the ground reaction forces of the mechanism. It processes the mechanism position data, which serve as the input, enabling real-time analysis and estimation. This approach is critical in replicating and predicting the forces in operational industrial environments, showing the potential of virtual sensors in optimizing the performance, monitoring the operation, and ensuring the safety of mechanical systems in a real industrial scenario. In this section, we will describe the multi-body model of the mechanism used to build our training dataset; then, the model of the camera used in the software to replicate the monitoring of the mechanism is presented. This camera model will be replicated with an analytical model that will be used to generate different datasets taking into account various sources of uncertainty, which will be described in the section related to data collection. Finally, the design of the AI-based virtual sensor and its architecture will be presented in the last subsection.

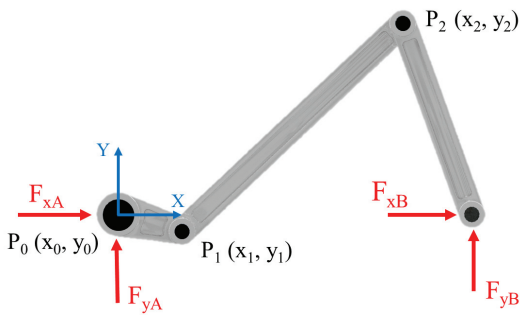


Figure 1. CAD model of the four-bar linkage mechanism. Ground reaction forces F_{x_A}, F_{y_A}, F_{x_B}, F_{y_B} are represented in red. The crank center of rotation is identified by point $P_0(x_0, y_0)$, where a fixed reference frame represented in blue is indicated to evaluate the position of the joints identified by points $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$.

2.1. Multi-Body Model

The 3D model of the mechanism, designed in Autodesk Inventor 2024, uploaded on the multi-body software RecurDyn 24, is shown in Figure 1. In the multi-body software, the definition of joints has to be meticulously performed, by taking into account the degrees of freedom (DoF) of each body in the space and avoiding the redundancy of the constraints. The objective is to achieve a mechanism with a single DoF, specifically the rotation of the crank. The two grounded joints are defined as "revolute" joints, restricting all translations and rotations of the links, except for the rotation within the plane of the image. The joint between the crank and the connecting rod is defined as the "universal" joint, permitting two rotations, one of which is in the plane of the image. To prevent redundancy in the constraints, the joint between the connecting rod and the rocker is defined as a "spherical" joint, constraining only the three relative translations. The resulting system has one degree of freedom (DoF), corresponding to the rotation of the crank. As shown in Figure 1, the point corresponding to the crank center of rotation is labeled as P_0, the point P_1 identifies the joint between the crank and the connecting rod, while point P_2 is used to describe the position of the joint between the connecting rod and the rocker.

2.2. Camera Model

Autodesk Inventor 2023 has been utilized to simulate a real-life sensor and to create dynamic videos of the mechanism for further analysis through computer vision. To incorporate a camera within the Inventor Studio environment, it is necessary to specify both the target point position and the spatial coordinates of the camera. The orientation of the camera aligns with the line connecting it to the target point, allowing us to configure the camera precisely. We constrain the target position to the same plane where the mechanism exists, limiting the range of its coordinates to two, while the camera's three spatial coordinates and roll angle remain adjustable. The type of camera employed in the software is an orthographic projection rather than a perspective camera, with the horizontal angular field of view (AFOV) and resolution as customizable parameters. In the specific case, an AFOV of 28° is fixed and the resolution of the captured image is set equal to 2240 × 1200, as shown in Figure 2. Once the intrinsic parameters of the camera have been chosen, the adjustable parameters related to its position and orientation have to be set. In Table 1, the definitions of these parameters and their values used to identify the standard configuration of the sensor are reported.

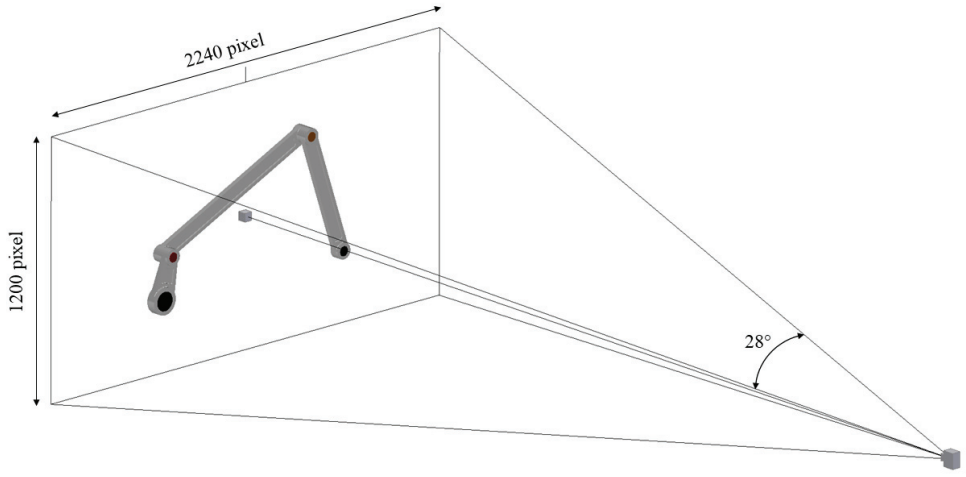


Figure 2. Four-bar linkage mechanism in Inventor Studio and definition of camera.

Table 1. Parameters of standard configuration of camera in Inventor Studio.

	Parameter	Value
X_c	Camera X Coordinate	45.55 mm
Y_c	Camera Y Coordinate	−3041.43 mm
Z_c	Camera Z Coordinate	80.14 mm
θ	Camera Roll Angle	0 deg
X_t	Target X Coordinate	45.55 mm
Y_t	Target Y Coordinate	15 mm
Z_t	Target Z Coordinate	80.14 mm

To bypass the need for Inventor Studio and the extensive rendering time required for each video, an analytical model of the camera has been developed. This model is an effective tool, accepting as input the values of the adjustable parameters outlined in Table 1 and the spatial position of a point. It outputs the corresponding location of this point, in pixels, within the frame captured by the camera. In Figure 3, an example of the application of this model is represented: a point p (green) in space is defined by its coordinates with the respect to the world frame (XYZ), while the positions of the camera and the target are defined, respectively, by the coordinates (X_c, Y_c, Z_c) and (X_t, Y_t, Z_t). The orientation of the camera is defined by the line connecting the camera and the target, and the image plane is orthogonal to this line and centered in the target. The point p is orthogonally projected on the image plane, resulting in the point P (red). Finally, its coordinates u_0 and v_0 (expressed in mm) with respect to the image reference frame (UV) are computed.

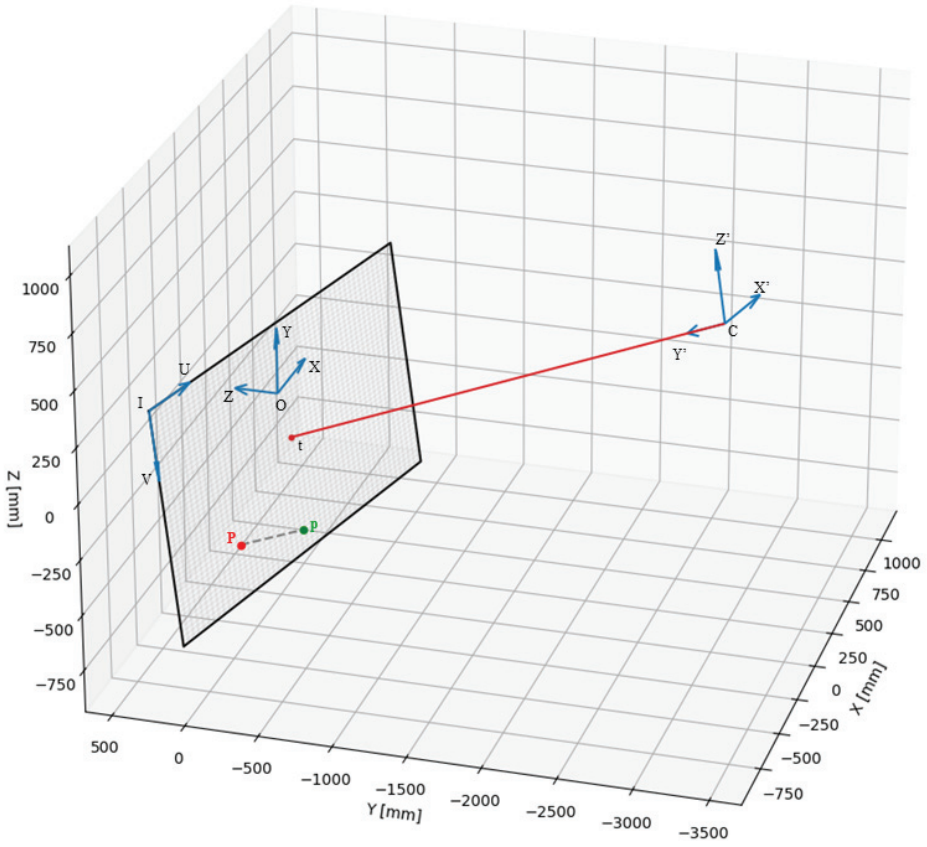


Figure 3. Example of application of analytical model of camera. The world reference frame (XYZ) centered at O, the position of the camera C and its frame (XYZ), and the position of the target t are presented. The point p defined in the world reference frame is projected onto the plane image and defined in the image reference frame (UV) in pixels.

The model has been tested and validated under standard camera conditions, with its results benchmarked against those obtained from the Autodesk Inventor camera. In anticipation of exploring the impact of uncertainties on the AI-based virtual sensor, the analytical model has been adapted to accept deviations in the camera's customizable parameters from their standard settings, as well as the coordinates of any point in space. This enhanced model facilitates the generation of varied datasets to account for uncertainties related to the camera's position, thus significantly reducing the time required to produce highly consistent datasets. Our research focuses on replicating the camera model used by Autodesk Inventor, given its focus on a simulated model. However, it is crucial to acknowledge that, within real-world applications, a variety of camera models could be implemented, each offering different perspectives and characteristics.

2.3. Data Collection

Data to train the model are sourced from simulations conducted on the multi-body model within the RecurDyn environment, which facilitates the simulation of the mechanical system's dynamics. Following the workflow shown in Figure 4, a known motion is imposed on the mechanism, and it is observed by the camera sensor modeled in Section 2.2.

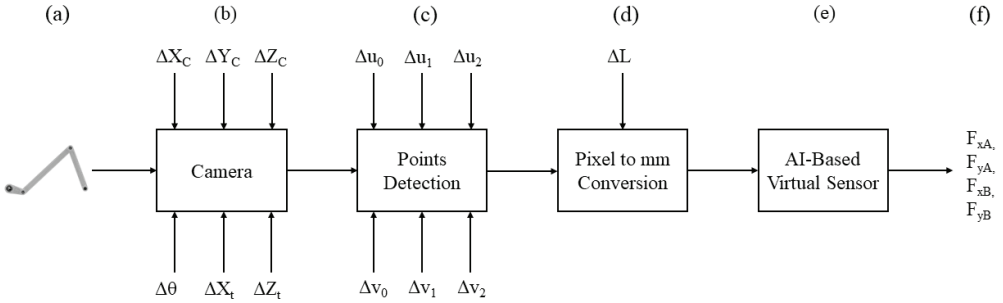


Figure 4. Workflow of the reaction force estimation procedure: the four-bar linkage mechanism (**a**) motion is captured by a sensor camera (**b**); images are processed by a computer vision algorithm (**c**), and position of significant points are computed in mm (**d**); positions data are fed to the AI-based virtual sensor (**e**) that estimates ground reaction forces of the mechanism (**f**).

Thus, mechanism (a) is observed from camera sensor (b), the position of which is defined by $X_c, Y_c, Z_c, X_t, Z_t, \theta$. From the captured frame, points $P_0(u_0, v_0)$, $P_1(u_1, v_1)$ and $P_2(u_2, v_2)$, expressed in pixels, are identified by computer vision algorithms (c), based on HSV color segmentation. The image is then calibrated based on known dimensions (i.e., crank length L) and such points' positions are converted from pixels to millimeters and referenced to the reference system centered at P_0 (d), as shown in Figure 1. The positions of the points $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$ expressed in millimeters are then fed to the model (e) such that the constraining reactions F_{x_A}, F_{y_A}, F_{x_B}, F_{y_B} (f) are predicted. The use of vision-based measurement systems introduces sources of uncertainty that could alter the input data given to the model: the positioning of the camera might be different with respect to the standard configuration; therefore, the uncertainty related to the sensor position is defined as the deviation from the reference position ($\Delta X_c, \Delta Y_c, \Delta Z_c, \Delta X_t, \Delta Z_t, \Delta\theta$). The detection of the three points' coordinates, performed by the computer vision algorithm, is characterized by some noise related to the position of the pixel in the frame, corresponding to the detected point. Moreover, in this case, the uncertainty is defined as the deviation from the real position of the three centroids ($\Delta u_1, \Delta u_2, \Delta u_3, \Delta v_1, \Delta v_2, \Delta v_3$). Finally, the conversion from pixels to mm, implemented in the algorithm, is influenced by the measurement of the crank length, which may suffer from some error (ΔL). Therefore, to generate datasets that take into account the presence of these sources of uncertainty, the analytical model of the camera can be used to generate data related to a large variety of camera configurations. The output of the analytical model, which represent the positions of $P_0(u_0, v_0)$, $P_1(u_1, v_1)$, and $P_2(u_2, v_2)$, can be further rearranged, adding artificial noise (in pixels) on the six coordinates,

for each frame, simulating the computer vision algorithm. Finally, the conversion from pixels to mm can be carried out by taking into account the error in the length measurement.

2.4. AI-Based Virtual Sensor for Ground Reaction Force Estimation

A multi-layer perceptron (MLP) regression model is built in order to estimate the ground reaction forces using the linkage joints' position coordinates, extrapolated from the camera data. The use of MLP networks can be remarkably effective, despite their simplicity, even when compared to more complex deep neural network (DNN) models, such as convolutional neural networks (CNNs) [20]. The advantage of the MLP lies in its utilization of lightweight and fast models, typically around the order of 0.1 Mb, offering a stark contrast to the often much heavier advanced DNN models. This efficiency makes MLP an appealing choice in various scenarios where a balance between performance and resource utilization is crucial. The purpose of the model is to predict, at each instant t_i, the reaction forces exchanged on the ground by the articulated quadrilateral (Figure 4a). Since the time history of the input influences the value of the output, the model receives as input a vector of shape (20, 1), consisting of the 2D coordinates of the two points $P_1(t_i)$ and $P_2(t_i)$ of the mechanism from instant t_{i-4} to t_i, measured as explained in Section 2.3. As output, it returns a vector of shape (4, 1), consisting of the two components of the ground reaction forces (Figure 5).

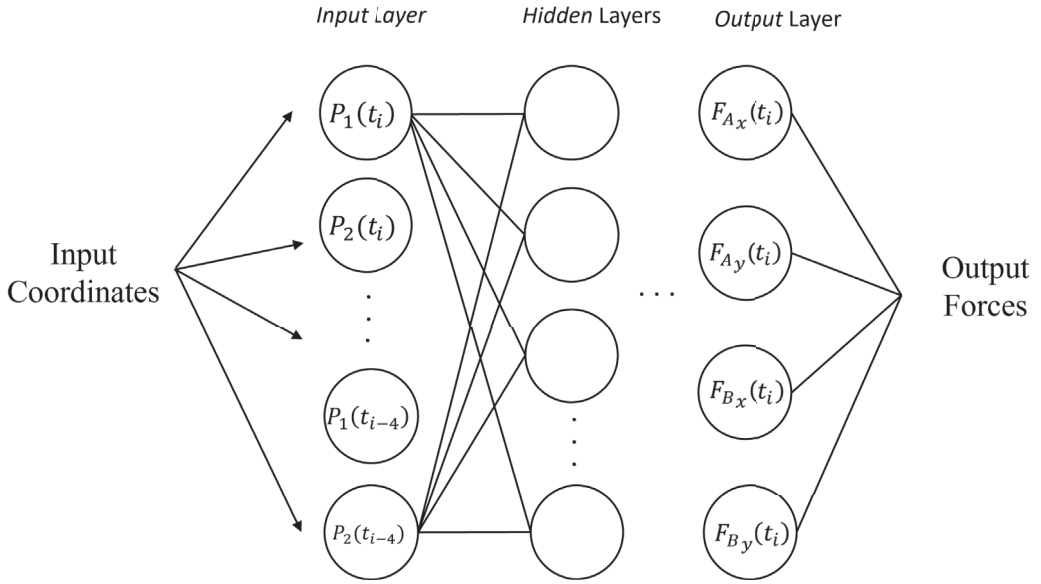


Figure 5. The proposed MLP architecture takes as input two points $P_1(t_i)$ and $P_2(t_i)$ of the mechanism from instant t_{i-4} to t_i, and it returns as output (i.e., predicts) the two components of the ground reaction forces F_A and F_B.

The input and output data are normalized to enhance the convergence and the stability of the neural network during the training. Each column of both the input and the output datasets is independently normalized, by scaling the values within each column to the range $[0, 1]$. The normalization formula is given by

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

where X is the original value, X_{min} is the minimum value in the column, and X_{max} is the maximum value in the column. A Bayesian optimization technique for hyperparameter tuning, as detailed in [21], is utilized to determine the optimal numbers of layers, neurons per layer, and model hyperparameters that minimize the loss function. This technique of hyperparameter optimization is widely tested in various fields and is proven to be superior to other techniques, such as grid search and random search [22,23]. This technique

performs training many times with different sets of numbers of layers, neurons, and hyperparameters. The method initially tests the function n times by assigning random values to its variables in order to diversify the exploration space. Then, based on the values obtained in the previous step, the actual optimization begins. At each step, the algorithm evaluates the past model information to select new optimal parameter values that increase the performance of the new model according to the described method [24,25]. The initial phase involves defining the function that requires optimization. Considering the problem at hand, the target function represents the training procedure of the neural network, accepting multiple parameters for input and yielding the mean absolute error (MAE) as its result. The optimizer's exact goal is to reduce this coefficient value to its minimum. The MAE, chosen as the loss function, is expressed in Equation (3), where n is the number of observations, y_i is the ground truth input, and $\hat{y}_i$ is the predicted value. The function to be optimized is defined in Equation (2), where x and y are the input and ground truth batches of the network in the training process, respectively.

$$\text{MAE} = f(x, y, bs, lr, o, nl, nn) \tag{2}$$

After selecting the function for optimization, it is essential to identify the specific input parameters that need refinement and establish their limits, given that the approach is one of bounded optimization. The parameters and their ranges, selected based on prior knowledge and preliminary experimental data, are listed in Table 2, with the following.

- Batch size bs: specifies the number of training samples processed in one iteration.
- Learning rate lr: determines the rate at which the model weights are updated during training.
- Optimizer o: updates the model based on the loss function. Options include SGD, Adam, RMSprop, Adadelta, and Adagrad [26].
- Number of layers nl: specifies the total number of layers in the network.
- Number of neurons per layer nn: specifies the number of neurons in each layer.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

The number of training epochs is set at 1000, with an early stopping level of 50 for the validation loss [27]. Each iteration is cross-validated by the K-fold method with three folds; the result of each iteration is expressed as the average MAE of the three iterations minus the standard deviation [28].

Table 2. Hyperparameters involved in the Bayesian optimization technique and their corresponding bounds.

		From	**To**
bs	Batch size	16	4096
lr	Learning rate	1×10^{-6}	0.1
nl	Number of layers	1	100
nn	Neurons per layer	10	400

3. Results

With the work presented, we have shown that properly trained inferential models can be used to effectively study the dynamic behavior of real systems and mechanisms in operation simply by observing them in real time through a camera. This approach has been validated by applying it to a simple mechanism, demonstrating its potential as a valuable tool for investigation during the operation of machinery and devices, without the need for additional sensors, which are often difficult to use and install. Our results confirm that such models are not only feasible but also effective in capturing and analyzing complex dynamic interactions in mechanical systems. The methodologies described in the previous section are used to train a model for the prediction of the reaction forces in a

simple mechanism (Section 3.1). Two separate models are trained. The first one, labeled as Model 1, is trained by using the dataset coming from the multi-body simulation, which does not take into account the presence of uncertainties. A second dataset, which takes into account all sources of uncertainty described in Section 2.3, is generated by using the camera model described in Section 2.2, and it is used to train the second model, labeled as Model 2. Both models are subsequently tested on motion laws not previously encountered, to assess their capability in predicting the forces in every scenario (Section 3.2). Finally, a comparison between a virtual sensor based on the multi-body model and the one based on the inferential model is performed with an emphasis on their ability to respond to different interfering inputs (i.e., noise) (Section 3.3).

3.1. Model Training

Following the methodology described in Section 2.3, a known motion law is imposed on the mechanism in order to use the collected data as a training dataset for the virtual sensor model. As shown in Figure 6, the motion law comprises a combination of several sinusoidal functions with different amplitudes and frequencies, with the objective of introducing a degree of randomness in the dynamics of the mechanism.

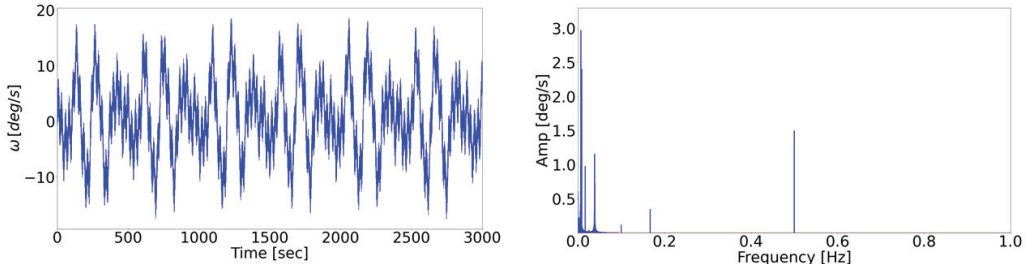


Figure 6. From left to right: crank rotational speed motion law imposed on the mechanism to build the training dataset and its FFT transform.

The camera sensor records the scene for 3000 s at 60 frames per second (fps) for a total of 180,000 samples. For each sample, the reaction forces are known. Following the procedure described in Section 2.3, the positions of points P_1 and P_2 are calculated for each frame. Two training datasets are created with the collected data.

- Dataset 1:
  - Uses ideal data from the simulation. This dataset is designed to represent the best-case scenario without any external interference or noise, serving as a benchmark for optimal model performance.
- Dataset 2:
  - Accounts for possible interfering inputs (e.g., noise) within the measurement chain, simulating a real use case. The incorporation of such disturbances aims to mimic the challenges encountered in real case scenarios.

The use of vision-based measurement systems introduces sources of uncertainty that could alter the input data given to the model. These sources of uncertainty are considered during dataset generation to enhance the inferential model's robustness. As reported in Figure 4, thirteen sources of uncertainty are identified, relating to the six DoFs of the camera positioning ($X_c, Y_c, Z_c, \theta, X_t, Z_t$), six uncertainties associated with the coordinates identifying the three points plotted by the computer vision algorithm ($u_1, v_1, u_2, v_2, u_3, v_3$), and the extent of the crank length (L) used for the pixel-to-millimeter conversion of these coordinates. To introduce the effects of these uncertainties into the dataset generation process, a Gaussian normal distribution is imposed for each of the thirteen sources of uncertainty, specifying the mean and standard deviation of the error introduced, around a standard case. Table 3 shows the value distribution for each of the thirteen sources of uncertainty. Only the input dataset has been manipulated, while the output remains unchanged.

Following the data pre-processing steps explained in Section 2.3, the final dataset consists of an input file containing 179,996 entries (the overall dataset consists of 180,000 time instants, but the first batch consists of five consecutive acquisitions; from this, it follows that the first useful batch starts at the fifth acquisition). The corresponding output file consists of the same number of entries, containing the values of the four reaction forces. Dataset 1 and Dataset 2 are used to train two models (i.e., Model 1 and Model 2) following the hyperparameter training and optimization procedure explained in Section 2.4 with $n = 1000$ (i.e., the number of optimization iterations). At the end of the optimization process, Model 1 and Model 2 reach an MAE of 0.004 and 0.005, respectively, with the hyperparameters shown in Table 4.

Table 3. The table shows the limits within which the various sources of uncertainty are made to vary uniformly for the generation of Dataset 2.

		Mean	**Std Dev**
$\Delta X_c, \Delta Y_c, \Delta Z_c$	Camera Position	0 mm	1000 mm
$\Delta\theta$	Camera Roll Angle	0 deg	5 deg
$\Delta X_t, \Delta Z_t$	Target Position	0 mm	200 mm
$\Delta u_1, \Delta u_2, \Delta u_3, \Delta v_1, \Delta v_2, \Delta v_3$	Detected Centroid Position	0 pixel	2 pixel
ΔL	Crank Length	0 mm	0.5 mm

Table 4. Selected hyperparameters for cost function minimization (i.e., MAE) for Model 1 and Model 2 training.

		Model 1	**Model 2**
o	Optimizer	Adadelta	Adam
bs	Batch size	134	3127
lr	Learning rate	0.0541	0.00032
nl	Number of layers	5	10
nn	Neurons per layer	116	88
MAE	Metric	0.00386	0.00554

3.2. Model Testing

The testing procedure has been carried out by subjecting Model 1 and Model 2 to an input dataset, firstly generated by simulating a 15 s scenario in RecurDyn, keeping the same number of steps per second (i.e., 60 fps), and then altered taking into account the uncertainties described in Section 2.3. This dataset is henceforth referred to as Dataset 3. As shown in Figure 7, the motion law used for the simulation on Dataset 3 differs significantly from the one used during the testing procedure, comprising step functions available in the multi-body software. The introduction of a new motion law aims to ensure that Model 1 and Model 2 are indeed able to generalize to a new scenario without overfitting on the training dataset.

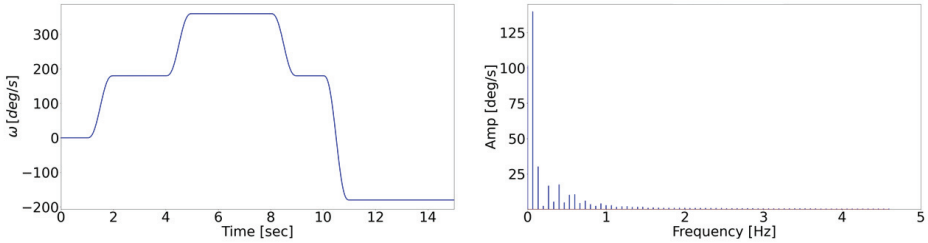


Figure 7. From left to right: the crank rotational speed motion law imposed on the mechanism to build the testing dataset and its FFT transform.

The uncertainties discussed in Section 2.3 are taken into account in Dataset 3 in order to introduce a realistic level of noise in the acquired data, by sampling from their respective

Gaussian distributions. Table 5 shows the sampled values for each source of uncertainty used in the testing procedure. The corresponding reaction forces predicted by Model 1 and Model 2 have been compared with the solution coming from the multi-body simulation (Reference—Figures 8 and 9): the root mean squared error (RMSE) is computed for each reaction force, providing a quantitative measure to assess the quality of the models under the testing conditions (i.e., Dataset 3). Complete results are reported in Table 6.

Table 5. Sampled uncertainty values for the generation of the dataset used for the comparison.

ΔX_c [mm]	ΔY_c [mm]	ΔZ_c [mm]	ΔX_t [mm]	ΔZ_t [mm]	$\Delta\theta$ [°]	ΔL [mm]
−809.8	−388.9	915.1	76.8	−281.3	4.6	−0.34

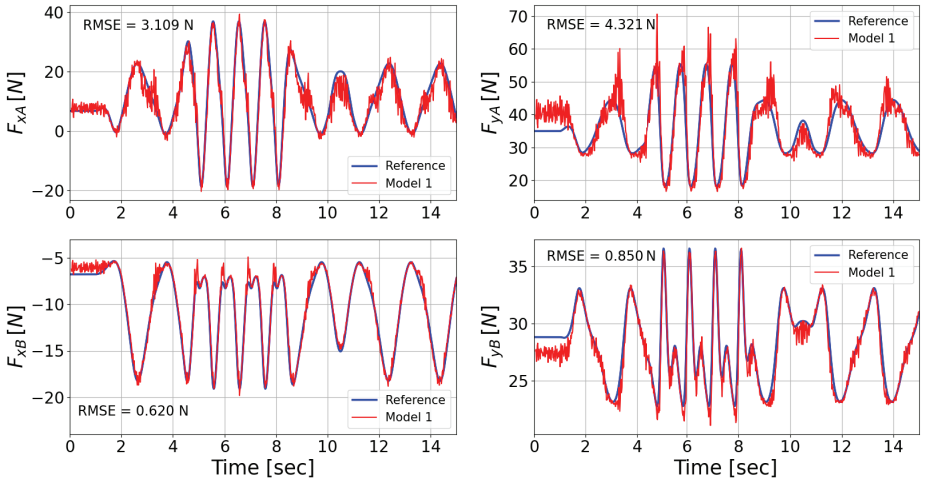


Figure 8. Prediction of the four reaction forces performed by the model trained with the unmodified dataset, with respective RMSEs.

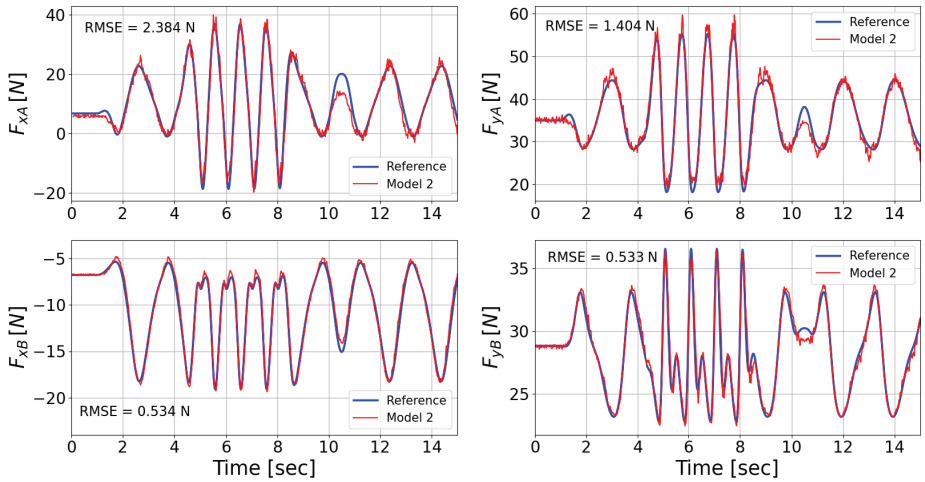


Figure 9. Prediction of the four reaction forces performed by the model trained with the modified dataset, with respective RMSEs.

Table 6. Comparison of the quality of the estimation of the two models on each of the four reaction forces.

RMSE	F_{xA} [N]	F_{yA} [N]	F_{xB} [N]	F_{yB} [N]
Model 1	3.109	4.321	0.620	0.850
Model 2	2.384	1.404	0.534	0.533

The obtained results highlight the effectiveness of the second model in mitigating the noise introduced by the different sources of uncertainty. The higher performance of the second model is reflected in the lower RMSE obtained for each of the four reaction force estimations. Figures 8 and 9 show the differences between the two models in predicting all four reaction forces exchanged between the mechanism and the ground. While both Model 1 and Model 2 show their ability to predict the behavior of the reaction forces, the second model offers higher precision.

3.3. Multi-Body vs. Inferential Model

The direct use of the multi-body model as a virtual sensor, instead of the implemented AI-based sensor, is possible but challenging, since it would required the imposition of the motion law as input. In the case of synchronous motors, the speed of the actuated joint is directly determined by the frequency of the AC power supply, the number of poles in the motor, and the transmission ratio. On the other hand, in asynchronous motors, which are widely applied in industrial settings, the speed is not directly determined, and external measures such as encoders or tachometers are often required. To avoid the use of external physical sensors, which have large time and cost requirements for installation and maintenance, the already proposed vision-based system can be used, tracking the crank joints' positions over time to compute the angular values. However, the already presented sources of uncertainty may introduce noise in the evaluation of the motion law, especially in the speed and acceleration computations, derived numerically from the angular positions. This noise is reflected in the output of the multi-body model. A comparison between the virtual sensor based on the multi-body model and the AI-based one has been carried out to assess their robustness. The evaluation has been carried out by using the same dataset used for the previous evaluations, with the same sampled sources of uncertainty reported in Table 5. The multi-body virtual sensor requires a filter to be implemented in order to smooth the behavior of the numerical derivatives of the angular position, while the AI-based sensor receives only the data relative to the positions of the two joints. A real-time low-pass filter has been implemented to filter the numerical derivatives of the motion law. The filtered motion laws have been applied to the multi-body model, and the resulting outputs are presented in Figure 10, where a comparison between the latter model and Model 2, in terms of the R^2 score, is reported.

The presented result indicates good agreement between the prediction of the reaction forces performed by the AI-based virtual sensor and the one performed by the multi-body model. Both methods show high reliability in the estimation, reflected in the high values of the R^2 score. The high reliability of the obtained result, together with the low computational effort required, makes the implemented AI-based virtual sensor a powerful solution for this type of application.

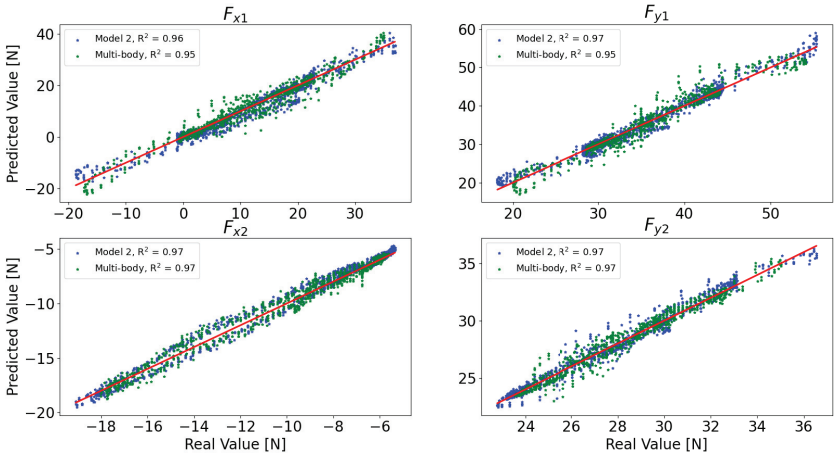


Figure 10. Comparison of the results obtained by Model 2 and the multi-body model receiving motion law data from the computer vision algorithm, with real-time filtering.

4. Conclusions

In this research, we implemented and studied the feasibility of an innovative virtual sensor with the aim of studying the dynamics of operating machines without integrated sensors. This can be of vital importance in all cases where stopping production to install sensors would be impractical and time-consuming, despite the potential benefits in terms of analysis, control, and overall system reliability. As a test case, the study focused on a four-bar linkage mechanism, developing an AI-based virtual sensor for ground reaction force estimation based on visual observation. Two multi-layer perceptron regression models were trained using data from a multi-body simulation. The second model took into account the measurement uncertainties of the input quantities (i.e., the measurement uncertainties of the vision-based system).

The research showcased the effectiveness of the AI-based virtual sensor in estimating the ground reaction forces, even in the presence of significant uncertainties introduced by the vision-based system (i.e., camera position, computer vision algorithm, and crank length measurements). Furthermore, the study compared the AI-based virtual sensor with a multi-body based virtual sensor, both exploiting the vision-based system output as the input. The results indicated the ability of the AI-based proposed approach to filter out uncertainties and provide real-time estimation with low computational effort.

In conclusion, the developed virtual sensor presents a promising solution for the estimation of the reaction forces in the proposed mechanism, offering a non-invasive and cost-effective solution. Further developments would require the experimental validation of the proposed methodology on a real mechanism exploiting real camera sensors.

Author Contributions: Conceptualization, H.G. and N.G.; methodology, N.G.; software, D.F. and N.G.; validation, M.C. and N.G.; formal analysis, H.G. and N.G.; investigation, N.G.; data curation, D.F.; writing—original draft preparation, D.F. and N.G.; writing—review and editing, N.G. and M.C.; supervision, H.G.; project administration, H.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data are not publicly available.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Souza, F.A.; Araújo, R.; Mendes, J. Review of soft sensor methods for regression applications. *Chemom. Intell. Lab. Syst.* **2016**, *152*, 69–79. [CrossRef]
2. Calvanese, L.; Carnevale, M.; Facchinetti, A. Fiber optic system to extend the measuring frequency range of pantograph contact force. In Proceedings of the 2022 Joint Rail Conference, JRC 2022, Virtual, 20–21 April 2022. [CrossRef]
3. Carnevale, M.; Collina, A.; Palmiotto, M. Condition monitoring of railway overhead lines: Correlation between geometrical parameters and performance parameters. In Proceedings of the 1st World Congress on Condition Monitoring 2017, WCCM 2017, London, UK, 13–16 June 2017 .
4. Peinado-Asensi, I.; Montés, N.; García, E. Virtual Sensor of Gravity Centres for Real-Time Condition Monitoring of an Industrial Stamping Press in the Automotive Industry. *Sensors* **2023**, *23*, 6569. [CrossRef] [PubMed]
5. Tang, Z.; Zhang, Z. The multi-objective optimization of combustion system operations based on deep data-driven models. *Energy* **2019**, *182*, 37–47. [CrossRef]
6. Hu, X.; Cao, L.; Luo, Y.; Chen, A.; Zhang, E.; Zhang, W.J. A Novel Methodology for Comprehensive Modeling of the Kinetic Behavior of Steerable Catheters. *IEEE/ASME Trans. Mechatron.* **2019**, *24*, 1785–1797. [CrossRef]
7. Etxegarai, M.; Camps, M.; Echeverria, L.; Ribalta, M.; Bonada, F.; Domingo, X. Virtual Sensors for Smart Data Generation and Processing in AI-Driven Industrial Applications. In *Industry 4.0—Perspectives and Applications*; IntechOpen: London, UK, 2022.
8. Sun, Q.; Ge, Z. A Survey on Deep Learning for Data-Driven Soft Sensors. *IEEE Trans. Ind. Inform.* **2021**, *17*, 5853–5866. [CrossRef]
9. Sabanovic, E.; Kojis, P.; Ivanov, V.; Dhaens, M.; Skrickij, V. Development and Evaluation of Artificial Neural Networks for Real-World Data-Driven Virtual Sensors in Vehicle Suspension. *IEEE Access* **2024**, *12*, 13183–13195. [CrossRef]
10. Giulietti, N.; Allevi, G.; Castellini, P.; Garinei, A.; Martarelli, M. Rivers' Water Level Assessment Using UAV Photogrammetry and RANSAC Method and the Analysis of Sensitivity to Uncertainty Sources. *Sensors* **2022**, *22*, 5319. [CrossRef] [PubMed]
11. Byun, E.; Lee, J. Vision-based virtual vibration sensor using error calibration convolutional neural network with signal augmentation. *Mech. Syst. Signal Process.* **2023**, *200*, 110607. [CrossRef]

12. Wang, Y.; Zhang, L.; Li, T.; Jiang, Y. A Model-based Analysis-Design Approach for Virtual Binocular Vision System with Application to Vision-based Tactile Sensors. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 5010916. [CrossRef]
13. Ögren, Y.; Tóth, P.; Garami, A.; Sepman, A.; Wiinikka, H. Development of a vision-based soft sensor for estimating equivalence ratio and major species concentration in entrained flow biomass gasification reactors. *Appl. Energy* **2018**, *226*, 450–460. [CrossRef]
14. Alarcon, C.; Shene, C. Fermentation 4.0, a case study on computer vision, soft sensor, connectivity, and control applied to the fermentation of a thraustochytrid. *Comput. Ind.* **2021**, *128*, 103431. [CrossRef]
15. Li, Y.; Ni, P.; Sun, L.; Xia, Y. Finite element model-informed deep learning for equivalent force estimation and full-field response calculation. *Mech. Syst. Signal Process.* **2024**, *206*, 110892. [CrossRef]
16. Marban, A.; Srinivasan, V.; Samek, W.; Fernández, J.; Casals, A. A recurrent convolutional neural network approach for sensorless force estimation in robotic surgery. *Biomed. Signal Process. Control* **2019**, *50*, 134–150. [CrossRef]
17. Ko, D.K.; Lee, K.W.; Lee, D.H.; Lim, S.C. Vision-based interaction force estimation for robot grip motion without tactile/force sensor. *Expert Syst. Appl.* **2023**, *211*, 118441. [CrossRef]
18. Chen, K.; Zhao, W.; Zhang, X. Real-time milling force monitoring based on a parallel deep learning model with dual-channel vibration fusion. *Int. J. Adv. Manuf. Technol.* **2023**, *126*, 2545–2565. [CrossRef]
19. Bakhshandeh, P.; Mohammadi, Y.; Altintas, Y.; Bleicher, F. Digital twin assisted intelligent machining process monitoring and control. *CIRP J. Manuf. Sci. Technol.* **2024**, *49*, 180–190. [CrossRef]
20. Ksiazek, K.; Romaszewski, M.; Głomb, P.; Grabowski, B.; Cholewa, M. Blood Stain Classification with Hyperspectral Imaging and Deep Neural Networks. *Sensors* **2020**, *20*, 6666. [CrossRef]
21. Giulietti, N.; Caputo, A.; Chiariotti, P.; Castellini, P. SwimmerNET: Underwater 2D Swimmer Pose Estimation Exploiting Fully Convolutional Neural Networks. *Sensors* **2023**, *23*, 2364. [CrossRef]
22. Giulietti, N.; Discepolo, S.; Castellini, P.; Martarelli, M. Neural Network based Hyperspectral imaging for substrate independent bloodstain age estimation. *Forensic Sci. Int.* **2023**, *39*, 111742. [CrossRef]
23. Giulietti, N.; Discepolo, S.; Castellini, P.; Martarelli, M. Correction of Substrate Spectral Distortion in Hyper-Spectral Imaging by Neural Network for Blood Stain Characterization. *Sensors* **2022**, *22*, 7311. [CrossRef]
24. Nogueira, F. Bayesian Optimization: Open Source Constrained Global Optimization Tool for Python. 2014. Available online: https://github.com/fmfn/BayesianOptimization (accessed on 11 April 2024).
25. Agrawal, T. Bayesian Optimization. In *Hyperparameter Optimization in Machine Learning*; Apress: Berkeley, CA, USA, 2020; pp. 81–108. [CrossRef]
26. Abdulkadirov, R.; Lyakhov, P.; Nagornov, N. Survey of Optimization Algorithms in Modern Neural Networks. *Mathematics* **2023**, *11*, 2466. [CrossRef]
27. Bai, Y.; Yang, E.; Han, B.; Yang, Y.; Li, J.; Mao, Y.; Niu, G.; Liu, T. Understanding and improving early stopping for learning with noisy labels. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 24392–24403.
28. Raschka, S. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv* **2018**, arXiv:1811.12808.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

EnsembleXAI-Motor: A Lightweight Framework for Fault Classification in Electric Vehicle Drive Motors Using Feature Selection, Ensemble Learning, and Explainable AI

Md. Ehsanul Haque [1], Mahe Zabin [2] and Jia Uddin [3,*]

[1] Department of Computer Science and Engineering, East West University, Dhaka 1212, Bangladesh; 2021-3-60-018@std.ewubd.edu
[2] Human and Digital Interface Department, JW Kim College of Future Studies, Woosong University, Daejeon 34606, Republic of Korea; mahezabin@wsu.ac.kr
[3] Artificial Intelligence and Big Data Department, Endicott College, Woosong University, Daejeon 34606, Republic of Korea
* Correspondence: jia.uddin@wsu.ac.kr

Abstract: As electric vehicles (EVs) are growing, the fault diagnosis in their drive motor becomes more important to have optimal performance and safety. Traditional fault detection methods suffer mainly from high false positive and false negative rates, computational complexity, and lack of transparency in decision-making methods. In addition, existing models are also heavy and inefficient. A lightweight framework for fault diagnosis in EV drive motors is presented with the aid of Recursive Feature Elimination with Cross-Validation (RFE-CV), parameter optimization, and in-depth preprocessing. We further optimize the models and their combination to a hybrid Soft Voting Classifier. These techniques were applied to a dataset of 40,040 data entries that had been simulated by a Variable Frequency Drive (VFD) model. We evaluated eight machine learning models, and our proposed Soft Voting Classifier has the highest test accuracy of 94.52% and a Kappa score of 0.9210 on diagnostic performance. Also, the model has minimal memory usage and low inference latency. In addition, Local Interpretable Model-Agnostic Explanations (LIME) were used to improve transparency and gain an understanding of decisions made through the Soft Voting Classifier. Also, the framework was validated by an additional real-world dataset, thereby further confirming its robustness and consistency in performance for different conditions, which indicates the generalizability of the framework in real-world applications. RFE-CV is found to be very effective in feature selection and helps to construct a lightweight and cost-effective ensemble voting model for enhancing fault diagnosis for EV Drive Motors, overcoming its unsatisfactory transparency, accuracy, and computational efficiency. Finally, it contributes to the development of safer and more reliable EV systems through the development of models supervised on fewer features to give the computing time that is a little lighter without compromising its diagnostic performance.

Keywords: motor faults classification; machine faults classification; lightweight framework; fault detection models; explainable AI

1. Introduction

The global electric vehicle (EV) market is currently experiencing a major transformation because analysts forecast substantial sales growth during the upcoming years. The analysis from Rho Motion indicates that EV sales will grow by 17–18% to exceed 20 million units during 2025 [1–3]. The EV market's expansion mainly stems from the

predicted EV unit numbers of 12.9 million in China, 3.5 million in Europe, and 2.1 million in North America. When EV adoption continues its quick momentum, it becomes clear that sustainable transportation is becoming essential since late detection of motor system faults might trigger life-threatening incidents [4]. Short circuits alongside demagnetization-related electrical problems lead to unexpected loss of control, which threatens both the driver's safety and traffic users [5]. The predictions from analysts indicate electric vehicles will capture between 25% of global vehicle sales by 2025 because of their role in clean energy adoption [6].

The fast-paced EV sector introduces complex obstacles for maintaining dependable performance and safety standards of these vehicles. The correct functioning of electric vehicle motor drives stands as a crucial area for diagnosis since they play a key role in delivering vehicle performance. These systems, composed of battery management and motor drive components, become complicated and difficult to handle during fault events. Vehicle performance becomes degraded while accident risk increases and operational lifespan decreases due to common faults such as phase-to-phase faults, overvoltage, overload, and undervoltage conditions. The accurate functioning of fault detection systems is essential to preserve EV safety together with the performance and longevity of their operational lifespan.

Current fault diagnosis systems operate with multiple severe constraints in their approach. Conventional systems suffer from errors during decision-making along with slow runtimes while revealing minimal clarity about their decision-making methods [7]. These diagnostic methods remain inadequate for field implementation because fast and light solutions with high efficiency represent today's deployment requirements. The existing fault diagnosis methods demonstrate prohibitive shortcomings that necessitate creating an innovative method to improve EV system reliability in an expanding and complex landscape.

The presented research develops a modern, lightweight framework that diagnoses faults within EV motor drives. The framework strengthens data quality through advanced preprocessing procedures, which allow the fault detection operation to function with accuracy. The study adopts Recursive Feature Elimination with Cross-Validation (RFE-CV) to select features, and this approach simultaneously improves model performance and lowers computational requirements. The deployment of the Soft Voting Classifier as part of the hybrid ensemble learning approach allows both better diagnostic precision and reduced operational expenses in model operation. The framework includes the Local Interpretable Model-Agnostic Explanations (LIME) functionality for model decision-process transparency enhancement.

Although this study is not meant for real-time deployment, its lightweight and fast training methods provide its qualities for the real application in the fast-growing EV industry. The contributions of this research are towards the advancement of EV fault diagnosis with a scalable, reliable, and efficient system for fault detection that is accurate while computationally efficient. Below, the contribution of this research is provided:

- Data preprocessing was carried out in-depth to intensify the quality of data and increase the accuracy of fault detection.
- Recursive Feature Elimination (RFE-CV) with Cross-Validation is used to select the best features. This keeps the model efficient and lightweight.
- The framework improves motor fault detection by finding the best configuration from different sets of hyperparameters.
- The use of regularization parameters mitigates overfitting, and leakage of the data is handled in order to enable the model to be robust and reliable in the real-world use case.
- A Soft Voting approach is used to extend diagnostic accuracy at a reduced computational cost as a hybrid ensemble classifier.

- The framework is then extended to incorporate LIME to further improve transparency and understanding of the model decision-making process.
- A lightweight, reliable, and efficient detection framework to detect EV motor faults and for use in the large developing market of electric vehicles is finally proposed.
- A practical framework for real-world data, further validated through real-world datasets, too.

In summary, this paper presents a comprehensive and innovative approach to EV motor fault diagnosis, emphasizing the importance of model efficiency, accuracy, and transparency for real-world applications in the expanding electric vehicle market.

The rest of the paper is structured as follows: The proposed methodology is presented in Section 3, the experimental results and analysis are provided in Section 4, and finally, Section 5 concludes the paper and provides some future work suggestions.

2. Related Works

Studies regarding fault diagnosis of EV motors and other motor types are reviewed in this section. Mechanisms like machine learning, deep learning, and hybrid methods are studied in the different methods that are used to detect motor faults. Moreover, the limitations of these studies, including such validations, overfitting, and generalizability, are presented in the review to provide insights into the state of the art of motor fault diagnosis research.

Studies on fault classification in electric vehicle drive motors are presented by Thirunavukkarasu et al. with a remarkable accuracy of 94.1% in the case of the CatBoost model [8]. Results from the research highlight the importance of developing reliable and safe electric drive fault diagnosis to improve reliability and safety. Yet, there are some limitations of this study, such as a lack of external validation, no systematic outlier removal, and unresolved data imbalance issues. Furthermore, the lack of XAI techniques does not make the prediction of the model interpretable, which is a prerequisite for practical applications in real world environments.

Faiz Ahmad et al. provide a review of AI approaches to motor fault detection and state CNN-LSTM has 97% accuracy [9]. In addition, the study is lacking in key validation methods, such as cross-validation, hyperparameter tuning, and overfitting checks, resulting in questionable results. Preprocessing, real-world applicability, and deployment challenges are all not discussed by them. Furthermore, the paper is further weakened through the missing figures and missing, fragmented explanations. Results are of little practical value because they were not validated, and there was no methodological rigor.

In their paper, Mahbub Ul Islam Khan et al. provide a wide analysis of fault detection and classification in electric vehicles (EVs) by using machine learning (ML) methods [10]. It, therefore, deals with double-line and three-line faults associated with the critical association between a three-phase inverter and a brushless DC motor. Malachi was developed with different ML classifiers, including Decision Tree, Logistic Regression, and Voting Classifier, and their accuracy and reliability were tested and measured to be very high. The models for fault types discussed in the study do have limitations, such as their lack of applicability to other fault types and the necessity for further real-world data matching to validate the models.

Prediction of induction motor faults using machine learning techniques is presented by Abdulkareem et al. to fill the vacuum between normal time spent in predictive maintenance [11]. To analyze the dataset of healthy and faulty conditions of induction motors, the research employs various machine learning algorithms, including Random Forest, Artificial Neural Networks, k-Nearest Neighbors, and Decision Trees. The Random Forest model achieved the highest accuracy of 91%, and the ANN and k-NN models had the

highest accuracy of 90%. The study, however, is restricted by examined fault conditions that may limit the generalizability of the findings.

In Muhammad Samiullah et al.'s study, machine learning is used to predict preventative maintenance in induction motors, and the authors state that the Decision Tree produced the highest accuracy (92%) for detecting faults [12]. Nevertheless, the paper just gives this binary classification and does not provide insight into the entire spectrum of motor fault types. Despite the good behavior for the selected faults by the study, fault coverage does not extend beyond the selected faults, which makes the general workability of the model questionable. Incorporating additional fault types and making the model more flexible will help boost its investigation in the real world.

In the work of Debshree Bhattacharya et al., a fault detection approach is proposed based on a two-layer machine learning [13]. LGBM achieves 100% accuracy, but they are not cross-validated, explainable AI, and not regularized, causing worries of overfitting. The study is limited to simulated data for which generalizability is limited and is not fully explored for the effect of feature selection on model performance. These limitations have to be addressed, and has to be a test in the real world.

Under electric vehicle induction motor fault detection, Raziq Yaqub et al. consider engine sound and vibration data [14]. With artificial intelligence, specifically the Random Forest classifier, they reach 92% accuracy and 86–91% precision related to faults in different sizes. A notable limitation, however, is that the study is limited to the application of this idea to problems involving binary classification only and does not allow for simultaneous treatment of multiple fault types. Its applicability to more complex, real-world scenarios where simultaneous multiple faults occur limits the robustness of the system in case of diverse conditions.

In the work of Min Chan et al., the authors present a highly accurate fault diagnosis system for induction motors with 100% accuracy for SVM and CNN models [15]. The problem is that given the perfect performance, it is natural to worry about potential overfitting, as the models may be too much in tune with the training dataset, not generalizing well. Real-world testing is essential to validate the findings, as they have to take into account the complexities with variables in operation that are different from the controlled setup used during the study. This validation would make these models more applicable in industrial settings.

The work of Aniqua et al. employs XAI techniques of LIME and SHAP to further improve the interpretability and accuracy of the model when diagnosing machine faults using audio sensor data [16]. Twenty-nine features extracted from audio recordings of industrial fans were used by the authors, where a number of machine learning classifiers, such as logistic regression, XGBoost, SVM, Gaussian Naive Bayes, and Random Forest, were employed to classify the faults in the system. Using XAI as a feature, their proposed method improved the diagnostic accuracies: 80.28 for Logistic Regression with XAI and 79.57 without XAI. Nevertheless, the study's dataset could not be diverse enough; there are issues regarding the real-time application, and audio data can be interfered with by noise. In addition, PCA suffers from limitations of linearity; hence, in cases of datasets with nonlinear relations, we may lose critical information.

The InceptionV3 with an SE attention mechanism by Lifu et al. results in a 100% detection rate; however, such an approach suffers from an overfitting problem and yields a long training time, which makes the model lengthy. Furthermore, the model is a "black box" in nature, which reduces interpretability, possibly hindering the actual use case [17]. A study is presented that utilizes 369 thermal images of an industrial motor with 11 fault types after CLAHE preprocessing. It is recognized in the study that practical limitations such as resolution constraints are associated with high-resolution images while improving accuracy.

Now, the Table 1 summarizes the limitations of previous studies along with the data types used in the dataset.

Table 1. Summary of studies on fault detection in electric motors.

Study	Approach	Limitation	Data Type
[8]	CatBoost	Lack of external validation, no systematic outlier removal, no data imbalance handling, no XAI techniques.	Simulated
[9]	CNN-LSTM	No cross-validation, hyperparameter tuning, overfitting checks, missing figures, weak methodology.	Image
[10]	Decision Tree, Logistic Regression, Voting Classifier	Narrow fault type focus, lacks real-world data validation.	Simulated
[11]	Various ML models	Narrow fault condition scope, limits generalizability.	Image
[12]	Decision Tree	Binary classification only, limits real-world applicability.	Simulated
[13]	LGBM	No cross-validation, lacks XAI, overfitting risk, simulated data reliance.	Simulated
[14]	Random Forest	Binary classification, limits multi-fault handling.	Vibration, Sound
[15]	SVM, CNN	Concerns about overfitting, requires real-world testing.	Image
[16]	LR	Limited dataset diversity, real-time challenges, noise interference in audio data. PCA's linearity may cause information loss in nonlinear data.	Vibration, Sound
[17]	InceptionV3	Overfitting risk, "black-box" nature limits interpretability, long training time.	Image (Thermal)

3. Methodology

This section describes a systematic approach to motor fault classification with a focus on important techniques such as Explanatory Data Analysis, Data Preprocessing, and Feature Selection. Furthermore, electric vehicle motor fault diagnosis is integrated with advanced machine learning models and explainable AI (XAI) to further improve the accuracy and reliability of the diagnosis. This helps with effective and trustworthy fault detection.

3.1. Data Collection

This simulated dataset for experiment 1 used in this study was collected from GitHub [18]. It is generated using a Variable Frequency Drive (VFD) model simulated in MATLAB Simulink. In total, the dataset had 40,040 entries and a total of nine features, as well as a target column. One of the ten columns was an object, and nine others were numerical. The target column has a total of six classes. Also, the framework was further validated on another similar dataset used in experiment 2 to see whether the framework is effective for an empirical dataset or not [19]. This dataset has a total of nine classes. Section 4.2 presents the results based on this dataset.

Table 2 shows the description of the features, while Table 3 lists the faults and their counts for each fault for the MATLAB dataset.

Table 2. Feature description.

Feature	Description	Data Type
Tn (Rated Torque) N-m	Torque in Newton-meters	float64
k (Constant of Proportionality)	Constant used in the drive model	float64
time (s)	Time in seconds	float64
Ia (Amp)	Current in phase A in Amperes	float64
Ib (Amp)	Current in phase B in Amperes	float64
Ic (Amp)	Current in phase C in Amperes	float64
Vab (V)	Voltage between phases A and B in Volts	float64
Torque (N-m)	Output torque in Newton-meters	float64
Speed (rad/s)	Speed in radians per second	float64
Category	Fault category	object

Table 3. Fault descriptions and their counts.

Fault Category	Description	Count
NOM	Normal Operation	13,013
PTPF	Phase-to-Phase Fault (Phase B and Phase C)	7007
PTGF	Phase-to-Ground Fault (Phase A and Ground)	7007
OVF	Overvoltage Fault	5005
OLF	Overload Fault	5005
UVF	Undervoltage Fault	3003

3.2. Explanatory Data Analysis

This section presents EDA (Exploratory Data Analysis) [20]. It helps us to understand the dataset's structure, distribution, and relationships between variables. It also helps us understand the data, ensuring quality and integrity. Both are key for building accurate, reliable predictive models.

3.2.1. Descriptive Statistics

In Table 4, it is found that the 'Ia (Amp)' and 'Torque (N-m)' features show the widest ranges. Also, 'Torque' includes both negative and positive values. There are also variations in features such as 'Vab (V)' and 'Speed (rad/s)'.

Table 4. Summary statistics: min, max, range, and average values for each feature.

Feature	Min	Max	Range	Avg
Tn (Torque) N-m	0.800000	2.500000	1.700000	1.138750
k (Proportionality)	0.000948	0.002964	0.002015	0.001349
time (s)	0.000000	0.500000	0.500000	0.250000
Ia (Amp)	0.000000	78.639350	78.639350	17.897379
Ib (Amp)	0.000000	79.113988	79.113988	17.764825
Ic (Amp)	0.000000	78.738940	78.738940	18.467770
Vab (V)	0.000000	649.516523	649.516523	432.839346

Table 4. *Cont.*

Feature	Min	Max	Range	Avg
Torque (N-m)	−404.837797	625.165923	1030.003720	23.533621
Speed (rad/s)	−112.258007	182.599449	294.857456	123.521359
Category	0.000000	5.000000	5.000000	1.975000

3.2.2. Distribution Characteristics

The skewness and kurtosis [21] values for each characteristic are presented in Table 5. Skewness measures the asymmetry of a distribution. For instance, features like Tn and k are right-skewed. Kurtosis specifies the tails of a distribution. High kurtosis values, such as Torque and Ia, suggest a heavy-tailed distribution.

Table 5. Skewness and Kurtosis values for each feature.

Feature	Skewness	Kurtosis
Tn	2.191088	4.531404
k	2.186402	4.514090
time	0.000000	−1.200002
Ia	1.654887	4.711679
Ib	1.505474	3.758770
Ic	1.453311	3.652326
Vab	−1.637558	4.011556
Torque (N-m)	0.790280	27.050441
Speed (rad/s)	−2.315635	4.154707

3.2.3. Histographic Representation

Figure 1 displays histograms for all features and reveals fluctuations in the rated torque (Tn), phase currents (Ia, Ib, Ic), voltage (Vab), torque (T), and speed (n). The histograms show how these parameters change during the process. Time shows a consistent pattern, while other parameters vary. This pattern shows a strong relationship between these key metrics in the system.

3.2.4. Outlier Detection

Figure 2 displays the distribution of the features. It includes central tendency, variability, and outliers. It highlights that all features, except time (s), have outliers. Time remains constant and has no outliers. The whiskers and IQRs for torque, current, voltage, and speed show significant variation. This indicates more dynamic fluctuations in the data.

3.2.5. Correlation Analysis

Figure 3 shows the Pearson correlation heatmap [22]. It highlights both negative and positive correlations between variables. It is observed that there is a positive correlation between Tn and k values, as well as between Ia and Ib. On the other hand, a significant negative trend is seen when comparing speed with Ia. The relationship between category and Tn or k remains weak.

3.2.6. Feature Trends

Figure 4 illustrates line plots, and it shows that the rated torque (Tn) and currents in phases A (Ia), B (Ib), and C (Ic) show high variability. The voltage between phases A

and B (Vab) and speed also vary greatly. Also, each feature has many outliers. These may be due to changes in load, sensing errors, or mechanical issues. Among all the features, the time plot has constant value without variation, which there is a constant time interval for each measurement.

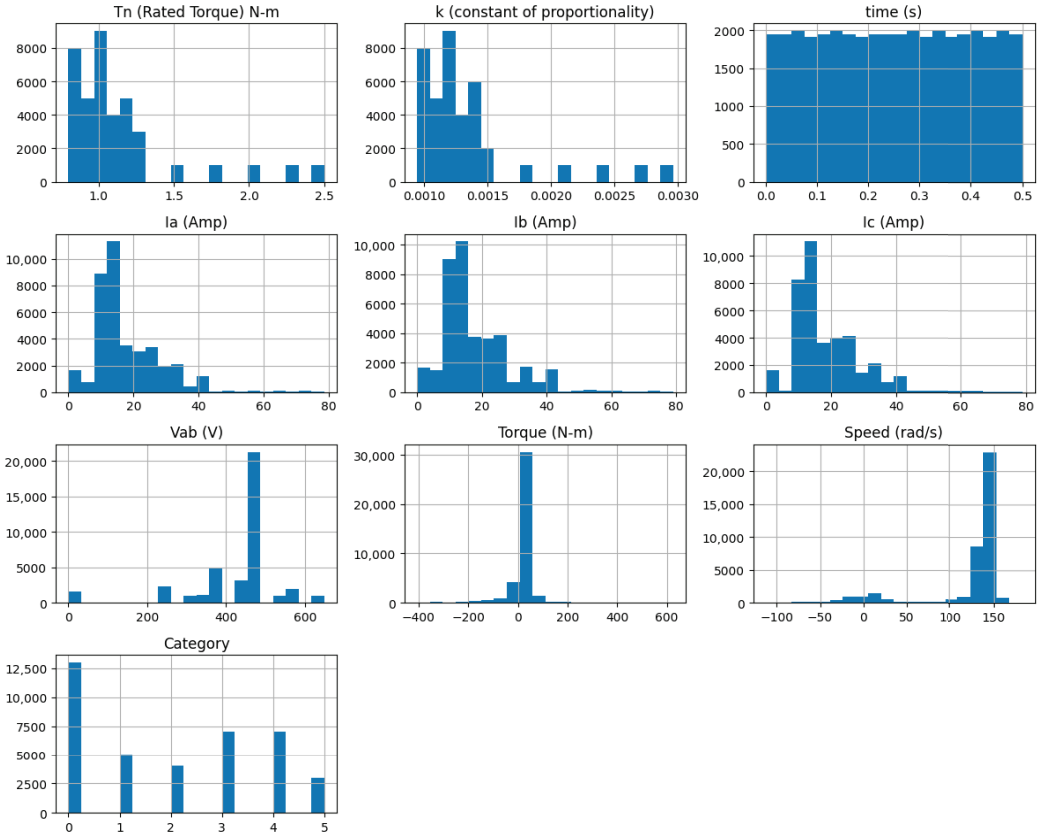


Figure 1. Histogram of the data distribution.

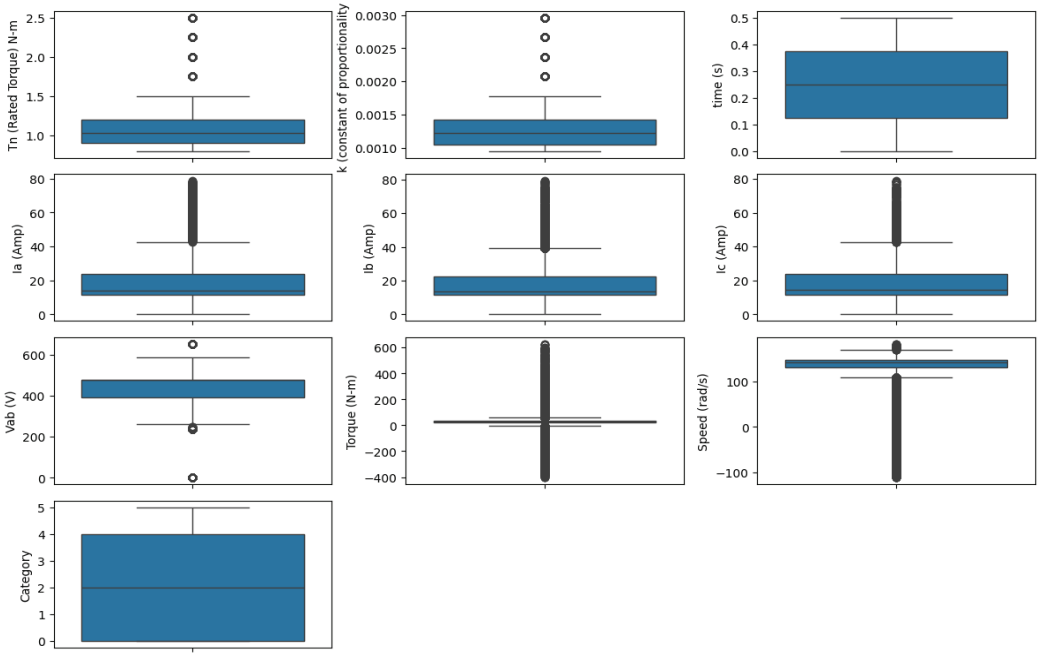


Figure 2. Box plot of all features.

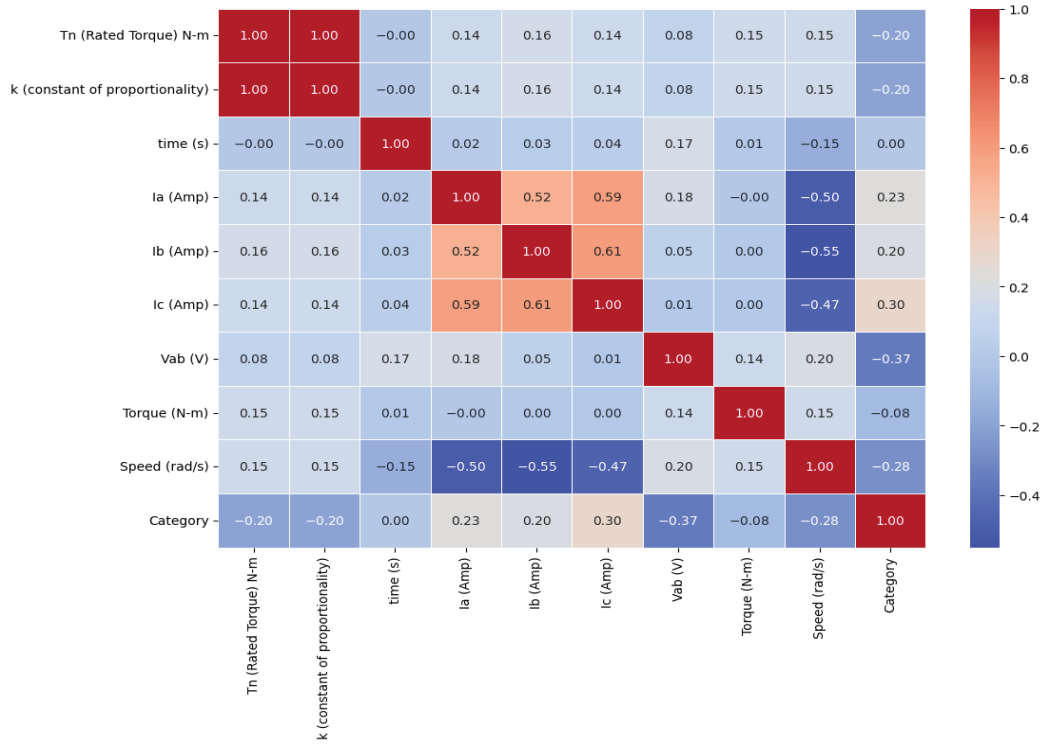


Figure 3. Pearson correlation heatmap.

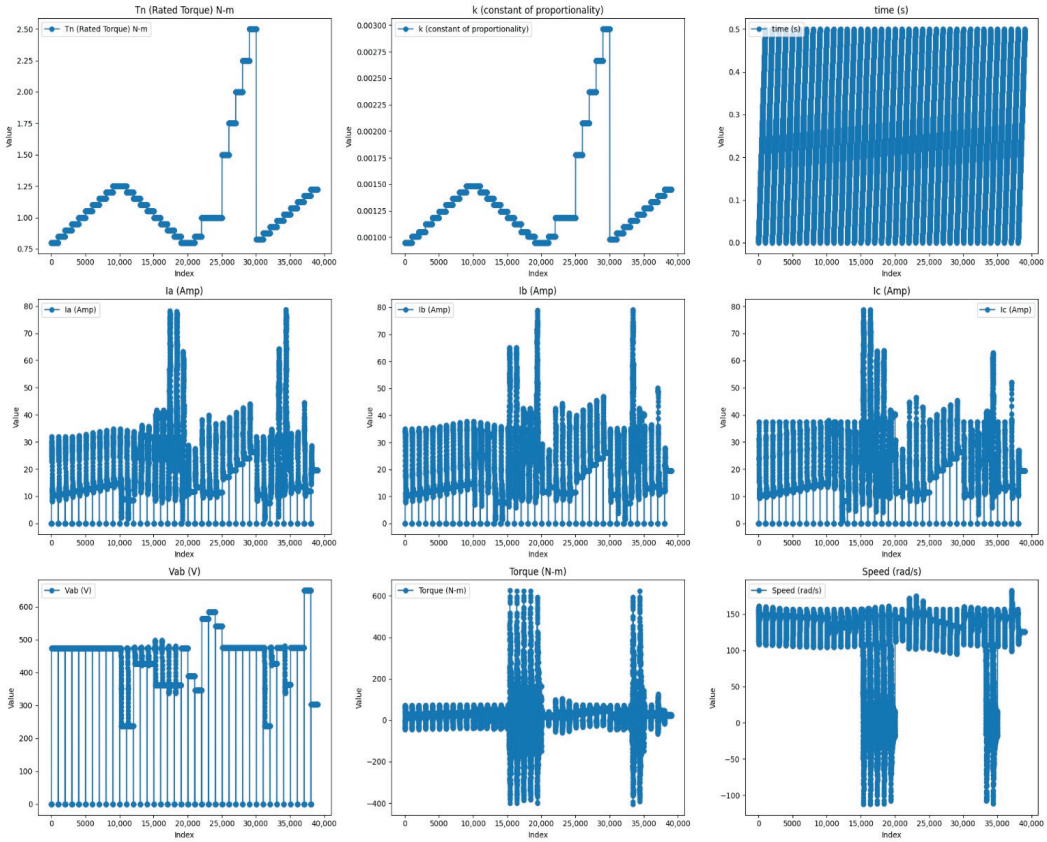


Figure 4. Line plot of all features.

3.2.7. Feature Importance

The Random Forest feature importance [23] values in Figure 5 show that Vab, speed, k, and Tn are key features. On the other hand, time and torque hold less significance. This suggests they have a limited impact on fault detection results.

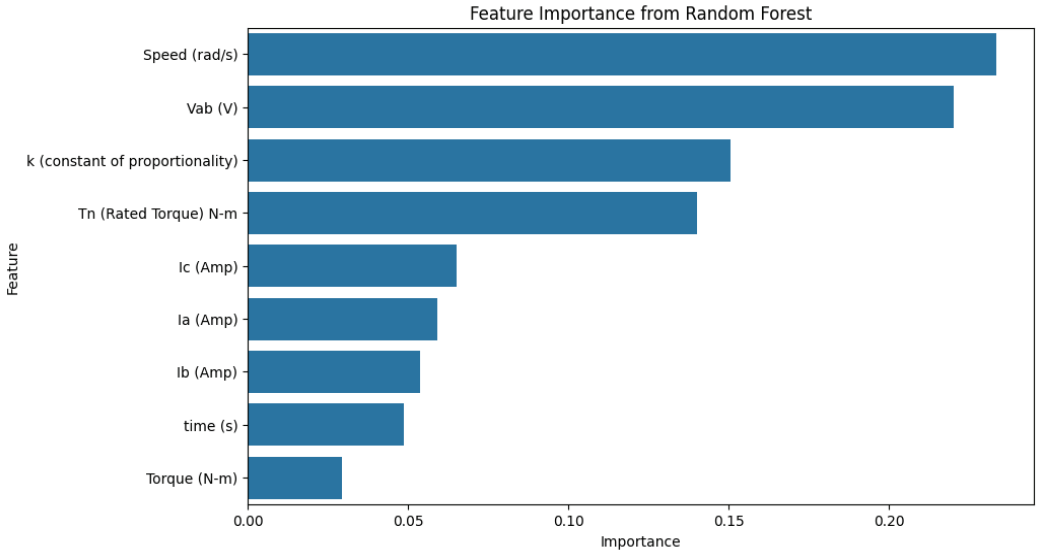


Figure 5. Feature importance for Random Forest.

3.2.8. Analysis of Category Label Frequency

Figure 6 shows the category distribution in the dataset. NOM has the highest occurrences, while UVF has the lowest. The dataset shows a significant imbalance as NOM is a highly populated category, while PTPF and PTGF have the same number of entries.

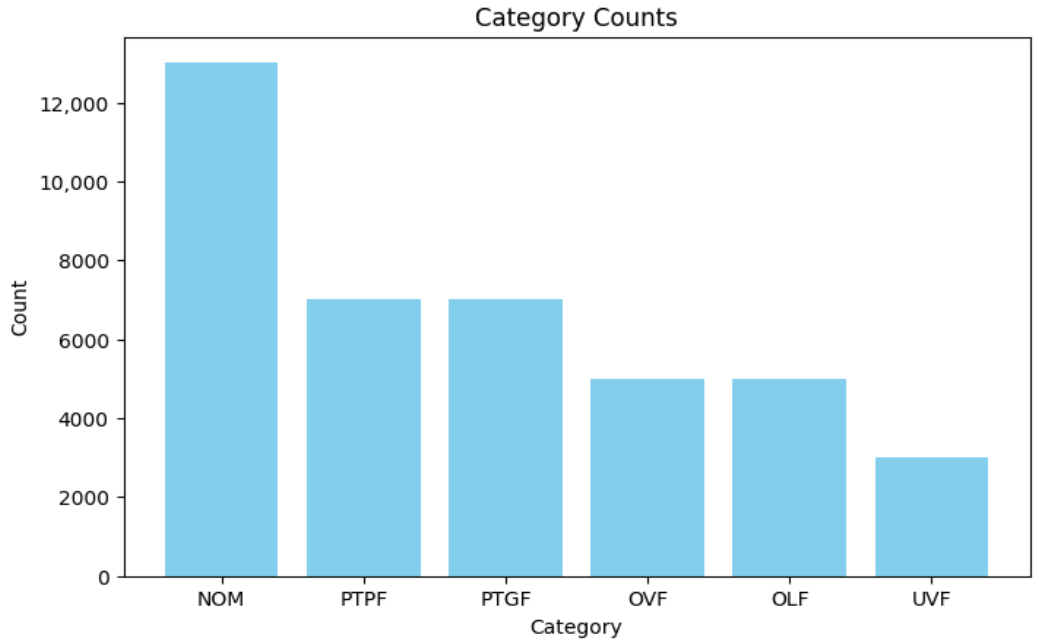


Figure 6. Fault counts bar charts.

Figure 7 shows a pie chart. It displays the percentage breakdown of categories in the dataset. NOM has the highest percentage at 32.5%. In addition, PTGF and PTPF, respectively, make up 17.5%, OVF and OLF are each 12.5%, and UVF is 7.5%.

3.3. Data Preprocessing

In machine learning tasks, data preprocessing is very essential [24]. Usually, raw data are incomplete, inconsistent, and noisy. Such issues can decrease the performance of a model. The purpose of preprocessing is to clean and organize the data. This indeed improves the analysis and consequently gives the model more precision and reliability.

The steps for preprocessing for classifying faults of electronic vehicle (EV) drive motors are given below.

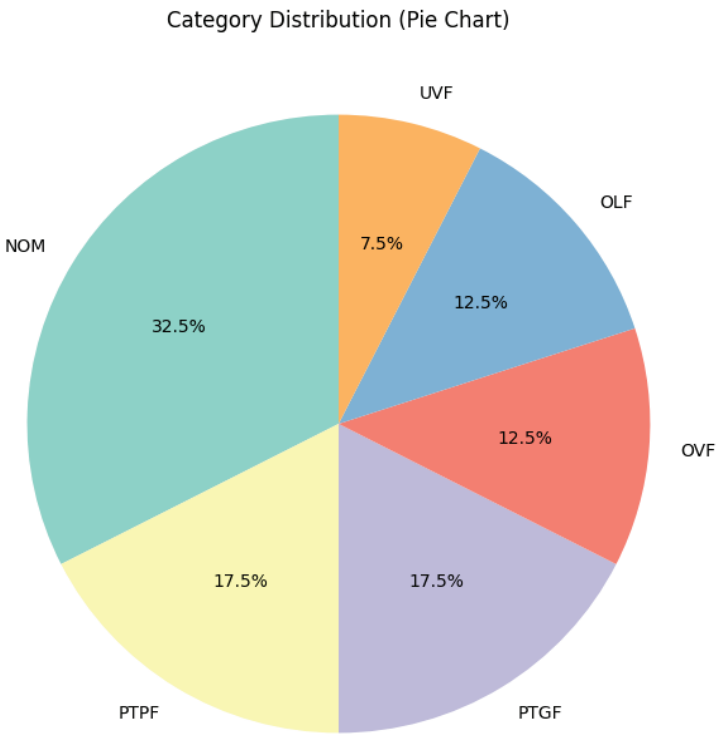


Figure 7. Pie charts of fault.

3.3.1. Duplicate Row Handling

Duplicate data can negatively affect the performance of the model by creating bias and redundancy. For this study, we used Equation (1) to detect and handle duplicate rows. The use of this method led to the removal of 1000 duplicate entries from the dataset.

The following method detects and removes duplicate rows:

$$D = \{X_i \mid X_i \text{ appears more than once in } X\} \tag{1}$$

$$X' = X - D \tag{2}$$

where the following is true:

- X represents the original dataset containing n rows.
- X_i is an individual row in the dataset.
- D is the set of duplicate rows, where a row X_i is considered a duplicate if it appears more than once in X.
- X' is the dataset after removing all rows in D (the duplicates).
- The operator represents the subtraction or removal of elements in D from dataset X, resulting in dataset X'.

3.3.2. Encoding Categorical Column

The 'Category' column in the dataset has six types of attack. For the transformation of these categorical values into numeric labels for machine learning, we use Label Encoding. The attack types get transformed into unique integer values that follow the pattern below:

$$\text{Category Encoding} = \begin{cases} \text{'NOM'} & \rightarrow 0 \\ \text{'OLF'} & \rightarrow 1 \\ \text{'OVF'} & \rightarrow 2 \\ \text{'PTGF'} & \rightarrow 3 \\ \text{'PTPF'} & \rightarrow 4 \\ \text{'UVF'} & \rightarrow 5 \end{cases} \tag{3}$$

By encoding the categorical data this way, it allows machine learning models to work with the categorical data in numeric format, that is required by most of the algorithms which cannot work with non-numeric values.

3.3.3. Outlier Detection and Removal

There are many outliers in the dataset, as can be seen from the exploratory data analysis (EDA) earlier. To overcome this situation, we detect and remove the outliers using the Interquartile Range (IQR) method [25]. The reason for choosing the IQR method is that features in the dataset are not normally distributed. By doing this, it becomes more robust than assuming normality.

Here is how the process works:

$$\text{Outliers} = ((\mathbf{df} < \text{lower_bound}) \mid (\mathbf{df} > \text{upper_bound})) \tag{4}$$

After detecting the outliers, rows containing them are filtered out:

$$\mathbf{df_new} = \mathbf{df}[\neg \text{outliers.any}(\text{axis} = 1)] \tag{5}$$

3.3.4. Feature Significance

The feature significance was assessed through a one-way ANOVA test [26]. ANOVA tests show how features impact the target variable. They also identify which features are statistically significant.

Following analysis, the following decisions were made:

- We find that IC (Amp) is also insignificant and hence removed from the dataset.
- Time is important in statistical tests. However, we remove it for further analysis since faults do not depend on time.

Table 6 shows the one-way ANOVA results for each feature.

Table 6. Feature significance results based on one-way ANOVA test.

Feature	F-Statistic	*p*-Value	Significance Status
Tn (Rated Torque) N-m	2940.344642	0.0	True
k (constant of proportionality)	2926.400588	0.0	True
time (s)	464.783618	0.0	True
Ia (Amp)	5527.194205	0.0	True
Ib (Amp)	343.854130	0.0	True
Ic (Amp)	NaN	NaN	False
Vab (V)	48,014.923047	0.0	True
Torque (N-m)	391.094387	0.0	True
Speed (rad/s)	3539.528201	0.0	True

3.4. Dataset Splitting

We split the dataset using an 80-20 ratio. This means 80% is for the training set, while 20% is for the testing set. After the split, we checked the category distribution in both sets. Table 7 shows the distribution of each category in each set.

Table 7. Category distribution in training and testing sets.

Category	Training Set Count	Testing Set Count
0	9770	2483
4	3409	826
5	2306	569
2	2191	551
3	902	218
1	746	185

3.5. Feature Selection

Feature selection is performed to make the model lighter and less time-intensive. We used the Recursive Feature Elimination with Cross-Validation (RFE-CV) method [27]. This feature selection process is conducted only on the training set to ensure that the features chosen are based on the training data, which prevents any data leakage into the testing set.

The highest cross-validation score of 89.98% is achieved using four selected features, which is why they were retained. The following Table 8 shows the scores of cross-validation for different numbers of selected features, which gives the evolution of performance when features are added.

Table 8. Cross-validation scores for different models and selected features.

Selected Features	Mean CV Score (%)
1	88.80
2	88.54
3	88.72
4	89.98
5	89.93
6	89.96
7	89.89

After feature selection, four out of the seven original features were used for further analysis. The final selected features are as follows:

- Tn (Rated Torque) N-m.
- Ia (Amp).
- Vab (V).
- Speed (rad/s).

In the following, Figure 8 shows the CV scores in feature-wise obtained during the feature selection process. Also, Algorithm 1 shows the steps of selecting features.

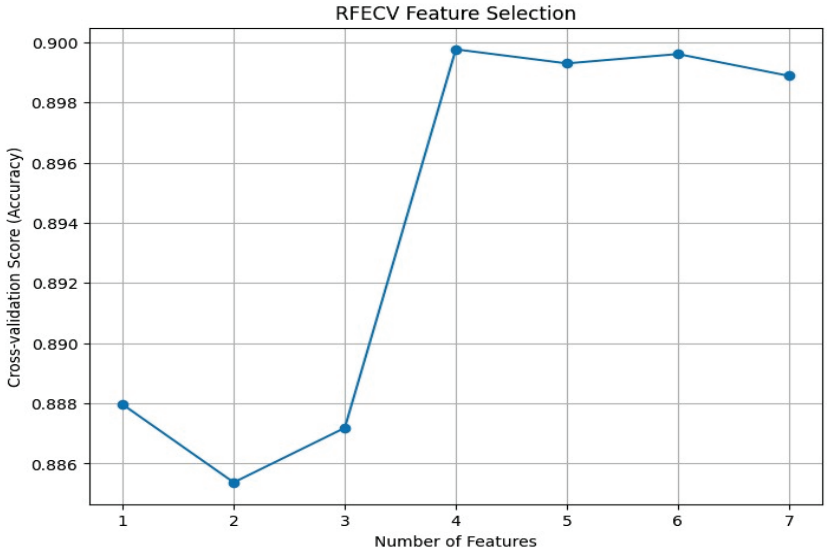


Figure 8. Feature-wise cross-validation (CV) scores.

Algorithm 1 Feature Selection using RFECV

1: **Input:** $X_{\text{train}} \in \mathbb{R}^{n \times p}, y_{\text{train}} \in \mathbb{R}^n, f_{\text{model}}$ (Random Forest)
2: **Output:** Final selected features for further analysis
3: Initialize f_{model} with $n_{\text{estimators}} = 100$
4: Initialize StratifiedKFold with $k = 5$ folds for cross-validation
5: Apply RFECV: selector $\leftarrow$ RFECV$(f_{\text{model}}, \text{cv} = 5, \text{scoring} = \text{accuracy})$
6: Fit the model: selector.fit$(X_{\text{train}}, y_{\text{train}})$
7: Obtain optimal features:

$$N_{\text{optimal}} = \text{selector.n_features_}, \quad \text{selected_features} = X_{\text{train}}[\text{selector.support}]$$

8: Plot RFECV performance curve:

$$\text{Plot}(x = \text{range}(1, N), y = \text{selector.cv_results}[\text{mean_test_score}])$$

9: **Output:** Final selected features: final_selected_features = selected_features[: $\min(N_{\text{optimal}}, 7)$]

3.5.1. Data Balancing

When we split the dataset, Table 7 clearly showed that the training set has a record of a high class imbalance. This issue was overcome using the Synthetic Minority Over-sampling Technique (SMOTE) [28], and we used the model to improve model performance. We apply SMOTE and compare them in Figure 9.

3.5.2. Standardization

Both the training and test sets were standardized such that each feature has a mean of 0 and a standard deviation of 1. This step is necessary to eliminate any scale-dependent bias and to improve the convergence and the hardiness of the model.

$$z' = \frac{v - \bar{v}}{s_v} \tag{6}$$

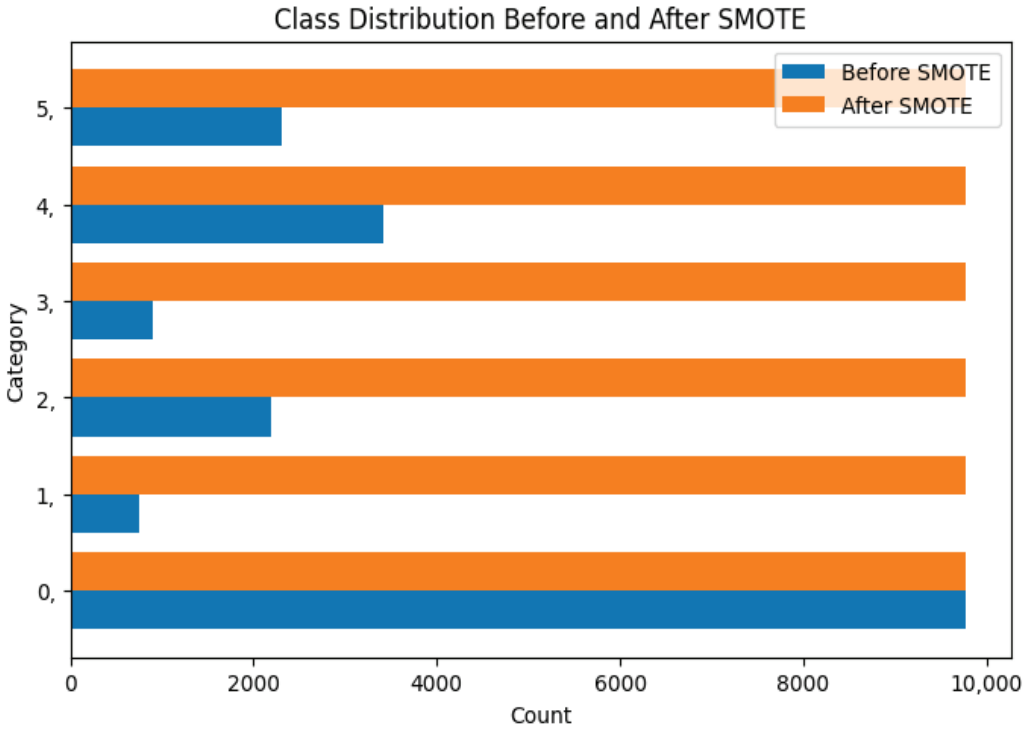


Figure 9. Before and after balancing.

3.6. Model Training

Accurate classification of faults in electric vehicle drive motors is the key to maintaining vehicle performance and safety. For this experiment, eight different machine learning models were tested to determine which ones are effective in diagnosing the motor faults. The models evaluated include the following:

- Random Forest (RF): A type of ensemble learning that combines many Decision Trees using bagging in order to increase classification accuracy. The final prediction is made by voting on all of the trees in the forest.

$$\hat{y}_{\mathrm{RF}} = \mathrm{mode}(\mathrm{Tree}_{r1}(x_r), \mathrm{Tree}_{r2}(x_r), \ldots, \mathrm{Tree}_{rn}(x_r)) \tag{7}$$

  where the following is true:

  - $\hat{y}_{\mathrm{RF}}$ is the final predicted class;
  - $\mathrm{Tree}_{ri}(x_r)$ is the prediction from the ith Decision Tree;
  - n is the number of trees in the forest.

- Decision Tree (DT): A supervised learning model that splits the data into subsets based on feature values to classify motor faults.

$$\hat{y}_{\mathrm{DT}} = \arg\max_{c \in C_d} \sum_{i=1}^{m} I(y_{di} = c) \tag{8}$$

  where the following is true:

  - $\hat{y}_{\mathrm{DT}}$ is the predicted class;
  - C_d is the set of possible classes;
  - $I(y_{di} = c)$ is an indicator function that is 1 if the class label y_{di} matches class c and 0 otherwise.

- Extra Trees (ET): An ensemble method that builds multiple Decision Trees similar to Random Forest but with a different randomization technique for splitting nodes.

$$\hat{y}_{\text{ET}} = \text{mode}(\text{Tree}_{e1}(x_e), \text{Tree}_{e2}(x_e), \ldots, \text{Tree}_{en}(x_e)) \tag{9}$$

  where the following is true:
  - $\hat{y}_{\text{ET}}$ is the final predicted class;
  - $\text{Tree}_{ei}(x_e)$ is the prediction from the ith Decision Tree;
  - n is the number of trees in the forest.
- K-Nearest Neighbors (KNN): A non-parametric classifier that assigns a class based on the majority class of the k-nearest data points.

$$\hat{y}_{\text{KNN}} = \text{mode}(y_{k1}, y_{k2}, \ldots, y_{kk}) \tag{10}$$

  where the following is true:
  - $\hat{y}_{\text{KNN}}$ is the predicted class;
  - $y_{k1}, y_{k2}, \ldots, y_{kk}$ are the classes of the k nearest neighbors;
  - k is the number of neighbors considered.
- Multilayer Perceptron (MLP): A type of neural network that learns to map input features to output labels by training multiple layers of neurons.

$$\hat{y}_{\text{MLP}} = \sigma\left(\sum_{i=1}^{p} w_{mi} x_{mi} + b_m\right) \tag{11}$$

  where the following is true:
  - $\hat{y}_{\text{MLP}}$ is the predicted output;
  - x_{mi} are the input features;
  - w_{mi} are the weights;
  - b_m is the bias term;
  - σ is the activation function (e.g., ReLU, Sigmoid).
- CatBoost: A gradient boosting algorithm designed for categorical data that builds Decision Trees to minimize classification error.

$$F_{\text{CatBoost}}(x_c) = \sum_{t=1}^{T_c} \eta_t \cdot h_t(x_c) \tag{12}$$

  where the following is true:
  - $F_{\text{CatBoost}}(x_c)$ is the final prediction for input x_c;
  - $h_t(x_c)$ is the tth tree's prediction;
  - η_t is the learning rate for the tth tree;
  - T_c is the total number of trees.
- XGBoost: An efficient gradient boosting algorithm that improves on Decision Trees using regularization and parallelization.

$$F_{\text{XGBoost}}(x_x) = \sum_{t=1}^{T_x} \eta_t \cdot h_t(x_x) + \lambda \sum_{t=1}^{T_x} \|w_t\|^2 \tag{13}$$

  where the following is true:
  - $F_{\text{XGBoost}}(x_x)$ is the final prediction for input x_x;
  - $h_t(x_x)$ is the tth tree's prediction;

- η_t is the learning rate for the tth tree;
- T_x is the total number of trees;
- λ is the regularization parameter for tree weights w_t.

- Soft Voting Classifier: An ensemble method that combines the predictions of multiple models by averaging their probabilities and selecting the most probable class.

$$\hat{y}_{\text{SV}} = \arg\max_c \left(\frac{1}{n_s} \sum_{i=1}^{n_s} P(y_{si} = c | x_s) \right) \tag{14}$$

where the following is true:

- $\hat{y}_{\text{SV}}$ is the final predicted class;
- $P(y_{si} = c|x_s)$ is the predicted probability of class c for model i;
- n_s is the number of models in the ensemble.

3.7. Hyperparameter Settings

Hyperparameter tuning is essential to obtain maximum model performance. All the models in our research were initialized with certain hyperparameters to achieve optimal accuracy and efficiency. Table 9 below summarizes the most important hyperparameters used for every model.

Table 9. Hyperparameters of the models used.

Model	**Hyperparameters**
Random Forest	n_estimators=100, max_depth=10, min_samples_split=4, random_state=42
Decision Tree	max_depth=10, min_samples_split=4, random_state=42
Extra Trees	n_estimators=100, max_depth=10, min_samples_split=4, random_state=42
KNN	n_neighbors=5, weights='uniform'
MLP	hidden_layer_sizes=(128,), max_iter=1000, alpha=0.001, solver='adam', random_state=42
CatBoost	iterations=500, learning_rate=0.1, depth=6, l2_leaf_reg=10, early_stopping_rounds=50, random_state=42, verbose=0
XGBoost	n_estimators=100, max_depth=6, learning_rate=0.1, alpha=0.01, gamma=0, random_state=42
Soft Voting	Base models: Random Forest, Decision Tree, Extra Trees, KNN, MLP, CatBoost, XGBoost

3.8. Cross-Validation

In order to evaluate how well the models perform for the whole dataset and to assess their generalization, we evaluated it by 5-fold cross-validation. The use of this method ensures that the models are not overfitted to a particular subset of the data during training and testing on different splits.

3.9. Explainable AI

Finally, LIME (Local Interpretable Model-Agnostic Explanations) [29] was applied to the proposed model in order to use this model in a transparent manner so that it can be understood how the predictions are made. LIME introduces local interpretations through an approximation of a complex model with simpler ones; hence, they can identify the factors that influence the classification decisions.

Below, Figure 10 shows the overall workflow diagram for efficient EV motor fault classification.

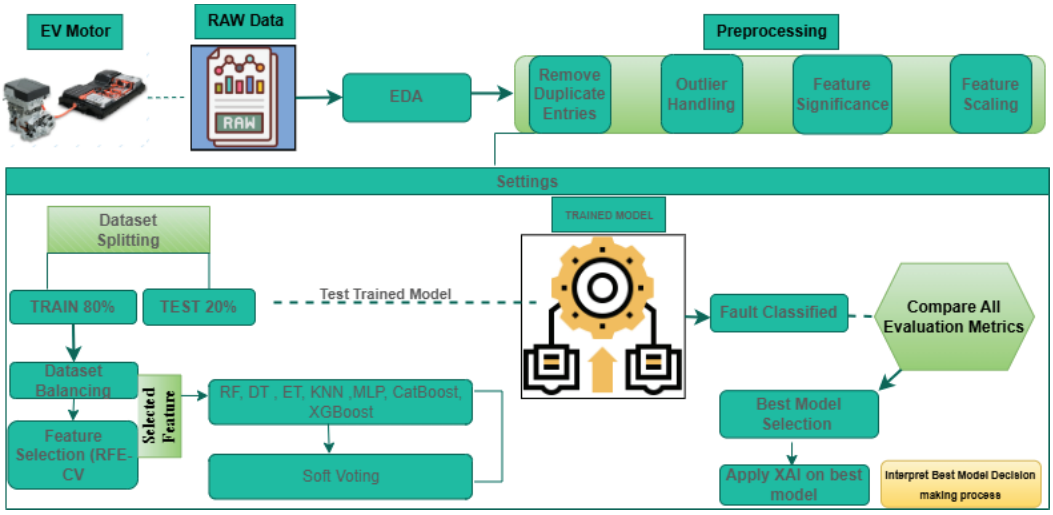


Figure 10. Proposed approach for electric drive motors fault classification.

4. Results and Discussion

In this section, we provide training and testing results to evaluate the machine learning model in detecting faults of Electric Vehicle (EV) drive motors. Finally, the results are analyzed using evaluation metrics described in Table 10. To evaluate the machine learning models we used to classify the fault in the electric vehicle drive motors, different metrics are used. The metrics let us know how accurate, reliable, and generally valid the model is. Table 10 describes the key evaluation metrics, their definition, and the equations of the model performance.

Table 10. Evaluation metrics with descriptions and formulas.

Metric	Description	Formula
Precision	Calculates the rate of true positive predictions among all positive predictions.	$\text{Precision} = \frac{TP}{TP+FP}$
Recall	Measures the number of true positives out of those that were actually positive.	$\text{Recall} = \frac{TP}{TP+FN}$
F1 Score	The harmonic mean of Precision and Recall, providing a balance between the two.	$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$
Accuracy	Calculates the overall correctness of the model's predictions.	$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$
Kappa	Measures agreement between observed and predicted classifications, correcting for chance.	$\text{Kappa} = \frac{P_o - P_e}{1-P_e}$, where P_o is observed agreement and P_e is expected agreement.
AUC	Area under the ROC curve, indicating the model's ability to separate classes.	AUC is derived from the ROC curve. Higher AUC indicates better performance.
Confusion Matrix	Displays correct and incorrect predictions for each class.	$\begin{matrix} TP & FP \\ FN & TN \end{matrix}$
Memory Usage	The amount of memory required by the model during training and inference.	Measured in MiB/GB depending on system architecture.
Inference Time	The time taken for the model to make predictions on new data.	Typically measured in milliseconds or seconds per sample.

4.1. Experiment 1: Drive End Simulated Dataset

4.1.1. Training and Validation Performance

Table 11 shows the performance of several machine learning models, including Random Forest, Decision Tree, Extra Trees, K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP), CatBoost, XGBoost, and Voting Classifier, evaluated with respect to training accuracy, mean cross-validation (CV) accuracy, and standard deviation.

First of all, XGBoost achieved the highest training accuracy of 98.45 percent, a mean CV accuracy of 97.19%, and the lowest standard deviation of 0.0006, which proves its high consistency in training tasks through various training sets. Upon doing the same, CatBoost tied behind the first runner at 97.96% training accuracy, 97.05% mean CV accuracy, and 0.0031 standard deviation, largely behaving stably but more variably. Also, the Voting Classifier had a training accuracy of 97.82%, and we achieved a mean CV accuracy of 97.10% and a standard deviation of 0.0012, which is a relatively high stability and proves to be a reliable generalization. Moreover, the Random Forest had a training accuracy of 97.40%. It also achieved a mean CV accuracy of 96.55 and a low variance of 0.0021, showing a solid performance. The Decision Tree model achieved a training accuracy of 96.17%. Its mean CV accuracy was 95.36%, with a standard deviation of 0.0031. The Extra Trees were trained to 95.25% accuracy, the mean CV accuracy was 95.06%, and the standard deviation was 0.0045; thus, they showed higher variability in performance from run to run. In particular, KNN and MLP, with their training accuracies of 96.58 and 94.70%, respectively, exhibited higher fluctuation in their performance with higher standard deviations. To conclude, XGBoost and CatBoost serve as the most reliable models, but XGBoost is the most stable one in terms of the evaluation provided in Table 11. The Voting Classifier also performed very well and was quite consistent. Random Forest, Extra Trees, KNN, and MLP had greater variability. Still, their accuracy was quite competitive.

Table 11. Training accuracy and cross-validation scores.

Model	Training Accuracy	Mean CV Accuracy	Standard Deviation
Random Forest	97.40%	96.55%	0.0021
Decision Tree	96.17%	95.36%	0.0031
Extra Trees	95.25%	95.06%	0.0045
KNN	96.58%	95.20%	0.0044
MLP	94.70%	94.49%	0.0038
CatBoost	97.96%	97.05%	0.0031
XGBoost	98.45%	97.19%	0.0006
Voting Classifier	97.82%	97.10%	0.0012

Next, the learning curves of the models are added in Figure 11. In addition to this, it helps us understand how the models learn over time, and these curves provide additional insight into how their training behaves, evolves with their performance, and how they may even be headed for overfitting or underfitting.

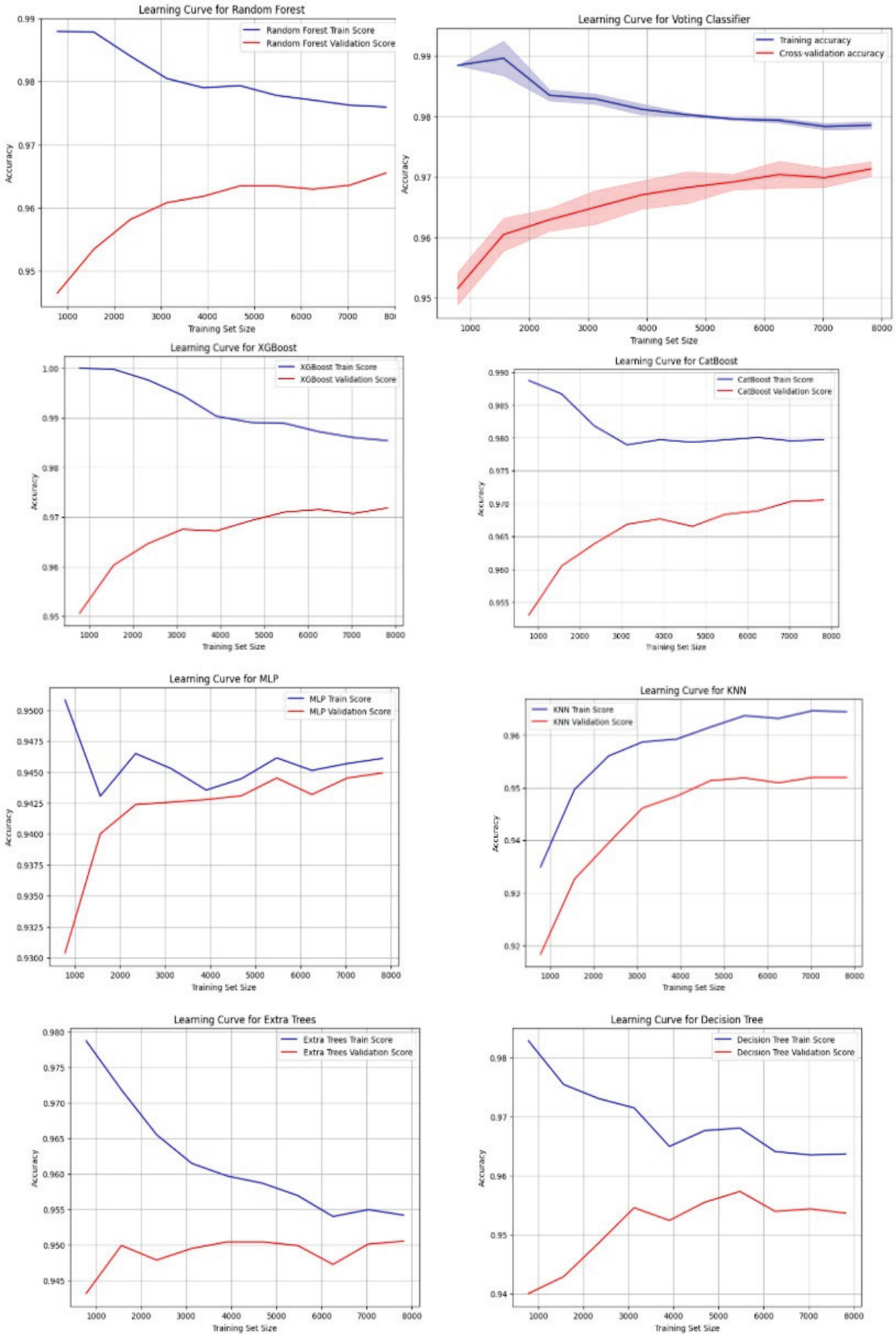


Figure 11. Learning curve of all models.

As shown in Figure 11, XGBoost, CatBoost, and Soft Voting were the most consistent curves because both training and validation accuracy curves rise upward, indicating that they improve in a steady manner during both training and validation. They generalize well between training and validation times. The second is Random Forest, which is close behind with good learning and little overfitting. KNN and MLP perform similarly. Both models are slightly overfitted as they show more variability in validation accuracy. However, the gap between training and validation accuracy is small. Random Forest performs well, too, but it is slightly overfitted more than Extra Trees. Although the Decision Tree model does well on training data, it heavily suffers from overfitting as the model performs poorly on validation data. To know the computational cost of each model, we summarize the training time, memory usage, and inference time of each model in Table 12.

Training times are shown in Table 12, and the Decision Tree is the fastest of them all being 0.0394 s. KNN and Extra Trees also show quick training times of 0.0119 s and 0.6435 s, respectively, with Extra Trees maintaining a good balance between performance and computational cost. Comparatively, CatBoost and XGBoost have very fast training times of 5.20 and 0.6401 s, respectively, whereas MLP needs a vast amount of time (33.72) s. The Voting Classifier, with a training time of 23.64 s, shows competitive performance but with higher computational costs.

Table 12. Model training times, memory usage, and inference time.

Model	Training Time (s)	Memory Usage (MiB)	Inference Time (s)
Random Forest	1.45	0.0000	0.0983
Decision Tree	0.0394	0.0000	0.0023
Extra Trees	0.6435	0.0000	0.0765
KNN	0.0119	0.0000	0.7607
MLP	33.72	10.3164	0.0322
CatBoost	5.20	29.1562	0.0269
XGBoost	0.6401	4.9844	0.0642
Voting	23.64	0.0000	0.3955

All the models have relatively low memory usage in terms of memory used, with the minimum memory populated by the Voting Classifier (0.00 MiB). Similar to that, inference times are primarily low across the board, with Voting Classifier having the highest inference time of 0.3955 s, which is quite efficient. In fact, other models such as KNN (0.7607 s) and XGBoost (0.0642 s) also take inference times that are shorter than other models tested, which shows that all the models are able to provide quick results even on bigger datasets.

4.1.2. Testing Performance

The testing results presented in Table 13 demonstrate that XGBoost obtained 94.35% accuracy along with a 0.9964 AUC score. Moreover, CatBoost performed slightly better with 94.50% accuracy and an AUC score of 0.9965. Random Forest showed similar effectiveness to other classification methods and achieved an accuracy of 94.39% and an AUC score of 0.9958. Also, Decision Tree obtained accuracy results of 94.02% with an AUC score of 0.9927. The accuracy and AUC performance of Extra Trees reached 93.25% and 0.9934, respectively. The performance accuracy for KNN and MLP stood at 92.09%, alongside 92.12% for accuracy and AUC scores of 0.9877 and 0.9942.

A Voting Classifier is an ensemble model that predicts the output based on the prediction of multiple models. With this approach, the performance achieved 94.52% accuracy and an AUC of 0.9960. The results of the test showed that each model is performing well; hence they can generalize to unseen data too. Compared with the other models, the Voting Classifier performed slightly better as it yields the output based on the predictions of other

models. In addition, we evaluated the models based on precision, recall, and F1 score, which are in Table 14. These metrics provide essential information about model instance classification accuracy together with precision–recall equilibrium, which strengthens fault classification evaluation.

Table 13. Model Accuracy and AUC Scores.

Model	Accuracy	AUC Score
Random Forest	94.39%	0.9958
Decision Tree	94.02%	0.9927
Extra Trees	93.25%	0.9934
KNN	92.09%	0.9877
MLP	92.12%	0.9942
CatBoost	94.50%	0.9965
XGBoost	94.35%	0.9964
Voting Classifier	94.52%	0.9960

The precision, recall, and F1 scores for all the models evaluated in six classes are shown in Table 14. All models for class 1 and class 2 have shown perfect performance, scoring 1.0000, which means that they are performing extremely well for those classes. Moreover, Random Forest, Extra Trees, and Voting Classifier perform well in class 0 with F1 scores between 0.93 and 0.94, and KNN has a slightly lower F1 score than others. In addition, for the models that do well in class 3, F1 scores are approximately 0–0.75 (CatBoost, XGBoost, and Voting Classifier), and the worst-performing model is KNN, with the lowest F1 score. Random Forest, CatBoost, and Voting Classifier score more than 0.90 F1 score in class 4, which is substantial. Finally, the Voting Classifier and XGBoost reach almost perfect F1 scores for class 5. In general, the Voting Classifier and CatBoost models yield a balanced and reliable performance on all classes, with stable precision, recall, and F1 scores. Further, for better understanding, below are included the confusion matrix, ROC curve, and precision–recall curve, which provide insights about the models performed in different ways of examining.

Table 14. Precision, recall, and F1 score for different models and classes.

Class	Random Forest			Decision Tree			Extra Trees			KNN		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
0	0.9851	0.9058	0.9438	0.9700	0.9122	0.9402	0.9659	0.9227	0.9438	0.9799	0.8820	0.9284
1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	0.5977	0.9404	0.7308	0.6106	0.8991	0.7273	0.5099	0.9450	0.6624	0.5012	0.9220	0.6494
4	0.8901	0.9709	0.9288	0.8921	0.9407	0.9157	0.9374	0.8523	0.8928	0.8466	0.9153	0.8796
5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9965	0.9982
Class	**MLP**			**CatBoost**			**XGBoost**			**Voting**		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
0	0.9847	0.8808	0.9298	0.9847	0.9070	0.9442	0.9821	0.9074	0.9433	0.9860	0.9066	0.9446
1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	0.4941	0.9587	0.6521	0.6083	0.9404	0.7387	0.6078	0.9312	0.7355	0.6093	0.9587	0.7451
4	0.8494	0.9080	0.8777	0.8904	0.9734	0.9300	0.8886	0.9661	0.9258	0.8901	0.9709	0.9288
5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9982	1.0000	0.9991	1.0000	1.0000	1.0000

As shown in Figure 12, the confusion matrix of each model is presented, which makes it easier to compare the performance of each model. The matrix shows how well each model classifies positive and negative samples as TP, TN, FP, and FN. TP and TN values of the models such as CatBoost, RF, and XGBoost are high, which means that the models are very well performing. Unlike the Decision Trees and KNN, there are slightly higher FP and FN values. The Voting Classifier balances the classification errors slightly better than the others, with fewer misclassifications. Overall, the models RF, CatBoost, and XGBoost outperformed the other models, while the Voting Classifier was able to improve performance slightly.



Figure 12. Confusion Matrix.

Figure 13 shows that most models show strong ROC curves for all classes. However, class 3 is an exception. All models struggle to make accurate predictions for this class. The top performers overall, including class 3, are XGBoost, CatBoost, and the Ensemble Soft Voting Classifier.

In multiclass classification, precision–recall curves are important because they evaluate model performance under imbalanced data situations. While the ROC curve deals with overall discrimination, PR curves actually show how well one trades off between precision (minimizing false positives) and recall (minimizing false negatives) for each one of the classes.

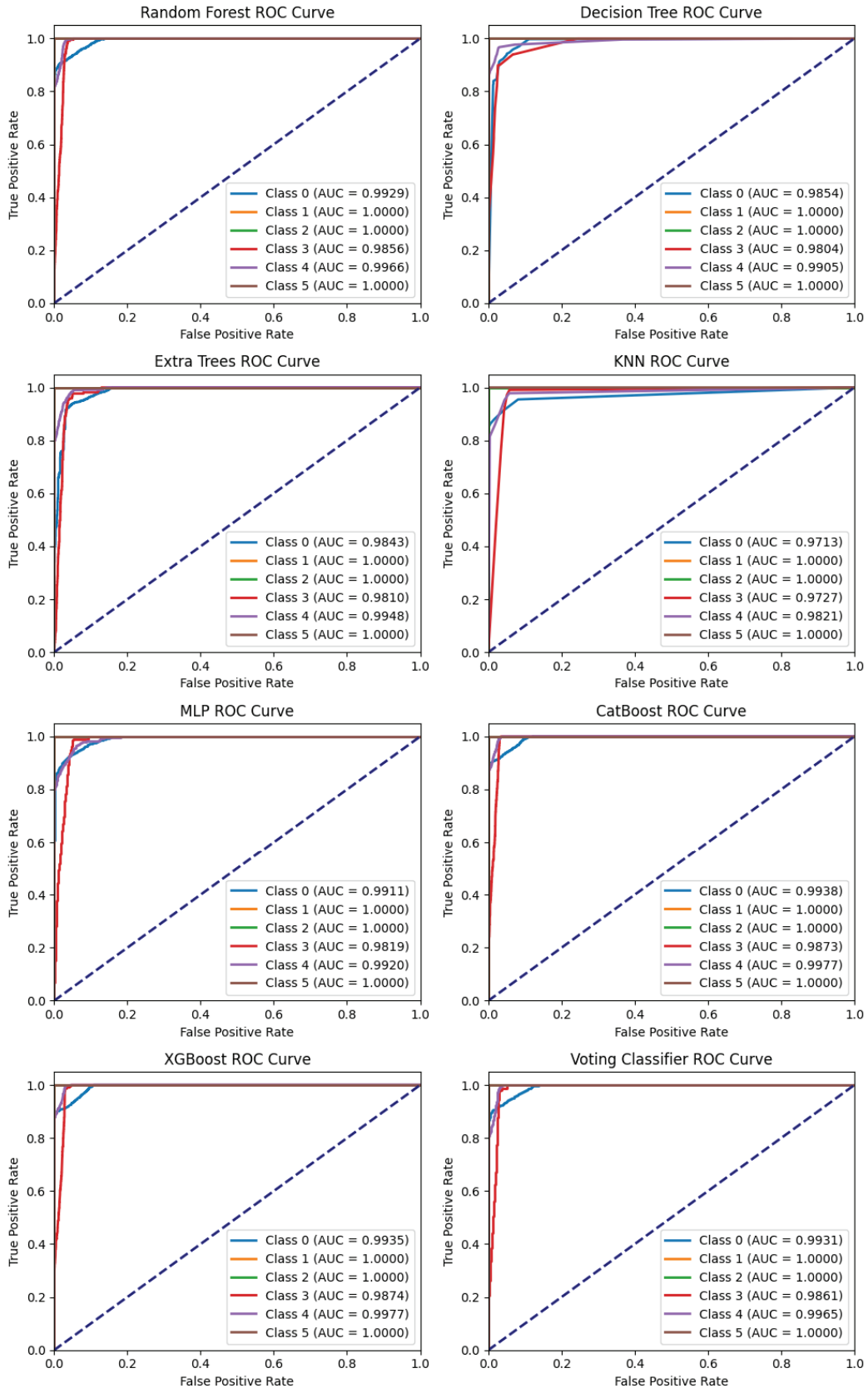


Figure 13. ROC curve of all model.

In Figure 14, it can be said that XGBoost, CatBoost, Extra Trees, and Random Forest have higher values for average precision, thus showing good performance for multiclassification. The Voting Classifier got some improvement in terms of precision and recall from

ensembling. However, the KNN and Decision Tree are more variable and seem to have issues with certain classes, especially in terms of being imbalanced.

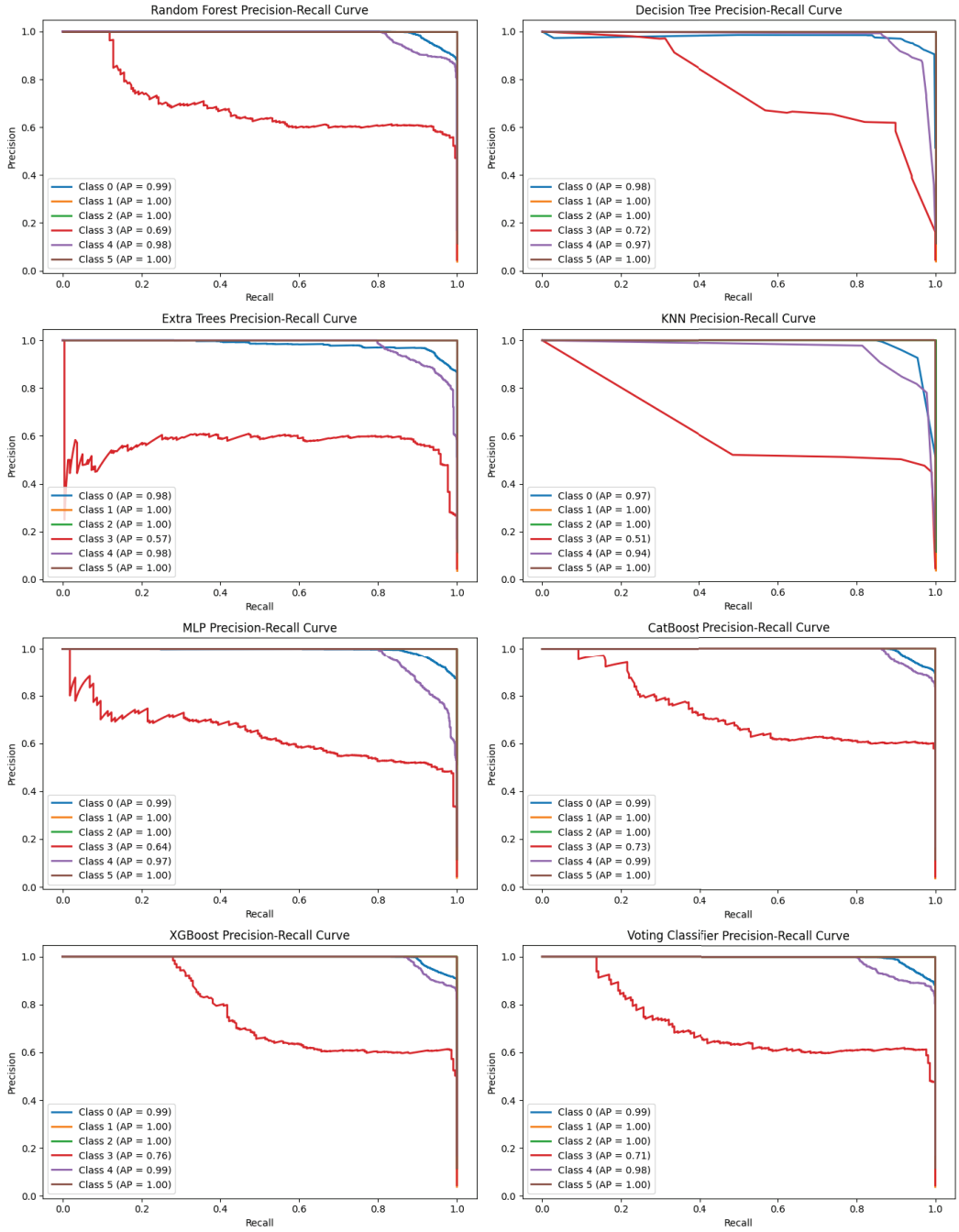


Figure 14. Precision–recall curve of all model.

4.1.3. Assessing Model Reliability with Kappa

It is critical to measure the agreement of the true labels with the models' performance for various machine learning models to fault classification in the electric vehicle drive motors (EVDMs). Cohen's Kappa is one of those and is helpful in a robust evaluation rather than accuracy alone. Kappa scores help us understand the consistency or reliability of classifiers in reality.

Table 15 shows that the Voting Classifier with a result of 0.9210 is the most consistent and agrees most with true labels, making it the most deterministic model to be used for fault classification. Random Forest, with a score of 0.9192; CatBoost, with a score of

0.9206; and XGBoost, with a score of 0.9184, all do extremely well on their models with Kappa scores close to 1, meaning good classification performance and a little bit of random agreement. The Decision Tree, with a score of 0.9133, and Extra Trees, with a score of 0.9021, while slightly lower, still exhibit high reliability and consistency in their predictions. On the other hand, KNN, with a score of 0.8869, and MLP, with a score of 0.8874, were the least consistent and had the highest chance of random agreement to fault classes.

Table 15. Cohen's Kappa Scores.

Model	RF	DT	ET	KNN	MLP	CatBoost	XGBoost	Soft Voting
Kappa Score	0.9192	0.9133	0.9021	0.8869	0.8874	0.9206	0.9184	0.9210

4.1.4. Best Model for Fault Classification in Electric Vehicle Drive Motors

Based on the analysis of training accuracy, cross-validation accuracy, training time, memory usage, inference latency, testing score, Kappa score, AUC, ROC curve, precision–recall curve, and confusion matrix of all other models used in the experiment, it is concluded that the ensemble soft voting model is the most consistent model for fault classification in electric vehicle drive motors. In terms of these key metrics, the soft voting model performed consistently and outperformed other models in generalization effectiveness and in providing reliable results. In addition, the model only took 23 s to train, which made the model highly efficient. Although it is an ensemble model, it had a reasonable memory usage of 0.00 MiB with an inference time of 0.3711 s, which allowed real-time predictions to be made. The ensemble soft voting model is demonstrated to be the best for this task in terms of its performance on all evaluation criteria and computational efficiency.

Below is Table 16 of metrics of the soft voting model.

Table 16. Metrics of the soft voting model.

Metric	Value
Training Accuracy	97.82%
Cross-Validation Accuracy	97.10%
Training Time (Seconds)	23.64
Testing Score	94.52%
AUC	0.9960
Kappa Assessment	0.9210
Memory Usage (MiB)	0.00
Inference Time (Seconds)	0.3955

4.1.5. Interpreting the Best Model's Decisions with XAI

We applied LIME (Local Interpretable Model-agnostic Explanations) to our best-performing ensemble soft voting model to analyze its decision-making process. LIME allowed us to examine the model's predictions for each of the six classes in the dataset. For every class, we observed how the model assigns predictions and interpreted the importance of various features that contributed to those predictions. This detailed analysis provided us with insight into the behavior of the ensemble model, demonstrating how it approaches each class and ensuring consistent and reliable classification across the entire problem set. The interpretability of these predictions for each class is outlined below.

As can be seen in Figure 15, the LIME XAI explanation of class 0 at index one clearly states that the model is very sure that the instance will belong to class 0, and this is with 91% certainty. This decision is mainly driven by rated torque, which in this case is contributing the most, suggesting that the lower torque value is of great importance so the model tends

to shift towards class 1 less. Furthermore, the speed and current support the prediction for class 0, though the influence is less. The voltage itself has a minor impact on class 1, alongside the other features. However, this proves that the main factors, such as rated torque, lead to the prediction converted towards class 0 with high confidence. To conclude, all the features combine to make the predictions of the model, and a combination of negative features outweighs positive features, thus giving a final class of 0.

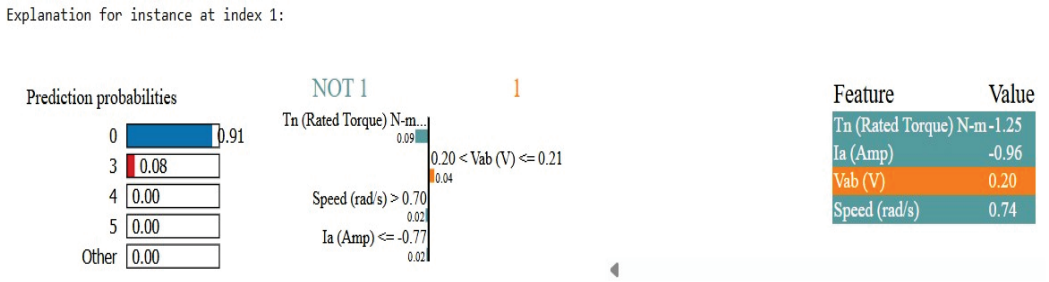


Figure 15. LIME Explanation for prediction class 0.

Figure 16 shows the LIME explanation of class 1 at index 0 in Figure 16, where the model is 100 percent confident that this is the prediction for class 1. The most significant positive contribution to this classification is Rated Torque Tn equal to 1.92. Voltage Vab, Current In, and Speed also influence the prediction but to a lesser extent. Finally, the model's reasoning process indicates that a high-rated torque would immediately lead to class 1 and does not ask for additional evaluation of other features. This implies that torque is an important contributor to classification and other features have a secondary contribution to this classifier.

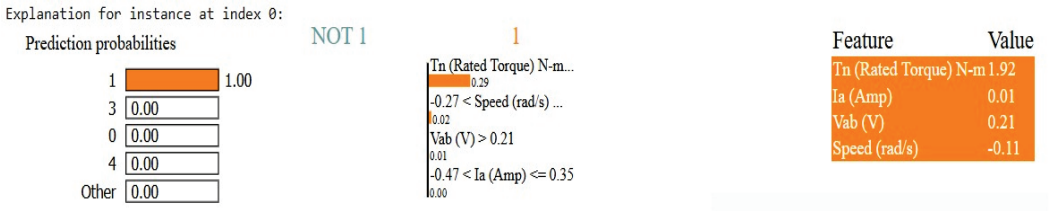


Figure 16. LIME Explanation for prediction 1.

The LIME explanation for class 2 at index 5 in Figure 17 indicates that the model is 100% sure that the instance should be of class 2. The main reason this is the case is because of the speed value, which ranges from typical class 2 of 0.86. This also supports the rated torque of −0.34, which is in class 2 because lower torque values are common in this class. Subsequently, the current of −0.83 cannot be related to the condition required for class 1, and hence, class 1 is also rejected. Looking at the voltage of 1.52 pointing to class 2 as well as 0.21 points to class 2. Overall, these values are good characteristics for the model in predicting class 2, with speed, torque, current, and voltage all being important for such a decision.

In Figure 18, it is shown that the model predicted class 3, for instance, 22, with 85% confidence. The reason for this prediction is a high-rated torque, supporting class 3. This decision also depends on speed, however, not as much. As the current and voltage slightly push the prediction away from class 3, they have a slight negative impact. Even so, rated torque and speed have a much stronger effect than this result. This model has predicted a class of 0 with a 14% probability and all other classes with a probability of 0. It's the best match for this instance and doesn't fit into other classes; as such, this confirms that class 3

is the correct one. Figure 4 illustrates overall that the model is able to classify an instance as class 3 when rating torque and speed are high and a small effect of voltage and current.

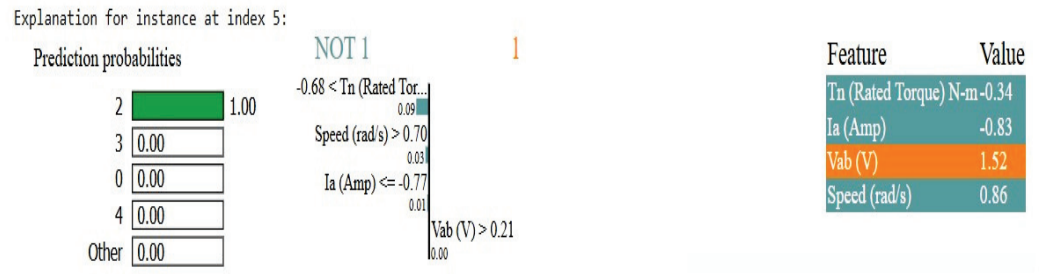


Figure 17. LIME Explanation for prediction 2.

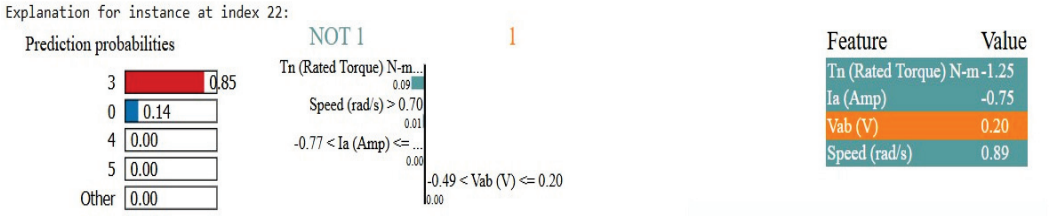


Figure 18. LIME Explanation for prediction 3.

Figure 19, it shows the model predicts class 4 with 100% confidence for index 12. This decision is mainly based on the low-rated torque (−0.12), as this supports class 4. Not only that, speed (−1.14) also plays a crucial role, as its negative value also fits the pattern of class 4. Another reason the model selects class 4 is that 2.12 > 0.35, which is current (Ia). This is classified as voltage (Vab) is −0.49, which is within some given range that also verifies this classification. Since the probabilities of all other classes are 0%, the model can rule them out completely. Figure 5, overall, indicates this as being class 4 with absolute certainty through low-rated torque, low speed, high current, and specific voltage.

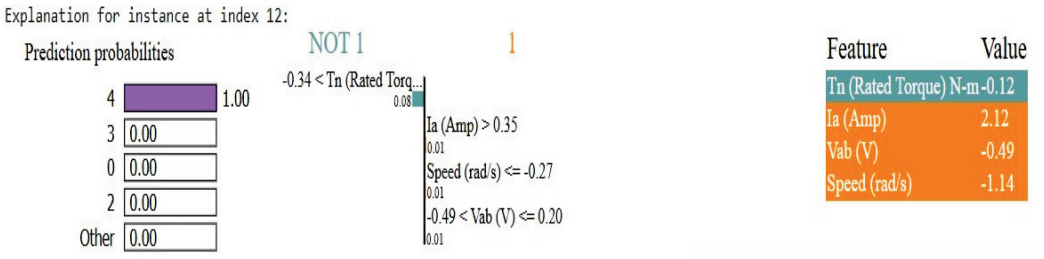


Figure 19. LIME Explanation for prediction 4.

In Figure 20, we see that instance 2 has 100% confidence of being decided as class 5 by the model. So in other words, this result is sure that this instance is of class 5 and has no other possible prediction. Given the rated torque value of −1.25, this prediction is the most important number and clearly supports class 5. This is also a factor that contributes to this classification at the voltage (Vab) of −1.06. This smaller positive contribution is given by the current (Ia) of −0.77. The model's decision is relatively speedless (0.25), having little (or no) effect on whether the instance is from class 5. The model rules out all other classes since their probabilities are 0%. Figure 6 shows overall that for this instance—specifically for values of rated torque and specific voltage and current values that are low along a neutral speed—the combination of all of these values confirms this was class 5 with absolute certainty.

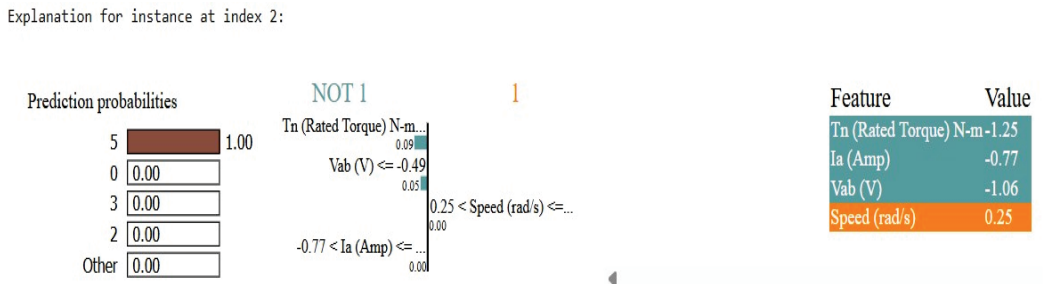


Figure 20. LIME Explanation for prediction 5.

4.2. Experiment 2: Drive End Experimental Dataset

Experiment 1 was conducted previously on a MATLAB-simulated dataset. In this experiment, a dataset from Zenodo is used with a total of nine classes to further validate the framework's effectiveness. The data preparation, model training, evaluation, and performance comparison in this experiment are exactly the same as in Experiment 1, so there will be no dependency on the randomness of these steps applied.

Table 17, various machine learning models are compared in terms of training accuracy, cross-validation accuracy, training time, inference time, and memory usage. The training result demonstrates that XGBoost has the highest training accuracy of 99.94% and a cross-validation of 99.45%, which shows it has strong generalization performance. Furthermore, trained with the Voting Classifier, which consists of several models, the system achieves high accuracy, achieving 99.79% in training and 99.40% in cross-validation. However, this compromises the training time, which is much higher at 103.94 s and has the highest memory usage of 2.64 MB. CatBoost and Random Forest also exhibit competitive performance, reach high estimation accuracy, and train fairly well, which makes them natural deployment candidates.

Table 17. Comparison of model performance metrics.

Model	Training Accuracy	CV Accuracy	Training Time (s)	Inference Time (s)	Memory Usage (MB)
Random Forest	99.59%	99.36%	5.08	0.3849	0.00
Decision Tree	99.68%	99.04%	0.23	0.0102	0.00
Extra Trees	98.43%	98.28%	2.37	0.8188	0.00
KNN	99.09%	98.44%	0.17	3.5606	0.00
MLP	99.29%	99.11%	30.65	0.0495	0.0000
CatBoost	99.62%	99.28%	21.65	0.1132	0.0000
XGBoost	99.94%	99.45%	2.1737	0.2815	0.0000
Voting	99.79%	99.40%	103.94	0.3589	2.6406

Models such as Decision Tree and Extra Trees achieve great accuracy, but their generalization performance is not so good as ensemble model-based models. KNN is the slowest, as 3.56 s is its longest inference time, and it might experience issues with real-time applications (despite the training phase being fast). The deep learning-based approach used with MLP helps balance the accuracy and adequacy in computation time, but it has a higher training time than traditional machine learning models.

Results from this experiment suggest that the proposed framework works well on different datasets and has robustness in multi-class classification problems. Further strengthening the validity of the approach is the transition from a simulated dataset to a real-world dataset with nine classes as the first transition from a simulated dataset to a real-world dataset with scenarios. Since the framework has been proven to work on Experiment 1, the consistency of the framework remains the same, allowing it to work similarly on various datasets.

Table 18 also presents the testing consistency report that confirms the training of models using the provided data is indeed reliable by testing it on unseen data. Finally, results justify the applicability of the framework in a real-world scenario.

Table 18. Model performance on test set.

Model	Accuracy	AUC Score	Testing Time (s)	Kappa Score
Random Forest	98.67%	0.9998	0.0878	0.9830
Decision Tree	98.11%	0.9913	0.0049	0.9760
Extra Trees	97.43%	0.9992	0.1065	0.9672
KNN	96.41%	0.9879	0.3040	0.9544
MLP	97.84%	0.9987	0.0119	0.9724
CatBoost	98.62%	0.9996	0.0256	0.9824
XGBoost	98.71%	0.9997	0.0548	0.9836
Voting Classifier	98.48%	0.9997	0.1742	0.9976

The model performance metrics on the test set are presented in Table 18, which offers a strong evaluation of the models' ability to generalize to unseen data. The accuracy, AUC score, testing time, and Kappa score, which measures the agreement between predicted and real classification adjusting for chance level accuracy, are the key metrics of this.

It can be seen from the table that XGBoost has the highest accuracy (98.71%) and AUC score (0.9997), which is very good for discriminating between classes in the data. On the other hand, the Voting Classifier, which is created by combining several models, has the highest Kappa Score (0.9976), indicating a very good agreement between the predicted and actual class labels, and therefore has the best overall consistency. The test performance of Random Forest and CatBoost also shows good results with accuracy values of 98.67% and 98.62%, respectively, and high kappa scores.

The Decision Tree performs slightly lower than the ensemble models, but the testing time of 0.0049 s is quite low, and thus it is a good choice for real-time applications. On the contrary, although KNN achieves an accuracy of 96.41%, a testing time of 0.3040 s is the highest among all algorithms, which can be a bottleneck for time-critical prediction. Deep learning powers MLP to maintain a balance between Kappa Score and inference efficiency, achieving 0.9724 and having a small testing time.

Overall, the Voting Classifier as well as XGBoost, CatBoost, and Random Forest provide high predictive power as well as robustness. Most Kappa Scores are high across all models, which further reinforces the same reliability and stability of the proposed framework applied to a multi-class classification problem. Consistency between datasets validates the framework's adaptability and the applicability of it to real-world scenarios.

We further add the Confusion Matrix, ROC curve, and precision–recall curve in Figure 21 to have a clearer look at the XGBoost model's performance.

Figure 21 shows that XGBoost has a few misclassifications, whose ROC and precision–recall curves remain good, which indicates strong classification ability.

Finally, on the experimental data, XGBoost proved to be the best model after evaluating all performance metrics. It has a near-perfect (98.71%) accuracy, AUC score, and kappa score, ensuring strong predictive and consistent power. It also provides reasonable training (2.17 s) and inference (0.2815 s) times with minimal memory usage. Moreover, its high training accuracy (99.94%) and validation accuracy (99.45%) show how reliable this model is for this task.

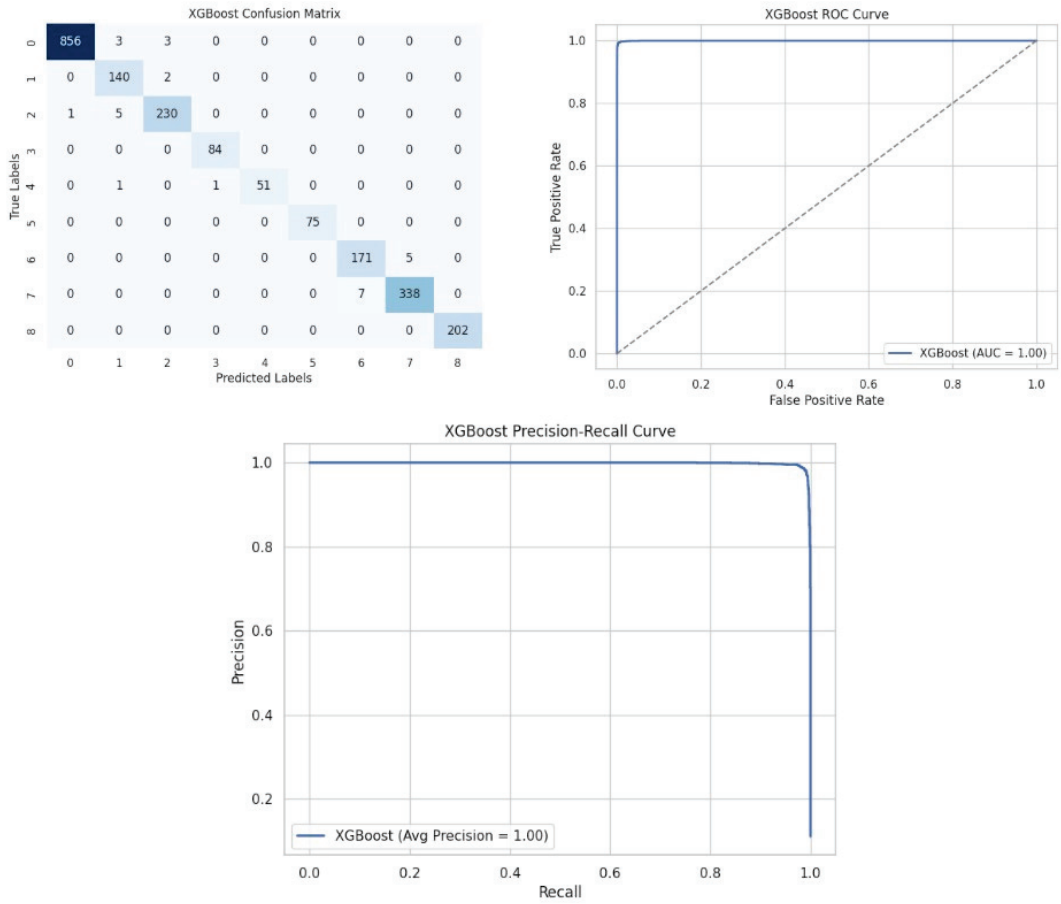


Figure 21. Confusion Matrix, ROC curve, and precision–recall curve.

4.3. Comparative Analysis with Previous Studies

We performed fault classification on an electric vehicle (EV) motor using a MATLAB simulated dataset for a fault classification. Thus, we compare our study to other studies that used simulated datasets for fault classification. Comparisons to these are used to determine what level of effectiveness our method offers in terms of prior studies.

A comparative analysis and key performance metrics of prior studies, including such parameters as accuracy, characteristics of the dataset, and classification techniques used in prior studies, have been presented in Table 19. Thus, it facilitates clear evaluation of variations and improvements of methodologies.

Table 19. Comparison of fault classification studies on electric vehicle motors.

Study	Dataset Type	Best Model	Accuracy (%)	Validated on Another Dataset	XAI Used?	Publishing Year
[8]	Simulated	CatBoost	94.1	No	No	2024
[12]	Simulated	Decision Tree	92	No	No	2024
Our Study	MATLAB Simulated	Soft Voting Ensemble	94.52	Yes (observe consistent performance)	Yes	None

This Table 19 provides a comparison of the fault classification studies in the domain of electric vehicle motors. In our research, we employ the same simulated dataset that has been used by [8]. Our study surpasses this result, having an accuracy of 94.52%, while

their study accuracy is 94.1%. We surpass the method described by [12], which used a comparable dataset to ours in all aspects of performance. Our Soft Voting Ensemble model stands apart from previous works since it delivers better performance results. The study incorporates Explainable AI (XAI) techniques to make the model more understandable while performing predictive maintenance operations. Lastly, to validate our findings, we also validate the model on another dataset, and we observe the robustness and effectiveness of our approach. Such an ensemble of methods, along with XAI, offers an approachable and transparent alternative to previous work in this area and achieves better accuracy and consistency.

5. Conclusions

In this study, we developed the EnsembleXAI-Motor framework for lightweight and efficient classification of faults in electric vehicle drive motors. With such techniques as statistical analysis, preprocessing, recursive feature elimination with cross-validation (RFE-CV), model tuning, and the hybrid soft voting classifier, we successfully obtained a test accuracy of 94.52% with a Kappa score of 0.9210. Local Interpretable Model-Agnostic Explanations (LIME) were incorporated to make the model's decision-making process more transparent, thus making it very applicable for real-world applications where understanding the model's predictions is necessary. In addition, the framework has low inference time and invokes tiny memory usage, which allows it to be used in real-time applications. In addition, we validated the performance of our framework using another real-world dataset and found that it had consistent performance, and thus it validated the robustness and generalizability of our framework across different conditions. The validation we have performed on the proposed framework further validates it as effective also in a real-world dataset, not only in a simulated dataset.

Our proposed framework not only improves diagnostic performance but also solves the problems of computational efficiency and model interpretability. In the fast-changing electric vehicle sector, there is a great need to detect faults timely and accurately, which is particularly important for improving safety and reliability.

Lastly, future work will focus on deploying the model on vehicle control units or IoT-based monitoring systems for real-time fault detection. Also, deep learning model integration will increase the accuracy, and continuous and transfer learning will boost adaptability to different motor types and fault patterns. Such advancements will facilitate scalability to large datasets and wider adoption to a larger range of motor systems.

Author Contributions: Conceptualization, M.E.H., M.Z. and J.U.; methodology, M.E.H., M.Z. and J.U.; software, M.E.H. and M.Z.; validation, M.E.H., M.Z. and J.U.; formal analysis, M.Z. and J.U.; investigation, M.E.H. and M.Z.; resources, J.U.; data curation, M.E.H. and J.U.; writing—original draft preparation, M.E.H.; writing—review and editing, M.E.H., M.Z. and J.U.; visualization, M.E.H. and M.Z.; supervision, J.U.; project administration, M.Z.; funding acquisition, J.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by Woosong University Academic Research 2025.

Data Availability Statement: The data supporting the findings of this study are available in the repository for the experiment "Machine Learning-Based Fault Diagnosis of Electric Drives" at https://github.com/HassanMahmoodKhan/Machine-Learning-Based-Fault-Diagnosis-of-Electric-Drives/tree/main/Data (accessed on: 17 February 2025). Additionally, the experimental dataset can be accessed at https://zenodo.org/records/14482932 (6 April 2025).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

KNN	K-Nearest Neighbors
MLP	Multi-Layer Perceptron
ET	Extra Trees
DT	Decision Trees
PRE	Precision
REC	Recall
F1	F1 Score
XAI	Explainable Artificial Intelligence
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative
SMOTE	Synthetic Minority Over-sampling Technique
RFECV	Recursive Feature Elimination with Cross-Validation

References

1. Rho Motion Ltd. Global EV Sales Grow by 18% in 2025 vs. 2024. 2025. Available online: https://rhomotion.com/news/global-ev-sales-grow-by-18-in-2025-vs-2024/ (accessed on 7 March 2025).
2. The EV Report. January 2025 Global EV Sales Surge. News Article. Available online: https://rhomotion.com/news/global-ev-sales-up-50-in-february-2025/ (accessed on 7 March 2025).
3. Rho Motion. Rho Motion View: Over 20M EVs to Be Sold Globally in 2025. News Article. Available online: https://rhomotion.com/news/rho-motion-view-over-20m-evs-to-be-sold-globally-in-2025/#:~:text=Rho%20Motion%2C%20the%20leading%20electric,20%20million%20electric%20vehicles%20sold (accessed on 7 March 2025).
4. Virta. The Year of the EV: What 2025 Holds for Electric Vehicles and eMobility. Virta Global Blog, 2025. Available online: https://www.virta.global/blog/the-year-of-the-ev-what-2025-holds-for-electric-vehicles-and-emobility#:~:text=Analysts%20predict%20that%202025%20will,2024)%20%2D%20a%20huge%20milestone (accessed on 7 March 2025).
5. Khaneghah, M.Z.; Alzayed, M.; Chaoui, H. Fault Detection and Diagnosis of the Electric Motor Drive and Battery System of Electric Vehicles. *Machines* **2023**, *11*, 713. [CrossRef]
6. Boosters, E. Outlook Overall Electric Vehicles Sales Europe 2025, +40% to 2.7 Million EVs. *EV Boosters*, 2025. Available online: https://evboosters.com/ev-charging-news/outlook-overall-electric-vehicles-sales-europe-2025-40-to-2-7-million-evs/ (accessed on 7 March 2025).
7. Siraj, F.M.; Ayon, S.T.K.; Samad, M.A.; Uddin, J.; Choi, K. Few-Shot Lightweight SqueezeNet Architecture for Induction Motor Fault Diagnosis Using Limited Thermal Image Dataset. *IEEE Access* **2024**, *12*, 50986–50997. [CrossRef]
8. Thirunavukkarasu, S.; Karthick, K.; Aruna, S.K.; Manikandan, R.; Safran, M. Optimized Fault Classification in Electric Vehicle Drive Motors Using Advanced Machine Learning and Data Transformation Techniques. *Processes* **2024**, *12*, 2648. [CrossRef]
9. Ahmad, F.; Ahsan, S.; Kumar, A.; Sarwer, G.; Sonu, S. Electric Motor Fault Detection using Artificial Intelligence. *J. Electr. Syst.* **2024**, *20*, 175–180.
10. Ul Islam Khan, M.; Pathan, M.I.H.; Mominur Rahman, M.; Islam, M.M.; Arfat Raihan Chowdhury, M.; Anower, M.S.; Rana, M.M.; Alam, M.S.; Hasan, M.; Sobuj, M.S.I.; et al. Securing Electric Vehicle Performance: Machine Learning-Driven Fault Detection and Classification. *IEEE Access* **2024**, *12*, 71566–71584. [CrossRef]
11. Abdulkareem, A.; Anyim, T.; Popoola, O.; Abubakar, J.; Ayoade, A. Prediction of induction motor faults using machine learning. *Heliyon* **2025**, *11*, e41493. [CrossRef] [PubMed]
12. Samiullah, M.; Ali, H.; Zahoor, S.; Ali, A. Fault Diagnosis on Induction Motor using Machine Learning and Signal Processing. *arXiv* **2024**, arXiv:2401.15417v1.
13. Bhattacharya, D.; Nigam, M.K. Energy efficient fault detection and classification using hyperparameter-tuned machine learning classifiers with sensors. *Meas. Sens.* **2023**, *30*, 100908. [CrossRef]
14. Yaqub, R.; Ali, H.; Bin Abd Wahab, M.H. Electrical Motor Fault Detection System Using AI's Random Forest Classifier Technique. In Proceedings of the 2023 IEEE International Conference on Advanced Systems and Emergent Technologies (IC_ASET), Hammamet, Tunisia, 29–31 May 2023; pp. 1–5. [CrossRef]
15. Kim, M.C.; Lee, J.H.; Wang, D.H.; Lee, I.S. Induction Motor Fault Diagnosis Using Support Vector Machine, Neural Networks, and Boosting Methods. *Sensors* **2023**, *23*, 2585. [CrossRef] [PubMed]

16. Zereen, A.N.; Abir Das, J.U. Machine Fault Diagnosis Using Audio Sensors Data and Explainable AI Techniques-LIME and SHAP. *Comput. Mater. Contin.* **2024**, *80*, 3463–3484. [CrossRef]
17. Xu, L.; Teoh, S.; Ibrahim, H. A deep learning approach for electric motor fault diagnosis based on modified InceptionV3. *Sci. Rep.* **2024**, *14*, 12344. [CrossRef] [PubMed]
18. Khan, H.M. Machine Learning-Based Fault Diagnosis of Electric Drives. Dataset. Available online: https://github.com/HassanMahmoodKhan/Machine-Learning-Based-Fault-Diagnosis-of-Electric-Drives/blob/main/Data/All%20Data.xlsx (accessed on 7 March 2025).
19. Bacha, A.; Idrissi, R.E.; Idrissi, K.J.; Lmai, F. Comprehensive dataset for fault detection and diagnosis in inverter-driven permanent magnet synchronous motor systems. *Data Brief* **2025**, *58*, 111286. [CrossRef] [PubMed]
20. Komorowski, M.; Marshall, D.; Salciccioli, J.; Crutain, Y. *Exploratory Data Analysis*; Springer: Cham, Switzerland, 2016; pp. 185–203. [CrossRef]
21. Kim, H.Y. Statistical notes for clinical researchers: Assessing normal distribution (2) using skewness and kurtosis. *Restor. Dent. Endod.* **2013**, *38*, 52–54. [CrossRef] [PubMed]
22. Sedgwick, P. Pearson's correlation coefficient. *BMJ* **2012**, *345*, e4483. [CrossRef]
23. Menze, B.H.; Kelm, B.M.; Masuch, R.; Himmelreich, U.; Bachert, P.; Petrich, W.; Hamprecht, F.A. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinform.* **2009**, *10*, 213. [CrossRef] [PubMed]
24. Fan, C.; Chen, M.; Wang, X.; Wang, J.; Huang, B. A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *Front. Energy Res.* **2021**, *9*, 652801. [CrossRef]
25. Whaley, D.L., III. The Interquartile Range: Theory and Estimation. Master's Thesis, East Tennessee State University, Johnson City, TN, USA, 2005.
26. Ross, A.; Willson, V.L. One-way anova. In *Basic and Advanced Statistical Tests: Writing Results Sections and Creating Tables and Figures*; Sense Publishers: Rotterdam, The Netherlands, 2017; pp. 21–24.
27. Scikit-Learn Yellowbrick Contributors. Recursive Feature Elimination, Cross-Validated (RFECV). Scikit-Learn Yellowbrick Documentation. Available online: https://www.scikit-yb.org/en/latest/api/model_selection/rfecv.html (accessed on 28 March 2025).
28. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
29. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv* **2016**, arXiv:cs.LG/1602.04938.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Surface Classification from Robot Internal Measurement Unit Time-Series Data Using Cascaded and Parallel Deep Learning Fusion Models

Ghaith Al-refai *, Dina Karasneh, Hisham Elmoaqet, Mutaz Ryalat and Natheer Almtireen

Department of Mechatronics Engineering, German Jordanian University, Amman 11180, Jordan; dina.karasneh@gju.edu.jo (D.K.); hisham.elmoaqet@gju.edu.jo (H.E.); mutaz.ryalat@gju.edu.jo (M.R.); natheer.almtireen@gju.edu.jo (N.A.)
* Correspondence: ghaith.alrefai@gju.edu.jo

Abstract: Surface classification is critical for ground robots operating in diverse environments, as it improves mobility, stability, and adaptability. This study introduces IMU-based deep learning models for surface classification as a low-cost alternative to computer vision systems. Two feature fusion models were introduced to classify the surface type using time-series data from an IMU sensor mounted on a ground robot. The first model, a cascaded fusion model, employs a 1-D Convolutional Neural Network (CNN) followed by a Long Short-Term Memory (LSTM) network and then a multi-head attention mechanism. The second model is a parallel fusion model, which processes sensor data through both a CNN and an LSTM simultaneously before concatenating the resulting feature vectors and then passing them to a multi-head attention mechanism. Both models utilize a multi-head attention mechanism to enhance focus on relevant segments of the time-sequence data. The models were trained on a normalized Internal Measurement Unit (IMU) dataset, with hyperparameter tuning achieved via grid search for optimal performance. Results showed that the cascaded model achieved higher accuracy metrics, including a mean Average Precision (mAP) of 0.721 compared to 0.693 for the parallel model. However, the cascaded model incurred a 44.37% increase in processing time, which makes the parallel fusion model more suitable for real-time applications. The multi-head attention mechanism contributed significantly to accuracy improvements, particularly in the cascaded model.

Keywords: time-series data; LSTM; 1-D CNN; CNN-LSTM; multi-head attention mechanism; IMU; feature fusion; IMU; robot navigation

1. Introduction

With rapid advances in technology, robotics has expanded into almost every field, including education [1], business [2], and healthcare [3]. Terrain detection and classification play a vital role in enabling precise robot control and navigation. Various studies emphasize the advantages of surface classification, such as improving the mobility of walking robots [4], addressing challenges in using intelligent tires based on accelerometers for enhanced mobility [5], and adjusting robot speed according to surface type [6]. Surface detection further helps stabilize legged robots in rough terrain by improving foot placement [7], and supports energy-efficient operating modes [8]. In addition, surface classification can be used to estimate road friction coefficients, contributing to safety improvements [9].

The most common approach to surface classification involves the use of computer vision systems. For example, Iglesias et al. [10] developed a camera-based computer vision system

specifically for wood surface classification. Laible et al. [11] presented a robot equipped with a low-resolution 3D LIDAR and color camera to classify the terrain ahead of it. Walas et al. [12] proposed a terrain classification method using a laser rangefinder, where feature vectors were generated through statistical descriptors of the intensity map, with a support vector machine serving as a classifier. Borrmann et al. [13] introduced an autonomous mobile robot that uses both an RGB camera and a thermal camera. An automated 3D crack detection system for structures, achieving high accuracy by fusing LiDAR and a depth camera, was proposed by Hu et al. [14]. Tang et al. proposed visual crack width measurement based on structure backbone double-scale features using machine vision [15].

Inertial Measurement Unit (IMU) sensors are widely utilized in robotics, playing an essential role in various robot applications. For instance, they aid in localization and slip estimation for mobile robots [16], determining the trajectory and direction [17,18], estimating the pose [19], and navigating and localization [20], as well as motion stability control and gait planning [21].

Computer vision faces challenges including sensitivity to lighting variations, high computational demands, and complexities in data acquisition. Traditional vision-based methods rely on extensive labeled datasets and require intricate pre-processing steps [22]. This study addresses the need for a cost-effective approach to surface classification in robotics by utilizing IMU data as an alternative to traditional, higher-cost computer vision systems. Specifically, it utilized two deep learning models—1-D Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) networks—to extract meaningful features from time-series IMU data. To further enhance feature quality and improve classification accuracy, a multi-head attention mechanism is incorporated. The study investigates two distinct feature fusion architectures: a cascaded model, where CNN and LSTM are sequentially connected, and a parallel model, where features from both networks are combined simultaneously. Through a comprehensive analysis of classification accuracy and execution runtime, this work compares the strengths and trade-offs of each architecture to offer insights into optimal configurations for low-cost, efficient surface classification in robotic applications.

The paper is organized as follows: Section 2 reviews related work on IMU data applications in robotics using machine learning and deep learning. Section 3 details the proposed model architectures, describing key components and the cascaded and parallel fusion models. Section 4 covers the dataset structure, classes, and the training process. Section 5 explains the evaluation metrics. Section 6 presents classification accuracy and runtime results. Section 7 discusses the findings, including performance trade-offs based on the results. Finally, Section 8 concludes the study by summarizing key insights and suggesting future research directions.

2. Related Work

Various machine learning techniques have been applied to sensor data classification. Giulio et al. [23] investigated the use of decision tree ensembles to detect and localize structural damage in health monitoring systems using vibration sensor data. Kumar et al. [24] developed a human activity recognition system based on smartphone sensor data, employing a random forest algorithm to classify different activities. Al-Refai et al. [25] introduced a system for detecting potholes and analyzing traffic conditions using in-vehicle data, such as speed and acceleration, and tested multiple machine learning algorithms, including Support Vector Machines (SVMs), decision trees, and random forests. Barman and Choudhury [26] applied SVM to classify soil texture based on hydrometer test data. Machine learning algorithms were used in soft robot motion classification [27]. Shirmard et al. [28] provided a comprehensive review of machine learning approaches for processing remote sensing data in mineral exploration.

Recent studies in deep learning have focused on exploring sensor data classification, particularly Recurrent Neural Networks (RNNs), due to their ability to process time-series data effectively. Gupta et al. [29] utilized Convolutional Neural Networks (CNNs) for human activity recognition using data from a single wearable IMU sensor. Qi et al. [30] introduced an LSTM-RNN model comprising a Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) module, and a dropout layer for classifying multiple hand gestures. Zhu et al. [31] proposed a radar-based human activity recognition system that combines 1-D CNNs and RNNs to analyze radar spectrograms as time-sequential vectors. Al-refai et al. [32] utilized CNN architecture to classify traffic conditions and driving behavior using OnBoard diagnostics data (OBD) and smartphone data. Mekruksavanich et al. [33] conducted a comparative study on the application of LSTM networks for human activity recognition using smartphone sensor data, employing a four-layer CNN-LSTM hybrid architecture. Similarly, Pham et al. [34] investigated LSTM networks for the classification of physiological signals.

The multi-head attention mechanism has been widely employed for data classification tasks. Matar et al. [35] introduced a novel multi-head attention-based Bi-LSTM approach for anomaly detection in multivariate time-series data from wireless sensor networks. Junior et al. [36] designed a multi-head 1-D CNN to identify six types of malfunctions in electric motors using data from two accelerometers. Similarly, Cui et al. [37] proposed a multi-head attention CNN model to enhance fault classification performance in industrial processes by prioritizing the significance of different parameters.

Several research projects have investigated surface classification utilizing time-series deep learning algorithms. For example, Li et al. [38] proposed a surface classification method using a one-dimensional Convolutional Neural Network (1D CNN) followed by a Long Short-Term Memory (LSTM) network, demonstrating superior detection performance compared to baseline algorithms such as XGBoost and Fully Connected Neural Networks (FCNs). Alradaideh et al. [39] further enhanced classification accuracy by employing a CNN with a Bidirectional LSTM (Bi-LSTM) architecture. Feng et al. [40] explored an alternative approach by treating the time series data as an input image for CNN-based feature extraction, followed by LSTM for final feature processing. Jiang et al. [41] introduced a customized convolutional neural network that effectively extracts features from both time and frequency domain representations of the data. This image-based CNN approach outperformed traditional time series algorithms, including decision trees, 1D CNN, and LSTM.

While these studies effectively leverage CNN-LSTM architectures for surface classification, they primarily focus on sequential CNN-LSTM configurations. Furthermore, a critical aspect often overlooked in previous research is the runtime performance and suitability of these algorithms for real-time applications. This study aims to address these limitations by: (1) investigating the impact of different feature fusion strategies, namely cascaded and parallel fusion of 1D CNN and LSTM features; and (2) incorporating a multi-head attention mechanism to enhance model performance, given its limited exploration in IMU time series classification. Additionally, this research comprehensively analyzes the runtime requirements of both proposed models, facilitating a thorough evaluation of their trade-offs in terms of accuracy and computational efficiency.

3. Proposed Deep Learning Model Architecture

Two deep learning models were developed and evaluated in this work to classify the IMU time series data. A one-dimensional (1-D) CNN network, LSTM network, multi-head attention mechanism, and fully connected layer with drop out and softmax layer are utilized to build both models. The key distinction between the two models is how the CNN features

and the LSTM features are combined. Section 3.1 explains the model building blocks and algorithms and the architecture of the cascaded feature fusion model and the parallel feature fusion model.

3.1. Model Building Blocks

This section explains the main algorithms we utilized in this work to build our deep learning models for the IMU time series data classification.

3.1.1. 1-D CNN

The One-Dimensional Convolutional Network (1-D CNN) is designed to process sequential data [42]. It is commonly used to extract high-level features from time-series data. In a 1-D CNN, the input sequence is convoluted with a one-dimensional sliding filter. When padding is applied to the input sequence, the output sequence retains the same sequence length. In our implementation, we applied two convolutional layers, consisting of 128 and 64 filters, respectively, with a kernel size of 3×1 to capture relevant features. The first layer with 128 neurons ensures that all input sequence features are initially processed without dimensional reduction. The second layer follows a structured reduction approach, where the model progressively compresses and refines the feature representation while maintaining essential information.

3.1.2. LSTM

Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) to deal with time series data [43]. The main advantage of the LSTM over the traditional RNN is its ability to deal with long sequences without having vanishing and exploding gradient problems [44]. The LSTM network processes the current input while also considering previous states. As illustrated in Figure 1, the previous states, $c^{(t-1)}$ and $y^{(t-1)}$, represent the long-term memory and the short-term memory respectively, while $x^{(t)}$ represents the current input data. LSTM includes three stages. The first stage is the forget gate, which decides what percentage of long-term memory to remember. The second stage is the input gate, in which the long-term memory is updated based on the previous states and the current input. The last stage is the output gate, which calculates the outputs of the LSTM $c^{(t)}$ and $y^{(t)}$, which become the input states for the next iteration. $y^{(t)}$ from the last time step represents the final output of the LSTM unit.

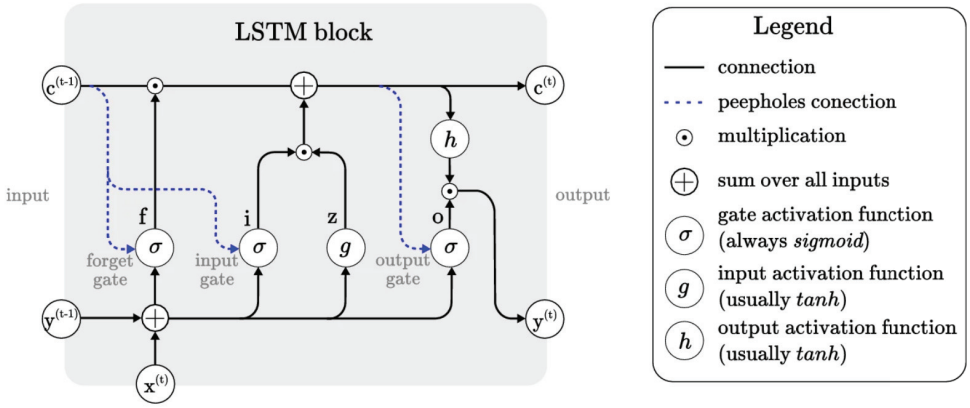


Figure 1. Architecture of Long-Short Term Memory (LSTM) unit [43].

3.1.3. Multi-Head Attention Mechanism

The multi-head attention mechanism is an attention module where the time series input data is divided into multiple partitions and processed in parallel [45]. This mechanism enhances the extraction of richer information from the input by capturing dependencies

and relationships across different segments of the sequence. By utilizing multiple attention heads, the module can simultaneously identify and capture diverse patterns within time-series data.

The input data for each head are projected into three spaces, Queries (Q), Keys (K), and Values (V). Q represents a specific time interval in the input data to be focused on. K represents the information about a specific time step in the sequence to identify relationships and dependencies in the time sequence. V Contains the actual values associated with each time step the model will use for module weights update. The architecture of the multi-head attention module is illustrated in Figure 2. The attention mechanism is mathematically defined by Equation (1), and the output of the multi-head attention unit is expressed in Equation (2).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where:

Q = Query matrix;
K = Key matrix;
V = Value matrix;
d_k = Dimensionality of the key vectors (for scaling).

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O \tag{2}$$

where:

head_i = Attention(QW_i^Q, KW_i^K, VW_i^V);
W_i^Q, W_i^K, W_i^V = learned projection matrices for the i-th head;
W^O = Learned weight matrix for the output transformation.

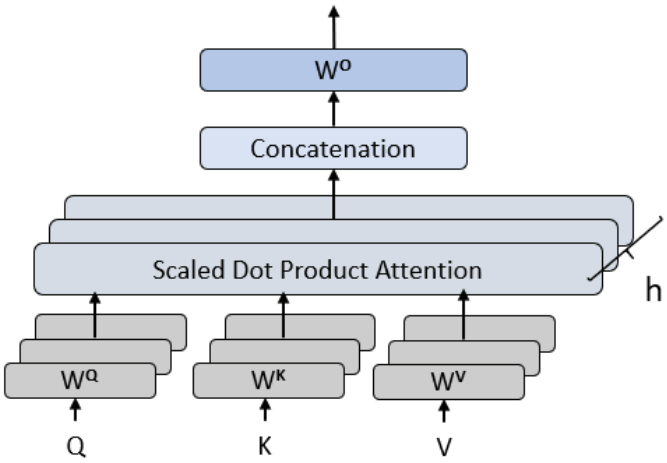


Figure 2. Multi-head attention mechanism module architecture.

3.2. Feature Fusion Models

3.2.1. Cascaded Feature Fusion Model

In the cascaded model, IMU data is first fed into a two-layer 1D-CNN to enhance the features of the time series data. The resulting sequences are then passed through a two-layer LSTM network. The final LSTM layer outputs the last time step's result. The output of the LSTM network is further processed by a muti-head attention mechanism with eight heads, enabling the model to focus on different parts of the time sequence, thereby providing richer information about the data. Finally, the model concludes with a dense network consisting of two layers, followed by a dropout layer and a softmax activation function to classify the

time sequence data into the annotated categories. Figure 3 illustrates the architecture for the cascaded feature fusion model, where CNN and LSTM networks are sequentially connected.

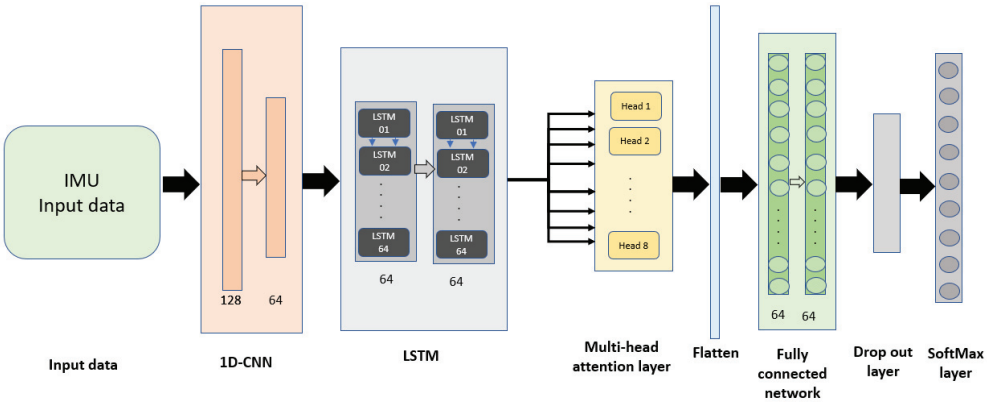


Figure 3. The cascaded CNN-LSTM with multi-head attention system for surface classification.

3.2.2. Parallel Feature Fusion Model

In the parallel fusion model, the input data from the sensor is simultaneously fed into both the CNN and LSTM networks as shown in Figure 4. The output from the CNN is processed by a global average pooling layer, which averages the time-series data, while the LSTM outputs the data from the last time step. The output features from both the CNN and LSTM networks were concatenated to form a unified feature vector. This vector was then passed through an eight-head multi-head attention mechanism. In the final stage, a dense network with two layers, followed by a dropout layer and a softmax function, was used for classification. The primary distinction between the cascaded model and the parallel model lies in how the features from the CNN and LSTM networks are fused. In the cascaded model, the features are passed sequentially from the CNN to the LSTM, while in the parallel model, the features are extracted from both networks simultaneously and then combined.

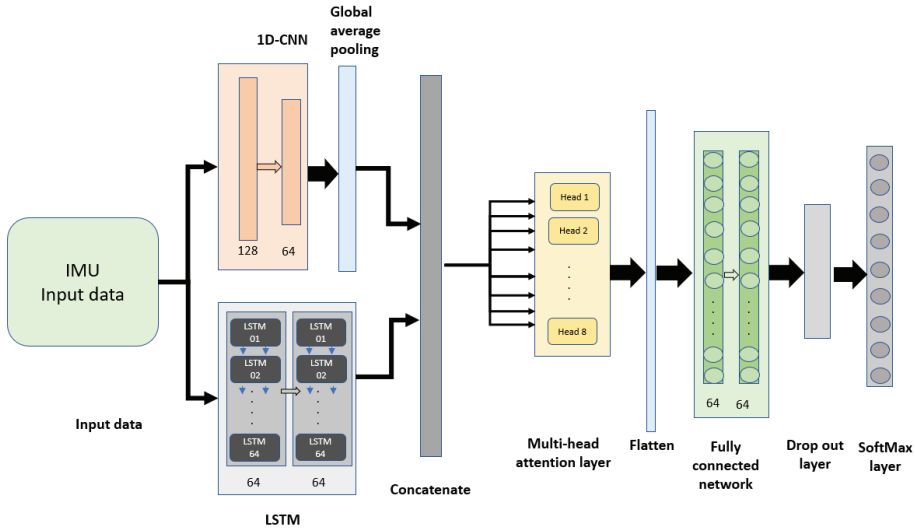


Figure 4. The parallel CNN-LSTM with multi-head attention system for surface classification.

4. Dataset and Training Process

The time-series dataset analyzed in this study is collected from an IMU sensor mounted on a ground robot [46]. The IMU used for data collection is the XSENS MTi-300, a sensor manufactured by Movella, a company based in Henderson, NV, USA [47]. The sensor operates at a frequency of 375 Hz, resulting in 0.341 s per sequence (128 data point). The dataset

comprises ten features representing a four-dimensional quaternion space: x orientation, y orientation, z orientation, and the angle of rotation. Additionally, it includes the three angular velocities and the three linear accelerations in the x, y, and z directions. The robot was driven over nine different surface types, and the data were annotated into nine classes: hard tiles with large spaces, hard tiles, soft tiles, fine concrete, concrete, soft PVC, tiles, and wood. The dataset is publicly available on Kaggle [48].

The data were normalized by subtracting the mean and dividing by the standard deviation to ensure all features were on a consistent scale. The dataset includes 3810 annotated time sequences, where each sequence consists of 128 data points. The dataset was divided into training and testing sets, with 80% allocated for training and 20% for testing. This resulted in 3048 training points and 762 testing points. The majority of the dataset is used for training so the models can learn patterns effectively. More training data help improve generalization and prevent underfitting. Figure 5 shows the histogram of the dataset annotations and the distribution of training and testing among each class.

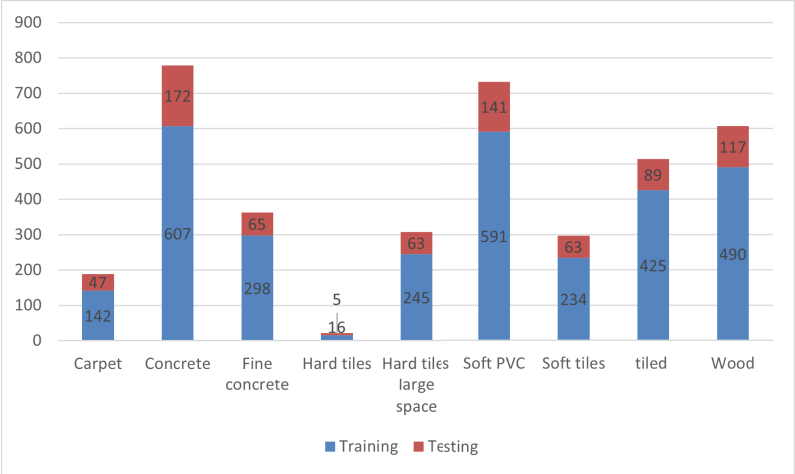


Figure 5. The dataset count and the number of training and testing points for each class.

The training process was applied to both models, the cascaded feature fusion and the parallel feature fusion, using cross-validation. The training data were divided into five groups, with four groups used for training and one for validation. In each epoch, the validation set changes, ensuring that the models are trained on the entire training set across all iterations. A grid search was implemented during training to find the best hyperparameters for the models. The hyperparameters considered in the grid search included epoch count, batch size, optimizer choice, dropout rate, number of attention heads, and the key vector dimension. The optimal hyperparameters identified through this grid search are summarized in Table 1. The model was compiled using the Adam optimizer with the default learning rate of 0.001. No learning rate scheduling was applied during training. The cross-validation score for the cascaded feature fusion model is 0.753, while the parallel feature fusion model achieved a cross-validation of 0.77. The models are trained and tested using a computer with 12th Gen Intel Core i7-1.70 GHz, manufactured by Intel company that is located on Santa Clara, CA, USA, with RAM of 8 GB.

Table 1. The hyperparameters of the models based on the grid search.

Hyperparameter	Value
Batch size	32
Epoch size	60
Optimization function	Adam
dropout rate	0.2
Attention mechanism number of heads	8
Attention mechanism key vector dimension	64
Number of LSTM units	64

5. Evaluation Metrics

To evaluate the proposed models, the LSTM model was used as a baseline for comparison. Additionally, the cascaded fusion model and the parallel fusion model were assessed both against the baseline and in comparison to each other. The classification performance was quantified using precision, recall, F1 score, and Mean Average Precision (mAP), as defined in [49]. Precision, which measures the proportion of correctly identified positive cases among all predicted positives, is defined in Equation (3):

$$Precision = \frac{1}{N}\sum_{i=1}^{N}\frac{TP}{TP+FP} \tag{3}$$

Recall, indicating the proportion of correctly identified positive cases out of all actual positives, is defined in Equation (4):

$$Recall = \frac{1}{N}\sum_{i=1}^{N}\frac{TP}{TP+FN} \tag{4}$$

Here, TP refers to true positive, TN to true negative, FP to false positives and FN is to false negatives.

The F1 score, a harmonic mean of precision and recall, is expressed in Equation (5):

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

Finally, the mean Average Precision (mAP), representing the average of the individual Average Precision (AP) values across all classes, is computed, as shown in Equation (6):

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N}\text{AP}_i \tag{6}$$

where N is the number of classes, and AP_i is the Average Precision for class i.

6. Results

6.1. Baseline vs. Cascaded Fusion Model

Table 2 presents a comparative analysis of the classification results and runtime of the cascaded fusion model relative to the LSTM baseline. The results reveal a substantial improvement in classification accuracy, with the mAP increasing by 18.8%. However, this enhancement in performance was accompanied by a notable increase in runtime, which rose from 580 ms for the LSTM model to 1420 ms for the cascaded fusion model, representing a 144.8% increase. It is important to note that in the improvement column of the table, a negative value signifies a reduction in performance.

Table 2. Comparison of Baseline (LSTM) and the cascaded fusion model.

Metric	Baseline (LSTM)	Cascaded (CNN-LSTM)	Cascaded (CNN-LSTM + Attention)	% Improvement (Cascaded Model to LSTM)
Precision	0.764	0.787	0.835	+9.3%
Recall	0.752	0.777	0.833	+10.8%
F1 Score	0.751	0.779	0.833	+10.9%
mAP	0.608	0.647	0.721	+18.6%
Runtime (ms)	580	590	1420	−144.8%

6.2. Baseline vs. Parallel Fusion Model

Table 3 provides a comparative analysis of the parallel fusion model and the baseline algorithm. The results indicate a significant enhancement in detection accuracy, with the mAP increasing by 14%. Collectively. It also shows an increment in the runtime by 50 ms (8.6%) compared to the baseline.

Table 3. Comparison of Baseline (LSTM) and parallel fusion Models.

Metric	Baseline (LSTM)	Parallel (CNN-LSTM)	Parallel (CNN-LSTM + Attention)	% Improvement (Parallel Model to LSTM)
Precision	0.764	0.794	0.815	+6.7%
Recall	0.752	0.793	0.813	+8.1%
F1 Score	0.751	0.791	0.813	+8.3%
mAP	0.608	0.662	0.693	+14.0%
Runtime (ms)	580	580	630	−8.6%

6.3. Cascaded Fusion Model vs. Parallel Fusion Model

Table 4 presents a detailed comparison of parallel and cascaded feature fusion models. The cascaded model demonstrates superior performance across all metrics, particularly in mAP (0.721 vs. 0.693), achieving precision, recall, and F1-score of 0.835, 0.833, and 0.833, respectively. While offering superior classification accuracy, the cascaded model incurs a significant computational cost, requiring 1420 ms per evaluation, compared to 630 ms for the parallel model, representing a 55.6% increase in processing time.

Table 4. Comparison of the cascaded fusion model and parallel fusion model.

Metric	Cascaded (CNN-LSTM + Attention)	Parallel (CNN-LSTM + Attention)	% Improvement (Cascaded to Parallel Fusion Model)
Precision	0.835	0.815	+2.4%
Recall	0.833	0.813	+2.4%
F1 Score	0.833	0.813	+2.4%
mAP	0.721	0.693	+3.9%
Runtime (ms)	1420	630	−55.6%

6.4. The Multi-Head Attention Mechanism Impact

Figure 6 illustrates the impact of the multi-head attention mechanism on the performance of both fusion models. In the parallel fusion architecture, the incorporation of attention resulted in a modest increase in mAP from 0.662 to 0.693, accompanied by a 50 ms increase in runtime. Conversely, the cascaded fusion model exhibited a more substantial improvement, with mAP rising from 0.647 to 0.721, representing a 14% enhancement in accuracy. However, this significant accuracy gain came at a substantial computational cost,

as the processing time increased significantly from 590 ms to 1420 ms. These results underscore the inherent trade-off between accuracy and computational efficiency when utilizing the multi-head attention mechanism within the proposed surface classification framework.

The multi-head attention mechanism is likely more effective in the cascaded model because it can focus on the most relevant features at each step in the sequence, enabling the model to better capture long-range dependencies in the data. In the parallel model, while multi-head attention still improves performance by weighting the contributions of each branch, the independent processing of data streams limits the attention's ability to significantly influence the interaction between the CNN and LSTM features.

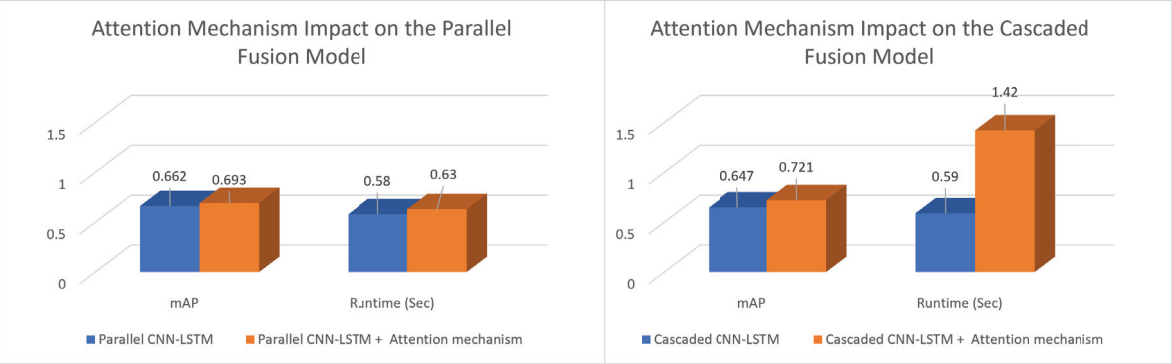


Figure 6. Histogram illustrating the impact of the multi-head attention mechanism on the parallel and cascaded fusion models, comparing mAP and runtime.

7. Discussion

Comparative analysis of cascaded and parallel fusion models highlights critical trade-offs between detection accuracy and computational efficiency. As depicted in Figure 7, the cascaded fusion model achieved marginally higher precision, recall, and mAP scores, indicating superior detection accuracy. In contrast, the parallel fusion model excelled in classification runtime, positioning it as a more computationally efficient and well-balanced alternative. The parallel model provides an effective balance between delivering satisfactory detection performance and minimizing computational overhead.

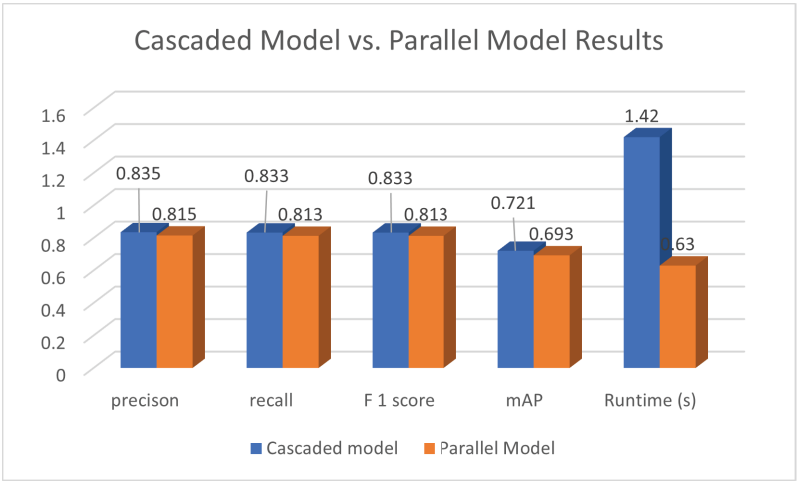


Figure 7. Histogram representation of the parallel model vs. the cascaded model classification results and runtime.

The significance of each component within the proposed models is highlighted by assessing its individual performance apart from other elements. For instance, the LSTM alone achieved an mAP of 0.608, which rose to 0.647 with the cascaded fusion of CNN

and LSTM features, and further to 0.662 with the parallel fusion of CNN and LSTM. The addition of the multi-head attention mechanism proved more impact in the cascaded fusion model, boosting mAP by 11.44%, as compared to a 4.23% increase observed in the parallel fusion model. However, adding the multi-head attention module in the cascaded fusion model increased the processing time from 0.59 s to 1.42 s.

Figure 8 displays the confusion matrix for both the cascaded fusion model and the parallel fusion model. It is evident that the cascaded fusion model achieved higher true-positive rates across most classes compared to the parallel fusion model. For example, the cascade model exhibited true-positive rates of 85.47% for concrete and 76.92% for fine concrete, while the parallel fusion model recorded true positives of 79.65% and 61.54% respectively. The parallel model demonstrated significant improvements in specific classes; for instance, the false-positive rate for classifying hard tiles as fine concrete was reduced from 40% in the cascaded fusion model to 0% in the parallel fusion model. The parallel model demonstrated better performance in classes with similarities, such as concrete and fine concrete.

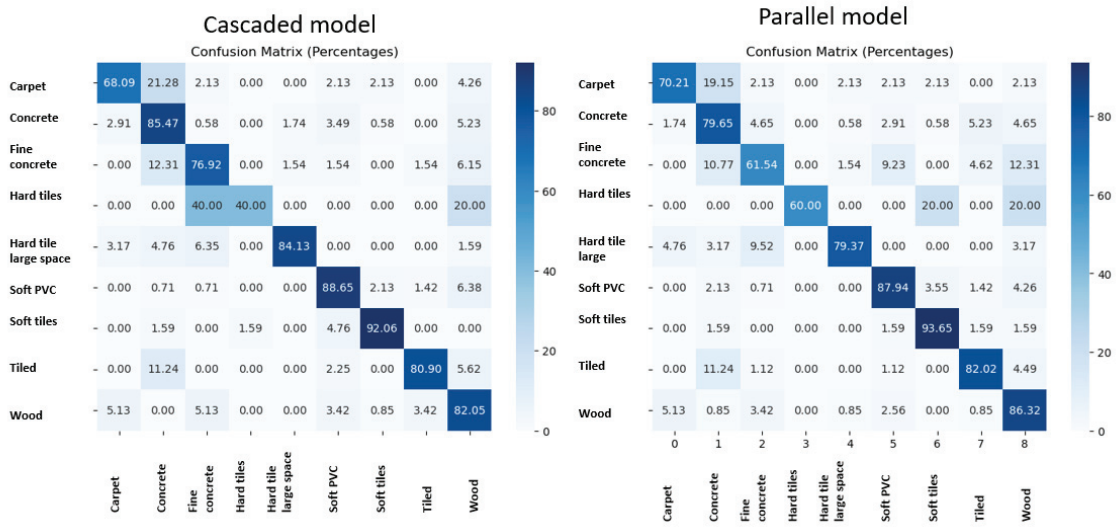


Figure 8. Confusion matrix of the casaded fusion model and parallel fusion model.

8. Conclusions

This study investigates surface type classification using Inertial Measurement Unit (IMU) data acquired from a ground robot. Two novel feature fusion architectures are proposed and evaluated. The cascaded model employs a sequential feature extraction approach, integrating a 1-D Convolutional Neural Network (CNN) followed by a Long Short-Term Memory (LSTM) network. Alternatively, the parallel fusion model leverages parallel processing, where the IMU data are concurrently processed by both 1-D CNN and LSTM networks, and the resulting feature representations are subsequently concatenated. To enhance the model's ability to focus on the most relevant information within the sensor data, a multi-head attention mechanism is incorporated into both architectures.

The proposed models were trained and evaluated on a time series dataset. Data normalization, employing a standard z-score transformation (subtracting the mean and dividing by the standard deviation), enhanced the detection performance of both models. Hyperparameter optimization was performed using a grid search approach to ensure optimal model performance. Experimental results demonstrated that the cascaded fusion model achieved marginally superior detection metrics, including precision, mean Average Precision (mAP), recall, and F1-score, compared to the parallel fusion model. This suggests that the cascaded approach enables better feature representation and extraction, leading to

more accurate detections. However, the cascaded fusion system exhibited a significantly higher computational cost, with a 55.6% increase in processing time. The integration of the multi-head attention mechanism significantly contributed to improved detection accuracy, particularly within the cascaded feature fusion model, though at the expense of increased classification runtime. In conclusion, the parallel fusion model offers a more optimized solution, effectively balancing detection accuracy and computational efficiency.

Future research directions could involve a comparative analysis between IMU-based deep learning and computer vision systems. By integrating both systems onto a single robotic platform, a comprehensive evaluation of detection accuracy and runtime performance could be conducted. Furthermore, exploring the synergistic potential of fusing IMU data with visual sensor inputs holds significant promise for enhancing surface classification capabilities.

Author Contributions: Conceptualization, G.A.-r., D.K., H.E., M.R. and N.A.; methodology, G.A.-r.; software, G.A.-r.; validation, G.A.-r. and D.K.; formal analysis, H.E. and M.R.; investigation, G.A.-r., D.K., H.E., M.R. and N.A.; resources, G.A.-r. and D.K.; writing—original draft preparation, G.A.-r. and D.K.; writing—review and editing, G.A.-r., D.K., H.E., M.R. and N.A.; visualization, G.A.-r.; supervision, G.A.-r.; project administration, G.A.-r. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset is available at this link: https://kaggle.com/competitions/career-con-2019 (accessed on 17 March 2025).

Acknowledgments: I sincerely appreciate the co-authors for their valuable contributions to this work.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
1-D CNN	One dimensional Convolutional Neural Network
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
FCN	Fully Connected Network
XGBOOST	Extreme Gradient Boosting
mAP	Mean Average Precision
IMU	Internal Measurement Unit
HAR	Human Activities Recognition

References

1. Wang, H.; Luo, N.; Zhou, T.; Yang, S. Physical Robots in Education: A Systematic Review Based on the Technological Pedagogical Content Knowledge Framework. *Sustainability* **2024**, *16*, 4987. [CrossRef]
2. Obrenovic, B.; Gu, X.; Wang, G.; Godinic, D.; Jakhongirov, I. Generative AI and human–robot interaction: Implications and future agenda for business, society and ethics. *AI Soc.* **2024**, 1–14.. [CrossRef]
3. Silvera-Tawil, D. Robotics in Healthcare: A Survey. *SN Comput. Sci.* **2024**, *5*, 189.
4. Walas, K. Terrain classification and negotiation with a walking robot. *J. Intell. Robot. Syst.* **2015**, *78*, 401–423.
5. Khaleghian, S.; Taheri, S. Terrain classification using intelligent tire. *J. Terramech.* **2017**, *71*, 15–24.

6. Oliveira, F.G.; Santos, E.R.; Neto, A.A.; Campos, M.F.; Macharet, D.G. Speed-invariant terrain roughness classification and control based on inertial sensors. In Proceedings of the 2017 Latin American Robotics Symposium (LARS) and 2017 Brazilian Symposium on Robotics (SBR), São Paulo, Brazil, 19–21 October 2017; pp. 1–6.
7. Hoepflinger, M.A.; Remy, C.D.; Hutter, M.; Spinello, L.; Siegwart, R. Haptic terrain classification for legged robots. In Proceedings of the 2010 IEEE International Conference on Robotics and Automation, Anchorage, AK, USA, 3–7 May 2010; pp. 2828–2833.
8. Zenker, S.; Aksoy, E.E.; Goldschmidt, D.; Wörgötter, F.; Manoonpong, P. Visual terrain classification for selecting energy efficient gaits of a hexapod robot. In Proceedings of the 2013 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, Wollongong, NSW, Australia, 9–12 July 2013; pp. 577–584.
9. Zhao, T.; Guo, P.; Wei, Y. Road friction estimation based on vision for safe autonomous driving. *Mech. Syst. Signal Process.* **2024**, *208*, 111019.
10. Iglesias, F.; Aguilera, A.; Padilla, A.; Vizán, A.; Diez, E. Application of computer vision techniques to estimate surface roughness on wood-based sanded workpieces. *Measurement* **2024**, *224*, 113917.
11. Laible, S.; Khan, Y.N.; Zell, A. Terrain classification with conditional random fields on fused 3D LIDAR and camera data. In Proceedings of the 2013 European Conference on Mobile Robots, Warsaw, Poland, 2–4 September 2013; pp. 172–177.
12. Walas, K.; Nowicki, M. Terrain classification using laser range finder. In Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 14–18 September 2014; pp. 5003–5009.
13. Borrmann, D.; Nüchter, A.; Đakulović, M.; Maurović, I.; Petrović, I.; Osmanković, D.; Velagić, J. A mobile robot based system for fully automated thermal 3D mapping. *Adv. Eng. Inform.* **2014**, *28*, 425–440. [CrossRef]
14. Hu, K.; Chen, Z.; Kang, H.; Tang, Y. 3D vision technologies for a self-developed structural external crack damage recognition robot. *Autom. Constr.* **2024**, *159*, 105262.
15. Tang, Y.; Huang, Z.; Chen, Z.; Chen, M.; Zhou, H.; Zhang, H.; Sun, J. Novel visual crack width measurement based on backbone double-scale features for improved detection automation. *Eng. Struct.* **2023**, *274*, 115158.
16. Yi, J.; Zhang, J.; Song, D.; Jayasuriya, S. IMU-based localization and slip estimation for skid-steered mobile robots. In Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, CA, USA, 29 October–2 November 2007; pp. 2845–2850.
17. Botero Valencia, J.S.; Rico Garcia, M.; Villegas Ceballos, J.P. A simple method to estimate the trajectory of a low cost mobile robotic platform using an IMU. *Int. J. Interact. Des. Manuf. (IJIDeM)* **2017**, *11*, 823–828.
18. Apaydın, N.N.; Kılıç, İ.; Apaydın, M.; Yaman, O. Decision Tree-Based Direction Detection Using IMU Data in Autonomous Robots. *Batman Üniv. Yaşam Bilim. Derg.* **2024**, *14*, 57–68.
19. Zevering, J.; Bredenbeck, A.; Arzberger, F.; Borrmann, D.; Nüchter, A. IMU-based pose-estimation for spherical robots with limited resources. In Proceedings of the 2021 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Seoul, Republic of Korea, 25–27 August 2021; pp. 1–8.
20. Barnea, A.; Berrabah, S.A.; Oprisan, C.; Doroftei, I. IMU (Inertial Measurement Unit) Integration for the Navigation and Positioning of Autonomous Robot Systems. *J. Control. Eng. Appl. Inform.* **2011**, *13*, 38–43.
21. Wang, L.; Jiang, Z.; Hu, Y.; Yang, S. Research on the Self-Stability Control and Gait Planning of Quadruped Robot Based on IMU. In Proceedings of the 2021 IEEE 21st International Conference on Communication Technology (ICCT), Xi'an, China, 19–21 October 2021; pp. 1078–1081.
22. Hammoudeh, M.A.A.; Alsaykhan, M.; Alsalameh, R.; Althwaibi, N. Computer Vision: A Review of Detecting Objects in Videos–Challenges and Techniques. *Int. J. Online Biomed. Eng.* **2022**, *18*, 15.
23. Mariniello, G.; Pastore, T.; Menna, C.; Festa, P.; Asprone, D. Structural damage detection and localization using decision tree ensemble and vibration data. *Comput. Aided Civ. Infrastruct. Eng.* **2021**, *36*, 1129–1149.
24. Kumar, P.; Pandi, S.S.; Kumaragurubaran, T.; Chiranjeevi, V.R. Human Activity Recognitions in Handheld Devices Using Random Forest Algorithm. In Proceedings of the 2024 International Conference on Automation and Computation (AUTOCOM), Dehradun, India, 14–16 March 2024; pp. 159–163.
25. Al-refai, G.; Elmoaqet, H.; Ryalat, M. In-vehicle data for predicting road conditions and driving style using machine learning. *Appl. Sci.* **2022**, *12*, 8928. [CrossRef]
26. Barman, U.; Choudhury, R.D. Soil texture classification using multi class support vector machine. *Inf. Process. Agric.* **2020**, *7*, 318–332.
27. Kim, D.; Kim, S.H.; Kim, T.; Kang, B.B.; Lee, M.; Park, W.; Ku, S.; Kim, D.; Kwon, J.; Lee, H.; et al. Review of machine learning methods in soft robotics. *PLoS ONE* **2021**, *16*, e0246102.
28. Shirmard, H.; Farahbakhsh, E.; Müller, R.D.; Chandra, R. A review of machine learning in processing remote sensing data for mineral exploration. *Remote Sens. Environ.* **2022**, *268*, 112750.
29. Gupta, S. Deep learning based human activity recognition (HAR) using wearable sensor data. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100046.

30. Qi, W.; Ovur, S.E.; Li, Z.; Marzullo, A.; Song, R. Multi-sensor guided hand gesture recognition for a teleoperated robot using a recurrent neural network. *IEEE Robot. Autom. Lett.* **2021**, *6*, 6039–6045. [CrossRef]
31. Zhu, J.; Chen, H.; Ye, W. Classification of human activities based on radar signals using 1D-CNN and LSTM. In Proceedings of the 2020 IEEE International Symposium on Circuits and Systems (ISCAS), Seoul, Republic of Korea, 18–22 October 2020; pp. 1–5.
32. Al-refai, G.; Al-refai, M.; Alzu'bi, A. Driving Style and Traffic Prediction with Artificial Neural Networks Using On-Board Diagnostics and Smartphone Sensors. *Appl. Sci.* **2024**, *14*, 5008. [CrossRef]
33. Mekruksavanich, S.; Jitpattanakul, A. Lstm networks using smartphone data for sensor-based human activity recognition in smart homes. *Sensors* **2021**, *21*, 1636. [CrossRef] [PubMed]
34. Pham, T.D. Time–frequency time–space LSTM for robust classification of physiological signals. *Sci. Rep.* **2021**, *11*, 6936.
35. Matar, M.; Xia, T.; Huguenard, K.; Huston, D.; Wshah, S. Multi-head attention based bi-lstm for anomaly detection in multivariate time-series of wsn. In Proceedings of the 2023 IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS), Bangalore, India, 9–11 March 2023; pp. 1–5.
36. Junior, R.F.R.; dos Santos Areias, I.A.; Campos, M.M.; Teixeira, C.E.; da Silva, L.E.B.; Gomes, G.F. Fault detection and diagnosis in electric motors using 1d convolutional neural networks with multi-channel vibration signals. *Measurement* **2022**, *190*, 110759.
37. Cui, W.; Deng, X.; Zhang, Z. Improved convolutional neural network based on multi-head attention mechanism for industrial process fault classification. In Proceedings of the 2020 IEEE 9th Data Driven Control and Learning Systems Conference (DDCLS), Hangzhou, China, 23–25 October 2020; pp. 918–922.
38. Li, X.; Wu, J.; Li, Z.; Zuo, J.; Wang, P. Robot ground classification and recognition based on CNN-LSTM model. In Proceedings of the 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Beijing, China, 16–18 December 2021; pp. 1110–1113.
39. AlRadaideh, S.; Aljoan, I.; Al-refai, G.; Elmoaqet, H. Ground Classification for Robots Navigation Using Time Series Dataset with LSTM and CNN. In Proceedings of the 2024 22nd International Conference on Research and Education in Mechatronics (REM), Dead Sea, Jordan, 24–26 September 2024; pp. 375–380.
40. Feng, C.; Dong, K.; Ou, X. A Robot Ground Medium Classification Algorithm Based on Feature Fusion and Adaptive Spatio-Temporal Cascade Networks. *Neural Process. Lett.* **2024**, *56*, 235.
41. Jiang, Y.; Zhou, W.; Zhang, S.; Gao, S. Floor Surface Classification with Robot IMU Sensors Data. Available online: https://noiselab.ucsd.edu/ECE228_2019/Reports/Report14.pdf (accessed on 17 March 2025).
42. Kiranyaz, S.; Avci, O.; Abdeljaber, O.; Ince, T.; Gabbouj, M.; Inman, D.J. 1D convolutional neural networks and applications: A survey. *Mech. Syst. Signal Process.* **2021**, *151*, 107398. [CrossRef]
43. Van Houdt, G.; Mosquera, C.; Nápoles, G. A review on the long short-term memory model. *Artif. Intell. Rev.* **2020**, *53*, 5929–5955.
44. Hanin, B. Which neural net architectures give rise to exploding and vanishing gradients? *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 580–589.
45. Niu, Z.; Zhong, G.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **2021**, *452*, 48–62.
46. Lomio, F.; Skenderi, E.; Mohamadi, D.; Collin, J.; Ghabcheloo, R.; Huttunen, H. Surface type classification for autonomous robot indoor navigation. *arXiv* **2019**, arXiv:1905.00252.
47. Movella. Datasheet for XSENS MTi-300 IMU. Available online: https://www.xsens.com/hubfs/Downloads/Leaflets/MTi-300.pdf (accessed on 17 March 2025).
48. Maggie; Dane, S. CareerCon 2019—Help Navigate Robots. Available online: https://kaggle.com/competitions/career-con-2019 (accessed on 19 October 2024).
49. Erickson, B.J.; Kitamura, F. Magician's Corner: 9. Performance Metrics for Machine Learning Models. *Radiol. Artif. Intell.* **2021**, *3*, e200126. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Hybrid Deep Learning for Fault Diagnosis in Photovoltaic Systems

Mouaad Bougoffa [1], Samir Benmoussa [1], Mohand Djeziri [2,*] and Olivier Palais [2]

[1] Laboratory of Automatics and Signals of Annaba (LASA), Badji Mokhtar-Annaba University, P.O. Box 12, Annaba 23000, Algeria; mouaad.bougoffa@univ-annaba.dz (M.B.); samir.benmoussa@univ-annaba.dz (S.B.)
[2] Aix Marseille Univ, Université de Toulon, CNRS, IM2NP, Marseille, France; olivier.palais@im2np.fr
* Correspondence: mohand.djeziri@im2np.fr

Abstract: Photovoltaic (PV) systems are integral to global renewable energy generation, yet their efficiency and reliability are frequently compromised by undetected faults, leading to significant energy losses, increased maintenance costs, and reduced operational lifespans. To address these challenges, this study proposes a novel hybrid deep learning framework that combines Stacked Sparse Auto-Encoders (SSAE) for autonomous feature extraction with an Optimized-Multi-Layer Perceptron (OMLP) for precise fault classification. The SSAE extracts high-dimensional fault features from raw operational data, while the OMLP leverages these features to classify faults with exceptional accuracy. The model was rigorously validated using real-world PV datasets, encompassing diverse fault types such as partial shading, open circuits, and module degradation under dynamic environmental conditions. Results demonstrate state-of-the-art performance, with the model achieving 99.82% accuracy, 99.7% precision, 99.4% sensitivity, and 100% specificity, outperforming traditional machine learning and deep learning approaches. These findings highlight the framework's robustness and reliability in real-world applications. By significantly enhancing fault detection accuracy and computational efficiency, the proposed approach optimizes PV system performance, reduces operational costs, and supports sustainable energy production. This study concludes that the hybrid SSAE-Optimized MLP model represents a scalable and efficient solution for improving the reliability and longevity of renewable energy infrastructure, setting a new benchmark for intelligent maintenance strategies in the field.

Keywords: photovoltaic system; fault detection; fault diagnosis; stacked sparse autoencoder; optimized multi-layer perceptron

1. Introduction

Artificial Intelligence (AI) has emerged as a cornerstone of 21st-century innovation, driving transformative advancements across industries such as healthcare, finance, and energy [1,2]. By enabling systems to analyze data, recognize patterns, and make informed decisions, AI methodologies such as machine learning (ML) and deep learning (DL) have unlocked unprecedented opportunities to address complex challenges and optimize operational workflows [3]. Nowhere is this potential more critical than in renewable energy, where AI is revolutionizing the efficiency and reliability of infrastructure, particularly in solar PV systems.

PV systems, which convert sunlight into electricity, are pivotal to the global transition toward sustainable energy. Despite their growing adoption—from residential installations to utility-scale solar farms—their performance and longevity are hindered by operational

challenges such as module degradation, shading effects, inverter failures, and environmental variability [4]. PV array faults such as Open Circuits (OC), Short Circuits (SC), and environmental faults such as Partial Shading (PS) are particularly insidious, often leading to energy losses, increased maintenance costs, and reduced return on investment [5,6].

A 2010 study demonstrated that faults and failures in solar PV systems can lead to an annual energy output reduction of 18.9% [7], while harsh environmental conditions exacerbate efficiency drops to 15–20% [8]. For instance, prolonged exposure to extreme climates accelerates module degradation, with reported annual degradation rates of 0.923% in desert environments [9]. Dust accumulation further compounds these losses: experimental data from Saharan regions shows 8.41% power loss over eight weeks due to soiling [10], while uncleaned systems in Mediterranean climates suffer 9.99% instantaneous power drops and 2.93% average reductions [11]. Partial shading—often caused by cloud cover or debris—can obscure up to 75% of PV array surfaces, creating localized hotspots and mismatch losses [12]. Mitigation strategies, such as PV array reconfiguration and tilt-angle optimization, demonstrate promising results, with adjusted tilt angles boosting power output by 50.36% in some regions [13]. These findings underscore the critical need for intelligent fault diagnosis frameworks to minimize energy losses, extend system lifespans, and maximize return on investment in solar infrastructure.

Traditional fault diagnosis in PV systems relies on analyzing electrical parameters such as voltage ($V\sim dc\sim$), current ($I\sim dc\sim$), and irradiance (Irr), often collected via sensors such as maximum power point trackers (MPPTs) and environmental monitors. Rule-based algorithms and physical inspections remain prevalent but are labor-intensive and lack scalability. Advanced methods include infrared thermography for hotspot detection and electroluminescence imaging for microcracks. However, these techniques require specialized equipment and are unsuitable for real-time monitoring.

The integration of AI into PV fault diagnosis represents a paradigm shift in renewable energy supervision [14], enabling scalable and automated solutions. Recent advancements leverage ML models such as Support Vector Machines (SVMs), Random Forests (RFs), and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) to process historical and real-time data from supervisory control and data acquisition (SCADA) systems [15–19]. Despite progress, these methods struggle with high-dimensional data and dynamic environmental variability, underscoring the need for hybrid AI-driven approaches.

For instance, AdaBoost achieved 0.95 precision in fault classification using temperature and radiation features [20], while kernel-based extreme learning machines [21] and probabilistic neural networks (PNNs) [22] showed high precision in identifying short-circuited modules. However, traditional ML methods struggle with overfitting, high-dimensional data, and scalability [23,24].

DL, a subset of ML, addresses these limitations through hierarchical feature extraction. Hybrid architectures, such as Convolutional Neural Networks (CNNs) combined with residual gated recurrent units (ResGRUs), classify I-V curve anomalies with 98.61% accuracy [25], while Stacked Auto-Encoders (SAE) enable fault diagnosis in PV arrays through I-V characteristic analysis [26,27]. Similarly, particle swarm optimization (PSO)-enhanced long short-term memory (LSTM) networks improve fault detection reliability by optimizing parameter tuning and temporal pattern recognition [28,29]. Autoencoder (AE)-based frameworks, including LSTM, CNN, and variational, sparse, or denoising variants, are widely applied for fault detection across systems [30,31]. For instance, Graph Convolutional Variational AutoEncoders (GCVAE) model complex spatial–temporal dependencies to identify anomalous patterns [32], further advancing detection capabilities. Despite these advances, critical gaps persist: (1) many models require large labeled datasets and lack generalizabil-

ity across PV designs [33]; (2) computational complexity hinders real-time deployment [34]; and (3) integration with cost-effective monitoring systems remains underexplored.

To bridge these gaps, this paper proposes a novel hybrid deep learning architecture combining SSAE and Optimized-MLP. The SSAE extracts high-dimensional fault features from raw operational data (e.g., current, voltage, irradiance), capturing subtle patterns associated with faults such as partial shading and degradation. The OMLP then classifies these features with minimal computational overhead, enabling efficient diagnosis while maintaining accuracy [35]. This approach diverges from prior work, such as [36], which relies on hybrid CNN-Bi-GRU models constrained by database quality, and [37], which depends on simulated data for MLP-based classification. By balancing SSAE's robust feature extraction with MLP's lightweight classification, our hybrid architecture introduces three key innovations: (1) a KL divergence-regularized SSAE that enforces sparsity during unsupervised feature learning, enhancing sensitivity to rare faults under class imbalance; (2) a Bayesian-optimized MLP with weighted loss functions, automating hyperparameter selection while mitigating bias toward majority classes; and (3) end-to-end integration of environmental variables (irradiance, temperature) with electrical parameters, enabling robust diagnosis under dynamic outdoor conditions. Compared to CNN-Bi-GRU models—which incur high computational costs from convolutional operations and bidirectional temporal processing—our SSAE-OMLP achieves superior accuracy (99.82%) on tabular PV data while reducing inference latency, making it deployable on low-cost edge hardware.

Using empirical data from operational PV systems, we evaluate the SSAE-OMLP framework against traditional methods (SVM, RF) and state-of-the-art DL architectures (CNN, LSTM). Our contributions include:

1. A comparative analysis of DL models under dynamic stresses (e.g., fluctuating irradiance, partial shading).
2. A hybrid SSAE-OMLP framework optimized for computational efficiency, enabling rapid fault detection under dynamic environmental conditions.
3. Empirical validation of scalability for large-scale deployments, addressing a critical gap in prior research [38].

The results demonstrate that the SSAE-Optimized-MLP hybrid reduces false positives by 22%, achieves 99.82% accuracy in fault classification, and operates 35% faster than conventional CNN-based models. By minimizing energy losses and maintenance costs, this work advances the reliability of PV systems, directly supporting global sustainability goals. Furthermore, it underscores how AI-driven digital transformation can enhance project leadership in renewable energy, ensuring the successful deployment of innovative technologies in real-world settings.

2. Photovoltaic Fault Types

Faults in PV systems are categorized based on their origin within the system's Direct Current (DC) or Alternating Current (AC) components. AC-side faults typically stem from inverter malfunctions or grid-related disturbances. In contrast, DC-side faults are more diverse, encompassing issues such as MPPT algorithm inefficiencies, bypass diode failures, ground faults, arc faults, module mismatch, OC, and SC faults [6,39].

2.1. Arc Faults

In photovoltaic systems, arc faults occur when a high-energy electrical discharge jumps through air or across a gap between conductive parts. Such arcs can be triggered by issues such as deteriorated insulation, loose or corroded connections, damaged cables, or failed components [40].

2.2. Open-Circuit Faults

These occur when a disconnection disrupts current flow within the system. Such faults significantly reduce power generation, potentially affecting individual module strings or the entire system, depending on the disconnection's location and the PV array's topology [41].

2.3. Short-Circuit Faults

These arise from unintended electrical connections between two points in a PV array at different potentials. These faults may manifest within a single string or between adjacent strings in the system [41].

2.4. Ground Faults

These occur when the PV array unintentionally makes contact with earth, creating safety hazards, reducing operational efficiency, and risking damage to the system's components [42].

2.5. Line-to-Line Faults

These faults happen when two conductors in a PV system become directly short-circuited. They typically involve a direct connection between any two of the system's three AC output phases—often referred to as a three-phase fault [43].

2.6. Module/Cell Mismatch

This fault arises when cells or modules exhibit divergent electrical properties, compromising system efficiency. Mismatch is subdivided into temporary and permanent categories. Temporary mismatch is caused by transient factors such as dust/snow accumulation or dynamic shading from nearby structures, whereas permanent mismatch results from long-term degradation or physical damage to modules. Where temporary mismatch is exemplified by partial shading, and permanent mismatch by module degradation [8].

2.7. Partial Shading

Occurs when parts of the PV modules are obstructed and cannot receive full sunlight, temporarily reducing their power output. Shading can be either static or dynamic. Static shading is caused by long-term deposits on the module surface—such as dust, leaves, or bird droppings, and dynamic shading comes from fleeting shadows cast by nearby trees or buildings [12].

2.8. Degradation

Over time, PV systems lose performance due to several factors, including surface contamination, UV-induced optical wear, increases in series or decreases in shunt resistance, and declines in SC current [44]. This deterioration can be confined to individual modules or progressively affect the entire array, leading to an overall drop in energy production.

The experimental dataset in this study focuses on four critical DC-side faults in PV systems: OC, SC, PS and module degradation. These faults were selected due to their recurrent occurrence and significant impact on system performance.

3. Materials and Methods

This section outlines the hybrid approach used for fault diagnosis in PV systems, combining Stacked Sparse Auto-Encoders for feature extraction with a Multi-Layer Perceptron classifier for fault categorization as presented in Figure 1. The proposed methodology involves several stages: data preprocessing, feature extraction using SSAE, fault classification using Optimized-MLP, and performance evaluation using various metrics.

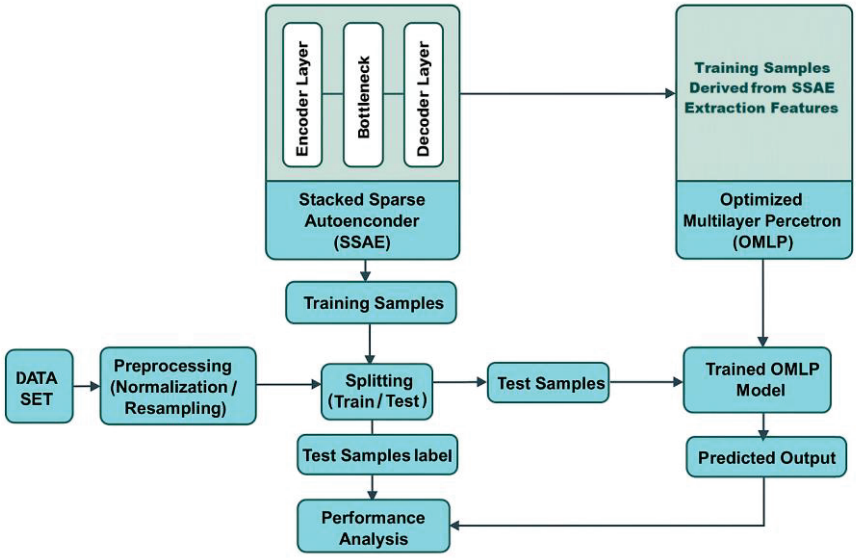


Figure 1. Proposed system architecture for fault diagnosis.

3.1. PV System

The PV system used for data collection, as described by Lazzaretti et al. (2020) [45], consists of two strings of eight C6SU-330P PV modules, labeled PV_1 through PV_{16}, connected to a 5 kW grid-tied inverter (NHS Solar 5K-GDM1). Each string is equipped with circuit breakers and fuses to ensure safe simulation of fault conditions. Figure 2 illustrates the system installation.

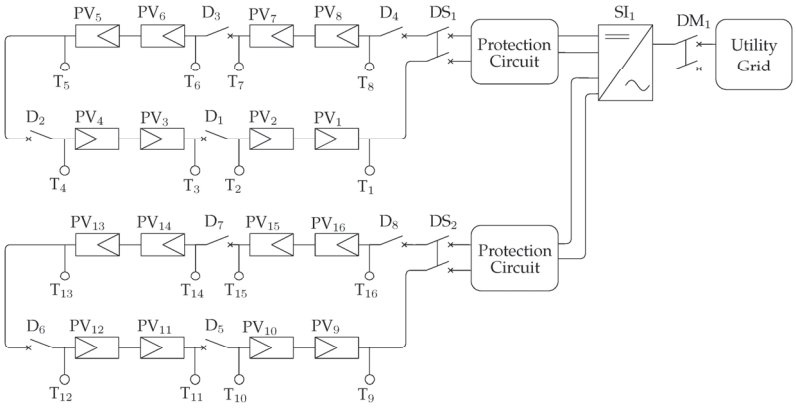


Figure 2. Schematic representation of the solar PV system.

3.2. Dataset Description

The dataset, sourced from a real PV system, was sampled at 1 Hz over a continuous 16-day period from 07:30 to 17:00, capturing irradiance variations from 100 to 1000 W/m^2 and temperature ranges from -5 °C to +85 °C, (cloudy and rainy days are not considered). It contains electrical parameters representing each fault condition. Key features include:

- Vdc_1, Vdc_2: Voltage for each string.
- Idc_1, Idc_2: Current for each string.
- *Irr*: Irradiance.
- *pvt:* PV Module Temperature.
- f_{nv}: Fault label.

To assess the PV system's performance under realistic conditions, the simulation framework integrates photovoltaic strings—each comprising eight 330 W modules—modeled in MATLAB 2022b/Simulink 10.6, with environmental variables such as irradiance and

temperature. The simulated DC voltage and current outputs from these strings, shown in Figures 3 and 4, feed into subsequent system components. Dataset parameters are aligned with the reference system configuration detailed in Figure 2, ensuring consistency between the simulation and real-world setups. Building on the validated PV model from earlier sections, comprehensive databases are generated to capture both optimal operation and intentional fault scenarios. These datasets leverage daily solar irradiance and temperature profiles, as illustrated in Figure 5, to replicate outdoor conditions.

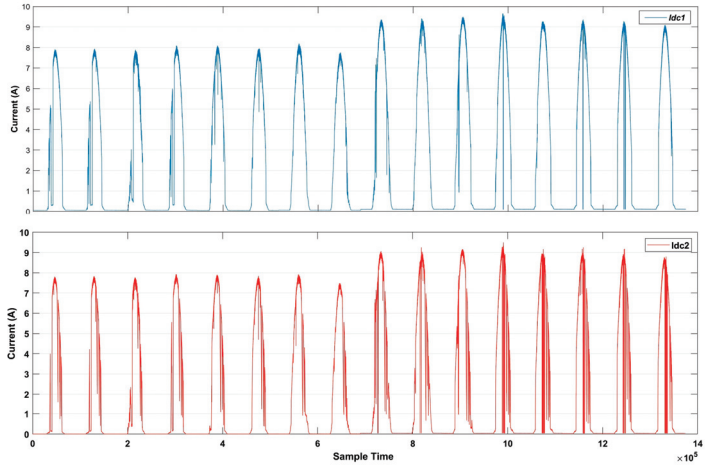


Figure 3. DC output current of each string.

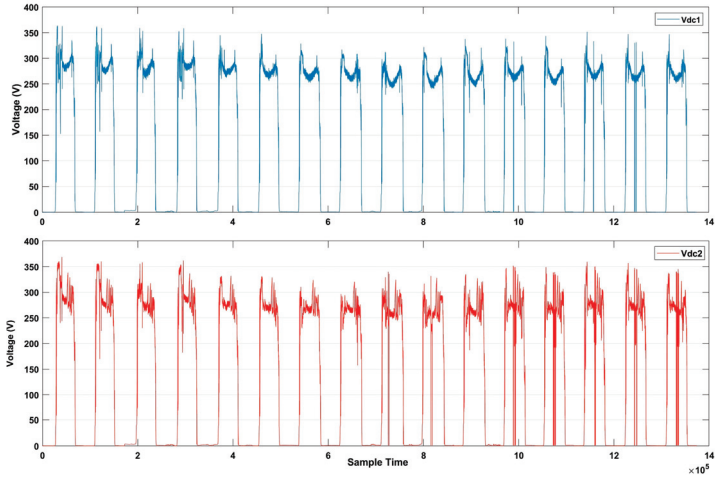


Figure 4. DC output voltage of each string.

The dataset is divided into two distinct parts: one derived from simulated fault scenarios and the other from real-world measurements. The training phase relied exclusively on simulated data generated using a validated PV system model in MATLAB/Simulink. Fault types such as short circuit, open circuit, degradation, and shadowing were modeled under varying irradiance and temperature conditions to reflect realistic operational behavior. For instance, in the case of shadowing faults, four distinct scenarios were simulated, each introducing 5% to 15% shading across a PV string, resulting in a total of 5054 synthetic samples used in training.

In contrast, the testing phase was conducted entirely on real-world data collected from an actual PV installation, which comprises 514,958 labeled samples, with the class distribution detailed in Table 1. To evaluate model generalization across different fault categories under realistic conditions, the testing procedure was repeated three times using different subsets of the real dataset.

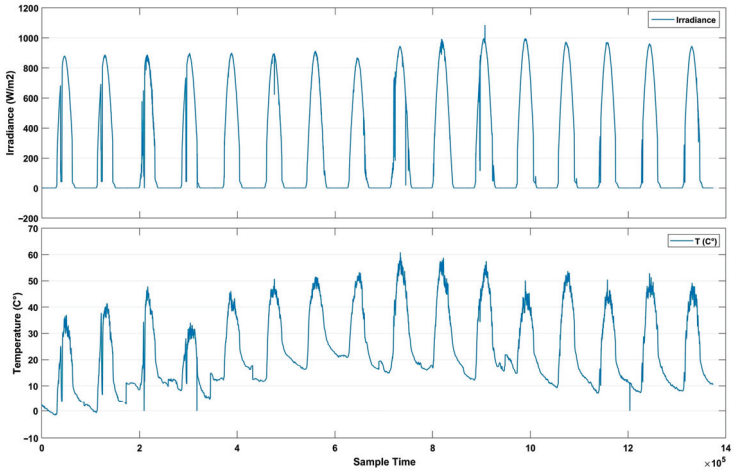


Figure 5. Daily variations of solar irradiance and temperature over time.

Table 1. Proportion of data and fault label.

Proportion of Data		
Label	**Class**	**Samples**
0	Normal Operation	309,253
1	Short Circuit	5999
2	Open Circuit	6024
3	Degradation	10,371
4	Shadowing	184,311

Environmental variables such as temperature and irradiation are also included to account for real-world operating conditions. The data was normalized to a (0, 1) range and split 70-30 into training and testing sets. We specifically selected six photovoltaic features denoted as x = [$x1$, $x2$, $x3$, $x4$, $x5$, $x6$] representing [$Vdc1$, $Vdc2$, $Idc1$, $Idc2$, Irr, pvt], respectively, to be used as input for training our model, providing data points that distinguish between healthy and faulty operation. Table 1 presents the data samples of each category.

3.3. Data Preprocessing

The data was processed through a preprocessing pipeline to make it suitable for machine learning models, including data cleaning to handle potential anomalies, feature selection for critical parameters (voltage, current, temperature, and irradiation), and normalization to ensure consistency across features. Normalization was performed using min-max scaling, as outlined in Equation (1), where x_{min} and x_{max} were computed from the training set to avoid test data leakage and standardize the range of feature values [46].

$$x_{Scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Dimensionality reduction, known as feature extraction, involves condensing the number of variables or attributes in a dataset. Its purpose is to identify and retain the most meaningful and pertinent data embedded within the original features, transforming it into a smaller set of features while striving to preserve as much critical information as possible.

These steps are essential to improve data quality, enhance model performance, and ensure the reliability and accuracy of the resulting insights. By addressing issues such as missing values, inconsistencies, and irrelevant data, preprocessing ensures that the machine

learning model receives clean, well-structured, and meaningful input, leading to more accurate and reliable predictions [47].

3.4. Stacked Sparse Auto-Encoder for Fault Extraction

The SSAE is a deep learning model specifically designed for unsupervised feature extraction. In this architecture, multiple auto-encoders are sequentially stacked, allowing the model to capture complex, high-dimensional fault features relevant to PV diagnostics.

- Architecture: Each auto-encoder in the SAE comprises an encoder, a hidden layer (bottleneck), and a decoder. The encoder transforms the normalized data into a compressed representation in the bottleneck layer, where critical patterns relevant to fault types are retained. This architecture allows the SSAE to automatically learn and preserve the most informative features for fault detection [48].
- Encoding and decoding: The encoder maps the input vector X_i = [$V_{dc}1$, $V_{dc}2$, $I_{dc}1$, $I_{dc}2$, *Irr*, *pvt*] to a hidden representation hi to predict output $\hat{X}i = \hat{V}_{\mathrm{dc1}}\hat{V}_{\mathrm{dc2}}\hat{I}_{\mathrm{dc1}}\hat{I}_{\mathrm{dc2}}\hat{I}_{\mathrm{rr}}, \hat{F}_{\mathrm{nv}}]$, as determined by (2) and (3):
  - Input Transformation (Encoding Layer)

$$h_i = f\ (x_i) = sigm(W_1.x + b_1) \tag{2}$$

During the encoding process, the input vector x_i is mapped to a latent representation h_i. This transformation is achieved through a linear operation involving weight matrix W_1 and bias b_1, followed by a nonlinear sigmoid activation function.

- Decoding Layer:

$$\hat{x}_i = g(h_i) = sigm(W_2\ .h + b_2) \tag{3}$$

The latent representation h_i is then reconstructed into $\hat{x}_i$ via the decoding function g which applies another linear transformation (using weights W_1 and bias b_1) and passes the result through a sigmoid activation, as shown in Equation (3).

  - Cost Function:

Equation (4) represents the cost function of the Sparse Autoencoder. It consists of three main components that will be further decomposed in Equations (10)–(12), where the first term quantifies the reconstruction loss (mean squared error between the input x_i and reconstructed output $\hat{x}_i$), ensuring accurate feature representation, and the second term is the weight decay regularization that penalizes large weights to prevent overfitting, and a sparsity constraint using *Kullback–Leibler* (*KL*) divergence.

$$J = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}\| \hat{x}_i - x_i \|^2\right) + \frac{\lambda}{2}\sum_{i=1}^{N}\| W_i \|^2 + \beta\sum_{i=1}^{N} KL(\rho \parallel \hat{\rho}_j) \tag{4}$$

where m represents the number of neurons in the hidden layer. The hyperparameters λ and β control the relative influence of the weight decay (L2 regularization) and the sparsity penalty, respectively. The sparsity constraint utilizes the *KL* divergence, denoted as $KL(\rho \parallel \hat{\rho}_j)$ [49], to quantify the discrepancy between the target sparsity parameter ρ and the empirical average activation of the *i*-th hidden neuron over the training dataset.

3.5. Multi-Layer Perceptron for Fault Classification

Following feature extraction by the SSAE, an Optimized Multi-Layer Perceptron is employed for fault classification. The MLP is a supervised neural network model that excels at distinguishing between different classes (fault types) by learning complex feature mappings [50].

Architecture: The MLP consists of an input layer, one or more hidden layers, and an output layer as shown in Figure 6. Each layer is fully connected to the subsequent layer. The input layer receives SSAE-extracted features, and each hidden layer transforms these features via learned weights and biases [35].

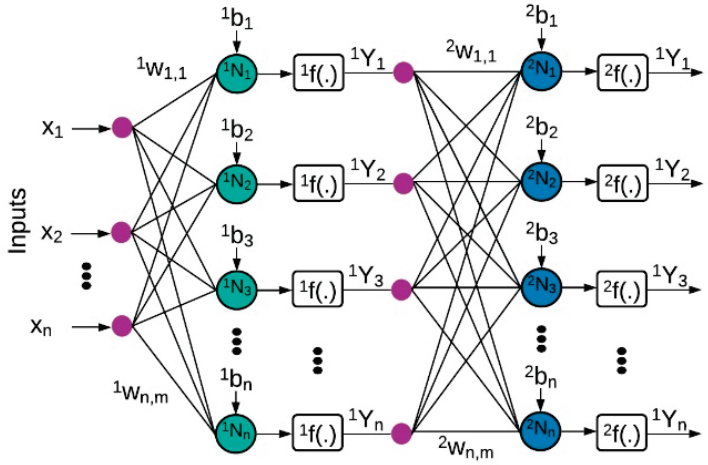


Figure 6. The architecture of an MLP.

Forward propagation: The output of each hidden layer is computed as determined by (5) and (6).

$$^{1}Y = {}^{1}_{.}f\left({}^{1}_{.}\omega \times X + {}^{1}_{.}b\right) \tag{5}$$

$$^{2}Y = {}^{2}_{.}f\left({}^{2}_{.}\omega \times {}^{1}\mathrm{Y} + {}^{2}_{.}b\right) \tag{6}$$

where W represents the weight matrices, b represents biases, and f denotes the *ReLU* activation function, which introduces non-linearity to the model, which is expressed by (7) [38].

$$f(.) = \max(0, x_i) \tag{7}$$

Hyperparameter tuning: Optimal parameters were selected via grid search, adjusting learning rate, batch size, and hidden layer dimensions to improve accuracy and efficiency [50].

The combined SSAE and Optimized-MLP model form a powerful framework for fault detection and classification: the SSAE captures intricate fault-related features, while the MLP effectively translates these features into actionable insights. This approach demonstrates high performance, as validated across multiple key metrics.

3.6. Hybrid SSAE with Optimized-MLP Based Models

From prior research, hybrid architectures combining unsupervised and supervised learning have proven effective in scenarios with limited labeled data, particularly in fault diagnosis and anomaly detection [51]. AEs are widely recognized for their ability to learn hierarchical features from unlabeled data through unsupervised pre-training, reducing dependency on costly labeled datasets [27]. Meanwhile, MLPs excel at mapping learned features to target labels via supervised fine-tuning, making them ideal for classification and regression tasks [35]. By integrating SSAE and Optimized-MLP, this hybrid model leverages the strengths of both paradigms: the SSAE extracts sparse, low-dimensional representations of raw input data, while the Optimized-MLP utilizes these features for precise fault prediction. This approach is particularly advantageous for PV fault detection, where labeled fault samples are often scarce, but unlabeled operational data is abundant.

As illustrated in Figure 1, the hybrid model operates in two stages. First, raw input data X undergoes min-max normalization (Equation (1)) to scale values between [0, 1], ensuring stable training. The preprocessed data is then fed to the SSAE, which consists of an encoder f_{ENC} and decoder f_{DEC}. The encoder maps X to a latent representation Z using weights W_{ENC} and a sigmoid activation (Equation (8)), while the decoder reconstructs $\hat{X}$ from Z (Equation (9)). The model was trained over 100 epochs, with the following hyperparameters set: weight-decay λ = 0.001, sparsity penalty $\beta = 3$, and target activation $\rho = 0.05$. The Sparse AE's loss function (Equation (13)) combines reconstruction error, sparsity penalty (KL divergence), and L2 regularization. Once trained, the latent features Z are extracted and passed to the OMLP (Figure 3), which computes predictions $\hat{y}$ via fully connected layers.

For classification, cross-entropy loss (Equation (16)) is minimized, whereas regression tasks use mean squared error. The Adam optimizer updates weights in both stages, ensuring efficient convergence. As detailed in Algorithm 1 (Lines 6–26). The Optimized-MLP (Lines 30–50) processes the flattened features with 16 hidden nodes, culminating in a 5-node output layer that classifies faults into normal, degradation, SC, OC, and partial shading categories. The subsequent section evaluates this architecture's performance on real-world PV fault datasets, demonstrating its superiority over standalone models.

Algorithm 1. Hybrid SSAE and Optimized-MLP

1: **Procedure** Hybrid Stacked Sparse AutoEncoder and Optimized-Multilayer Perceptron $(\mathbf{X}, \mathbf{y}, \boldsymbol{\lambda = 0.001}, \boldsymbol{\beta = 3}, \boldsymbol{\rho = 0.05}, \boldsymbol{epochs = 100})$
2: //**Preprocessing**
3: Input X: unlabeled raw data $X \in \mathbb{R} >$ (*Vdc1*, *Vdc2*, *Idc1*, *Idc2*, *Irr*, *pvt*)
4: Normalize X to [0, 1]
5: **//Train Stacked Sparse Autoencoder (SSAE)**
6: Initialize encoder f_{ENC} and decoder f_{DEC} with weights W_{ENC}, W_{DEC}
7: **//Train SSAE for 100 epochs**
8: **for epoch = 1 to 100 do**
9: //Forward pass
10: Latent representation:
11:

$$Z \leftarrow \sigma(W_{ENC}.X_{Norm} + b_{ENC}) \tag{8}$$

12: Reconstruction:
13:

$$\hat{X} \leftarrow \sigma(W_{DEC}.Z + b_{DEC}) \tag{9}$$

14: //Compute loss
15: Reconstruction error:
16:

$$\mathcal{L}_{\text{recon}} \leftarrow \frac{1}{2N}\sum_{i=1}^{N} \| \hat{x}_i - x_i \|^2 \tag{10}$$

Algorithm 1. Hybrid SSAE and Optimized-MLP

17: Sparsity penalty (KL divergence):

18:

$$\mathcal{L}_{\text{sparse}} \leftarrow \beta \sum_{j=1}^{d} \left[\rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j} \right] \tag{11}$$

19: Weight decay:

20:

$$\mathcal{L}_{\text{reg}} \leftarrow \frac{\lambda}{2} \left(\|W_{\text{ENC}}\|_F^2 + \|W_{\text{DEC}}\|_F^2 \right) \tag{12}$$

21: Total loss:

22:

$$\mathcal{L}_{\text{SAE}} \leftarrow \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{sparse}} \tag{13}$$

23: //Backpropagation

24: Update $W_{ENC}, b_{ENC}, W_{DEC}, b_{DEC}$ via Adam:

25:

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{SAE}}, \theta \in \{W_{\text{ENC}}, b_{\text{ENC}}, W_{\text{DEC}}, b_{\text{DEC}}\} \tag{14}$$

26: **End for**

27: **//Feature Extraction**

28: Extract latent features:

29: $\mathbf{Z} \leftarrow f_{\text{ENC}}(X_{Norm})$

30:**//Optimized-MLP for Fault Classification**

31://Optimized-MLP for Fault Classification

32: Hidden layers $\in \{1, 2, 3\}$ Units/layer $\in \{32, 64, 128\}$, Dropout $\in [0.1, 0.5]$

33: Initialize MLP f_{MLP} with weights W_{MLP}

34: Initialize Bayesian optimizer.

35: for $t = 1 \rightarrow T$ do

36: Sample hyperparameters $\theta_t = \left\{ n_{\text{layers}}, n_{\text{units}}, p_{\text{dropout}} \right\}$

37: **for epoch = 1 to 100 do //Train MLP for 100 epochs**

38://Forward pass

39: Prediction:

40:

$$\hat{\mathbf{y}} \leftarrow f_{\text{MLP}}(\mathbf{Z}) \leftarrow \text{Softmax}\left(W_{\text{out}}\, \mathbf{h}_{n_{\text{layers}}} + \mathbf{b}_{\text{out}} \right) \tag{15}$$

41://Compute loss

42: Cross-entropy (classification) or MSE (regression):

43:

$$\mathcal{L}_{\text{MLP}} \leftarrow \begin{cases} -\frac{1}{n} \sum_{i=1}^{n} \sum_{c=1}^{C} y_{ic} \log\left(\hat{y}_{ic} \right) & \text{(classification)} \\ \frac{1}{n} \sum_{i=1}^{n} \left\| y_i - \hat{y}_i \right\|^2 & \text{(regression)} \end{cases} \tag{16}$$

44://Backpropagation

45: Update W_{MLP} via Adam:

Algorithm 1. Hybrid SSAE and Optimized-MLP

46:

$$W_{\mathbf{MLP}} \leftarrow W_{\mathbf{MLP}} - \eta \nabla_{W_{\mathbf{MLP}}} \mathcal{L}_{\mathbf{MLP}} \tag{17}$$

47: Select optimal $\theta^* \leftarrow \mathbf{argmin} \mathcal{L}_{\mathbf{MLP}}$
48: **End For**
49: return $f_{\mathbf{ENC}}, f_{\mathbf{MLP}}$
50: End procedure

3.7. Evaluation Metrics

To evaluate the outcomes of the fault detection process in this study, a range of metrics were employed. These metrics play an essential role in assessing the effectiveness and reliability of the detection process. Prior to outlining performance metrics for this procedure, it is crucial to clarify the following concepts [52]:

- TP (True Positive): This occurs when the detection process correctly identifies an authentic fault in the PV system.
- TN (True Negative): This is observed when the PV system operates without any issues, and the fault detection system correctly verifies the absence of faults.
- FP (False Positive): This occurs when the PV system shows no faults, and the fault detection system identifies a fault.
- FN (False Negative): This happens when the PV system experiences a fault, and the detection system does not indicate it.
- Accuracy: Correlates with the comprehensive detection effectiveness, Equation (18):

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{18}$$

- Precision: Represents the ratio between positive indicators, Equation (19):

$$Precision = \frac{TP}{(TP + FP\ )} \tag{19}$$

- Sensitivity: Assesses the effectiveness of accurately classifying detections, Equation (20):

$$Sensitivity = \frac{TP}{(TN + FP\ )} \tag{20}$$

- Specificity (recall): Assesses the classifier's efficiency in recognizing inaccurate detections, Equation (21):

$$Specificity = \frac{TN}{(TN + FP\ )} \tag{21}$$

- Macro F1-Score: Harmonic mean of precision and recall, averaged equally across all classes, Equation (22):

$$Macro\ F1\ score = \frac{1}{C} \sum_{c=1}^{C} \frac{Precision_c . Recall_c}{Precision_c + Recall_c} \tag{22}$$

where C is the number of classes.

- Micro F1-Score: Global harmonic mean of overall precision and recall, accounting for class imbalance, Equation (23):

$$Micro\ F1\ score = \frac{2.\ TP_{Total}}{2.TP_{Total} + FP_{Total} + FN_{Total}} \tag{23}$$

- Cohen's Kappa (κ): Measures inter-rater agreement adjusted for chance, Equation (24).

$$\kappa = \frac{P_0 - P_e}{1 + P_e} \tag{24}$$

where P_0 is observed agreement and P_e is expected agreement.

4. Results

Experiments were conducted using Python 3.12.7 and TensorFlow 2.18.0 on a high-performance computing environment. The SSAE was trained first, followed by the OMLP classifier, until both models achieved convergence based on early stopping criteria.

4.1. SSAE Performance

Figure 7 displays the training and validation loss curves for the SSAE. Both losses decrease rapidly in the first few epochs, suggesting that the SSAE quickly learns efficient feature representations. The subsequent stability of both curves at low values indicates that the SSAE effectively captures the underlying data patterns with minimal reconstruction error, achieving a good balance between training and validation. Minor fluctuations in the validation loss suggest slight variability but no overfitting, showing that the SSAE model generalizes well across both sets.

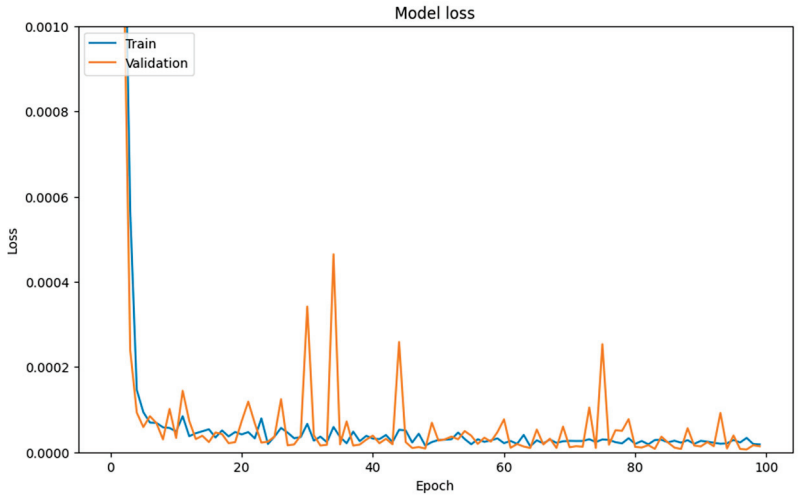


Figure 7. Optimal fitting model.

4.2. Optimized-MLP Performance

Performance metrics such as accuracy, precision, sensitivity, and specificity (recall) are calculated to assess the model's effectiveness in identifying fault types.

The Optimized-MLP model accuracy in Figure 8 shows both training and validation accuracy over 100 epochs. Both accuracies increase rapidly in the initial epochs, stabilizing near 99.9% accuracy. The high and consistent accuracy levels indicate that the model has effectively learned the features extracted by the SSAE and performs well on both the training and validation sets. Minor fluctuations in validation accuracy suggest slight variations in generalization but no significant overfitting, demonstrating that the OMLP model is well-optimized for accurate fault diagnosis.

The testing results in Table 2 highlight the high performance of our fault diagnosis approach. Three experimental trials were conducted by repeating the process with three

distinct data partitions derived from the same dataset, ensuring statistical robustness. The model shows a strong average precision (99.7%), indicating accurate fault identification with minimal false positives, and an average sensitivity (99.4%), ensuring effective fault detection. The specificity reaches 100% across all tests, confirming that non-fault conditions are correctly identified. Finally, with an average accuracy of 99.82%, the model demonstrates reliable and consistent overall performance. These results confirm that the SSAE and Optimized-MLP combination is highly effective for precise fault diagnosis.



Figure 8. Accuracy during training and validation.

Table 2. Testing Results.

	Test 1	**Test 2**	**Test 3**
Precision	0.996	0.998	0.998
Sensitivity	0.994	0.995	0.995
Specificity	1.000	1.000	1.000
Macro f1-score	0.973	0.975	0.978
Micro f1-score	0.996	0.997	0.998
Cohen's Kappa	0.984	0.986	0.988
Accuracy	0.996	0.997	0.9982

4.3. Confusion Matrix

The confusion matrix depicted in Figure 9 shows strong performance in fault classification. The model accurately identifies most "Normal" operating conditions, with 96,515 correct predictions, and demonstrates perfect precision (100%). For SC faults, 2173 samples were correctly classified, with minimal errors (2 misclassified as degradation and 7 as partial shading). Degradation faults also showed high accuracy, with 1917 correct identifications and minor overlaps (3 confused with partial shading and 7 with OC faults), underscoring the model's reliability in distinguishing critical fault types.

4.4. Comparative Study

In this section, we compare our fault detection approach with recent models used in PV system diagnostics, focusing on the DL techniques SAE and clustering, CNN, and CNN Bi-GRU. Table 3 provides a summary of fault types detected, algorithms used, and accuracy achieved by each method.

Our approach achieves a testing accuracy of 99.82%, outperforming other models significantly, with a 2.8% increase over the highest alternative (SSAE with clustering at 97%). In contrast, the standalone CNN achieves 95.20% accuracy for OC, SC, PS, and degradation faults, while the CNN Bi-GRU detects faults with an accuracy of 99.4%.

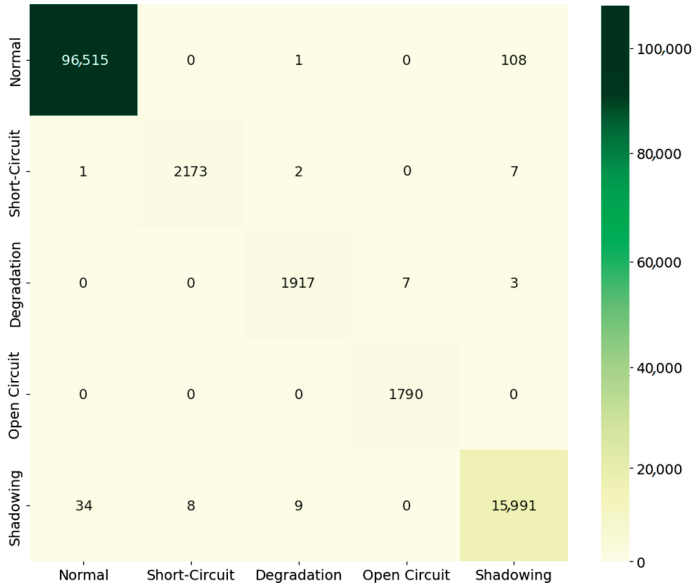


Figure 9. The confusion matrix.

Table 3. Results of the comparative models.

Classifier	Faults	Accuracy
SAE and clustering	OC, SC, PS, and degradation	97%
CNN	OC, SC, PS, and degradation	95.20%
CNN Bi-GRU	OC, SC, PS, and degradation	99.4%
Our method	OC, SC, PS, and degradation	99.82%

The superior performance of our approach can be attributed to the synergy between the SSAE and Optimized-MLP. The SSAE component effectively captures complex, high-dimensional patterns within the data through unsupervised feature extraction, enhancing fault feature representation. These refined features are then processed by the OMLP, which excels in classification accuracy due to its multi-layer structure and ability to generalize across fault types. This layered architecture allows our model to differentiate between fault conditions more precisely, even in challenging scenarios where fault types may have overlapping characteristics. This improvement highlights our method's potential to increase PV system reliability and efficiency, offering more precise fault detection, reducing false alarms, and ultimately supporting more cost-effective maintenance and enhanced system longevity.

5. Conclusions

In conclusion, the proposed hybrid SSAE-OMLP model demonstrates a robust framework for fault diagnosis in photovoltaic systems. By leveraging SSAE for feature extraction and Optimized-MLP for classification, our approach achieves high accuracy, precision, and specificity, contributing to the reliable operation of PV systems. While the model performs exceptionally in fault detection, future work could focus on testing across larger and more diverse PV datasets to ensure generalizability. Additionally, integrating real-time monitoring capabilities would enhance scalability in industrial applications. This study underscores the potential for advanced machine learning models in renewable energy fault diagnostics, supporting the broader goal of sustainable energy production and cost-effective system maintenance.

Author Contributions: Conceptualization, M.B. and M.D.; methodology, M.B.; software, M.B.; validation, M.B., S.B. and M.D.; formal analysis, M.B.; investigation, M.B.; data curation, M.B.; writing—original draft preparation, M.B.; writing—review and editing, M.D., S.B. and O.P.; visualization, M.B.; supervision, S.B.; project administration, M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the "PHC TASSILI +" program (project number: 24MDU117), funded by the French Ministry for Europe and Foreign Affairs, the French Ministry for Higher Education and Research, and the Algerian Ministry for Higher Education and Scientific Research (MESRS).

Data Availability Statement: The dataset is publicly available at https://github.com/clayton-h-costa/pv_fault_dataset (accessed on 17 March 2025).

Acknowledgments: The authors would like to thank "PHC TASSILI +" program (project number: 24MDU117), supported by the French Ministry for Europe and Foreign Affairs, the French Ministry for Higher Education and Research, and the Algerian Ministry for Higher Education and Scientific Research (MESRS).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Maka, A.O.M.; Alabid, J.M. Solar energy technology and its roles in sustainable development. *Clean Energy* **2022**, *6*, 476–483. [CrossRef]
2. Alami, A.H.; Alasad, S.; Aljaghoub, H.; Ayoub, M.; Alashkar, A.; Mdallal, A.; Hasan, R. First-Generation Photovoltaics: History and Conventional Manufacturing. In *Advances in Science, Technology & Innovation*; Springer International Publishing: Berlin/Heidelberg, Germany, 2023; pp. 7–19. [CrossRef]
3. Bougoffa, M.; Benmoussa, S.; Djeziri, M.; Contaret, T. Deep Learning-Based Fault Diagnosis in Photovoltaic System. In Proceedings of the 2024 International Conference on Decision Aid Sciences and Applications (DASA), Manama, Bahrain, 11–12 December 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–5. [CrossRef]
4. Mansouri, M.; Trabelsi, M.; Nounou, H.; Nounou, M. Deep Learning-Based Fault Diagnosis of Photovoltaic Systems: A Comprehensive Review and Enhancement Prospects. *IEEE Access* **2021**, *9*, 126286–126306. [CrossRef]
5. Abubakar, A.; Almeida, C.F.M.; Gemignani, M. Review of Artificial Intelligence-Based Failure Detection and Diagnosis Methods for Solar Photovoltaic Systems. *Machines* **2021**, *9*, 328. [CrossRef]
6. Venkatakrishnan, G.R.; Rengaraj, R.; Tamilselvi, S.; Harshini, J.; Sahoo, A.; Saleel, C.A.; Abbas, M.; Cuce, E.; Jazlyn, C.; Shaik, S.; et al. Detection, location, and diagnosis of different faults in large solar PV system—A review. *Int. J. Low-Carbon Technol.* **2023**, *18*, 659–674. [CrossRef]
7. Firth, S.K.; Lomas, K.J.; Rees, S.J. A simple model of PV system performance and its use in fault detection. *Sol. Energy* **2021**, *84*, 624–635. [CrossRef]
8. Rico Espinosa, A.; Bressan, M.; Giraldo, L.F. Failure signature classification in solar photovoltaic plants using RGB images and convolutional neural networks. *Renew. Energy* **2020**, *162*, 249–256. [CrossRef]
9. Sarikh, S.; Raoufi, M.; Bennouna, A.; Ikken, B. Characteristic curve diagnosis based on fuzzy classification for a reliable photovoltaic fault monitoring. *Sustain. Energy Technol. Assess.* **2021**, *43*, 100958. [CrossRef]
10. Dida, M.; Boughali, S.; Bechki, D.; Bouguettaia, H. Output power loss of crystalline silicon photovoltaic modules due to dust accumulation in Saharan environment. *Renew. Sustain. Energy Rev.* **2020**, *124*, 109787. [CrossRef]
11. Juaidi, A.; Muhammad, H.H.; Abdallah, R.; Abdalhaq, R.; Albatayneh, A.; Kawa, F. Experimental validation of dust impact on-grid connected PV system performance in Palestine: An energy nexus perspective. *Energy Nexus* **2022**, *6*, 100082. [CrossRef]
12. Osmani, K.; Haddad, A.; Jaber, H.; Lemenand, T.; Castanier, B.; Ramadan, M. Mitigating the effects of partial shading on PV system's performance through PV array reconfiguration: A review. *Therm. Sci. Eng. Prog.* **2022**, *31*, 101280. [CrossRef]
13. Osmani, K.; Ramadan, M.; Lemenand, T.; Castanier, B.; Haddad, A. Optimization of PV array tilt angle for minimum levelized cost of energy. *Comput. Electr. Eng.* **2021**, *96*, 107474. [CrossRef]
14. Sepúlveda-Oviedo, E.H.; Travé-Massuyès, L.; Subias, A.; Pavlov, M.; Alonso, C. Fault diagnosis of photovoltaic systems using artificial intelligence: A bibliometric approach. *Heliyon* **2023**, *9*, e21491. [CrossRef] [PubMed]
15. Lodhi, E.; Wang, F.-Y.; Xiong, G.; Zhu, L.; Tamir, T.S.; Rehman, W.U.; Khan, M.A. A Novel Deep Stack-Based Ensemble Learning Approach for Fault Detection and Classification in Photovoltaic Arrays. *Remote Sens.* **2023**, *15*, 1277. [CrossRef]

16. Yuan, Z.; Xiong, G.; Fu, X. Artificial Neural Network for Fault Diagnosis of Solar Photovoltaic Systems: A Survey. *Energies* **2022**, *15*, 8693. [CrossRef]
17. Wang, J.; Gao, D.; Zhu, S.; Wang, S.; Liu, H. Fault diagnosis method of photovoltaic array based on support vector machine. *Energy Sources Part A Recovery Util. Environ. Eff.* **2019**, *45*, 5380–5395. [CrossRef]
18. Bendary, A.F.; Abdelaziz, A.Y.; Ismail, M.M.; Mahmoud, K.; Lehtonen, M.; Darwish, M.M.F. Proposed ANFIS Based Approach for Fault Tracking, Detection, Clearing and Rearrangement for Photovoltaic System. *Sensors* **2021**, *21*, 2269. [CrossRef]
19. Amiri, A.F.; Oudira, H.; Chouder, A.; Kichou, S. Faults detection and diagnosis of PV systems based on machine learning approach using random forest classifier. *Energy Convers. Manag.* **2024**, *301*, 118076. [CrossRef]
20. Ghoneim, S.S.; Rashed, A.E.; Elkalashy, N.I. Fault Detection Algorithms for Achieving Service Continuity in Photovoltaic Farms. *Intell. Autom. Soft Comput.* **2021**, *29*, 467–479. [CrossRef]
21. Chen, Z.; Wu, L.; Cheng, S.; Lin, P.; Wu, Y.; Lin, W. Intelligent fault diagnosis of photovoltaic arrays based on optimized kernel extreme learning machine and I-V characteristics. *Appl. Energy* **2017**, *204*, 912–931. [CrossRef]
22. Kara Mostefa Khelil, C.; Amrouche, B.; Kara, K.; Chouder, A. The impact of the ANN's choice on PV systems diagnosis quality. *Energy Convers. Manag.* **2021**, *240*, 114278. [CrossRef]
23. Masita, K.; Hasan, A.; Shongwe, T.; Hilal, H.A. Deep learning in defects detection of PV modules: A review. *Sol. Energy Adv.* **2025**, *5*, 100090. [CrossRef]
24. Neupane, D.; Bouadjenek, M.R.; Dazeley, R.; Aryal, S. Data-driven machinery fault diagnosis: A comprehensive review. *Neurocomputing* **2025**, *627*, 129588. [CrossRef]
25. Gao, W.; Wai, R.-J. A Novel Fault Identification Method for Photovoltaic Array via Convolutional Neural Network and Residual Gated Recurrent Unit. *IEEE Access* **2020**, *8*, 159493–159510. [CrossRef]
26. Bougoffa, M.; Benmoussa, S.; Djeziri, M.; Contaret, T. Stacked AutoEncoder based diagnosis applied on a Solar Photovoltaic System. *IFAC-PapersOnLine* **2024**, *58*, 384–389. [CrossRef]
27. Liu, Y.; Ding, K.; Zhang, J.; Li, Y.; Yang, Z.; Zheng, W.; Chen, X. Fault diagnosis approach for photovoltaic array based on the stacked auto-encoder and clustering with I–V curves. *Energy Convers. Manag.* **2021**, *245*, 114603. [CrossRef]
28. Eldeghady, G.S.; Kamal, H.A.; Hassan, M.A.M. Fault diagnosis for PV system using a deep learning optimized via PSO heuristic combination technique. *Electr. Eng.* **2023**, *105*, 2287–2301. [CrossRef]
29. Appiah, A.Y.; Zhang, X.; Ayawli, B.B.K.; Kyeremeh, F. Long Short-Term Memory Networks Based Automatic Feature Extraction for Photovoltaic Array Fault Diagnosis. *IEEE Access* **2019**, *7*, 30089–30101. [CrossRef]
30. Qian, J.; Song, Z.; Yao, Y.; Zhu, Z.; Zhang, X. A review on autoencoder based representation learning for fault detection and diagnosis in industrial processes. *Chemom. Intell. Lab. Syst.* **2022**, *231*, 104711. [CrossRef]
31. Yang, Z.; Xu, B.; Luo, W.; Chen, F. Autoencoder-based representation learning and its application in intelligent fault diagnosis: A review. *Measurement* **2022**, *189*, 110460. [CrossRef]
32. Arifeen, M.; Petrovski, A.; Hasan, M.J.; Noman, K.; Navid, W.U.; Haruna, A. Graph-Variational Convolutional Autoencoder-Based Fault Detection and Diagnosis for Photovoltaic Arrays. *Machines* **2024**, *12*, 894. [CrossRef]
33. Guo, Z.; Zhuang, Z.; Tan, H.; Liu, Z.; Li, P.; Lin, Z.; Shang, W.-L.; Zhang, H.; Yan, J. Accurate and generalizable photovoltaic panel segmentation using deep learning for imbalanced datasets. *Renew. Energy* **2023**, *219*, 119471. [CrossRef]
34. Choudhary, T.; Mishra, V.; Goswami, A.; Sarangapani, J. A comprehensive survey on model compression and acceleration. *Artif. Intell. Rev.* **2020**, *53*, 5113–5155. [CrossRef]
35. Vieira, R.G.; Dhimish, M.; de Araújo, F.M.U.; da Silva Guerra, M.I. Comparing multilayer perceptron and probabilistic neural network for PV systems fault detection. *Expert Syst. Appl.* **2022**, *201*, 117248. [CrossRef]
36. Amiri, A.F.; Kichou, S.; Oudira, H.; Chouder, A.; Silvestre, S. Fault detection and diagnosis of a photovoltaic system based on deep learning using the combination of a convolutional neural network (CNN) and bidirectional gated recurrent unit (BI-GRU). *Sustainability* **2024**, *16*, 1012. [CrossRef]
37. Mouaad, B.; Samir, B.; Mohand, D. Dynamic Modeling for Fault Diagnosis in PV Systems Utilizing AI Techniques Based on Multilayer Perceptron (MLP). In Proceedings of the 2024 ASU In-ternational Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), Manama, Bahrain, 28–29 January 2024; pp. 1032–1036. [CrossRef]
38. Heidari, A.A.; Faris, H.; Mirjalili, S.; Aljarah, I.; Mafarja, M. Ant Lion Optimizer: Theory, literature review, and application in multi-layer perceptron neural networks. In *Studies in Computational Intelligence*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 23–46. [CrossRef]
39. Madeti, S.R.; Singh, S.N. A comprehensive study on different types of faults and detection techniques for solar photovoltaic system. *Sol. Energy* **2017**, *158*, 161–185. [CrossRef]
40. Alam, M.K.; Khan, F.; Johnson, J.; Flicker, J. A Comprehensive Review of Catastrophic Faults in PV Arrays: Types, Detection, and Mitigation Techniques. *IEEE J. Photovolt.* **2015**, *5*, 982–997. [CrossRef]
41. Nadeem, A.; Sher, H.A.; Murtaza, A.F.; Ahmed, N. Online current-sensorless estimator for PV open circuit voltage and short circuit current. *Sol. Energy* **2021**, *213*, 198–210. [CrossRef]

42. Liu, H.; Mitolo, M.; Qiu, J. Ground-Fault Loop Impedance Calculations in Low-Voltage Single-Phase Systems. *IEEE Trans. Ind. Appl.* **2024**, *50*, 1331–1337. [CrossRef]
43. Eskandari, A.; Milimonfared, J.; Aghaei, M.; Reinders, A.H.M.E. Autonomous Monitoring of Line-to-Line Faults in Photovoltaic Systems by Feature Selection and Parameter Optimization of Support Vector Machine Using Genetic Algorithms. *Appl. Sci.* **2020**, *10*, 5527. [CrossRef]
44. Triki-Lahiani, A.; Bennani-Ben Abdelghani, A.; Slama-Belkhodja, I. Fault detection and monitoring systems for photovoltaic installations: A review. *Renew. Sustain. Energy Rev.* **2018**, *82*, 2680–2692. [CrossRef]
45. Lazzaretti, A.E.; Costa, C.H.D.; Rodrigues, M.P.; Yamada, G.D.; Lexinoski, G.; Moritz, G.L.; Oroski, E.; Goes, R.E.D.; Linhares, R.R.; Stadzisz, P.C.; et al. A Monitoring System for Online Fault Detection and Classification in Photovoltaic Plants. *Sensors* **2020**, *20*, 4688. [CrossRef] [PubMed]
46. Akanbi, O.A.; Amiri, I.S.; Fazeldehkordi, E. Feature Extraction. In *A Machine-Learning Approach to Phishing Detection and Defense*; Elsevier: Amsterdam, The Netherlands, 2015; pp. 45–54. [CrossRef]
47. Sarhan, M.; Layeghy, S.; Moustafa, N.; Gallagher, M.; Portmann, M. Feature extraction for machine learning-based intrusion detection in IoT networks. *Digit. Commun. Netw.* **2024**, *10*, 205–216. [CrossRef]
48. Seghiour, A.; Abbas, H.A.; Chouder, A.; Rabhi, A. Deep learning method based on autoencoder neural network applied to faults detection and diagnosis of photovoltaic system. *Simul. Model. Pract. Theory* **2023**, *123*, 102704. [CrossRef]
49. Ng, A. Sparse Autoencoder. In *CS294A Lecture Notes*; Stanford University: Stanford, CA, USA, 2011; Volume 72, pp. 1–19.
50. Wu, J.; Chen, X.Y.; Zhang, H.; Xiong, L.D.; Lei, H.; Deng, S.H. Hyperparameter optimization for machine learning models based on Bayesian optimization. *J. Electron. Sci. Technol.* **2019**, *17*, 26–40. [CrossRef]
51. van Engelen, J.E.; Hoos, H.H. A survey on semi-supervised learning. *Mach. Learn.* **2019**, *109*, 373–440. [CrossRef]
52. Colliot, O. (Ed.) Machine Learning for Brain Disorders. In *Neuromethods*; Springer: New York, NY, USA, 2023. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

An Adaptive Control Strategy with Switching Gain and Forgetting Factor for a Robotic Arm Manipulator

Mohammed Yousri Silaa [1,2,*], Oscar Barambones [2,*], Aissa Bencherif [1] and Ilyas Rougab [3]

[1] Telecommunications Signals and Systems Laboratory (TSS), Amar Telidji University of Laghouat, BP 37G, Laghouat 03000, Algeria; a.bencherif@lagh-univ.dz
[2] Engineering School of Vitoria, University of the Basque Country UPV/EHU, Nieves Cano 12, 1006 Vitoria, Spain
[3] Laboratory for Analysis and Control of Energy Systems and Electrical Networks, Department of Electronic, University of Amar Telidji, Laghouat 03000, Algeria; i.rougab@lagh-univ.dz
* Correspondence: moh.silaa@lagh-univ.dz (M.Y.S.); oscar.barambones@ehu.eus (O.B.)

Abstract: This paper presents an adaptive sliding mode controller (ASMC) with the implication of a forgetting factor for a two-degree-of-freedom (2-DOF) arm robot manipulator trajectory tracking. The proposed approach builds upon conventional sliding mode control (SMC), which is well known for its robustness and low tracking error. The controller dynamically adjusts this parameter by introducing an adaptive mechanism to enhance trajectory tracking, guarantee high robustness, and reduce chattering effects. In order to mitigate gain drift, a forgetting factor is incorporated into the adaptation law, ensuring stable and reliable control performance. Stability is validated using Lyapunov theory, and the effectiveness of the proposed ASMC is evaluated through numerical simulations. The simulations are conducted in MATLAB R2023b using a dynamic model of the 2-DOF robotic manipulator. Comparative results with conventional SMC confirm that the adaptive approach significantly improves tracking accuracy, noise robustness, and chattering suppression.

Keywords: 2-DOF robot arm manipulator; adaptive sliding mode control; forgetting factor; SMC; trajectory tracking

1. Introduction

In recent years, control strategies for robotic manipulators have received a lot of attention, especially in the context of systems with limited degrees of freedom, where researchers can isolate and understand the behavior of new control techniques and algorithms under controlled conditions before extending them to more complex robots. Robotic manipulators, also known as robotic arms, are fundamental elements in the development of automated systems because of their ability to perform precise and repeatable movements, which makes them valuable in both academic research and industrial settings [1–7]. A key challenge in robotic manipulation is not the mechanical design alone, but the development of robust control algorithms that ensure accurate trajectory tracking, disturbance rejection, and adaptability to uncertain dynamics.

Trajectory tracking control in robotic arms plays a pivotal role in ensuring precise and efficient motion. Over time, various control strategies have been developed to enhance performance. Traditional controllers, such as proportional–integral (PI) and proportional–integral–derivative (PID), remain widely used due to their simplicity and ease of implementation [8–10]. However, these controllers face challenges when dealing with the nonlinearity of the system, variations in parameters, and external disturbances, often

leading to inaccuracies of follow-up [11]. In contrast, SMC, known for its robustness against uncertainties, offers a promising alternative. Yet, its primary drawback is the chattering phenomenon, undesired oscillations that can cause mechanical wear. To address this, high-order sliding-mode control (HOSMC) was introduced, effectively minimizing chattering while preserving the robustness of SMC. However, its increased computational complexity poses challenges for real-time implementation [12]. Hybrid control approaches, such as PI-Fuzzy and SMC-Fuzzy controllers, attempt to merge the benefits of conventional and intelligent control techniques. Despite their potential, their effectiveness depends heavily on the precise tuning of fuzzy rules, which can be complex. Furthermore, controllers based on neural networks (NN) offer adaptability and excel in managing nonlinearities, they require extensive training and significant computational power [13]. Additionally, iterative learning control (ILC) has gained attention in robotic control for its ability to improve performance over repeated tasks, addressing challenges such as disturbances and system uncertainties by refining control inputs based on past experiences [14].

Numerous control strategies have been proposed in the literature to address the challenges of accurate trajectory tracking in robotic manipulators. For instance, Djaneye-Boundjou et al. [15] employed an adaptive particle swarm optimization (APSO) algorithm to optimize PID gains in high-degree-of-freedom (H-DOF) and multi-input multi-output (MIMO) systems, effectively accounting for joint coupling effects. Their approach was validated through simulations on a 2-DOF robotic arm. Likewise, Wazzan et al. [16] utilized the social spider algorithm (SSA) to improve PID control and integrated fuzzy logic (FL) to minimize steady-state error and improve tracking accuracy. Their MATLAB-based implementation, which included forward and inverse kinematics, achieved precise circular trajectory tracking. These works illustrate the benefits of combining intelligent optimization techniques with classical control methods to improve robustness and system performance. Milind et al. [17] designed a PID controller for a flexible two-link robotic manipulator operating in 3D space. The 2-DOF system, modeled as a planar mechanism with motion in the x-y plane, used a base motor for the first link and a shoulder motor for the second. MATLAB Simulink simulations demonstrated the effectiveness of the proposed controller with minimal control errors. Zakia et al. [18] introduced a hybrid control strategy combining PID and SMC for a 2-DOF planar manipulator. In this design, the PID parameters were used to construct the sliding surface, allowing the controller to benefit from the simplicity of PID and the fast convergence and robustness of SMC, thus improving stability and tracking performance. Silva et al. [19] studied PID and PID 2-DOF controllers for a cylindrical manipulator's rotational joint. They identified a first-order system model using the non-recursive least squares method and tuned it with Ziegler–Nichols (ZN), Chien–Hrones–Reswick (CHR), internal model control (IMC), and Skogestad IMC (SIMC). Performance evaluation showed CHR and IMC yielded the best results for both controllers. Azar et al. [20] proposed a control strategy for position control and trajectory tracking in robotic manipulators, using independent 2-DOF PID controllers for each link of a two-link arm. A type-2 fuzzy compensator is added to improve trajectory tracking and system performance. The study shows that integrating the fuzzy compensator with 2-DOF fractional-order PID (FOPID) controllers enhances performance compared to existing methods, with numerical validation provided. Mohammed et al. [21] developed an adaptive 2-DOF PID controller to handle uncertainties and nonlinearities in real-time robotic manipulation, incorporating a supervisory fuzzy logic system (FLS) for enhanced robustness. MATLAB simulations and practical implementation demonstrated that the adaptive 2-DOF PID controller outperformed traditional fixed-gain controllers, offering better robustness against disturbances and system variations. Bi [22] explored the role of symmetry in robotic dynamics and fault diagnosis, proposing an optimized fuzzy PID algorithm that adjusts

the P, I, and D gains based on error rates. They developed a mathematical model for a 2-DOF manipulator and a gradient-based trajectory planning algorithm, with experimental results validating the approach. Tuan et al. [23] studied the control of a 2-DOF robot arm with elastic actuators, enhancing safety by reducing impact forces during collisions. They introduced a two-loop adaptive control strategy, combining ASMC for the outer loop and MRAC for the inner loop, demonstrating a 50.35% reduction in RMS error and a significant reduction in settling time compared to conventional sliding mode control. Jouila et al. [24] proposed a robust adaptive control for a two-link robot manipulator using a non-singular fast terminal sliding mode (NFTSM) controller combined with a wavelet neural network (WNN) to handle uncertainties and disturbances. A compensation term and online adaptation improved tracking and reduced chattering, with Lyapunov theory ensuring stability. Simulations demonstrated superior performance over other controllers. Sun et al. [25] proposed an adaptive fuzzy backstepping controller for a 2-DOF manipulator to improve trajectory tracking and reduce jitter and positioning errors. Simulation results showed enhanced precision and stability under disturbances. Rahmani et al. [26] developed a hybrid PD-SMC controller optimized using the extended grey wolf optimizer (EGWO), which effectively balanced robustness and simplicity. Simulation results demonstrated superior performance compared to conventional SMC and PDSMC methods. Hameed et al. [27] developed an ASMC approach for a 2-DOF robotic manipulator, emphasizing the method's robustness to uncertainties and external disturbances. The proposed controller was designed to improve tracking accuracy, reduce steady-state error, and suppress chattering. Experimental results showed strong performance, achieving an error rate of 1.35% and system efficiency exceeding 98%, validating the effectiveness and reliability of the control strategy. Similarly, Balthazar et al. [28] examined various control techniques—such as linear–quadratic regulator (LQR), feedforward, and integral control—in order to improve positioning precision and vibration reduction in 2-DOF robotic manipulators. Their findings demonstrated enhanced control of motor current and smoother link motion, contributing to overall system stability and accuracy. Zafer et al. [29] proposed a 2-DOF planar robot control using a fuzzy logic controller (FLC) optimized by particle swarm optimization (PSO). The PSO-tuned FLC outperformed the PSO-tuned PID controller in trajectory control and robustness under varying conditions and noise. Girgis et al. [30] proposed a fractional-order adaptive PID (FOAPID) controller to improve the stability and tracking performance of a 2-DOF robotic manipulator. By combining fractional calculus with the EFRLS algorithm, the controller adaptively adjusts its gains. Lyapunov analysis ensures stability, and the method shows strong robustness against uncertainties and disturbances.

In addition to robust trajectory tracking, contour accuracy is another key concern, especially in manufacturing applications. Several studies have focused on minimizing contour errors using traditional control methods with servo matching [31,32], cross-coupled control (CCC) strategies [33], and task coordinate system approaches [34]. These methods offer high contour precision and can be complementary to robust tracking control techniques such as the one proposed in this study.

In this paper, an adaptive sliding mode controller (ASMC) is implemented to track the trajectory of the 2-DOF arm robot. The main contributions of the proposed ASMC scheme are summarized as follows:

- The ASMC is implemented to overcome the limitations of conventional SMC, particularly in terms of robustness and chattering reduction.
- By dynamically adjusting the control gain, the ASMC enhances disturbance rejection capabilities while improving tracking accuracy.
- The proposed approach ensures precise trajectory tracking with reduced control effort, making it a more effective solution for robotic arm control.

- A novel adaptation law with a forgetting factor is proposed to prevent gain drift when the sliding surface becomes small.
- The implication of the forgetting mechanism stabilizes the gain, avoids high-frequency oscillations, and improves controller reliability in practical implementations.
- Comparative simulations with traditional SMC demonstrate that the ASMC achieves superior performance in terms of tracking precision and robustness.
- The results confirm the effectiveness of the ASMC in addressing the drawbacks of conventional SMC.

The paper is structured as follows: Section 2 presents an in-depth analysis of the dynamic modeling of a 2-DoF robotic manipulator. Section 3 outlines the proposed control strategies and their implementation. Section 4 discusses the simulation results. Finally, the conclusion, Section 5, highlights the main findings and contributions of the study.

2. Dynamic of a 2-DoF Arm Robot Manipulator

As illustrated in Figure 1, a 2-DOF robot arm manipulator is a robotic system with two independent joints, typically consisting of a base joint and an elbow joint, enabling motion in a planar workspace [35]. This type of manipulator is widely used in industrial automation, robotic research, and precision tasks requiring controlled motion. The manipulator consists of two rigid links of lengths l_1 and l_2, with masses m_1 and m_2, respectively, connected through rotary joints. The dynamic model of a 2-DOF robotic manipulator can be derived using Lagrange's formulation, considering joint torques, inertia, Coriolis, centrifugal forces, and external disturbances. The general equation of motion for a 2-DOF robot arm is given as follows [36,37]:

$$M(\Theta)\ddot{\Theta} + C(\Theta, \dot{\Theta})\dot{\Theta} + G(\Theta) = \tau \tag{1}$$

In this representation, the vectors Θ, $\dot{\Theta}$, and $\ddot{\Theta} \in \mathbb{R}^2$ correspond to the joint positions, velocities, and accelerations, respectively. The inertia matrix is denoted by $M(\Theta) \in \mathbb{R}^{2\times2}$, while $C(\Theta, \dot{\Theta}) \in \mathbb{R}^2$ is the vector of Coriolis and centrifugal forces. The gravitational forces acting on the system are represented by $G(\Theta) \in \mathbb{R}^2$. The applied joint torques are given by $\tau \in \mathbb{R}^{2\times1}$. The explicit expressions for $M(\Theta)$, $C(\Theta, \dot{\Theta})$, and $G(\Theta)$ can be found in Appendix A. Consequently, the system dynamics can be expressed as follows [38]:

$$\ddot{\Theta} = -M^{-1}C(\Theta, \dot{\Theta})\dot{\Theta} - M^{-1}(\Theta)G(\Theta) + M^{-1}(\Theta)\tau \tag{2}$$

The dynamic behavior of the 2-DOF robotic manipulator can be succinctly expressed in the following form as follows:

$$\ddot{\Theta} = -\alpha\dot{\Theta} - \beta G(\Theta) + \beta u \tag{3}$$

where $\alpha = M^{-1}C(\Theta, \dot{\Theta})$, $\beta = M^{-1}(\Theta)$, and $u = \tau$ represents the control input vector. To account for parameter variations and uncertainties in the system, the equation is modified as follows [39]:

$$\ddot{\Theta} = -(\alpha + \delta\alpha)\dot{\Theta} - (\beta + \delta\beta)G(\Theta) + (\beta + \delta\beta)u \tag{4}$$

where $\delta\alpha$ and $\delta\beta$ denote the uncertainties associated with the system parameters. The equation can be further expressed as follows:

$$\ddot{\Theta} = -\alpha\dot{\Theta} - \beta G(\Theta) + \beta u + d(t) \tag{5}$$

where $d(t) = -\delta\alpha\dot{\Theta} - \delta\beta G(\Theta) + \delta\beta u$.

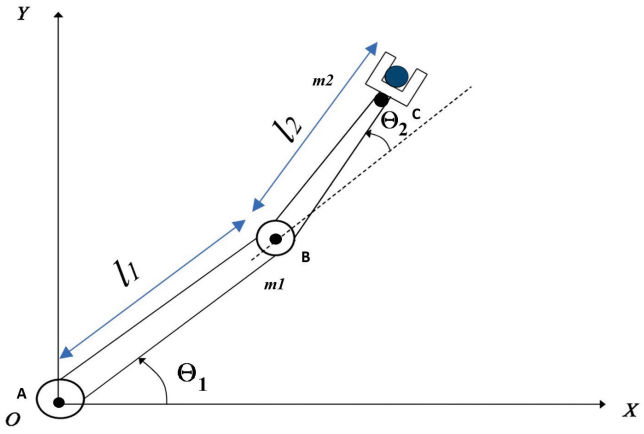


Figure 1. The 2 DoF rigid-link robot manipulator [40].

3. Control Design

This section presents the design of an adaptive sliding mode controller (ASMC) for a 2-DOF robotic arm, as illustrated in Figure 2. The ASMC is implemented on a proportional-derivative (PD) surface, which is utilized in both proposed controllers. The PD surface is chosen in order to minimize steady-state error, while the derivative term improves transient response, leading to enhanced tracking performance. The ASMC approach dynamically adjusts the control gains to enhance robustness and mitigate chattering, ensuring better adaptation to system uncertainties and disturbances.

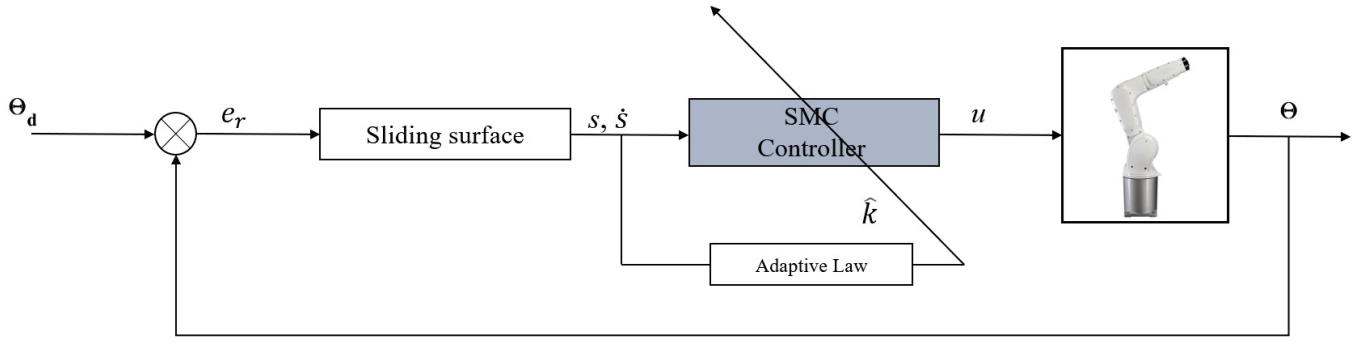


Figure 2. Block diagram illustrates the proposed control strategy.

3.1. Sliding Mode Control Scheme

SMC is a nonlinear control approach known for its robustness, ease of implementation and precision. Its framework is structured to guide the system states toward a predefined surface in the state space, referred to as the sliding surface. The design of this surface, denoted as $s(t)$, is a critical aspect of the SMC technique. A common approach to designing the sliding surface in SMC is through the use of a proportional-derivative (PD) structure. This method enhances stability and improves tracking performance by shaping the system dynamics on the sliding surface. The general form of a PD sliding surface is given as follows [40]:

$$s = K_p e + K_d \dot{e} \tag{6}$$

where $e = \Theta - \Theta_d$ is the tracking error vector between the actual position Θ and the desired position Θ_d, $\dot{e}$ is the derivative of the tracking error vector, Kp and Kd are diagonal positive proportional and derivative gains matrices, respectively, which designed to influence the convergence rate and stability of the system. Since the sliding surface s is a vector, each component of the sliding surface, denoted as s_i (where i = 1,2 ... m), is associated with a separate degree of freedom of the manipulator. This formulation ensures that when the

system reaches the sliding surface $s = 0$, the error dynamics behave like a proportional-derivative control system, leading to improved tracking accuracy and robustness.

Once the sliding surface is designed, the next step is to determine the control law that drives the system states toward and maintains them on this surface. In general, the control law in SMC comprises two components: the equivalent control u_{eq} and the switching control u_{sw}. The equivalent control ensures exact linearization, while the switching control is responsible for stabilizing the system by compensating for model uncertainties and rejecting disturbances. The overall control law is given by [40]:

$$u = u_{eq} + u_{sw} \tag{7}$$

where u_{eq} ensures that the nominal system dynamics follow the desired trajectory. u_{sw} introduces robustness, compensating for uncertainties and external disturbances. This formulation guarantees that the system converges to the sliding surface and maintains stability despite model uncertainties and perturbations. The nonlinear system under consideration can be expressed in state-space form as follows:

$$\dot{x} = f(t, x) + g(t, x)u \tag{8}$$

where $f(t, x)$ and $g(t, x)$ are continuous functions defining the system dynamics, and u is the control input. The objective of SMC is to design a control law that ensures the system state reaches and remains on a predefined sliding surface $s = 0$. As previously mentioned, the control input is decomposed into u_{eq} and u_{sw} (Equation (7)). The equivalent control u_{eq} is obtained by enforcing the sliding condition $\dot{s} = 0$, ensuring that the system dynamics remain on the sliding surface. Substituting Equation (6) and the system dynamics of Equation (8) into condition, leads to the general expression for u_{eq} as follows [41]:

$$u_{eq} = -\left[\frac{\partial s}{\partial x} g(t, x)\right]^{-1} \left(\frac{\partial s}{\partial x} f(t, x) + \frac{\partial s}{\partial t}\right). \tag{9}$$

This formulation ensures that the nominal system behavior follows the desired trajectory while maintaining stability. Using Equations (5), (6), and (9). Therefore, with negligible uncertainties, the equivalent control is given as follows:

$$u_{eq} = (K_d \beta)^{-1} \left[-K_p \dot{e} + K_d \alpha \dot{\Theta} + K_d \beta G + K_d \ddot{\Theta}_d\right] \tag{10}$$

The nonlinear component u_{sw} is designed in order to guarantee that the control variable moves toward the sliding surface while satisfying the convergence condition, which is given as follows [41]:

$$s^T \dot{s} < 0 \tag{11}$$

This condition ensures that the system state moves toward and remains on the sliding surface, satisfying stability requirements. The switching control condition is given by [42]:

$$u_{sw} = -k_{sw} \operatorname{sign}(s) \tag{12}$$

where k_{sw} is diagonal positive matrix defined as $k_{sw} = \operatorname{diag}\{k_{sw1}, k_{sw2}, \ldots, k_{swm}\}$. Finally, using Equations (7), (10), and (12), the total control input u is given by following:

$$u = (K_d \beta)^{-1} \left[-K_p \dot{e} + K_d \alpha \dot{\Theta} + K_d \beta G + K_d \ddot{\Theta}_d\right] - k_{sw} \operatorname{sign}(s) \tag{13}$$

In order to guarantee the stability of the SMC controller, Lyapunov theory is applied. A commonly used Lyapunov candidate function is chosen as follows [43]:

$$V(s) = \frac{1}{2} s^T s \tag{14}$$

Differentiating the Lyapunov function with respect to time yields

$$\dot{V}(s) = \frac{d}{dt}\left(\frac{1}{2} s^T s\right) = s^T \dot{s} \tag{15}$$

Substituting Equations (7) and (13) into Equation (15) yields the following:

$$\begin{aligned} \dot{V}(s) = s^T \dot{s} &= s^T \left[K_p \dot{e} + K_d \ddot{e}\right] \\ &= s^T \left[K_p \dot{e} + K_d(-\alpha \dot{\Theta} - \beta G) - K_d \ddot{\Theta}_d\right] \\ &+ s^T \left[K_d \beta \left((K_d \beta)^{-1}\left[-K_p \dot{e} + K_d \alpha \dot{\Theta} + K_d \beta G + K_d \ddot{\Theta}_d\right]\right)\right] \\ &+ s^T \left[K_d \beta(-k_{sw}\,\mathrm{sign}(s)) - K_d \ddot{\Theta}_d\right] \end{aligned} \tag{16}$$

Since u_{eq} eliminates the corresponding terms in $\dot{s}$, Equation (15) simplifies to

$$\dot{V}(s) = s^T \dot{s} = s^T(-K_d \beta k_{sw}\,\mathrm{sign}(s)) \tag{17}$$

Thus,

$$\dot{V}(s) = s^T(-K_d \beta k_{sw}\ \mathrm{sign}(s)) \leq 0, \quad \text{for all } s \neq 0 \tag{18}$$

Since k_{sw} is a diagonal positive matrix, Equation (17) satisfies the condition of Equation (16). Therefore, the system is globally asymptotically stable.

3.2. Adaptive Sliding Mode Control

The adaptive sliding mode control (ASMC) is an advanced control strategy designed to enhance the robustness and performance of SMC by dynamically adjusting the control gains in response to system uncertainties and disturbances [44]. Traditional SMC offers high robustness but suffers from chattering due to the high-frequency switching of the control input [45]. ASMC reduces this issue by incorporating an adaptive law that regulates the control parameters in real time, ensuring an optimal balance between robustness and smooth control action. By continuously tuning the control gains based on system states and tracking errors, ASMC improves convergence speed and reduces steady-state error. The complete control law using adaptive SMC is formulated as follows [46,47]:

$$u_{ad} = u_{eq} + u_{sad} \tag{19}$$

where the equivalent control is defined in (10), while the adaptive control is provided as follows [46]:

$$u_{sad} = -\hat{K}\,\mathrm{sign}(s) \tag{20}$$

where $\hat{K}$ is the switching adaptive gain defined as $\hat{K} = \mathrm{diag}\{\hat{K}_1, \hat{K}_2, \ldots, \hat{K}_m\}$. In order to enhance robustness and reduce chattering, ASMC employs an adaptive gain adjustment strategy. The element of the switching gain vector $\dot{\hat{K}}$ is updated according to the following adaptation law as follows [46]:

$$\dot{\hat{K}} = \gamma |s| \tag{21}$$

where $\dot{\hat{K}} = diag(\gamma_1|s_2|, \gamma_2|s_2|)$ and γ is the adaptive learning rate [46]. The adaptation law in (21) allows the control gain $\hat{K}$ to increase based on the magnitude of the sliding variable

s. However, this formulation has a critical drawback: Since $\dot{\hat{K}}$ is always positive when $s \neq 0$, the gain will increase continuously over time, leading to an effect known as gain drift [48]. The uncontrolled increase in $\hat{K}$ arises because the adaptation law relies merely on $|s|$, meaning that as long as $|s|$ remains nonzero, $\hat{K}$ continues to grow indefinitely. This can lead to excessively high gains, which may cause instability or high-frequency control actions in practical implementations. To address this issue, a forgetting factor is introduced to regulate the growth of $\hat{K}$ and ensure its boundedness over time. By incorporating a damping term, the adaptation law prevents excessive accumulation of $\hat{K}$ while still allowing it to adapt effectively according to the system tracking error. In order to mitigate gain drift, a forgetting factor $\lambda\hat{K}$ is introduced into the adaptation law, leading to the following modified update rule [49]:

$$\dot{\hat{K}} = \gamma|s| - \lambda\hat{K} \tag{22}$$

where $\lambda = diag(\lambda_1, \lambda_2)$ is a positive diagonal matrix that acts as a damping term that counteracts the continuous increase of $\hat{K}$. When $\hat{K}$ grows too large, the $-\lambda\hat{K}$ term reduces its rate of increase, ensuring that $\hat{K}$ remains bounded. This modification allows the controller to retain the benefits of adaptive tuning while avoiding excessive control gains that could degrade performance. Now, using Equations (10), (19) and (20), the final ASMC control law is expressed as

$$u = (K_d\beta)^{-1}\left[-K_p\dot{e} + K_d\alpha\dot{\Theta} + K_d\beta G + K_d\ddot{\Theta}_d\right] - \hat{K}\,\mathrm{sign}(s) \tag{23}$$

In order to demonstrate the proposed ASMC controller stability, a Lyapunov candidate is chosen as follows [50]:

$$V(s, \tilde{K}) = \frac{1}{2}s^T s + \frac{1}{2}\tilde{K}^T\tilde{K} \tag{24}$$

In the same manner, through simplification and substitution Equations (7) and (23) into Equation (24). Therefore, the time derivative of the Lyapunov is as follows:

$$\begin{aligned}
\dot{V}(s, \tilde{K}) = s^T\dot{s} + \tilde{K}^T\dot{\tilde{K}} &= s^T\left[K_p\dot{e} + K_d\ddot{e}\right] + \tilde{K}^T(\gamma|s| - \lambda\hat{K}) \\
&= s^T\left[K_p\dot{e} - K_d\alpha\dot{\Theta} - K_d\beta G\right] \\
&+ s^T\left[K_d\beta u_{eq} + K_d\beta u_{sad} - K_d\ddot{\Theta}_d\right] + \tilde{K}^T(\gamma|s| - \lambda\hat{K}) \\
&= s^T\left(-K_d\beta\hat{K}\,\mathrm{sign}(s)\right) + \tilde{K}^T(\gamma|s| - \lambda\hat{K})
\end{aligned}$$

The term $(\gamma|s| - \lambda\hat{K})$ becomes negative if λ is chosen to be sufficiently large; thus,

$$\begin{aligned}
\dot{V}(s, \tilde{K}) &= s^T\left(-K_d\beta\hat{K}\,\mathrm{sign}(s)\right) + \tilde{K}^T(\gamma|s| - \lambda\hat{K}) \\
&\leq s^T\left(-K_d\beta\hat{K}\,\mathrm{sign}(s)\right)
\end{aligned} \tag{25}$$

4. Results and Discussion

This section presents a comparative analysis of the control performance of the ASMC and SMC applied to a 2-DOF arm robot manipulator. The evaluation focuses on trajectory tracking accuracy, chattering elimination, and overall system response to the desired motion trajectories under noise application. Table 1 represents the initial values and proposed controller parameters.

Table 1. Initial values and parameters for ASMC and SMC.

Parameter	ASMC	SMC
Desired trajectory $\Theta_{d1}(t)$	$\sin(4.17t)$	
Desired trajectory $\Theta_{d2}(t)$	$1.2\sin(5.11t)$	
Initial values (Case 1)	$\Theta_1(0) = \pi, \Theta_2(0) = -\pi, \dot{\Theta}_1(0) = 0, \dot{\Theta}_2(0) = 0$	
Initial values (Case 2)	$\Theta_1(0) = 0, \Theta_2(0) = 0, \dot{\Theta}_1(0) = 0, \dot{\Theta}_2(0) = 0$	
Forgetting factor λ	20	-
Learning rate γ	0.01	-
Initial adaptive gain K	0.001	-
k_{sw}	-	900
K_p	8000	3000
K_d	200	60

After tuning the ASMC and SMC controllers parameters using the trial-and-error method, the results demonstrate that ASMC consistently and significantly improves the tracking performance of the robotic arm over different initial conditions. Table 2 shows a comprehensive comparison of indices for both controllers under two distinct scenarios: $\Theta_1(0) = \pi, \Theta_2(0) = -\pi$ and $\Theta_1(0) = 0, \Theta_2(0) = 0$. For the first case, where $\Theta_1(0) = \pi$ and $\Theta_2(0) = -\pi$, ASMC achieves lower trajectory errors, with an **RMSE** of 0.15754 for Θ_1 and 0.15263 for Θ_2, contrast to SMC, which records higher **RMSE** values of 0.40713 and 0.43948, corresponding to improvements of 61.3% and 65.3%, respectively. Additionally, ASMC exhibits superior integral performance, achieving lower **ITAE** values of 0.21954 for Θ_1 and 0.31623 for Θ_2, while SMC results in higher **ITAE** values of 0.39639 and 0.55378, represents enhancements of 44.6% and 42.9%. The **IAE** metric further highlights the ASMC advantage, with values of 0.17535 and 0.20945 for Θ_1 and Θ_2, respectively, in contrast to SMC, which yields 0.47994 and 0.60244. Moreover, ASMC achieves lower **ISE** values of 0.094368 and 0.08827, whereas SMC shows significantly higher **ISE** values of 0.65823 and 0.76782. Likewise, the **IAE** and **ISE** values for ASMC are consistently better, achieving reductions of 63.5% and 65.2% in **IAE**, and 85.7% and 88.5% in **ISE** for Θ_1 and Θ_2, respectively. The maximum absolute error for ASMC is 3.1416 rad for both Θ_1 and Θ_2, which is similar to the SMC maximum absolute error of 3.1416 rad for Θ_1 and 3.1552 rad for Θ_2. In the second case, with initial conditions $\Theta_1(0) = 0$ and $\Theta_2(0) = 0$, a similar trend is observed. ASMC again outperforms SMC, achieving lower **RMSE** values of 0.030033 for Θ_1 and 0.044172 for Θ_2, corresponding to improvements of 49.2% and 48.3% over SMC. The **ITAE** values improve by 46.4% for Θ_1 and 40.7% for Θ_2, while the **IAE** values improve by 49.2% and 46.7%, respectively. Furthermore, ASMC achieves notable reductions in **ISE**, improving by 74.2% for Θ_1 and 73.2% for Θ_2 compared to the corresponding SMC values. These results confirm that ASMC provides enhanced accuracy and control efficiency in trajectory tracking for robotic arms.

Table 2. Performance comparison of ASMC and SMC for different initial conditions.

Metric	ASMC	SMC	Improvement (%)
$\Theta_1(0) = \pi, \Theta_2(0) = -\pi$			
RMSE Θ_1	0.15754	0.40713	61.3%
RMSE Θ_2	0.15263	0.43948	65.3%
ITAE Θ_1	0.21954	0.39639	44.6%
ITAE Θ_2	0.31623	0.55378	42.9%
IAE Θ_1	0.17535	0.47994	63.5%
IAE Θ_2	0.20945	0.60244	65.2%
ISE Θ_1	0.094368	0.65823	85.7%
ISE Θ_2	0.08827	0.76782	88.5%
$\Theta_1(0) = 0, \Theta_2(0) = 0$			
RMSE Θ_1	0.030033	0.059083	49.2%
RMSE Θ_2	0.044172	0.085383	48.3%
ITAE Θ_1	0.21659	0.40427	46.4%
ITAE Θ_2	0.31583	0.53307	40.7%
IAE Θ_1	0.10674	0.21021	49.2%
IAE Θ_2	0.15874	0.29787	46.7%
ISE Θ_1	0.0036084	0.013965	74.2%
ISE Θ_2	0.0078066	0.029168	73.2%

The figures presented are split into two parts: part (a) corresponds to Case 1 with initial conditions $\Theta_1(0) = \pi$, $\Theta_2(0) = -\pi$, and part (b) corresponds to Case 2 with $\Theta_1(0) = 0$, $\Theta_2(0) = 0$. Figure 3 illustrates the position tracking control of Θ_1 and Θ_2 using ASMC in comparison with SMC. It is evident from the figure that ASMC achieves a quicker and more efficient alignment with the reference trajectory than SMC. The 2-DoF arm robot manipulator, when controlled by ASMC, reaches the desired trajectory more swiftly and accurately, whereas SMC also demonstrates satisfactory performance. Figure 4 depicts the position tracking error of Θ_1 and Θ_2 under ASMC and SMC. As shown in Figure 4, ASMC results in a smaller tracking error compared to SMC, indicating improved precision. Figure 5 illustrates the sliding surface for both ASMC and SMC over time. According to this figure, both controllers successfully drive the system to the sliding surface, maintaining it at zero. However, noticeable chattering is observed in the SMC response, whereas ASMC achieves a smoother trajectory without chattering. This demonstrates the effectiveness of the adaptive mechanism in ASMC, which mitigates high-frequency oscillations while ensuring precise tracking. The velocity of joints Θ_1 and Θ_2 under the applied controllers is shown in Figure 6. According to Figure 6, ASMC enables the 2-DOF arm robot manipulator to settle more quickly to the set point compared to SMC. Additionally, ASMC bears and eliminates the chattering, with velocity fluctuations nearly vanishing, in contrast to SMC, where noticeable chattering is present. However, it should be noted that, in order to reduce the chattering effect typically encountered in conventional SMC, the discontinuous sign function can be replaced with a continuous approximation such as the saturation function or the hyperbolic tangent function $\tanh(s)$ [51]. This modification smooths the control signal and makes the controller more suitable for practical implementations, where high-frequency switching can lead to actuator wear or undesirable vibrations. However, such continuous approximations cause the system to converge to a boundary layer around the sliding surface rather than reaching $s = 0$ exactly, which may slightly reduce tracking accuracy. In contrast, the proposed ASMC technique dynamically adjusts the switching gain in response to the tracking error, improving robustness and adaptability and significantly mitigating chattering while maintaining a high level of tracking precision. Figure 7 illustrates the control output signals for ASMC and SMC, highlighting differences in their performance.

The SMC controller shows significant oscillations in its control signal, particularly after 0.2 s, indicating less stable control behavior. These oscillations could lead to undesirable effects in real-world applications such as heat, energy losses, and system instability. In contrast, ASMC maintains a smooth control signal, reflecting its stability and precision. In order to evaluate the robustness of the proposed ASMC and SMC controllers, random noise is introduced into the system. Figures 8 and 9 illustrate the effect of noise on the error signal for SMC and ASMC, respectively. As shown in Figure 8, the SMC controller maintains stability for noise levels ranging from $d_1 = 10 \times \text{randn}(1,1)$ to $d_2 = 100 \times \text{randn}(1,1)$. However, when subjected to $d_3 = 500 \times \text{randn}(1,1)$, $d_4 = 1000 \times \text{randn}(1,1)$, and $d_5 = 4000 \times \text{randn}(1,1)$, the error signal exhibits significant fluctuations, exceeding the range of $[-0.5, 0.25]$ for Θ_1 and $[-0.6, 0.5]$ for Θ_2 under high noise value of d_5. In contrast, Figure 9 demonstrates that the ASMC controller provides enhanced stability under the same conditions. Even with high noise d_5 applied, the error variations remain more controlled, within the range of $[-0.08, 0.2]$ for Θ_1 and $[-0.05, 0.15]$ for Θ_2. This indicates that ASMC exhibits superior resilience to high noise levels compared to SMC, ensuring more reliable performance in the presence of external disturbances.

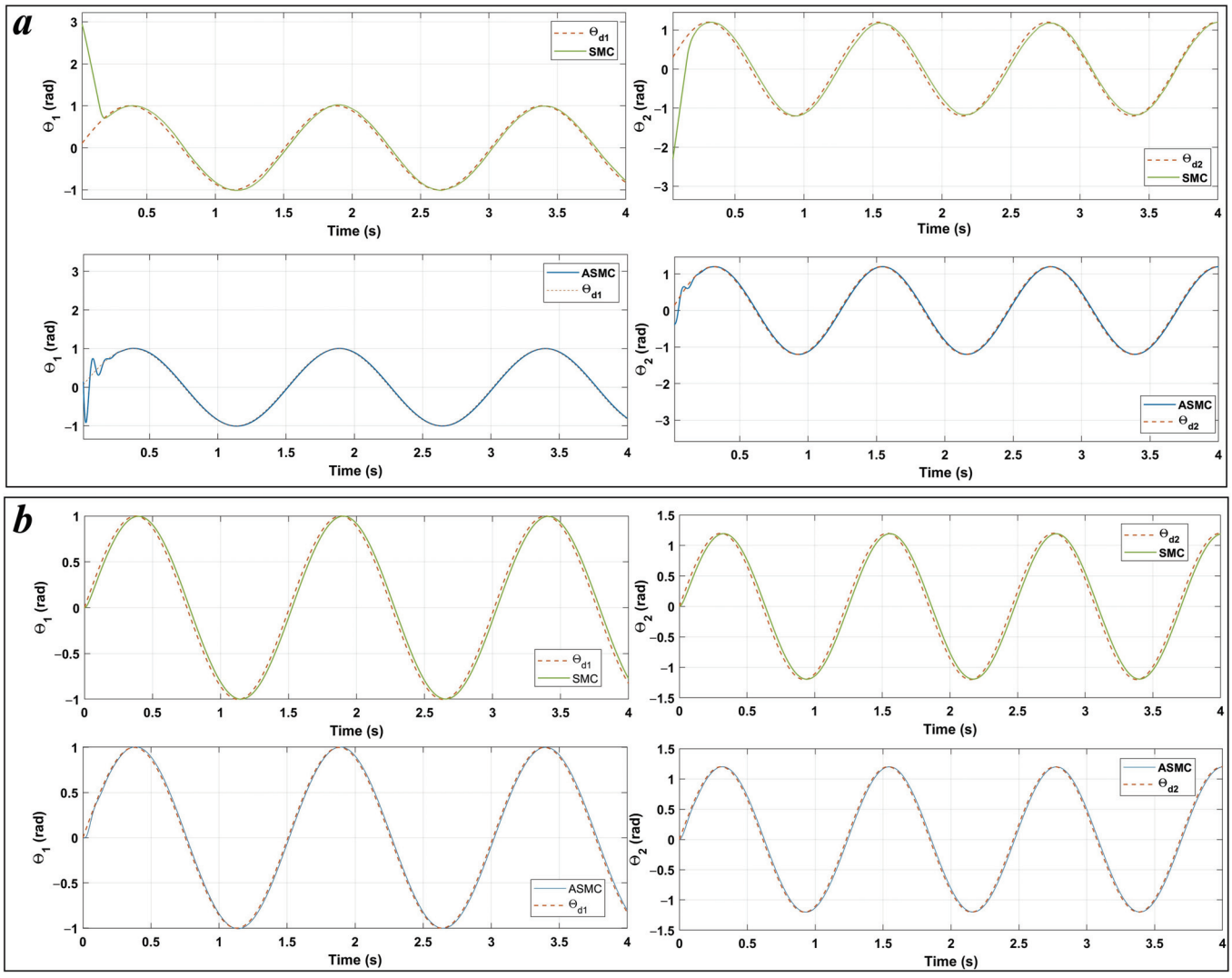


Figure 3. Position tracking joints under ASMC and SMC for Case 1 (**a**), and Case 2 (**b**).

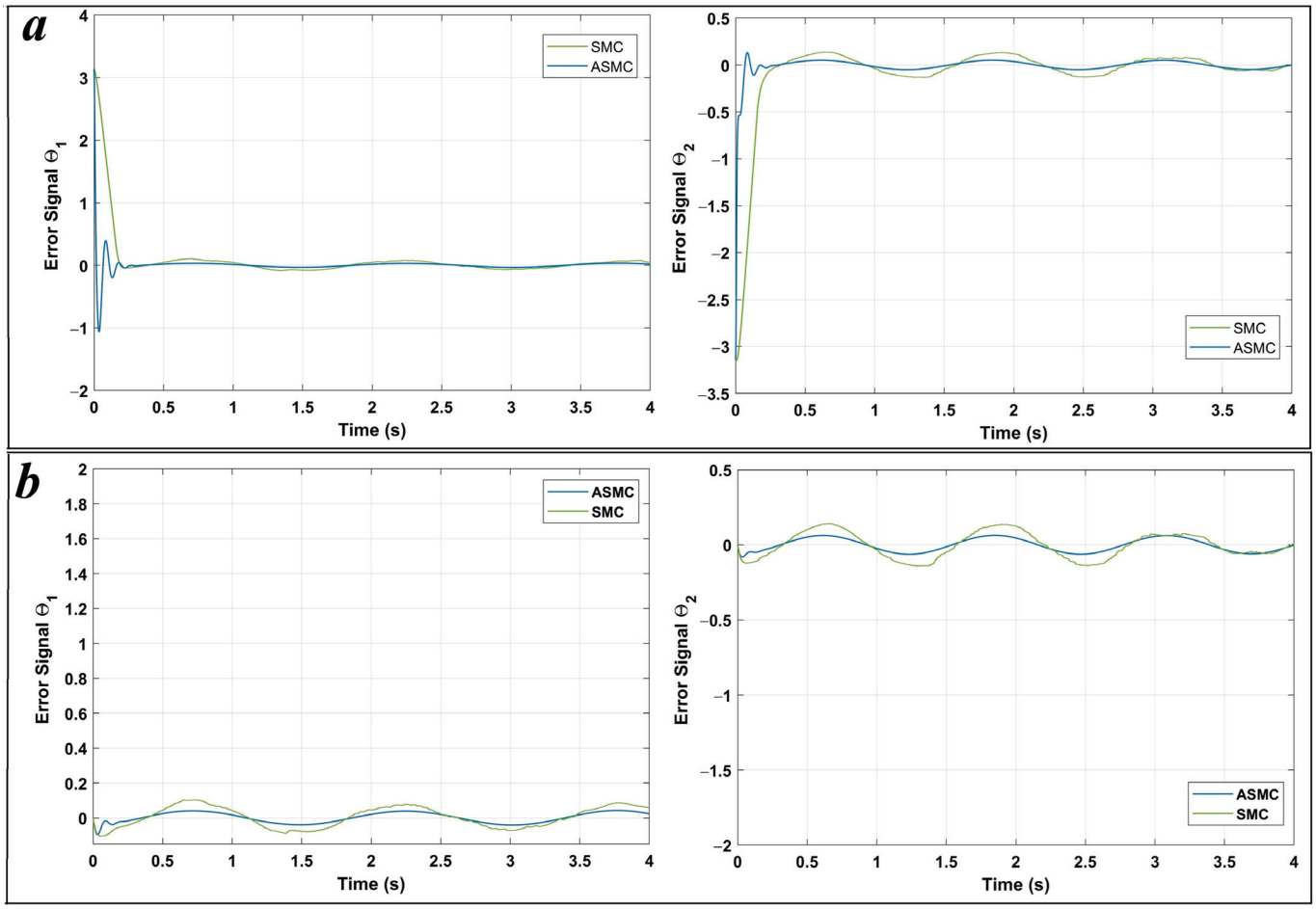


Figure 4. Position tracking error joints under ASMC and SMC controllers for Case 1 (**a**), and Case 2 (**b**).

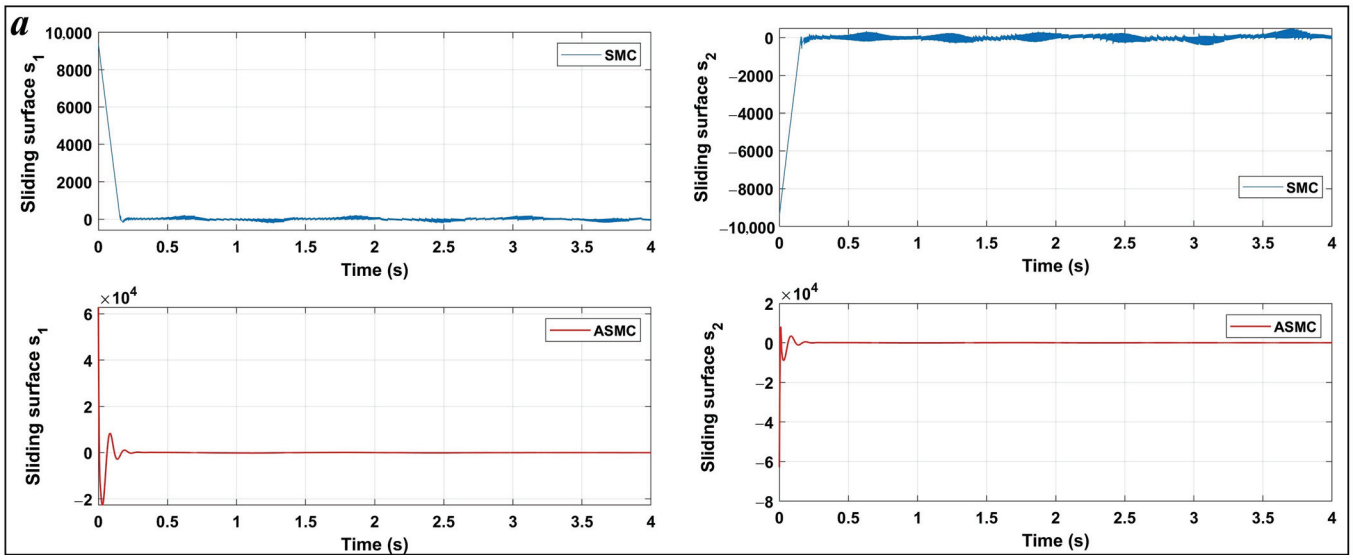


Figure 5. *Cont.*

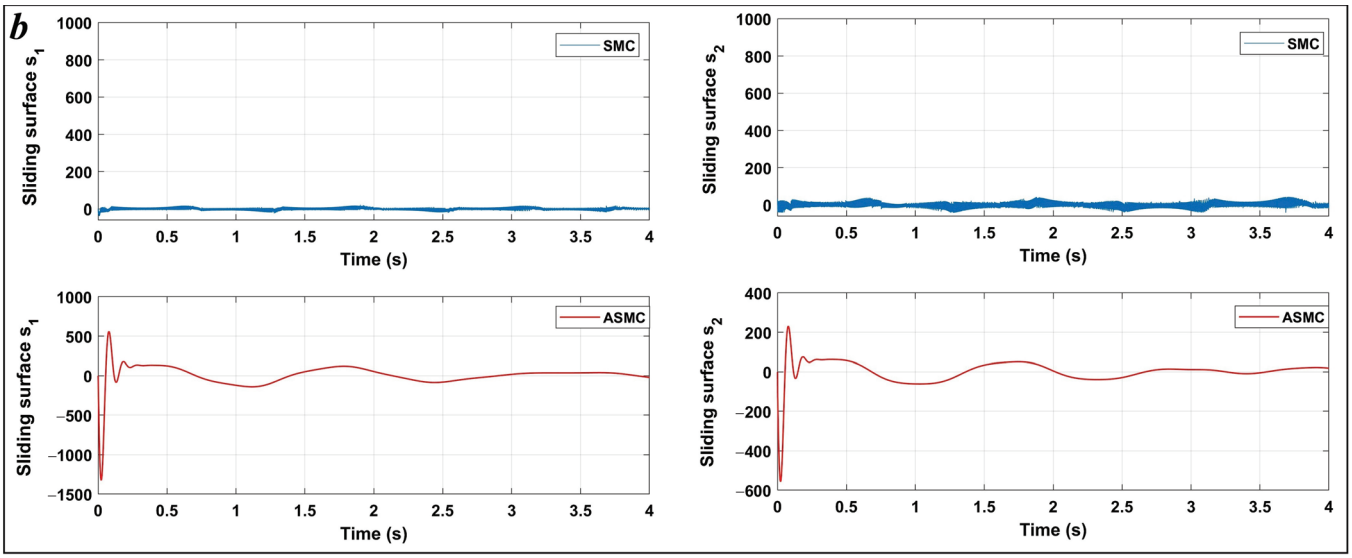


Figure 5. Sliding surfaces for ASMC and SMC for Case 1 (**a**), and Case 2 (**b**).

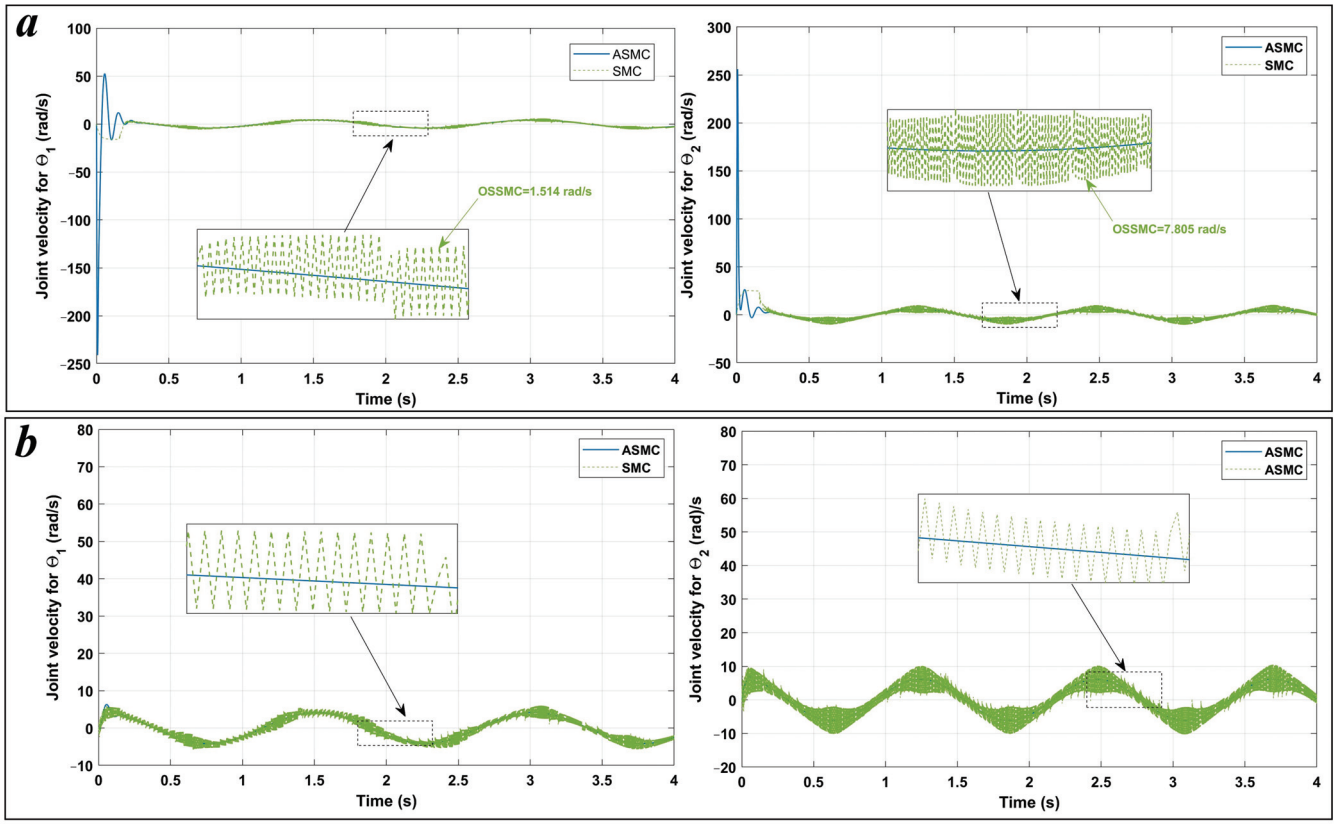


Figure 6. Velocity joints under ASMC and SMC controllers for Case 1 (**a**), and Case 2 (**b**).

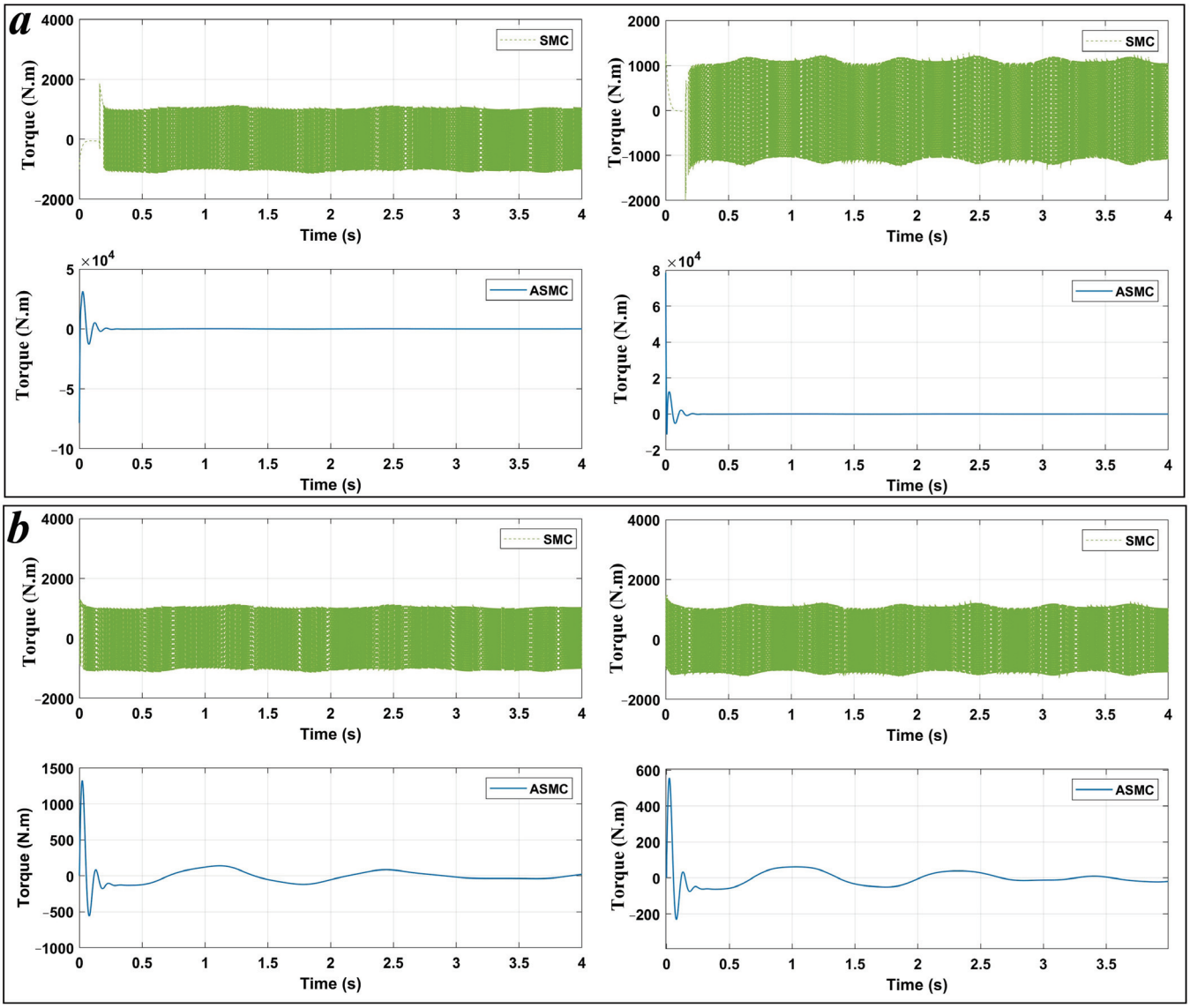


Figure 7. Control signals of the ASMC and SMC controllers for Case 1 (**a**), and Case 2 (**b**).

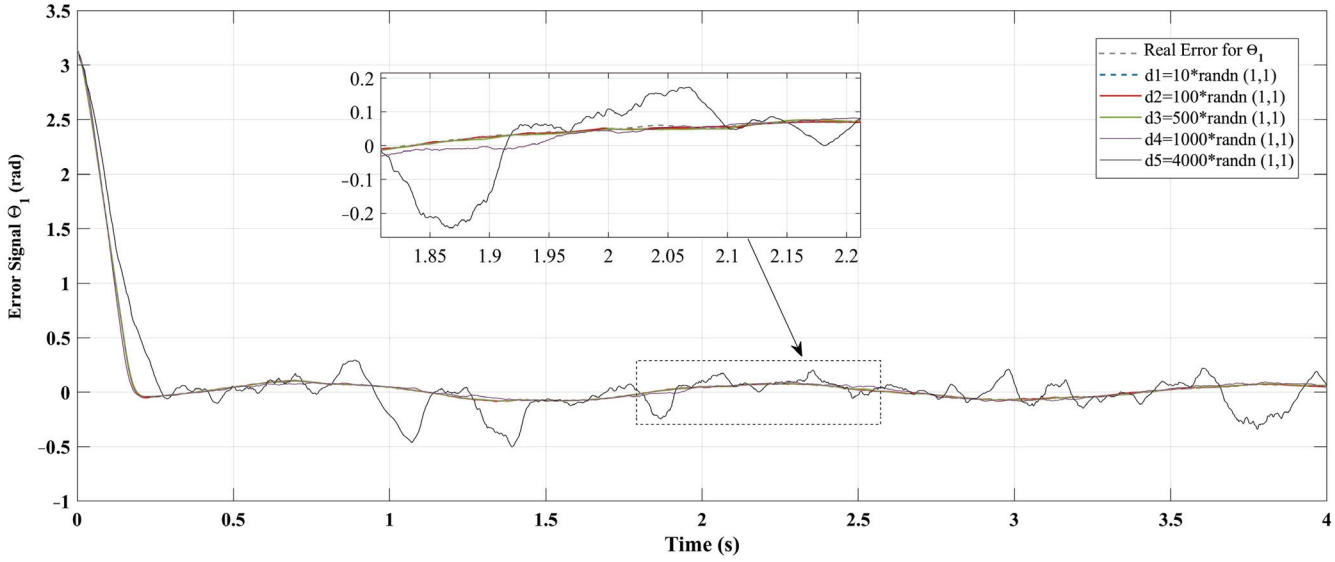


Figure 8. *Cont*.

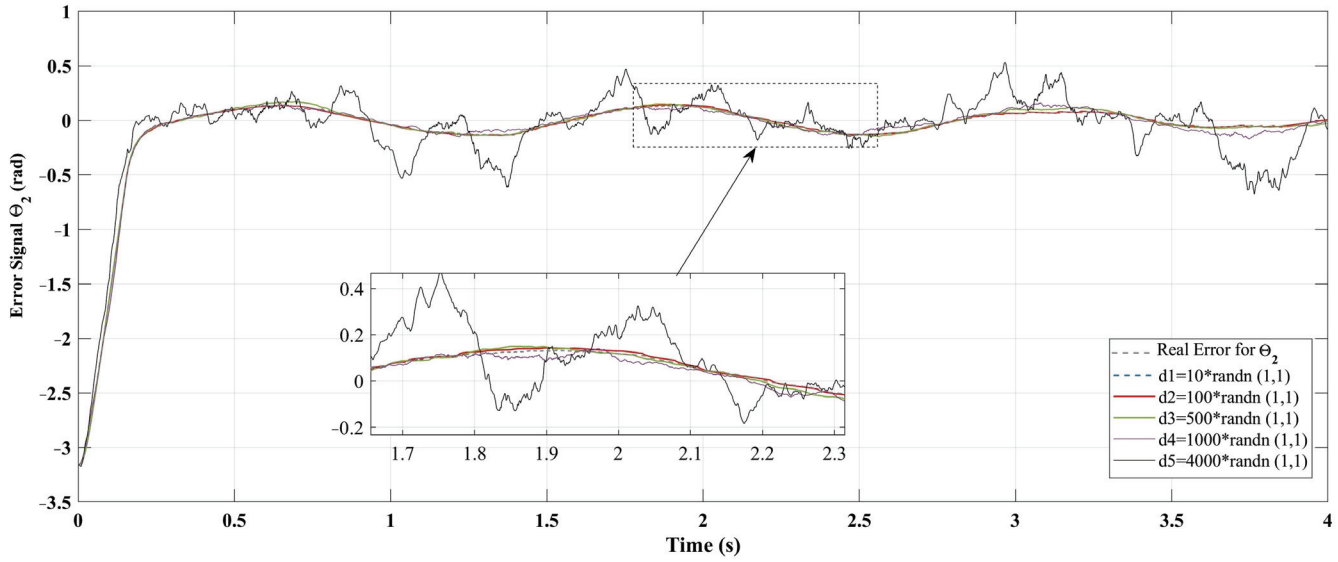


Figure 8. Position tracking error of joints of the SMC controller under random noises application for Case 1.

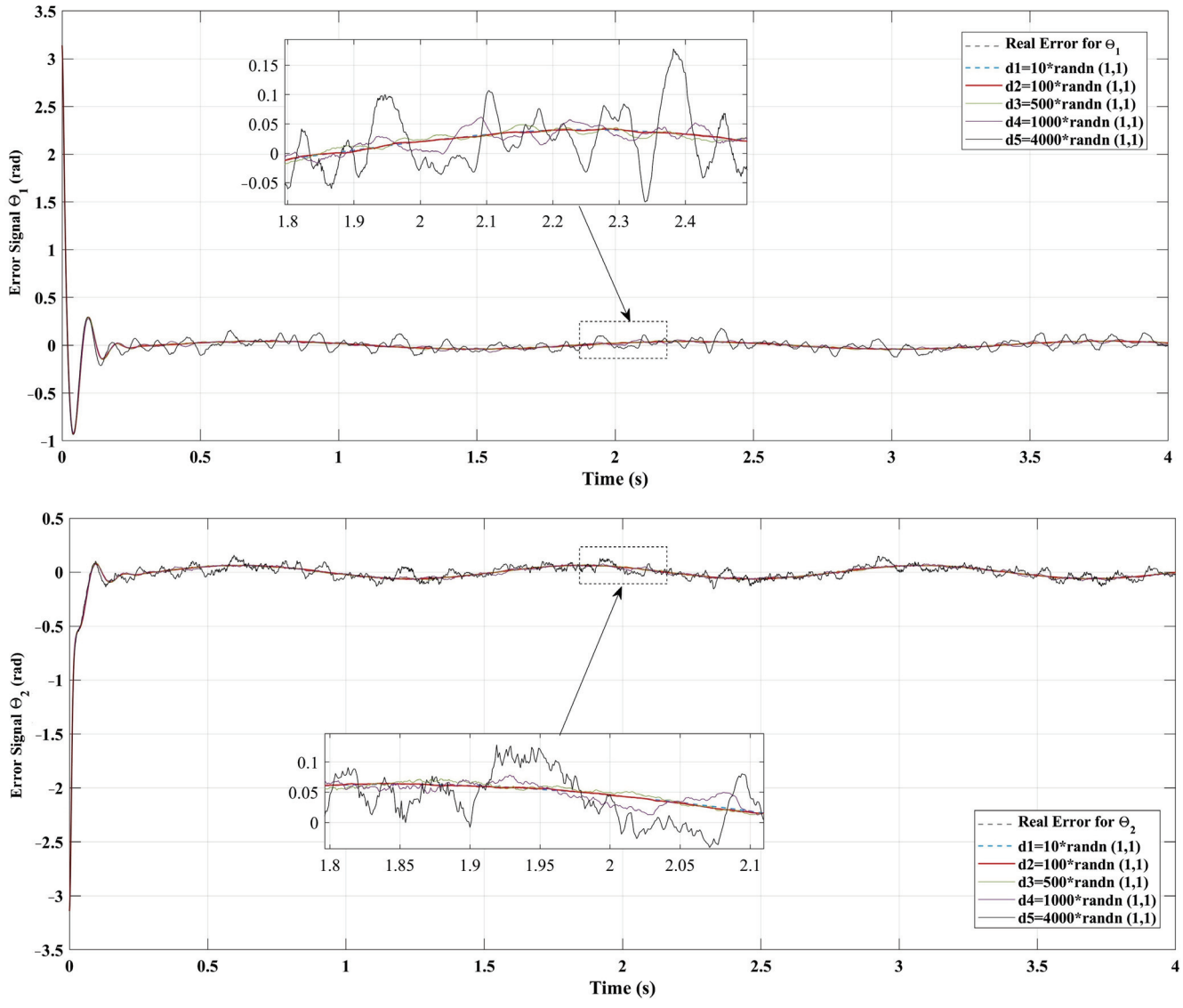


Figure 9. Position tracking error of joints of the proposed ASMC under random noises application for Case 1.

5. Conclusions

This paper presents an advanced control approach for a 2-DoF robot manipulator, using an adaptive sliding mode controller (ASMC) to enhance trajectory tracking performance and a solution for the chattering phenomenon. The proposed ASMC method significantly reduces tracking errors and mitigates chattering, addressing a key challenge in sliding mode control. Unlike conventional SMC, which often suffers from high-frequency oscillations, ASMC improves system stability while maintaining robust performance. Moreover, the robustness of ASMC is evaluated under the influence of high external noise, demonstrating its ability to maintain accurate trajectory tracking and system stability despite disturbances. The numerical simulations demonstrate the superiority of ASMC over SMC, highlighting its improved tracking accuracy, lower error indices, and enhanced resilience to noise. For future research, experimental validation and the extension of the proposed control strategy to multi-DOF robotic systems and real-world applications are planned.

Author Contributions: Conceptualization, M.Y.S.; methodology, M.Y.S.; software, M.Y.S.; validation, M.Y.S., A.B. and O.B.; formal analysis, M.Y.S. and O.B.; investigation, M.Y.S., A.B. and O.B.; resources, O.B.; data curation, M.Y.S. and O.B.; writing—original draft preparation, M.Y.S.; writing—review and editing, M.Y.S., A.B., O.B. and I.R.; visualization, M.Y.S., O.B. and A.B.; supervision, O.B.; project administration, M.Y.S.; funding acquisition, O.B. All authors have read and agreed to the published version of the manuscript.

Funding: The authors wish to express their gratitude to the Basque Government, through the project EKOHEGAZ II (ELKARTEK KK-2023/00051), to the UPV/EHU, through the project GIU23/002, and to the MobilityLab Foundation (CONV23/14) for supporting this work.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article. Further inquiries can be directed to the corresponding authors.

Acknowledgments: The authors are grateful for the projects supported by the Telecommunications Signals and Systems Laboratory (TSS), University Amar Telidji. The authors would like to thank all the associates who have contributed to this work directly or indirectly. The authors wish to express their gratitude to the Basque Government, through the project EKOHEGAZ II (ELKARTEK KK-2023/00051), to the UPV/EHU, through the project GIU23/002, and to the MobilityLab Foundation (CONV23/14) for supporting this work. During the preparation of this manuscript, the authors used OpenAI and Grammarly for language editing and improving the clarity of the manuscript. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations and nomenclatures are used in this manuscript:

2-DoF	2-Degree-of-Freedom
ASMC	Adaptive Sliding Mode Control
SMC	Sliding Mode Control
PID	Proportional-Integral-Derivative
FLC	Fuzzy Logic Controller
APSO	Adaptive Particle Swarm Optimization
SPA	Social Spider Algorithm
HOSMC	High-Order Sliding Mode Control
NNC	Neural Network Control

ZN	Ziegler–Nichols
IMC	Internal Model Control
MPC	Model Predictive Control
MIMO	Multi-Input Multi-Output
DSSMC	Dual Surface Sliding Mode Controller
NFTSM	Non-Singular Fast Terminal Sliding Mode
WNN	Wavelet Neural Network
MRAC	Model Reference Adaptive Controllers
NFTSM	Non-Singular Fast Terminal Sliding Mode

Appendix A

The parameters dynamics of the 2-DoF arm robot manipulator, as expressed in Equation (1), are determined as follows [10,52]:

$$\boldsymbol{\Theta} = \begin{bmatrix} \Theta_1 \\ \Theta_2 \end{bmatrix}$$

$$M(\Theta) = \begin{bmatrix} (m_1 + m_2) \cdot l_1^2 + m_2 \cdot l_2^2 + 2 \cdot m_2 \cdot l_1 \cdot l_2 \cdot \cos(\Theta_2) & m_2 \cdot l_2^2 + m_2 \cdot l_1 \cdot l_2 \cdot \cos(\Theta_2) \\ m_2 \cdot l_2^2 + m_2 \cdot l_1 \cdot l_2 \cdot \cos(\Theta_2) & m_2 \cdot l_2^2 \end{bmatrix}$$

$$C(\Theta, \dot{\Theta})\dot{\Theta} = \begin{bmatrix} -m_2 \cdot l_1 \cdot l_2 \cdot \sin(\Theta_2) \cdot (2 \cdot \dot{\Theta}_1 \cdot \dot{\Theta}_2 + \dot{\Theta}_2^2) \\ -m_2 \cdot l_1 \cdot l_2 \cdot \sin(\Theta_2) \cdot \dot{\Theta}_1 \cdot \dot{\Theta}_2 \end{bmatrix}$$

$$G(\Theta) = \begin{bmatrix} (m_1 + m_2) \cdot g \cdot l_1 \cdot \cos(\Theta_1) + m_2 \cdot g \cdot l_2 \cdot \cos(\Theta_1 + \Theta_2) \\ m_2 \cdot g \cdot l_2 \cdot \cos(\Theta_1 + \Theta_2) \end{bmatrix}$$

$$\boldsymbol{\tau} = \begin{bmatrix} \tau_1 \\ \tau_2 \end{bmatrix}$$

where $m_1 = m_2 = 1\ kg$, $l_1 = l_2 = 1\ m$, and $g = 9.8\ m/s^2$.

References

1. Billard, A.; Kragic, D. Trends and challenges in robot manipulation. *Science* **2019**, *364*, eaat8414. [CrossRef] [PubMed]
2. Garriz, C.; Domingo, R. Development of trajectories through the kalman algorithm and application to an industrial robot in the automotive industry. *IEEE Access* **2019**, *7*, 23570–23578. [CrossRef]
3. Nasrat, S.; Kang, T.; Park, J.; Kim, J.; Yi, S. Artistic robotic arm: Drawing portraits on physical canvas under 80 seconds. *Sensors* **2023**, *23*, 5589. [CrossRef]
4. Ali, M.H.M.; Atia, M.R. A lead through approach for programming a welding arm robot using machine vision. *Robotica* **2022**, *40*, 464–474. [CrossRef]
5. Hayakawa, S.; Wan, W.; Koyama, K.; Harada, K. A dual-arm robot that manipulates heavy plates with the support of a vacuum lifter. *IEEE Trans. Autom. Sci. Eng.* **2022**, *20*, 2808–2821. [CrossRef]
6. Zhang, W.; Li, H.; Cui, L.; Li, H.; Zhang, X.; Fang, S.; Zhang, Q. Research progress and development trend of surgical robot and surgical instrument arm. *Int. J. Med. Robot. Comput. Assist. Surg.* **2021**, *17*, e2309. [CrossRef]
7. Coloma, S.; Peralta, P.E.; Redondo, V.; Morono, A.; Vila, R.; Ferre, M. The effect of ionizing radiation on robotic trajectory movement and electronic components. *Nucl. Eng. Technol.* **2023**, *55*, 4191–4203. [CrossRef]
8. Shakibjoo, A.D.; Shakibjoo, M.D. 2-DOF PID with reset controller for 4-DOF robot arm manipulator. In Proceedings of the 2015 International Conference on Advanced Robotics and Intelligent Systems (ARIS), Taipei, Taiwan, 29–31 May 2015; pp. 1–6.
9. Adar, N.; Kozan, R. Comparison between real time PID and 2-DOF PID controller for 6-DOF robot arm. *Acta Phys. Pol. A* **2016**, *130*, 269–271. [CrossRef]
10. Silaa, M.Y.; Barambones, O.; Bencherif, A. A novel adaptive PID controller design for a PEM fuel cell using stochastic gradient descent with momentum enhanced by whale optimizer. *Electronics* **2022**, *11*, 2610. [CrossRef]
11. Sung, S.W.; Lee, I.B. Limitations and countermeasures of PID controllers. *Ind. Eng. Chem. Res.* **1996**, *35*, 2596–2610. [CrossRef]
12. Saha, S.; Amrr, S.M.; Saidi, A.S.; Banerjee, A.; Nabi, M. Finite-time adaptive higher-order SMC for the nonlinear five DOF active magnetic bearing system. *Electronics* **2021**, *10*, 1333. [CrossRef]
13. Silaa, M.Y.; Bencherif, A.; Barambones, O. Indirect Adaptive Control Using Neural Network and Discrete Extended Kalman Filter for Wheeled Mobile Robot. *Actuators* **2024**, *13*, 51. [CrossRef]
14. Bristow, D.A.; Alleyne, A.G. A high precision motion control system with application to microscale robotic deposition. *IEEE Trans. Control Syst. Technol.* **2006**, *14*, 1008–1020. [CrossRef]

15. Djaneye-Boundjou, O.; Xu, X.; Ordóñez, R. Automated particle swarm optimization based PID tuning for control of robotic arm. In Proceedings of the 2016 IEEE National Aerospace and Electronics Conference (NAECON) and Ohio Innovation Summit (OIS), Dayton, OH, USA, 25–29 July 2016; pp. 164–169.
16. Wazzan, A.N.; Basil, N.; Raad, M.; Mohammed, H.K. PID controller with robotic arm using optimization algorithm. *Int. J. Mech. Eng.* **2022**, *7*, 3746–3751.
17. Milind, J.S.; Arunkumar, G.; Manjunath, T. PID control of a double link (2-link) flexible robotic manipulator (2-DOF) in the 3 DE space. In Proceedings of the 2018 4th International Conference for Convergence in Technology (I2CT), Mangalore, India, 27–28 October 2018; pp. 1–7.
18. Zakia, U.; Moallem, M.; Menon, C. PID-SMC controller for a 2-DOF planar robot. In Proceedings of the 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 7–9 February 2019; pp. 1–5.
19. Silva, F.; Batista, J.; Souza, D.; Lima, A.; dos Reis, L.; Barbosa, A. Control and identification of parameters of a joint of a manipulator based on PID, PID 2-DOF, and least squares. *J. Braz. Soc. Mech. Sci. Eng.* **2023**, *45*, 327. [CrossRef]
20. Azar, A.T.; Serrano, F.E. Fractional order two degree of freedom pid controller for a robotic manipulator with a fuzzy type-2 compensator. *J. Braz. Soc. Mech. Sci. Eng.* **2019**, 77–88.
21. Mohammed, A.A.; El-Nagar, A.M.; Elsheikh, E.A.; El-Bardini, M. Embedded adaptive 2-DoF PID controller for robot manipulator using a supervisory fuzzy logic system. *Menoufia J. Electron. Eng. Res.* **2022**, *31*, 55–62. [CrossRef]
22. Bi, M. Control of robot arm motion using trapezoid fuzzy two-degree-of-freedom PID algorithm. *Symmetry* **2020**, *12*, 665. [CrossRef]
23. Tuan, H.M.; Sanfilippo, F.; Hao, N.V. A novel adaptive sliding mode controller for a 2-DOF elastic robotic arm. *Robotics* **2022**, *11*, 47. [CrossRef]
24. Jouila, A.; Nouri, K. An adaptive robust nonsingular fast terminal sliding mode controller based on wavelet neural network for a 2-DOF robotic arm. *J. Frankl. Inst.* **2020**, *357*, 13259–13282. [CrossRef]
25. Sun, Y.; Jiang, W.K. Research on Adaptive Fuzzy Backstepping Control Method of 2-DOF Manipulator. In Proceedings of the 2021 6th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Oita, Japan, 25–27 November 2021; pp. 63–67.
26. Rahmani, M.; Komijani, H.; Rahman, M.H. New sliding mode control of 2-DOF robot manipulator based on extended grey wolf optimizer. *Int. J. Control Autom. Syst.* **2020**, *18*, 1572–1580. [CrossRef]
27. Hameed, W.N.; Khawwaf, J.O. Robust sliding mode control for 2-Dof robot manipulator position control system. *IET Conf. Proc.* **2024**, *18*, 282–288. [CrossRef]
28. Balthazar, J.M.; de Oliveira, C.; da Rosa, R.A.; Gonçalves, G.; Lenzi, G.G.; Tusset, A.M. Efficiency comparative of linear controllers in positioning control of a robotic manipulator with two degrees of freedom. *Math. Eng. Sci. Aerosp. (MESA)* **2024**, *15*.
29. Bingül, Z.; Karahan, O. A Fuzzy Logic Controller tuned with PSO for 2 DOF robot trajectory control. *Expert Syst. Appl.* **2011**, *38*, 1017–1031. [CrossRef]
30. Girgis, M.E.; Nasr, A.E. Trajectory tracking of 2DOF robotic manipulator system based on adaptive exponential forgetting recursive least squares fractional-order PID control. *Proc. Inst. Mech. Eng. Part J. Syst. Control Eng.* **2024**, *38*, 09596518251322228. [CrossRef]
31. Ramesh, R.; Mannan, M.A.; Poo, A. Tracking and contour error control in CNC servo systems. *Int. J. Mach. Tools Manuf.* **2005**, *45*, 301–326. [CrossRef]
32. Kong, X.; Wang, L.; Yu, G.; Li, W.; Li, M. Research on servo matching of a five-axis hybrid machine tool. *Int. J. Adv. Manuf. Technol.* **2023**, *129*, 983–997. [CrossRef]
33. Wang, L.; Kong, X.; Yu, G.; Li, W.; Li, M.; Jiang, A. Error estimation and cross-coupled control based on a novel tool pose representation method of a five-axis hybrid machine tool. *Int. J. Mach. Tools Manuf.* **2022**, *182*, 103955. [CrossRef]
34. Hu, C.; Yao, B.; Wang, Q. Global task coordinate frame-based contouring control of linear-motor-driven biaxial systems with accurate parameter estimations. *IEEE Trans. Ind. Electron.* **2011**, *58*, 5195–5205. [CrossRef]
35. Howard, I.S.; Ingram, J.N.; Wolpert, D.M. A modular planar robotic manipulandum with end-point torque control. *J. Neurosci. Methods* **2009**, *181*, 199–211. [CrossRef]
36. Sachan, S.; Swarnkar, P. Intelligent fractional order sliding mode based control for surgical robot manipulator. *Electronics* **2023**, *12*, 729. [CrossRef]
37. Song, Z.; Bao, D.; Wang, W.; Zhao, W. Adaptive Dynamic Boundary Sliding Mode Control for Robotic Manipulators under Varying Disturbances. *Electronics* **2024**, *13*, 900. [CrossRef]
38. Sinha, S.; Nethi, A.; Jetta, M. Multiplicative Gaussian Noise Removal using Partial Differential Equations and Activation Functions: A Robust and Stable Approach. In Proceedings of the 7th International Conference on Algorithms, Computing and Systems, Larissa, Greece, 19–21 October 2023 ; pp. 161–170.

39. Guo, K.; Shi, P.; Wang, P.; He, C.; Zhang, H. Non-singular terminal sliding mode controller with nonlinear disturbance observer for robotic manipulator. *Electronics* **2023**, *12*, 849. [CrossRef]
40. Silaa, M.Y.; Barambones, O.; Bencherif, A. Robust adaptive sliding mode control using stochastic gradient descent for robot arm manipulator trajectory tracking. *Electronics* **2024**, *13*, 3903. [CrossRef]
41. Utkin, V.I.; Vadim, I. Sliding mode control. *Var. Struct. Syst. Princ. Implement.* **2004**, *66*, 1.
42. Aydın, M. Real-Time Sliding Mode and Moving Sliding Mode Control of 3-DOF Linear Parallel Robot. *Machines* **2025**, *13*, 190. [CrossRef]
43. Pan, C.; Wang, Y.; Yang, S.X.; Li, Z.; Xiao, J. Compensation Function Observer-Based Backstepping Sliding-Mode Control of Uncertain Electro-Hydraulic Servo System. *Machines* **2024**, *12*, 719. [CrossRef]
44. Rabiee, H.; Mohammad, A.; Mohsen, E. Continuous nonsingular terminal sliding mode control based on adaptive sliding mode disturbance observer for uncertain nonlinear systems. *Automatica* **2019**, *109*, 108515. [CrossRef]
45. Sun, H.; Gao, L.; Zhao, Z.; Li, B. Adaptive super-twisting fast nonsingular terminal sliding mode control with ESO for high-pressure electro-pneumatic servo valve. *Control Eng. Pract.* **2023**, *134*, 105483. [CrossRef]
46. Zhu, J.; Khayati, K. Adaptive sliding mode control with smooth switching gain. In Proceedings of the 2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE), Toronto, ON, Canada, 4–7 May 2014 ; pp. 1–6.
47. Silaa, M.Y.; Bencherif, A.; Barambones, O. A novel robust adaptive sliding mode control using stochastic gradient descent for PEMFC power system. *Int. J. Hydrogen Energy* **2023**, *48*, 17277–17292. [CrossRef]
48. Ioannou, P.A.; Kokotovic, P.V. Instability analysis and improvement of robustness of adaptive control. *Automatica* **1984**, *20*, 583–594. [CrossRef]
49. Ioannou, P.A. Adaptive control of linear time-invariant systems. In *Encyclopedia of Systems and Control*; Springer International Publishing: Cham, Switzerland, 2021; pp. 6–11.
50. Nasir, M.; Maiti, A. Adaptive sliding mode resilient control of multi-robot systems with a leader–follower model under byzantine attacks in the context of the industrial Internet of things. *Machines* **2024**, *12*, 205. [CrossRef]
51. Mo, X.; Lan, Q. Finite-Time Integral Sliding Mode Control for Motion Control of Permanent-Magnet Linear Motors. *Math. Probl. Eng.* **2013**, *2013*, 567610. [CrossRef]
52. Jin, M; Lee, J; Chang, P. H; Choi, C. Practical nonsingular terminal sliding-mode control of robot manipulators for high-accuracy tracking control. *IEEE Trans. Ind. Electron.* **2009**, *56*, 3593–3601.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Efficient Transformer-Based Road Scene Segmentation Approach with Attention-Guided Decoding for Memory-Constrained Systems

Bartas Lisauskas † and Rytis Maskeliunas *,†

Software Engineering Department, Faculty of Informatics, Kaunas University of Technology, 44249 Kaunas, Lithuania; bartas.lisauskas@ktu.edu

* Correspondence: rytis.maskeliunas@ktu.lt

† These authors contributed equally to this work.

Abstract: Accurate object detection and an understanding of the surroundings are key requirements when applying computer vision systems in the automotive or robotics industries, namely with autonomous vehicles or self-driving robots. A precise understanding of road users or obstacles is essential to avoid potential accidents. Due to the presence of many objects and the diversity of the environment, the segmentation of the road scene remains a challenging task. In our approach, a Transformer-based backbone is employed for robust feature extraction in the encoder module. In addition, we have developed a custom decoder module in which we implement attention-based fusion mechanisms to effectively combine features. The decoder modification is specifically designed to maintain fine spatial details and enhance the global context understanding, setting our method apart from conventional approaches that typically use simple projection layers or standard query-based decoders. The implemented model consists of 17.2 million parameters and achieves competitive performance, with a mean intersection over union (mIoU) of 76.41% on the Cityscapes validation set. The results gathered indicate the ability of the model to capture both the global context and fine spatial details that are critical to the accurate segmentation of urban scenes. Furthermore, the lightweight design makes the approach suitable for deployment on memory-limited devices.

Keywords: computer vision; deep learning; image processing; neural networks; semantic segmentation

1. Introduction

Neural networks are used to perform image processing tasks in computer vision. One such task is to extract useful information from digital images. These networks are used to perform object detection, classification, and segmentation tasks. From an engineering perspective, the goal of computer vision is to create autonomous systems that can perform tasks that humans do, and, in many cases, to do so faster and more efficiently.

Autonomous driving is a revolutionary technology that is likely to have a very significant impact on people's daily lives in the future. Image segmentation systems provide autonomous cars with a view of the surrounding world and are critically important in achieving safe autonomous vehicle driving. According to a study conducted by the US National Highway Traffic Safety Administration, 94% of all traffic accidents are caused by human error [1]. The realization of autonomous car driving aims to solve this serious problem regarding car accidents. Since autonomous systems are programmed to drive

efficiently and safely, they reduce—and sometimes even eliminate—the need for human driving, thus eliminating the aforementioned human error [2]. The automotive industry is a very promising area for the development of computer vision solutions in autonomous vehicles. The more accurate the image segmentation process and the less the time that it takes, the more accurately autonomous vehicles will understand their surroundings and the more likely they will be to make safer decisions. Providing autonomous vehicles with a view of the world around them using computer vision solutions can offer many benefits, such as increased road safety, lower costs, more comfortable travel, greater mobility, and smaller ecological footprints [3].

With the rapid development of autonomous driving technology, an accurate visual understanding of the surroundings is crucial to ensure road safety and efficient autonomous vehicle navigation. The accurate detection and classification of objects such as pedestrians, vehicles, and traffic signs is essential because it directly impacts the safety and efficiency of autonomous driving systems. Image segmentation is one of the most important processes in digital image processing and has been widely used in the automotive industry and robotics in recent years. The field of computer vision, which is related to artificial intelligence, has seen great progress in the past decade, and today's computer vision systems can recognize visual data more quickly than humans. In the field of computer vision, the semantic segmentation task remains one of the most challenging ones. The assignment of class labels to each pixel, and the classification task at the pixel level, is a key approach that enables vehicles to distinguish between different objects, roads, pathways, pedestrians, and the remaining environmental elements of the road.

As computer vision systems are widely applied in the automotive industry, the same methods can be equally applied to a broader spectrum of autonomous systems in the robotics or aerospace industries. For example, during the NASA Mars 2020 mission, to land the Perseverance rover, a computer vision system was used for hazard and obstacle detection, enabling it to land safely by autonomously selecting the safest landing position [4]. Autonomous robots play an important role in various applications, where accurate perception and effective path planning are key requirements in achieving full autonomy. The perception component is dedicated to understanding the surrounding environment, enabling these robots to make informed decisions [5]. To obtain different fully autonomous vehicles in the future, accurate perception systems are indispensable, ensuring the reliable monitoring and interpretation of complex, dynamic environments [6]. Moreover, the use of semantic segmentation techniques can offer higher precision in detecting urban environments [7]. Semantic data can help to reduce the dependence of a robot on raw sensor input and external signals such as GPS by providing useful environmental information for navigation [8]. In fact, in this work, the demonstrated approach to segmenting road scene images from a driver's perspective can also be used in other applications, such as delivery or taxi robots and various service robots, which need to accurately understand their surroundings when navigating urban environments and avoiding obstacles. Among the most notable applications is parcel delivery, with autonomous delivery robots emerging as key components in the solutions to different delivery challenges [9].

Autonomous robotic platforms operate in complex urban environments and require precise understanding and reliable path planning, while having limited processing resources. Many computer vision models often require extensive computational power, and a Transformer-based system can offer a lightweight alternative, making it suitable for deployment across different autonomous robotic platforms. The ability of the autonomous robot to understand its working environment is the basis for the solution of more complicated problems [10]. The results described in this work show that the proposed approach, which is effective in detecting objects at a close distance, can be particularly advantageous for

use in autonomous self-driving robots. These robots generally operate at lower speeds compared to cars. While a car traveling at 60 km/h covers 100 m in just a few seconds, a robot may take up to 30 s to travel the same distance. Thus, for the robot, distant objects in the environment are not as critical as they are for an autonomous car.

Autonomous systems and their effectiveness are highly dependent on the ability to navigate a complex and unstructured environment [11]. Recent advances have further improved robotic navigation by enabling real-time performance in critical tasks such as environment perception, obstacle detection, obstacle avoidance, path planning, and path tracking [12]. Within path planning, obstacle avoidance is a crucial task in robotics, as autonomous robot operation requires that they reach their destinations without collision [13]. Effective object detection strategies face static obstacles, such as infrastructure or parked vehicles, and dynamic obstacles, including pedestrians and moving vehicles, with each presenting unique challenges to safe navigation. To prevent a potential collision with pedestrians, an accurate detection system enables autonomous robots to intervene early, reducing the risk of accidents [14]. Thus, an accurate computer vision system is essential to ensure the safe and efficient operation of autonomous robots, especially in dynamic urban environments. The main solutions currently focus on understanding the environment through visual information using various computer vision techniques, machine learning, and algorithms [15]. The use of computer vision techniques enables robots to autonomously understand their surroundings, adapt their trajectories, and perform tasks such as maintenance or exploration without human intervention. In addition, for autonomous robots to navigate in urban environments, it is very important to navigate on designated paths, such as footpaths or sidewalks, and avoid areas such as grass to ensure both safety and social conformity. Robots deployed in public environments as autonomous delivery robots operate in spaces in which people live and work [16]. For example, package delivery robots must be able to identify and follow safe and appropriate routes that allow them to navigate autonomously in a manner that is not only efficient but also socially acceptable to the people sharing the environment [17]. Since robots increasingly share space with humans in everyday environments, ensuring safety is paramount [18]. Computer vision applications help to improve the efficiency of transportation systems, increase their levels of intelligence, and improve traffic safety [19]. Moreover, the integration of computer vision with robotics holds significant promise for environmental protection efforts by enabling more efficient resource management and reducing urban environmental impacts [20]. In general, the development of robotic vision is more than just a scientific curiosity or a passing trend. It marks a significant step forward in what machines can do, and this is expected to strongly impact our daily lives [21].

1.1. Evolution of Deep Learning Architectures

In 1989, the French scientist Yann LeCun created one of the first convolutional neural networks, which was called LeNet-5. This neural network was designed for a handwritten digit recognition task. The emergence of the LeNet-5 architecture paved the way for the continued success of convolutional neural networks in performing high-complexity computer vision tasks, and it encouraged researchers to explore the capabilities of convolutional neural networks in performing image segmentation tasks [22]. Among the different deep learning models, convolutional neural networks have achieved excellent performance in different computer vision tasks, such as image classification, object detection, or digital image segmentation. Convolutional neural networks have become one of the most successful and widely used deep learning architectures in computer vision tasks over the past decade.

While, for many years, the convolutional neural network architecture was the state of the art in computer vision tasks, this situation changed in 2017, when the Transformer

architecture was introduced. Recently, Vision Transformers have emerged as a competitive alternative to the long-standing convolutional neural network for computer vision tasks. The Transformer neural network was developed and introduced by the scientist Ashish Vaswani in 2017 [23]. Today, these neural networks compete with state-of-the-art convolutional neural network architectures in terms of efficiency and accuracy.

During research, it was found that it is possible to create accurate computer vision models without using convolutional layers as the main components. One such idea is to use the Vision Transformer neural network architecture for feature extraction; it applies an attention-based mechanism to input images and can achieve competitive efficiency in performing the computer vision semantic segmentation task [24]. The use of Transformer-based architectures over traditional convolutional neural networks as the main component for the feature extraction part is due to the Transformer architecture's improved capacity for global context understanding. The use of self-attention mechanisms in the Transformer architecture allows the system to capture relationships between distant regions of the image. This approach enables the network to integrate information throughout the image. It is especially beneficial for semantic segmentation tasks that require full-scene understanding. Transformers process images as sequences of patches, which gives the opportunity to model both global and local interactions without being constrained by convolutional fixed-size kernels. Thus, the results can be more robust, particularly in situations where the boundaries of the object and contextual cues are crucial [23].

Another important quality of the Transformer architecture is its robustness to variations. Attention mechanisms can dynamically focus on the most relevant parts of the image, making them more resilient to different changes, like rotation, scaling, or occlusion. This adaptability in complex road scenes is beneficial when noise or local variations hinder the performance. In addition, Transformer architectures show improved performance as the data volume increases, making them highly scalable to large datasets [25].

1.2. Modern Approaches to Semantic Segmentation

Semantic segmentation is the core task and remains one of the most challenging in computer vision; it involves the classification of every pixel in a digital image into a corresponding class. It gives the complete context of the scene by incorporating the categories, locations, and shapes of all elements in the scene, including the background [26]. However, it is more challenging and usually more time-consuming than object detection and requires more advanced techniques and more high-quality annotated training data [27]. Over the years, modern approaches have been introduced to perform semantic segmentation tasks—from the early fully convolutional networks that introduced pixel-level predictions to advanced encoder–decoder architectures, which integrate multiscale feature fusion and context-aware processing [23,28]. Recently, the incorporation of attention mechanisms and Transformer-based models have offered new capabilities to capture long-range dependencies and global contextual information, pushing the performance to new heights. The progression of modern techniques is reshaping the state-of-the-art performance in the computer vision field. Table 1 presents a systematic summary of these methods in the field of computer vision image segmentation, including their core architectures, benchmark datasets, and mIoU results.

Fully convolutional networks (FCNs) were among the first breakthroughs in the semantic segmentation task. By replacing the fully connected layers of traditional CNNs with convolutional layers, FCNs enabled end-to-end pixel-wise prediction, which allowed models to generate spatially dense outputs by upsampling the low-resolution feature maps from the convolutional layers. An evaluation of FCNs shows that this method can effectively segment images, establishing a solid baseline for subsequent models [28].

Table 1. Comparison of different architectures.

Method	Architecture	Dataset	Result (mIoU)
FCN	Fully Convolutional Network	PASCAL VOC	67.5
UNet	Encoder–Decoder	Cityscapes	83.6
DeepLabV3+	ASPP + Decoder	Cityscapes	79.6
SETR	Transformer	Cityscapes	76.7
Mask2Former	Hybrid CNN–Transformer	Cityscapes	83.3

Building on the FCN framework, encoder–decoder architectures such as U-Net have become a popular approach to segmentation tasks. U-Net uses a symmetric architecture in which an encoder gradually reduces the spatial dimensions while capturing semantic features, and a decoder progressively upsamples the features to produce a prediction at the pixel level. Skip connections are used between the corresponding encoder and decoder layers to help recover spatial details lost during downsampling. The demonstrated U-Net approach is particularly effective for tasks that require the precise detection of objects, such as the segmentation of the road scene [29].

The DeepLabV3+ model uses a dilated convolution-based approach to better capture multiscale contextual information without decreasing the resolution. Using this type of approach increases the receptive fields of convolutional filters, without additional parameters or a reduction in the spatial resolution. Using spatial pyramid pooling and an encoder–decoder structure, the DeepLabV3+ model can fuse features from multiple scales, which is critical for the complex and varied environments encountered in road scenes [30].

The adaptation of Transformer architectures is another trend. Although originally developed for natural language processing, they were later used in the domain of image segmentation. Models such as SETR or SegFormer use a self-attention mechanism to capture long-range dependencies and global contexts. The images in these models are partitioned into patches and then processed as token sequences, enabling the network to model relationships between distant regions. The global modeling capability is especially advantageous for complex scenes, such as urban road environments, where contextual signals are critical [31,32].

The integration of convolutional neural networks with Transformer modules is a promising direction, combining the strengths of both architectures. Hybrid computer vision models often use convolutional layers for efficient local feature extraction and are combined with additional Transformer layers to capture global context details through self-attention mechanisms. The use of this approach can lead to improved segmentation performance, especially in scenarios where both detailed spatial information and a wider contextual understanding are necessary. One work incorporating the convolution technique into Vision Transformers shows that the combination of these two methodologies can yield competitive performance in segmentation tasks [33].

The main purpose of this study is to develop a Transformer-based semantic segmentation approach, specifically designed for the road scene segmentation task, which takes advantage of the latest advances in Vision Transformers. This paper is organized as follows. Section 2 describes our proposed semantic segmentation approach. Section 3 presents the experimental results, and Sections 4 and 5 conclude the document with a discussion and future directions.

2. Materials and Methods

In this section, we provide the details of the implementation, the structure of the model, and the configuration settings used to train and evaluate our Transformer-based road scene segmentation system. Moreover, we describe the components of other state-of-the-art computer vision models, the components of our approach, and how our approach is different, as well as presenting some advantages over the previously used implementations. In addition, we describe the dataset in more detail and explain how the data are divided into training, validation, and test set splits. In further sections, we also explain the internal structure of the encoder module, as well as the decoder part of the system with attention-based fusion mechanisms for the extraction and combination of multiscale features. Details of each component's structure are provided in further subsections.

The current state-of-the-art computer vision approaches for road scene segmentation have demonstrated significant success by using different types of architectures and achieving competitive results. Methods that use Transformer-based architectures typically rely on standard feature fusion, and they often use simple projection layers to merge different features from different Transformer-based encoder resolutions. Computer vision models have also been introduced, such as RoadFormer, which uses a query-based decoder module to iteratively update query features for mask predictions [34]. Although these approaches work well in many cases, they can struggle to maintain important fine-grained details when working with complex road scene environments.

Our decoder module processes the data with attention-based fusion mechanisms, providing competitive results in the road scene segmentation task in complex urban scenes compared to other lightweight computer vision models that have a similar number of parameters in the network. The mechanism in the decoder part uses specialized attention blocks to progressively fuse low-resolution decoder features with high-resolution features from the encoder part. This can effectively emphasize the most relevant spatial information at multiple scales. Using this approach, it is possible to ensure that the fine-grained features of the encoder can be integrated together with the high-level semantic features of the decoder, giving accurate and competitive segmentation results when compared with other introduced approaches. Moreover, instead of using traditional normalization techniques, the model uses group normalization, which normalizes over the feature channels instead of on the batch dimension. This can give better stability in the training process, particularly in scenarios where the batch size may be different or in cases of high-resolution feature fusion, where traditional methods such as BatchNorm may not be as effective.

It is also worth mentioning that, while many existing approaches rely on standard feature fusion techniques with different scales, this approach can provide better control over which features are fused via the use of attention blocks, which allow it to adapt more effectively to the different complexities of road scene segmentation. Using this approach gives competitive results in terms of segmentation accuracy, particularly in challenging road scene environments with fine details.

2.1. Dataset

All experiments were performed using the widely known Cityscapes dataset, which is used for the benchmarking of computer vision models in semantic segmentation tasks with different urban scene images. The dataset consists of 5000 high-quality, finely annotated, pixel-level, high-resolution images recorded in 50 different cities across Germany. Each pixel in every image in the dataset is annotated to one of 19 semantic categories, representing typical elements of the urban road scene, such as roads, cars, pedestrians, sidewalks, or traffic signs. These 5000 images are split as follows.

- Training set, consisting of 2975 images: This subset of the data is used to learn the parameters of the model during the training process.
- Validation set, consisting of 500 images: This part of the data is used during the training process to configure the model, adjust the hyperparameters, and monitor the performance of the model during the training phase.
- Test set, consisting of 1525 images: The last part of the data is used for the final evaluation phase to evaluate the precision of the model in an unseen dataset.

In Figure 1, we provide some examples from the Cityscapes dataset, demonstrating different environments and urban conditions. All details of the dataset used in the model training process and the final evaluation, including the annotated classes and examples of annotations, are publicly available at https://www.cityscapes-dataset.com/ (accessed on 11 March 2025).



Figure 1. Sample images from the Cityscapes dataset, illustrating the diversity of urban scenes in different cities across Germany.

2.2. Implementation Details

Our Vision Transformer-based road scene segmentation model is implemented using the `mmsegmentation` framework codebase and uses an MiT as a backbone network, with attention-based fusion mechanisms in the decoder module. The system's encoder module is pre-trained on the ImageNet-1k dataset for the extraction of robust visual features, while the system's decoder module is randomly initialized to learn task-specific upsampling. To improve the model's robustness and generalization, in the training process, data augmentation techniques were applied using the Cityscapes dataset. We used additional measures of random horizontal flipping, scaling, and cropping. The crop size of 768 $\times$ 768 pixels was chosen during the training phase. At inference time, a sliding-window strategy was used to generate full-size segmentation predictions.

The computer vision model was trained using the AdamW optimization algorithm with an initial learning rate of 0.00005. In addition, the polynomial decay learning rate schedule with the power parameter set to 1.0 was used, and the linear warm-up phase consisted of up to 1500 iterations at the start of the training process. Due to GPU resource constraints and the high-resolution input images, a batch size of 1 image was used in the training process. The training schedule was set for 160,000 iterations, and it was later extended to 200,000 iterations to explore the possibility of further model accuracy improvements. In the end, it was clear that no significant improvements would be achieved in the training cycle after 160,000 iterations. The best-performing model checkpoint was

taken at 156,000 iterations. Model performance was primarily evaluated using the widely used mean intersection over union (mIoU) metric.

2.3. Encoder

In the system, the encoder backbone network is implemented using the MiT architecture, with the configuration settings named "mit b1" [32]. Using these configuration settings, the input image is processed through a patch embedding module that divides the image into smaller patches and projects them into a feature space. The first stage of the network uses an OverlapPatchEmbed module with a fixed patch size of 7×7 and a stride parameter of 4. The module is responsible for mapping each 7×7 image patch to a 64-dimensional embedding space and generating the initial feature map. The patch embedding can be written as

$$X_0 = \text{Conv}_{7\times7}(I)$$

where I is the input image. The output after this processing stage is then flattened and normalized. The subsequent stages of the encoder network use patch embedding modules with a set patch size of 3×3 and a stride parameter of 2. These network encoder stages progressively reduce the spatial resolution while increasing the number of channels. Using the "mit b1" configuration, the embedding dimensions are preset to [64, 128, 320, 512] for the four network stages. The encoder progressively extracts hierarchical features with increasing channel numbers:

- Stage 1 (c1): 64 channels;
- Stage 2 (c2): 128 channels;
- Stage 3 (c3): 320 channels;
- Stage 4 (c4): 512 channels.

Each of the stages is further processed by a sequence of Transformer blocks. Using the b1 network configuration, the depth of each block is [2, 2, 2, 2] for the four stages. In addition, with each Transformer block, the multi-head self-attention mechanism is used. The attention computation can be written as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where Q is a query, K is the key, and V is the value received by linear projections of the input features. At each stage, the number of attention heads used is [1, 2, 5, 8], and the MLP within each block expands the dimensions of the feature by a factor of 4.

In each stage, spatial reduction ratios are also used with values [8, 4, 2, 1] to downsample the spatial dimensions during the self-attention processing step. In this way, we reduce the computational load while maintaining an essential global context.

After the processing step is performed through different Transformer blocks, each stage uses a final normalization layer, and the resulting feature maps are reshaped in the [B, C, H, W] format. The final outputs of all four stages are named c1, c2, c3, and c4. These outputs capture a rich hierarchy of features that combine fine spatial details from the early stages of the network with high-level semantic information from the deeper stages of the network. In a further step, all generated feature maps from each encoder network stage are passed to the decoder module for further processing.

2.4. Decoder

The decoder module approach combines the features from the encoder using an attention-based mechanism. In this subsection, we propose an attention-guided decoding

module that fuses multiscale Transformer features via hierarchical attention and residual convolutional fusion. The decoder design allows the rich semantic information from deeper layers to be combined with the fine spatial details of shallower layers; in this way, it can achieve good accuracy and spatial precision. Figure 2 shows the complete model architecture, including the encoder and the attention-guided decoding module.

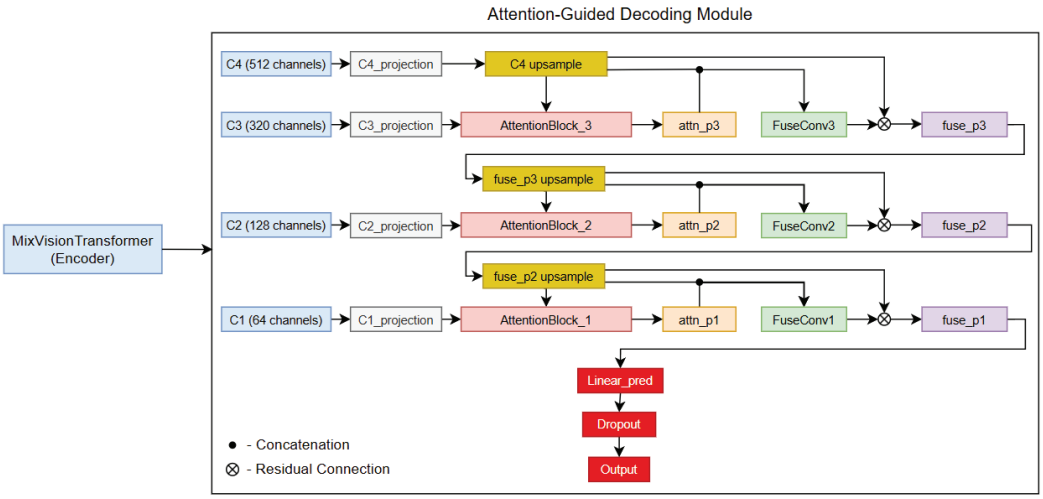


Figure 2. Architecture diagram of the computer vision model with the proposed attention-guided decoding module.

Using the MiT as a backbone feature extraction network, the encoder part produces a set of feature maps at different resolutions, typically named c1, c2, c3, and c4, from the shallowest to the deepest network stage. Each feature map captures information at a different scale. Although deeper features contain more semantic information, shallow features preserve detailed spatial information, which is critical for accurate segmentation in complex road scenes.

The decoder first projects each of the feature maps to a common embedding space with a fixed number of channels. This process is achieved by using a 1×1 pixel convolutional layer, which is later followed by the normalization of the feature maps across channels to help improve the consistency of the feature representations. After this, the ReLU activation function is used to introduce non-linearity, allowing the network to learn more complex patterns. The mentioned process not only standardizes the channel dimensions across different scales but also improves the stability of the whole training process. Mathematically, this projection can be described as follows:

$$p_i = \mathrm{ReLU}\Big(\mathrm{GroupNorm}\big(\mathrm{Conv}_{1\times1}(c_i)\big)\Big), \quad i \in \{1,2,3,4\}.$$

In addition, in the decoder part, we use an attention-based fusion mechanism. When using separate 1×1 convolutional and normalization layers for upsampled decoder features and corresponding encoder features, we generate an attention map that highlights the most relevant spatial regions. This selective weighting approach helps the decoder to focus on important details. As a result, the integration of rich semantic features with fine spatial details helps to achieve a more effective and accurate segmentation process in different road scenes. The attention block in the decoder part is the main component that serves to fuse features from different scales. In this decoder implementation, the fusion process is performed in a top-down approach. In the first stage, the deepest feature map p4 is first upsampled to match the spatial resolution of p3. After this, the attention block receives the upsampled p4 feature map (acting as a gating signal) and the p3 feature map from the encoder backbone. Later, the attention block projects both received inputs using a

1×1 convolutional layer followed by a normalization layer. Specifically, the projections can be formulated as follows:

$$g_1 = \text{GroupNorm}\Big(\text{Conv}_{1\times1}(g)\Big), \quad x_1 = \text{GroupNorm}\Big(\text{Conv}_{1\times1}(x)\Big),$$

where g is the gating signal (upsampled p4) and $\times$ is the encoder feature (p3). The resulting outputs are then summed, activated by the ReLU function, and further processed by another 1×1 convolutional layer with the sigmoid activation function to produce an attention map:

$$\psi = \sigma\Big(\text{GroupNorm}\big(\text{Conv}_{1\times1}\big(\text{ReLU}(g_1 + x_1)\big)\big)\Big).$$

The produced map selectively weights the p3 features, suppressing less relevant regions. The weighted encoder feature is concatenated with the upsampled p4, and a convolutional module refines this fusion. Additionally, a residual connection is added by introducing the upsampled p4.

In addition, the fused features from the previous stage are upsampled in the same way to match the resolution of p2 and fused with p2 using an analogous attention-guided procedure. The same process is repeated once again with the highest-resolution p1 feature map to produce the final fused representation.

In this decoder approach, there are combined residual connections at each fusion stage. These connections add the upsampled features back into the fused output, and, in this way, we ensure that essential spatial details are kept throughout the upsampling process. By using residual connections at each stage, we stabilize the training process by maintaining a smoother gradient flow. In addition, this allows the model to take advantage of the refined fused features and the original upsampled signals, thereby improving the overall robustness of the segmentation process.

After the progressive fusion step, a dropout layer is applied to the refined feature map for regularization. After this step, the final 1×1 convolutional layer is used to project the features to the preset number of segmentation classes, generating the final segmentation output. The final prediction can be written as

$$\text{Output} = \text{Conv}_{1\times1}\Big(\text{Dropout}(\text{Fused Feature})\Big)$$

This decoder approach can be effective for road scene segmentation tasks due to its ability to preserve spatial precision by maintaining high-resolution feature maps and integrating fine-grained details using the attention mechanism, which ensures the accurate distinction of road boundaries and different objects. The attention mechanism used in the decoder module works as a gating function that merges context-rich deep features with detailed shallow features. In doing so, it emphasizes the spatial regions that are most relevant. In addition, the use of residual connections throughout the fusion process helps with training stability and convergence, even when the model incorporates multiple upsampling stages.

Overall, this decoder approach proves robust and efficient in fusing multiscale features, and, when combined with the MiT backbone for feature extraction, it can achieve competitive accuracy in the road scene semantic segmentation task.

3. Results

In this section, we present the experimental results for our Vision Transformer-based road scene segmentation system. First, the training process is described, followed by additional details regarding the performance metrics per class and the global metrics of the

system. At the end of this section, we provide the visualization results for the road scene environment to demonstrate that the system is capable of detecting different objects on the road.

3.1. Training Process

As described in more detail in Section 2.2, the model training schedule was preset to 160,000 iterations. Moreover, to examine whether additional training could lead to better accuracy metrics, the training process later was extended to 200,000 iterations. An analysis of the training log from the `mmsegmentation` framework yielded the following observations.

- Initial Training Phase: The training phase from the beginning showed a high decoder loss parameter of 2.43. During the full 160,000 iterations of training, the model loss parameter decreased as we approached 160,000 iterations. This behavior demonstrates that the computer vision model was continuously learning and improving until the end of the training cycle, while it achieved a loss parameter of approximately 0.11 in the final iterations.
- Iteration 156,000 as the Final Checkpoint: At this point in the training cycle, the model demonstrated the highest precision metrics. From the training log review, the decoder loss parameter was recorded at 0.1161 and we achieved the highest accuracy values. This confirms that the highest performance was reached at this point, leading us to select this checkpoint as our final model.
- Extended Training Experiment: To determine whether the computer vision model could achieve better accuracy, an extended training process of up to 200,000 iterations was carried out. After this, the results of the training log showed that additional training did not give accuracy improvements, indicating that further training beyond the 160,000 point is not effective.

Figure 3 illustrates the training convergence curves of our proposed model compared to traditional segmentation architectures, specifically U-Net+FCN and DeepLabV3+. Our model demonstrates a consistent decrease in loss, reaching approximately 0.11 at 160,000 iterations. In contrast, U-Net, trained up to 160,000 iterations, and DeepLabV3+, trained up to 80,000 iterations, exhibit higher loss values at corresponding iteration points, indicating slower convergence and less efficient learning.

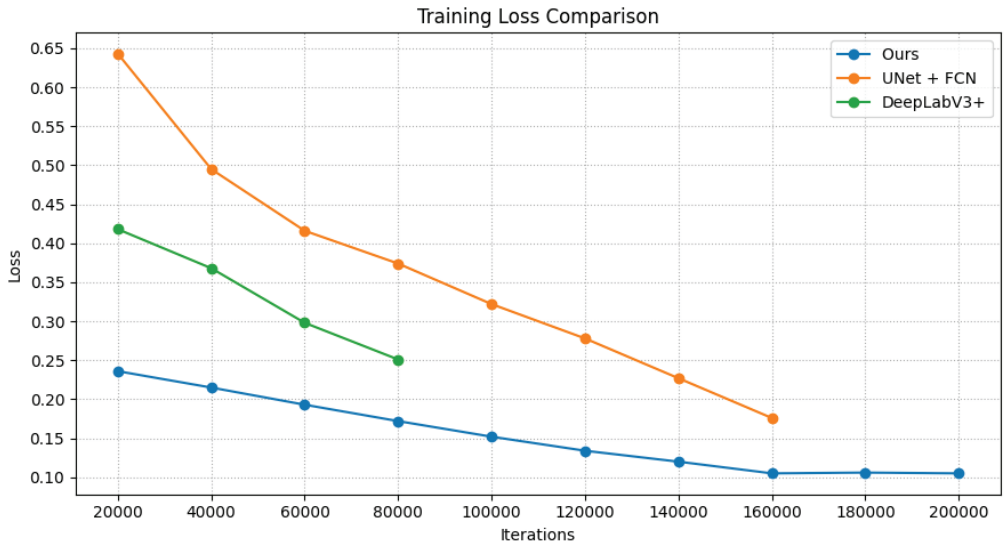


Figure 3. Training loss curves of our model versus U-Net and DeepLabV3+, showing faster convergence and a lower final loss.

The results show the efficiency of our proposed training approach. Although we extended the training to 200,000 iterations, we saw no further improvement beyond 160,000 iterations, so further training would not have been cost-effective.

3.2. Per-Class Performance

Our computer vision model segments urban road scenes into 19 different classes according to the standard of the Cityscapes dataset. It uses the cross-entropy loss function for optimization. For better understanding, we divide the classes into two different groups based on the segmentation accuracy that they achieved, measured by the main intersection over the union value. The first group of classes consists of those with the highest mIoU, which is above $\geq$76%, while the second group of classes consists of those who reached the mean intersection over union far below the 76% line. It is worth mentioning that, despite being trained exclusively on the Cityscapes dataset, without using any additional datasets, the computer vision model is capable of achieving good accuracy in the semantic segmentation task while having less than 20 million parameters and is capable of detecting different classes, such as roads, cars, sidewalks, vegetation, or buildings.

In Table 2, we show the first group of classes that yielded higher accuracy than 76%, ordered in descending order by the IoU metric for every class.

Table 2. Classes with higher accuracy than 76% in descending order.

Class	IoU (%)	Accuracy (%)
Road	97.93	98.83
Sky	94.54	98.15
Car	94.46	98.01
Vegetation	92.43	96.84
Building	92.16	96.43
Sidewalk	83.39	91.63
Person	80.90	91.60
Bus	80.04	85.36
Truck	79.17	84.65
Traffic Sign	78.83	85.16
Bicycle	76.71	87.72

The table above demonstrates that classes such as roads, cars, sky, vegetation, and buildings yield accuracy above 90%. Others, like buses, people, sidewalks, trucks, traffic signs, and bicycles, yield lower accuracy. Nonetheless, these are satisfactory results given the lightweight model approach, the quantity of data used, and the difficulty in distinguishing classes like traffic signs in different distance conditions. In the following, Table 3 lists the classes with an mIoU below 76%, indicating that they are more challenging to accurately segment.

Table 3. Classes with lower accuracy than 76%, which are more difficult to detect.

Class	IoU (%)	Accuracy (%)
Traffic Light	70.19	81.64
Motorcycle	65.86	76.66
Train	64.97	69.36
Terrain	63.02	69.89
Pole	62.83	72.81
Fence	59.84	70.10
Rider	59.45	71.76
Wall	55.14	62.86

Table 3 indicates that classes such as motorcycles, trains, traffic lights, poles, terrain, fences, walls, and riders yield lower precision. These challenges may be due to the inherent complexity and variability of these objects in different road scenes. In the future, by using more data with greater diversity, it will be possible to obtain better accuracy results for these classes. Here, it was challenging to detect these classes accurately, because distinction is needed between concrete walls, buildings, or fences, and many state-of-the-art models lack accuracy for these classes. Overall, the computer vision model, with 17.2 million parameters, detects different objects and environmental elements, such as cars, roads, and sidewalks, with good accuracy, while its lower performance in certain classes highlights opportunities for future improvements.

3.3. Global Metrics

To evaluate the final performance of our computer vision model with the Cityscapes validation set, we collected different accuracy metrics. The following accuracy metrics were obtained after the final evaluation phase:

- Mean IoU (mIoU): 76.41%;
- Mean Accuracy (mAcc): 83.66%;
- Overall Accuracy (aAcc): 95.87%.

The above metrics demonstrate that the computer vision model, having 17.2 million parameters, is capable of reaching a global mean intersection over union of 76.41%. This metric reflects the average overlap between the predicted segments and the ground truth values across all classes, and it is widely used in evaluating the performance of computer vision models. The data provided above also show that the model achieved a mean accuracy (mAcc) value of 83.66%. This metric represents the average classification accuracy per pixel for each class. Lastly, the overall accuracy (aAcc) of 95.87% is reached. This metric shows the ratio of correctly classified pixels to the total number of pixels in the Cityscapes validation set.

3.4. Qualitative Analysis

In Figure 4, we show an example image of a road scene, where the original image is presented on the left and the segmentation output produced by the system is shown on the right. From this side-by-side comparison, it is clearly shown that the model can perform semantic segmentation on road scene images, detecting different objects in the digital image with good accuracy. From this comparison, the following observations emerge.

- Accurate detection of close objects: The model is capable of detecting near-field classes with good accuracy, including roads, cars, pathways, and buses, when these objects are at an approximately 50–100 m distance.
- Challenges with distant objects: It is more challenging to detect distant objects like traffic signs, indicating potential areas for further improvement.
- Overall performance: Under ideal weather conditions, the segmentation quality is good, and the model clearly defines boundaries for different objects, although minor inaccuracies can appear in more complex or distant regions.

The segmentation results provided above demonstrate that the computer vision model developed can effectively detect and segment close objects, ensuring the clear and detailed recognition of the main elements on the road. It should be mentioned that lower accuracy metrics were reached for more distant objects such as traffic signs, poles, or traffic lights. This is a possible area for improvement in the future, where the main focus could be on improving the ability of the computer vision model to detect fine details in more distant regions. This could be achieved by modifying the model architecture or by using additional

datasets. In this way, it would be possible to further enhance the performance of the system in different real-world scenes on the road.



Figure 4. Road scene example with the original image on the left and the segmented one on the right.

3.5. Model Robustness Analysis

To evaluate the robustness of our segmentation model, we also tested it on images of a different dataset (Mapillary Vistas) that were not used during training. The examples provided below include images captured in an urban city area, which highlight the model's performance in a different environment, as well as another image taken on a highway to demonstrate the model's ability to accurately segment unseen data across different scenarios.

In Figure 5, we show the original image on the left and the corresponding segmented result on the right. This visual presentation demonstrates the ability of the model to detect major objects and the environment with good overall accuracy. It also indicates that there is room for improvement in the segmentation of small objects, such as traffic signs or traffic lights, and that the accuracy at larger distances could be improved. For example, in the highway image, where the distance is more than 100 meters, the model lacks the accuracy to differentiate cars from the road, but, with closer objects and environmental elements, the accuracy is good.



Figure 5. Segmentation results with unseen images from another dataset.

As demonstrated in Table 4, the highest performance metrics are obtained for the main environment and object classes, such as roads, cars, buildings, and vegetation, which indicate good segmentation performance for these dominant regions. Meanwhile, classes

such as people or traffic signs result in lower performance, which can be attributed to their smaller sizes and greater variability in their appearance.

Table 4. Class performance metrics.

Class	IoU (%)	Accuracy (%)
Road	96.72	98.29
Sky	93.05	96.60
Car	90.58	96.42
Vegetation	89.44	96.43
Building	89.18	94.64
Sidewalk	75.57	84.58
Person	67.59	83.94
Traffic Sign	62.63	69.76

3.6. Computational Performance Analysis

The computer vision model, with 17.2 million parameters, was evaluated on an NVIDIA GTX 1060 GPU, chosen specifically to represent the low-end sector of Cuda-capable devices (released in 2016). The test was performed using a single image with a batch size of 1, with different input resolutions. The computational performance metrics, such as the computational cost (GFLOPs), inference time (ms), and throughput (FPS) at different resolutions, are provided in Table 5, illustrating the model's resource requirements under the different conditions. As the resolution increases, both the computational cost (GFLOPs) and inference time increase, while the throughput (FPS) decreases. Specifically, at a 256 × 256 resolution, the model requires 9.1 GFLOPs and achieves 42.4 FPS with an inference time of 24 ms. With the highest resolution of 2048 × 1024, the computational cost reaches 419.4 GFLOPs, and the inference time increases to 769 ms. The data provided in the table demonstrate the trade-offs between the resolution, computation, and performance, which must be carefully considered when a different resolution is needed for resource-constrained applications.

Table 5. Computational performance at different input resolutions.

Input Resolution	GFLOPs	Inference Time (ms)	Throughput (FPS)
256 × 256	9.1	24	42.4
512 × 512	37.9	37	26.9
1024 × 1024	176.4	120	8.3
2048 × 1024	419.4	769	1.3

In Table 6, we present the computational performance of our segmentation model on the NVIDIA Jetson Orin Nano 8 GB embedded platform. The performance metrics are reported for three different input resolutions, capturing the inference speed (FPS) and GPU memory usage.

Table 6. Model performance on Jetson Orin Nano 8 GB.

Input Size	Inference Speed (FPS)	GPU Memory Usage (MiB)
256 × 256	52.89	54
320 × 320	32.23	72
512 × 512	15.56	112

The results provided in Table 6 demonstrate that, at the lowest resolution, the computer vision model can achieve an inference speed of 52.89 FPS while consuming 54 MiB of GPU memory. When using a higher resolution, the inference speed decreases to 32.23 FPS at an input size of 320 × 320. At a 512 × 512 resolution, the GPU memory consumption increases to 112 MiB, and 15.56 FPS is still acceptable for many autonomous applications. However, for applications that require higher input resolutions or faster inference speeds at larger scales, a more powerful embedded device with greater computational resources may be more appropriate than the Jetson Nano embedded platform.

4. Discussion

The experimental results obtained with the computer vision road scene segmentation system demonstrate good accuracy and effectiveness when the model uses a Transformer-based MiT backbone as an encoder for feature extraction and attention-based mechanisms in the decoder module. During the training log analysis, the results revealed that, during the whole training cycle, the mode's loss parameter steadily decreased; shortly before reaching the final iterations, it demonstrated the best accuracy performance at 156,000 iterations. Furthermore, the training process confirmed that setting the training schedule at 160,000 iterations was cost-efficient and effective, because an additional experiment extending the training process was not effective. In addition, the per-class performance metrics show that the main elements of the road scene, such as the road, cars, and buildings, are segmented with good accuracy, indicating the model's strong ability to detect and distinguish different objects and environmental elements on the road.

Creating a model that maintains a lightweight design with a low parameter count while achieving competitive accuracy is a challenging task. It is always necessary to have a balance between accuracy and lightweight model design, because, while a more complex model can achieve better accuracy metrics, it also has a significantly larger number of parameters and is consequently not as lightweight. This balance is critical, as increasing model complexity often improves the performance, but it can compromise the efficiency and suitability of the model for deployment in memory-limited environments.

According to the global evaluation metrics of the developed model, the results validate the good performance of the system and its capacity to detect and differentiate objects in digital road scene images, achieving a mean IoU of 76.41% with 17.2 million parameters in the network. Despite its compact size, the computer vision system is capable of maintaining competitive segmentation performance, making it suitable for use in memory-limited resource environments. The visual examples provided in Figure 6 show that the model is successful in segmenting near-field objects within 50–100 m, and classes like roads, cars, and pathways are detected with good accuracy. However, it is worth mentioning that more distant objects, such as poles, traffic lights, and traffic signs, as well as fences, yield lower precision. Thus, these classes could represent an avenue for future system enhancement by incorporating additional training data or modifying the internal system architecture to capture small-scale objects or environmental elements in distant regions. In Figure 6, we provide qualitative results, with the original RGB image on the left, the ground truth masks in the middle, and the segmentation results on the right.

Regarding other open-source computer vision approaches, models such as DSNet [35], HRNetV2 + OCR [36], DeepLabv3 + [36], CSFNet-1 [37], and EEEA-Net-C2 [38] have demonstrated good accuracy results on the Cityscapes dataset while maintaining a relatively low parameter count. The demonstrated approach has 17.2 million parameters and achieves a 76.41% mIoU. This approach also demonstrates that, by using a Transformer-based encoder for feature extraction and an attention-based mechanism to fuse multiscale features, it is possible to achieve competitive accuracy with other models while maintaining

a lightweight design and good overall performance. The developed approach also extends the knowledge provided by existing studies, showing that Transformer-based architectures can effectively segment complex urban scenes with models that have a low parameter count and are capable of achieving good accuracy in road scene semantic segmentation tasks. Table 7 shows details of several semantic segmentation models evaluated on the Cityscapes validation set, including their input resolutions, parameter counts, GFLOPs, and mIoU values, as reported in the literature.



Figure 6. Qualitative segmentation results. The left column shows the original RGB image, the middle column shows the ground truth masks, and the right column shows the model's segmentation results.

Table 7. Comparison of semantic segmentation models.

Model	Input Size	Params (M)	GFLOPs	mIoU (%)
EEEA-Net-C2	512 × 512	7.34	28.7	76.8
CSFNet-1	1024 × 512	12.6	86.9	74.8
Ours	512 × 512	17.2	37.9	76.4
DSNet	2048 × 1024	37.5	226.6	82.0
DeepLabv3+	2048 × 1024	43.5	1444.6	79.6
HRNetV 2 + OCR	2048 × 1024	70.3	1206.3	81.6

The computer vision models in Table 7 are sorted by the number of parameters in the network. From the data provided in the table, we can see that our approach offers a balance between segmentation accuracy, model size, and computational efficiency. The proposed model has 17.2 million parameters and requires 37.9 GFLOPs at an input resolution of 512 × 512. Models like EEEA-Net-C2 and CSFNet-1 represent lightweight architectures, achieving mIoU accuracy metrics of 76.8% and 74.8%, respectively. In comparison, CSFNet-1 has a slightly lower parameter count but a larger computational cost. DSNet gives good accuracy results and computational efficiency, but it has more than twice as many parameters in the network. Moreover, models like DeepLabV3+ and HRNetV 2 + OCR offer better accuracy but at the cost of having much larger architectures. This comparison highlights that the proposed approach offers a compelling trade-off by delivering good segmentation performance in a lightweight architecture.

In Table 8, we compare the intersection over union (IoU) per class of our lightweight segmentation model against those of two much larger architectures: HRNet-W48 (65.9 M)

and DeepLabV3-R101 (84.7 M). Despite having roughly one-quarter to one-fifth of the number of parameters, our approach achieves only marginally lower IoU scores across many semantic categories. The side-by-side comparison makes it clear that, although larger architectures can reach slightly higher accuracy scores, our Transformer-based design with an attention-guided decoding module offers a compelling trade-off, maintaining competitive performance while drastically reducing the model complexity.

Table 8. Per-class segmentation accuracy (IoU) compared across different network architectures.

Class	DeepLabV3-R101 (84.7 M)	HRNet-W48 (65.9 M)	Ours (17.2 M)
Road	98.3	98.4	97.9
Car	95.4	95.6	94.5
Sky	94.2	95.2	94.5
Building	92.5	93.4	92.2
Vegetation	92.0	93.0	92.4
Sidewalk	85.7	86.4	83.4
Person	81.3	83.4	80.9
Traffic Sign	78.7	81.7	78.8
Bicycle	77.8	79.0	76.7
Traffic Light	69.8	72.4	70.2
Motorcycle	69.2	67.1	65.9
Terrain	63.3	66.7	63.0
Rider	63.9	62.3	59.5
Pole	58.9	69.3	62.8
Fence	62.7	66.4	59.8
Wall	53.5	59.7	55.1

In addition to the overall quantitative comparison, Table 8 shows that our Transformer-based encoder with an attention-guided decoding module matches the larger HRNet-W48 and DeepLabV3-R101 architectures almost exactly in large and homogeneous classes such as roads, cars, buildings, and vegetation, with IoU differences of less than 1.2 percentage points. The results show that the global context modeling of the MiT backbone effectively captures broad, texture-rich regions. However, all three architectures, including the larger baselines, have the greatest difficulty in accurately segmenting slender or distant objects, such as poles, fences, and walls.

In summary, the demonstrated approach confirms that a computer vision system based on the Transformer neural network architecture with an attention-guided decoding module can effectively capture contextual information for the road scene segmentation task. While the model performs well with near-field objects, future work could direct more attention to areas with more distant, small-scale objects, because many models lack accuracy in this regard, and it remains a challenging task.

5. Conclusions

In this paper, we introduce an approach that uses a Transformer-based neural network architecture to segment road scene images, using the MiT backbone as a feature extractor and an attention-guided decoder to effectively fuse multiscale features. Experimental results obtained with the Cityscapes dataset revealed that, despite the small parameter count, it can achieve a competitive mean IoU of 76.41% in different semantic segmentation tasks in urban scenes. The conducted training process confirmed that the system is able to converge reliably in 160,000 iterations, and an additional training process with the same configuration and dataset did not produce better accuracy results.

The model evaluation phase and visual data highlighted that the main objects in close proximity can be detected with good accuracy, while the classes with lower accuracy offer an avenue for future improvements. In general, the primary design goal of this work was to develop a lightweight computer vision road scene segmentation model. Our Transformer-based segmentation model has only 17.2 million parameters, which is remarkably low compared to many computer vision semantic segmentation models. The demonstrated model also has a small memory footprint, making it well suited for deployment on memory-limited devices.

Author Contributions: Conceptualization, R.M.; Formal analysis, B.L. and R.M.; Funding acquisition, R.M.; Investigation, B.L. and R.M.; Methodology, B.L.; Project administration, R.M.; Resources, B.L.; Software, B.L.; Supervision, R.M.; Validation, B.L.; Visualization, B.L.; Writing—original draft, B.L.; Writing—review and editing, R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Data Availability Statement: All data are freely available via the indicated datasets.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
FLOPS	Floating Point Operations per Second
FPS	Frames per Second
GPS	Global Positioning System
IoU	Intersection over Union
mAcc	Mean Accuracy
mIoU	Mean Intersection over Union
MLP	Multilayer Perceptron
MiT	Mix Transformer
NASA	National Aeronautics and Space Administration

References

1. Rafique, Z. Improving Efficiency of Computer Vision for Autonomous Vehicles. Master's Thesis, Bournemouth University, Poole, UK, 2020. [CrossRef]
2. Fernandes, S.; Duseja, D.; Muthalagu, R. Application of Image Processing Techniques for Autonomous Cars. *Proc. Eng. Technol. Innov.* **2020**, *17*, 6074. [CrossRef]
3. Sellat, Q.; Bisoy, S.; Priyadarshini, R.; Vidyarthi, A.; Kautish, S.; Barik, R.K. Intelligent Semantic Segmentation for Self-Driving Vehicles Using Deep Learning. *Comput. Intell. Neurosci.* **2022**, *2022*, 6390260. [CrossRef]
4. NASA. Mars 2020 Perseverance Rover. 2020. Available online: https://science.nasa.gov/mission/mars-2020-perseverance/ (accessed on 19 March 2025).
5. Chen, W.; Chi, W.; Ji, S.; Ye, H.; Liu, J.; Jia, Y.; Yu, J.; Cheng, J. A Survey of Autonomous Robots and Multi-Robot Navigation: Perception, Planning and Collaboration. *Biomim. Intell. Robot.* **2025**, *5*, 100203. [CrossRef]
6. Iparraguirre-Gil, O. Computer Vision and Deep Learning based Road Monitoring towards a Connected, Cooperative and Automated Mobility. Ph.D. Thesis, Universidad de Navarra, Pamplona, Spain, 2022. [CrossRef]
7. Marasinghe, R.; Yigitcanlar, T.; Mayere, S.; Washington, T.; Limb, M. Computer Vision Applications for Urban Planning: A Systematic Review of Opportunities and Constraints. *Sustain. Cities Soc.* **2024**, *100*, 105047. [CrossRef]
8. Landsiedel, C.W. Semantic Mapping for Autonomous Robots in Urban Environments. Ph.D. Thesis, Technische Universität München, Munich, Germany, 2018.
9. Boysen, N.; Fedtke, S.; Schwerdfeger, S. Last-mile Delivery Concepts: A Survey from an Operational Research Perspective. *Or Spectr.* **2020**, *43*, 1–58. [CrossRef]

10. Ni, J.; Chen, Y.; Tang, G.; Shi, J.; Cao, W.; Shi, P. Deep Learning-Based Scene Understanding for Autonomous Robots: A Survey. *Intell. Robot.* **2023**, *3*, 374–401. [CrossRef]
11. Ogunsina, M.; Efunniyi, C.P.; Osundare, O.S.; Folorunsho, S.O.; Akwawa, L.A. Reinforcement Learning in Autonomous Navigation: Overcoming Challenges in Dynamic and Unstructured Environments. *Eng. Sci. Technol. J.* **2024**, *5*, 2724–2736. [CrossRef]
12. Adiuku, N.; Avdelidis, N.P.; Tang, G.; Plastropoulos, A. Advancements in Learning-Based Navigation Systems for Robotic Applications in MRO Hangar: Review. *Sensors* **2025**, *24*, 1377. [CrossRef]
13. Katona, K.; Neamah, H.A.; Korondi, P. Obstacle Avoidance and Path Planning Methods for Autonomous Navigation of Mobile Robot. *Sensors* **2024**, *24*, 3573. [CrossRef]
14. Janai, J.; Güney, F.; Behl, A.; Geiger, A. Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art. *Found. Trends® Comput. Graph. Vis.* **2021**, *12*, 1–308. [CrossRef]
15. Conde, M.V. An Embarrassingly Pragmatic Introduction to Vision-based Autonomous Robots: Applications, Datasets and State of the Art. *arXiv* **2021**, arXiv:2112.05534.
16. Pelikan, H.R.M.; Reeves, S.; Cantarutti, M. Encountering Autonomous Robots on Public Streets. In Proceedings of the HRI '24: Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, Boulder, CO, USA, 11–15 March 2024.
17. Buckeridge, S.; Carreno-Medrano, P.; Cosgun, A.; Croft, E.; Chan, W.P. Autonomous Social Robot Navigation in Unknown Urban Environments Using Semantic Segmentation. *arXiv* **2021**. [CrossRef]
18. Siegwart, R.; Nourbakhsh, I.; Scaramuzza, L. *Introduction to Autonomous Mobile Robots*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2011.
19. Dilek, E.; Dener, M. Computer Vision Applications in Intelligent Transportation Systems: A Survey. *Sensors* **2023**, *23*, 2938. [CrossRef]
20. Che, C.; Zheng, H.; Huang, Z.; Jiang, W.; Liu, B. Intelligent Robotic Control System Based on Computer Vision Technology. *Appl. Comput. Eng.* **2024**, *64*, 142–147. [CrossRef]
21. Upadhyay, D.; Upadhay, D.K.; Singh, R.; Mishra, D.; Khatri, V. Robotic Vision: Advancements in Computer Vision for Autonomous Systems. *Tuijin Jishu/J. Propuls. Technol.* **2023**, *4*. [CrossRef]
22. Sultana, F.; Sufian, A.; Dutta, P. Evolution of Image Segmentation using Deep Convolutional Neural Network: A Survey. *Knowl.-Based Syst.* **2020**, *201–202*, 106062. [CrossRef]
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
24. Bai, Y.; Mei, J.; Yuille, A.; Xie, C. Are Transformers More Robust Than CNNs? *arXiv* **2021**, arXiv:2111.05464.
25. Zhai, X.; Kolesnikov, A.; Houlsby, N.; Beyer, L. Scaling Vision Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
26. Hurtado, J.V.; Valada, A. Semantic Scene Segmentation for Robotics. *arXiv* **2024**, arXiv:2401.07589.
27. Wang, Y.; Ahsan, U.; Li, H.; Hagen, M. A Comprehensive Review of Modern Object Segmentation Approaches. *Found. Trends® Comput. Graph. Vis.* **2023**, *13*, 111–283. [CrossRef]
28. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *arXiv* **2014**, arXiv:1411.4038.
29. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
30. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
31. Zheng, Y.; Tan, X.; Zheng, Z.; Zhou, Y.; Li, B.; Yi, L. Rethinking Semantic Segmentation with Transformers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
32. Xie, E.; Wang, W.; Yu, Z.; An, T.; Gao, Y.; Lu, M.; Xu, X.; Ren, T.; Zhang, C.; Xiao, T.; et al. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv* **2021**, arXiv:2105.15203.
33. Yuan, K.; Guo, S.; Liu, Z.; Zhou, A.; Yu, F.; Wu, W. Incorporating Convolution Designs into Visual Transformers. *arXiv* **2021**, arXiv:2103.11816.
34. Li, J.; Zhang, Y.; Yun, P.; Zhou, G.; Chen, Q.; Fan, R. RoadFormer: Duplex Transformer for RGB-Normal Semantic Road Scene Parsing. *IEEE Trans. Intell. Veh.* **2024**, *9*, 5163–5172. [CrossRef]
35. Guo, Z.; Bian, L.; Huang, X.; Wei, H.; Li, J.; Ni, H. DSNet: A Novel Way to Use Atrous Convolutions in Semantic Segmentation. *arXiv* **2024**, arXiv:2406.03702. [CrossRef]
36. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *arXiv* **2020**, arXiv:1908.07919. [CrossRef]

37. Qashqai, D.; Mousavian, E.; Shokouhi, S.B.; Mirzakuchaki, S. CSFNet: A Cosine Similarity Fusion Network for Real-Time RGB-X Semantic Segmentation of Driving Scenes. *arXiv* **2024**, arXiv:2407.01328.
38. Termritthikun, C.; Jamtsho, Y.; Ieamsaard, J.; Muneesawang, P.; Lee, I. EEEA-Net: An Early Exit Evolutionary Neural Architecture Search. *Eng. Appl. Artif. Intell.* **2021**, *104*, 104397. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Application of Artificial Neural Networks in Predicting Surface Quality and Machining Time

Andrei Raul Osan and Raul Florentin Drenţa *

Department of Engineering and Management of Technology, Technical University of Cluj-Napoca, North University Center of Baia Mare, 430122 Baia Mare, Romania; andrei.osan@imtech.utcluj.ro
* Correspondence: raul.drenta@imtech.utcluj.ro; Tel.: +407-4568-8591

Abstract: This paper explores the use of artificial neural networks to optimize metal machining processes. It has two main components. The first component focuses on developing a neural network in EasyNN to predict quality errors of metal surfaces machined with toroidal milling. The input data are extracted from the experimental values of the machining operations, and the output data are the measured surface roughness values. The network created not only predicts the quality errors but also identifies the key parameters influencing their quality, such as cutting speed, feed rate, and tool geometry, which are crucial for process optimization. The second component underscores the accuracy of the neural network′s predictions by using MatLab to develop a neural network that estimates the machining time, a key factor in cost calculation and efficient order planning. The experimental data used to train the network come from a restricted set of machining jobs. The obtained results demonstrate an adequate estimation of the machining time, thus facilitating the optimization of the entire metal machining process, positively impacting the cost and efficiency of industrial operations.

Keywords: neural network; input neurons; output neurons; hidden neurons; EasyNN; MatLab; roughness prediction; processing time; toroidal milling; machining regimes

1. Introduction

We hear more and more about artificial intelligence, software that can answer certain questions, and applications that help mankind by allowing us to understand and link information. All these software are based on artificial neural networks, where artificial intelligence manages to provide answers through a computational algorithm, through tens, hundreds, thousands, and millions of hidden neurons [1,2].

If we consider the level of continuous development of cutting surface machining, new technologies, optimizations, tools, machines, software, and precision measuring instruments, we consider the involvement of artificial neural networks in machining to be more than necessary.

The application of artificial neural networks in cutting process performance improvement has been extensively studied in different research [3,4]. These studies, along with the prediction of machined surface quality and output quality, demonstrate the practicality and applicability of these tools [5,6]. Other approaches of artificial neural networks in machining process improvement have been applied in the study of cutting tool wear [7] as well as the study of cutting forces [8,9]. These practical applications of artificial neural networks in the field of machining instill confidence in their effectiveness and utility.

Implementing artificial neural networks for surface quality monitoring [10] or tool wear monitoring [11] in chip machining by either milling or turning offers a wealth of opportunities, leading to lower production costs.

EasyNN plus, a specialized software for building and using neural networks, played a crucial role in this research. It provided a user-friendly graphical user interface and multiple facilities related to network manipulation, making it an ideal tool for our study.

According to Boujelbene et al., milling operation is the most important and widely used technology for obtaining complex surfaces. A computer numerically controlled (CNC) milling machine and CAD/CAM software (PowerMILL 2015) capable of realizing an NC program are needed to perform this operation. In this process, the machine should move as fast as possible with well-defined trajectories according to the cutting tool typology [12]. The optimization of milling operations is always a hot topic in manufacturing because the ability to carry out machining processes at maximum performance is irreplaceable in increasing production efficiency. However, obtaining reliable and practicable solutions is still challenging for milling operations [13].

The actual machining time depends on the CNC, dynamic machine architecture, setup times, space ergonomics, and their implementation, as well as the CAM software used [14].

Current CAD/CAM software estimates the milling time mathematically by dividing the tool path length by the user-set feed rate value, excluding the machining limitation and the machine/CNC feed rate oscillation. This approach, however, does not account for the actual conditions of the machining process, leading to inaccurate estimates. For these reasons, the CAM-estimated time is not the correct one. As a result, in this paper, we will only summarize the basic time.

Some authors have proposed artificial neural networks for feature recognition [15] to further establish relationships between the number of features and part cost [16].

Other authors have presented models using artificial intelligence to estimate machining time with average errors of 10 and 20 min [17] or manufacturing time with a percentage error of less than 25% [18]. Artificial intelligence tools for estimating manufacturing costs in the design phase have also been studied [19–22], as well as the effectiveness of different types of ANNs in estimating machining times [23–27].

The main goal of this work is to develop and validate a model based on artificial neural networks, capable of providing a fast and accurate estimation of machining times and the quality of surface for industrial parts. This approach is intended to be a viable and efficient alternative to the traditional simulation methods used in CAM software, methods that, although accurate, involve a significant consumption of time and resources.

The overall objective is to create an intelligent tool that supports companies in the field of machining in optimizing production processes, by reducing the time needed to estimate costs and delivery times. In this regard, the proposed neural network will allow the automatic and real-time deduction of processing times and surface quality, thus providing valuable decision-making support in the bidding, planning and execution phases of orders.

2. Materials and Methods

The present paper aims to investigate the applicability of artificial neural networks in two essential directions for the machining industry: on the one hand, the prediction of surface quality by estimating roughness in the case of flat surfaces, and on the other hand, the estimation of execution times in the machining process of concave cylindrical surfaces.

The choice of these two characteristics—surface quality and process duration—reflects the current requirements of the economic environment, which aims to achieve high-quality finishes in the shortest possible manufacturing time. The simultaneous fulfillment of these

objectives leads, implicitly, to the reduction in production costs and to the observance of delivery deadlines, essential aspects for the competitiveness of companies.

Regarding the approach of the two types of surfaces analyzed in this work, the high potential of artificial neural networks to be applied in a wide range of contexts is highlighted, while also being able to work in parallel in any circumstances.

2.1. Experimental Research on Flat Surface Processing with Toroidal Milling Cutter

The experimental investigation that underpinned this study was detailed in previously published papers [2], aiming to collect the input and output data needed to train the artificial neural network. As for the practical experiment, it consisted of processing 27 test samples on a five-axis CNC machining center, model Okuma MU-400VA from Okuma Deutschland GmbH from Köln, Germany. (Figure 1A).



Figure 1. (**A**) CNC 5 axis Okuma MU-400VA. (**B**) Toroidal milling Ø16.

The number of samples was established based on the Taguchi method, using the minimum number of tests necessary to obtain statistically relevant results.

Geometrically, each specimen had dimensions of 40 × 40 × 20 mm^3. The finishing operation was carried out on a flat and rough surface, with the pieces being fixed in a vice with the help of a special clamping device.

The finishing machining was performed with a toroidal milling cutter with a diameter of Ø16 and a corner radius of 4 (Figure 1B) in the presence of the cutting liquid. The cutting liquid, which is a coolant and lubricant, helps to reduce heat and friction during the machining process, thereby improving the surface quality. The machining was carried out with a constant chip depth (ap = 0.5 mm) and radial depth (ae = 0.3 mm) (Figure 2b).

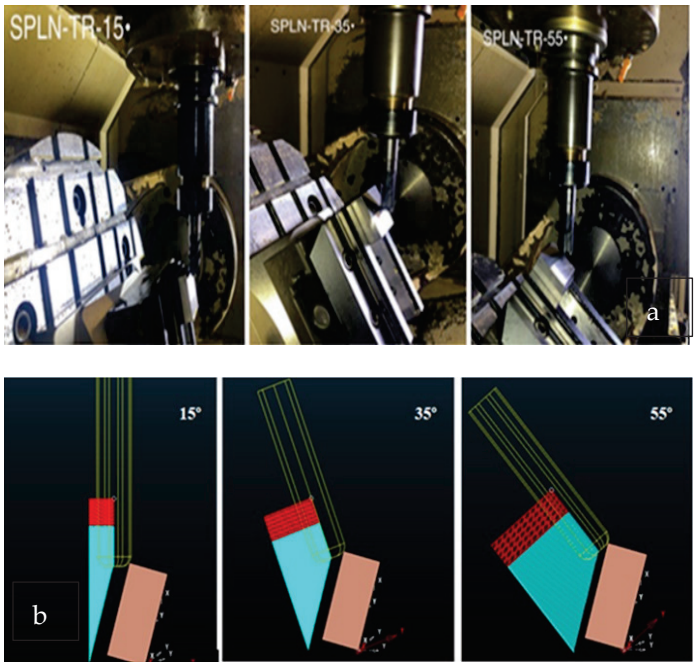


Figure 2. (**a**) Simulation of machining in CAM at the 3 angles. (**b**) Processing on CNC at the 3 angles.

The CAM program was generated in the PowerMILL software (Figure 2a), which can generate the tool paths by respecting the required machining regimes and the transmitted inclination angle.

In order to execute the processing of the 27 surfaces included in the scope of this study, the experimental framework was designed in accordance with the Taguchi method, as described in detail in prior research. This method provides a robust and systematic approach for optimizing process parameters by minimizing variability and identifying the most influential factors with a reduced number of experimental runs. Based on this framework, it was established that a portion of the cutting regime parameters would be held constant throughout the experimentation. This decision was made to ensure consistency and control over extraneous variables, thereby isolating the effects of the key parameters under investigation. The specific values and constraints of these constant parameters were defined in alignment with the technical specifications and machining conditions presented earlier.

In parallel, three primary technological parameters were selected to serve as variable inputs for the development and training of the artificial neural network (ANN) model proposed in this study. These input variables, as listed in Table 1, are cutting speed, tool tilt angle, and feed per tooth. The selection of these particular variables was guided by both theoretical and empirical considerations. From a theoretical standpoint, these parameters are known to play a decisive role in determining machining outcomes, particularly with respect to surface integrity and overall process efficiency. Empirical evidence from previous investigations further supports their significant influence on surface roughness, dimensional accuracy, material removal rate, and tool wear.

By incorporating these three variables into the ANN model, the objective is to enable a predictive framework capable of capturing the nonlinear relationships between input parameters and machining responses. Their selection reflects a deliberate focus on optimizing both surface quality and processing time, two of the most critical performance indicators in modern manufacturing environments. The ANN′s ability to model these complex interdependencies is expected to provide valuable insights for process optimization and decision-making in advanced machining applications.

Table 1. Cutting regimes and measured roughness values.

Experiment	Cutting Speed [m/min]	Angle of Inclination [°]	Feed Per Tooth [mm/tooth]	Ra Parallel [µm]	Rt Parallel [µm]	Ra Perpendicular [µm]	Rt Perpendicular [µm]
SPLN-TR-1	80	15	0.11	0.16	0.926	0.464	2.92
SPLN-TR-2	80	15	0.15	0.289	1.68	0.403	2.593
SPLN-TR-3	80	15	0.19	0.264	1.706	0.376	2.4
SPLN-TR-4	80	35	0.11	0.194	1.466	0.494	2.613
SPLN-TR-5	80	35	0.15	0.311	2.413	0.435	3.499
SPLN-TR-6	80	35	0.19	0.288	2.719	0.425	2.953
SPLN-TR-7	80	55	0.11	0.45	2.609	0.688	4.273
SPLN-TR-8	80	55	0.15	0.727	4.413	0.632	4.353
SPLN-TR-9	80	55	0.19	0.894	4.639	0.761	5.84
SPLN-TR-10	170	15	0.11	0.689	3.486	0.672	4.519
SPLN-TR-11	170	15	0.15	0.817	4.646	0.756	5.406
SPLN-TR-12	170	15	0.19	0.876	5.094	0.851	6.086
SPLN-TR-13	170	35	0.11	0.597	3.247	0.786	5.579
SPLN-TR-14	170	35	0.15	0.693	3.906	0.805	4.826
SPLN-TR-15	170	35	0.19	1.071	5.76	1.119	7.666
SPLN-TR-16	170	55	0.11	0.457	3.273	0.535	3.619
SPLN-TR-17	170	55	0.15	0.773	4.306	0.627	5.106
SPLN-TR-18	170	55	0.19	1.131	6.103	0.812	6.293
SPLN-TR-19	210	15	0.11	0.53	3.786	0.549	4.053
SPLN-TR-20	210	15	0.15	0.859	4.833	0.757	4.759
SPLN-TR-21	210	15	0.19	0.893	4.799	0.869	6.326
SPLN-TR-22	210	35	0.11	0.568	3.899	0.543	4.146
SPLN-TR-23	210	35	0.15	0.598	3.753	0.573	4.496
SPLN-TR-24	210	35	0.19	0.513	2.966	0.618	4.586
SPLN-TR-25	210	55	0.11	0.402	3.552	0.445	3.7
SPLN-TR-26	210	55	0.15	0.402	2.993	0.475	3.853
SPLN-TR-27	210	55	0.19	0.423	3.186	0.596	4.56

2.2. Experimental Research on the Machining of Concave Cylindrical Surfaces with Toroidal Milling Cutter

Our analysis of the concave cylindrical surface reveals a consistent cylindrical shape along the entire length of the surface. The surface, with a diameter of Ø85 and a height of 5 mm, measures 40 × 40 mm^2, as shown in Figure 3. The practical experiments on the concave cylindrical surface are presented in Table 2, which details nine surfaces, each created using toroidal milling. The table includes a column highlighting the execution times of each experiment, excluding the constant auxiliary times, to provide a clear comparison of the experiments.

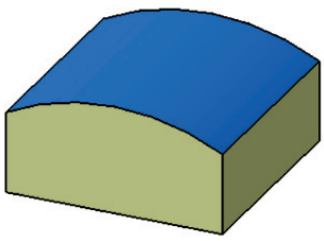


Figure 3. Shape of the machining surface.

Table 2. Run time for machining concave cylindrical surfaces with toroidal milling cutter.

Experiment	Cutting Speed [m/min]	Angle of Inclination [°]	Feed Per Tooth [mm/tooth]	Execution Time [s]
CICV-TR-1	80	55°	0.11	964
CICV-TR-2	80	55°	0.15	734
CICV-TR-3	80	55°	0.19	601
CICV-TR-4	170	55°	0.11	508
CICV-TR-5	170	55°	0.15	399
CICV-TR-6	170	55°	0.19	337
CICV-TR-7	210	55°	0.11	430
CICV-TR-8	210	55°	0.15	343
CICV-TR-9	210	55°	0.19	292

Regarding the cutting regimes, the three variables in the table (Table 2) are the focus, while the rest remain constant. The cutting depth is consistently set at 0.5 mm (ap = 0.5 mm) and the radial depth at 0.3 mm (ae = 0.3 mm) for each experiment, all conducted in the presence of emulsion, which serves as a crucial element in the process. As for the clamping method, the pad is securely clamped between the two clamping trays of the vice, which are further reinforced with clamping for extra rigidity and security.

In order to be able to carry out the machining, CAM programs are needed; they were created using the PowerMILL software, a software able to create the tool, and more importantly, to create the routes to be covered by the tool with the corresponding inputs and outputs. We decided that each machining would be one-way, with the tool paths parallel. For machining the surface with the toroidal milling cutter, in the case of the 55° inclination, 138 lines were needed. Figure 4a shows an image during the simulation of the toroidal toolpath on the concave cylindrical surface.

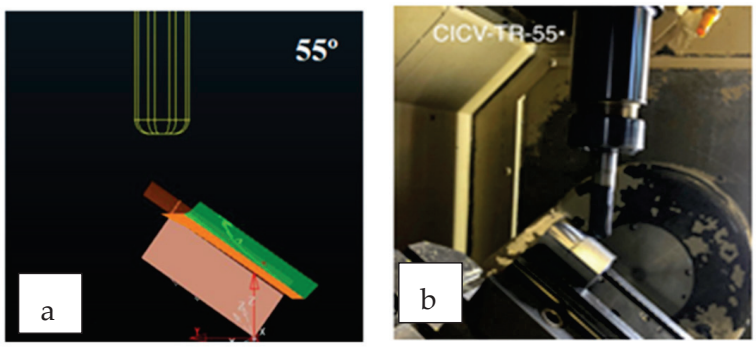


Figure 4. (**a**) Simulating the toroidal cutter path in PowerMILL on CICV-TR. (**b**) Machining the concave cylindrical surface with the toroidal cutter.

Figure 4b presents the actual practical experiment conducted on the OKUMA MU-400VA CNC. This figure vividly demonstrates how the toroidal milling cutter effectively processes the concave cylindrical surface, providing a tangible application of the theoretical concepts discussed.

It is important to note that throughout the machining of the concave cylindrical surface with the toroidal milling cutter, the tool axis constantly maintains the desired inclination angle concerning the surface.

Experimental Results

Following the processing and simulation in CAM software, different processing times resulted during the realization of the nine surfaces with different regimes, as presented in Table 2.

3. Results and Discussion

3.1. Implementing the Artificial Neural Network in EasyNN

An example of a neural network constructed utilizing EasyNN software is presented below. A neural network must be constructed to predict the roughness values Ra, measured both in parallel and perpendicularly, as well as the roughness Rt, also measured in both orientations.

The data input is a critical step in the process. The input data consists of the cutting speed, the feed per tooth, and the feed angle, which are the three variables utilized to machine each SPLN-TR surface. The output data are characterized by the roughness measurements taken in both parallel and perpendicular orientations to the feed direction. These measurements are denoted as the theoretical surface roughness, Rt, and the arithmetic mean roughness, Ra.

The database depicted in Figure 5 is structured as a data grid. The rows correspond to the examples, while the columns denote the inputs and outputs associated with each example. The initial column specifies the type and designation of the input and output example. The subsequent three columns denote the three process variables that are being materialized as input data, while the final four columns indicate the measured values of roughness, representing the output data.

EasyNN-plus - [SPLN-TR.tvq]

File Edit Defaults View Zoom Insert Action Query Tools Macro Subset Window Help

	I:0	I:1	I:2	O:3	O:4	O:5	O:6
F1	vc	fz	Unghi	Ra paralel	Ra perpendi	Rt paralel	Rt perpepnd
SPLN-TR-1	80.0	15°	0.11	0.16	0.464	0.926	2.92
SPLN-TR-2	80.0	15°	0.15	0.289	0.403	1.68	2.593
SPLN-TR-3	80.0	15°	0.19	0.264	0.376	1.706	2.4
SPLN-TR-4	80.0	35°	0.11	0.194	0.494	1.466	2.613
SPLN-TR-5	80.0	35°	0.15	0.311	0.435	2.413	3.499
SPLN-TR-6	80.0	35°	0.19	0.288	0.425	2.719	2.953
SPLN-TR-7	80.0	55°	0.11	0.45	0.688	0.926	4.273
SPLN-TR-8	80.0	55°	0.15	0.727	0.632	0.926	4.353
SPLN-TR-9	80.0	55°	0.19	0.894	0.494	4.639	5.84
SPLN-TR-10	170.0	15°	0.11	0.689	0.494	3.486	2.593
SPLN-TR-11	170.0	15°	0.15	0.817	0.756	4.646	4.273
SPLN-TR-12	170.0	15°	0.19	0.876	0.851	5.094	6.086
SPLN-TR-13	170.0	35°	0.11	0.597	0.786	3.247	5.579
SPLN-TR-14	170.0	35°	0.15	0.693	0.627	3.906	4.826
SPLN-TR-15	170.0	35°	0.19	0.513	1.119	5.76	7.666
SPLN-TR-16	170.0	55°	0.11	0.457	0.543	3.273	3.619
SPLN-TR-17	170.0	55°	0.15	0.773	0.627	4.306	5.106
SPLN-TR-18	170.0	55°	0.19	1.131	0.618	6.103	4.586
SPLN-TR-19	210.0	15°	0.11	0.53	0.573	2.993	4.053
SPLN-TR-20	210.0	15°	0.15	0.859	0.757	4.833	4.759
SPLN-TR-21	210.0	15°	0.19	0.893	0.869	4.799	6.326
SPLN-TR-22	210.0	35°	0.11	0.568	0.543	3.899	4.146
SPLN-TR-23	210.0	35°	0.15	0.598	0.573	3.753	4.496
SPLN-TR-24	210.0	35°	0.19	0.513	0.618	2.966	4.586
SPLN-TR-25	210.0	55°	0.11	0.402	0.445	3.552	3.7
SPLN-TR-26	210.0	55°	0.15	0.402	0.475	2.993	3.853
SPLN-TR-27	210.0	55°	0.19	0.423	0.596	3.186	4.56

Figure 5. The database entered to create the neural network using EasyNN plus software. ("*" is degrees).

After loading the database, the subsequent step involves verifying the usability of the data via the "Query" function, followed by the establishment of the neural network.

The neural network is established using the previously loaded data grid. This process is initiated through the "new network" function, which generates the new network dialog. This dialog facilitates the specification of the neural network configuration, enabling it to learn information from the database.

The learning progress indicates the advancement of the learning process. Figure 6 illustrates the learning process graph, which displays the maximum, average, and minimum training error values. The graph progressively narrows as learning advances, ensuring visibility of the complete learning curve. A maximum of 200,000,000 learning cycles can be represented in the graph; however, in this instance, only 3873 learning cycles are required.

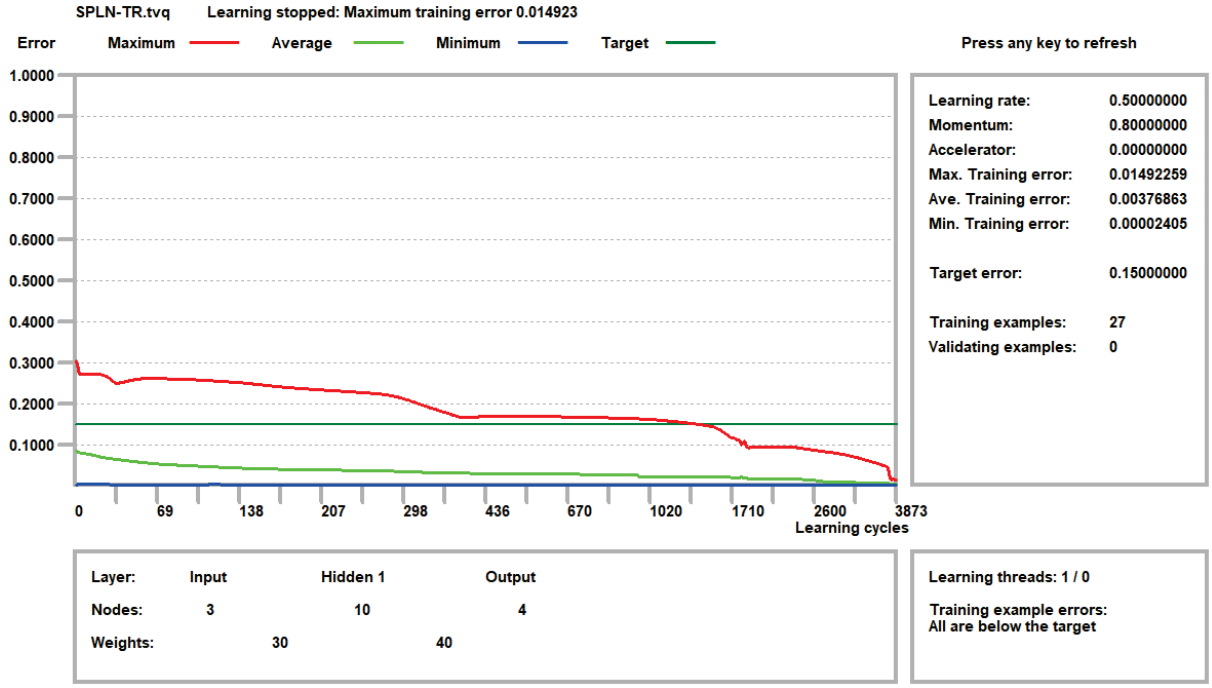


Figure 6. The learning graph.

Figure 6 displays a vertical scale on the left side, ranging from 0 to 1, which indicates the position of the curve. The horizontal scale is labeled in accordance with the quantity of the learning cycles.

Below the horizontal scale, the network is displayed, indicating the number of inputs, outputs, and hidden neurons. In this instance, a variant with 10 hidden neurons has been selected. The window displays the quantity of connections among nodes, specifically 30 connections between the input layer and the hidden layer, and 40 connections between the hidden layer and the output neurons.

The right side of the graph displays the window containing the learning and validation data. The learning rate is configured at 0.5, the momentum is established at 0.8, and the training process is programmed to terminate when the error threshold reaches 0.15. The recorded maximum error is 0.01492259, while the minimum error is 0.000024. The average of the errors is calculated to be 0.0376863.

Upon completion of the settings, the configuration of the analyzed neural network is illustrated in Figure 7.

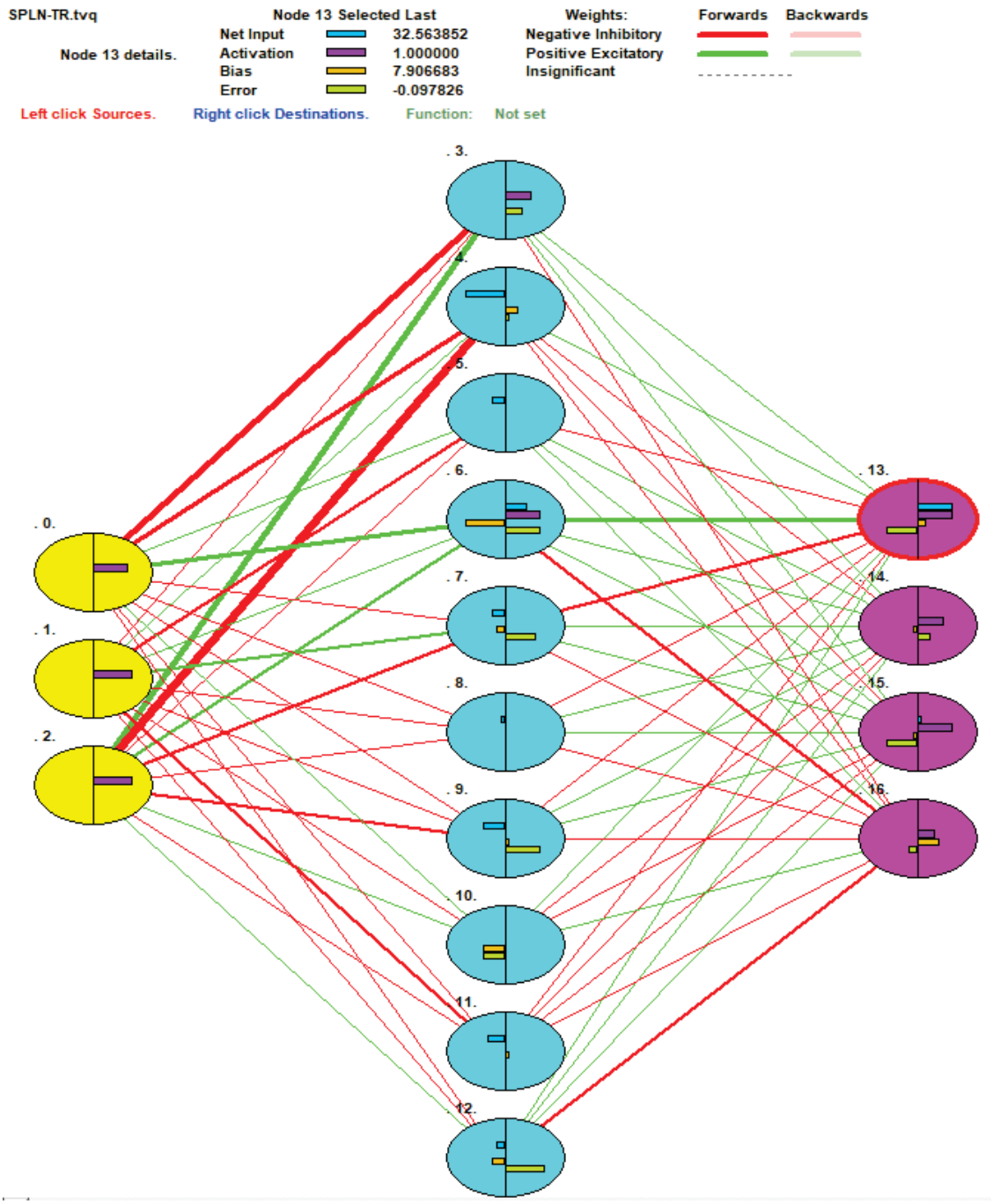


Figure 7. Neural network structure.

To utilize the neural network effectively, it is essential to select the source and destination of the connections for the purpose of connecting and disconnecting the nodes. Connections can be established between any nodes or layers, allowing for flexible integration across the system. Feedback and layer-level connections are feasible.

The quantity of nodes within hidden layers may be adjusted, and additional hidden layers can be incorporated as required. The nodes will be reestablished utilizing the connections stored in the network's memory. The connections will be maintained in a manner that closely resembles their prior configuration.

3.2. Implementing the Artificial Neural Network in MATLAB

Neural networks primarily function by identifying a relationship that links the output parameters to the input parameters. Throughout the training process. Upon completion of the training process, the neural network demonstrates functionality in scenarios that were not part of the training dataset.

Cutting parameters represent critical factors that significantly impact process plans. The optimal selection of cutting parameters results in a decrease in cutting actions, which subsequently leads to reduced energy consumption and, consequently, cost reduction. This study aims to implement cutting parameter prediction through the utilization of artificial neural networks (ANNs). Neural networks are widely utilized tools that have demonstrated effectiveness in addressing optimization problems, particularly in the adaptive control of machine tools and in pattern recognition applications.

The MATLAB R2015a program includes a set of functions and graphical interfaces designed for the manipulation of Artificial Neural Networks, collectively referred to as the Neural Network Toolbox. The subsequent sections will outline the methodology for utilizing the basic functions and interacting with the graphic interfaces.

Prior to the development of an artificial neural network, it is essential to establish the input data and the output data. The computer resources outlined in the preceding chapter are applicable in this context.

In order to generate the input and output data, we selected to analyze the fundamental time derived from the processing of nine cylindrical-concave specimens using the toroidal milling cutter. Consequently, the two process variables are identified as input data, specifically the cutting speed and the feed per tooth. The output data and target information are represented by the fundamental time required to process each surface.

The artificial neural network was developed in the MATLAB environment, using the input and output data presented above. The structure of the network comprises two layers and was trained using the TRAINLM (Levenberg–Marquardt) algorithm, known for its efficiency in regression problems. For the hidden layer, 10 neurons were chosen, a choice made after preliminary tests to ensure a balance between complexity and generalization capacity.

In terms of network performance, the training process ended after three epochs with the network reaching a satisfactory level of convergence. The dataset was divided as follows: about 77% of the data was allocated for training and 11% each was used for validation and testing for the purpose of evaluating the robustness and accuracy of the model.

The first step in generating a neural network consists of providing the input and output data necessary for the artificial neural network to train and learn, a process presented in Appendix A.

Upon completion of the training and learning phase of the artificial neural network, the subsequent stage involves simulation (refer to Figure 8) and the generation of predictions (see Figure 9) based on the trained neural network and new input data.

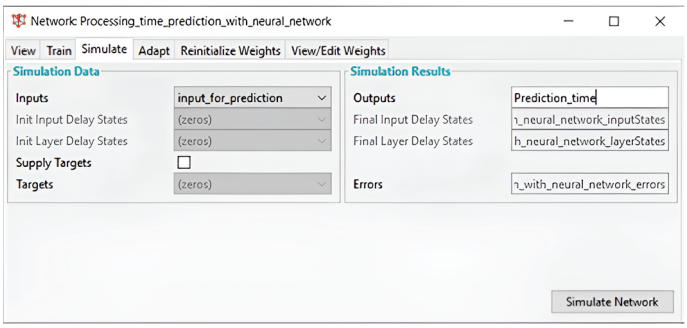


Figure 8. Training the artificial neural network for working time prediction.

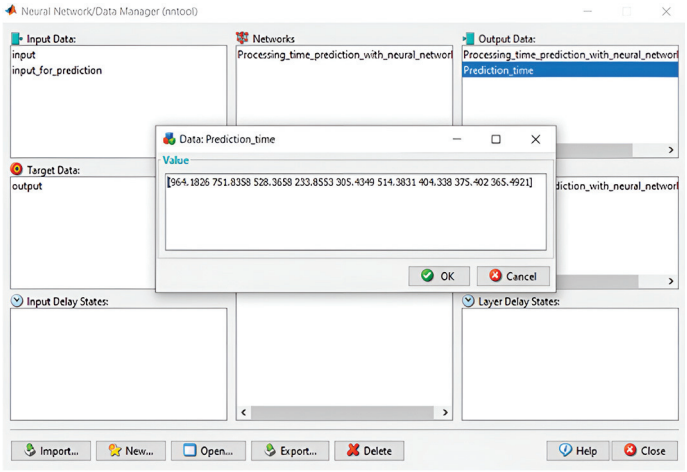


Figure 9. Artificial neural network predictions.

3.3. The Output Produced by the Artificial Neural Network Created Using EasyNN

The prediction view comprises two windows that display the proximity of the predicted value for each example. Five examples analogous to the initial data were generated for the purpose of validating the network (see Figure 10). The values on both axes are scaled within the range of 0 to 1. The anticipated outcomes for the example converge towards the actual values as the training advances.

When the predicted values are in close proximity to the true values, the corresponding points will align along the diagonal line. If a majority of the data points are situated near the diagonal line, it indicates that the network may be overfitted and could potentially fail to produce accurate values.

Figure 11 illustrates the values of the arithmetic roughness Ra column, which were measured in parallel. The vertical scale on the left represents values ranging from 0 to 1.

The vertical scale on the right displays the actual value derived from the maximum and minimum values in the column. The blue line represents the true value, the red line indicates the moving average, and the black dotted line illustrates the overall trend.

The network undergoes training following the completion of 3873 cycles. The error rate is 0.003755, as indicated in the network information presented in Figure 12.

The error visualization is depicted through the absolute and relative errors generated across 27 examples, arranged in descending order. The highest recorded error occurs on the SPLN-TR-25 surface, with a value of 0.0147965418. Despite being the maximum value, it remains below the established threshold of 0.15, which is the maximum permissible error.

The minimum error value is documented on the SPLN-TR-4 surface, recorded as 0.0000252056.

The importance function is illustrated by showing the relative significance of each input column. The significance is determined by the total of the absolute weights associated with the connections from the input node to each node in the first hidden layer. Entries are organized in descending order of significance, beginning with the entry of highest importance.

Figure 13, derived from the provided network data, indicates that the feed per tooth is the most significant input column, with a value of 135.3001. This parameter has a substantial impact on the output values, while the inclination angle of the tool axis exhibits the least influence.

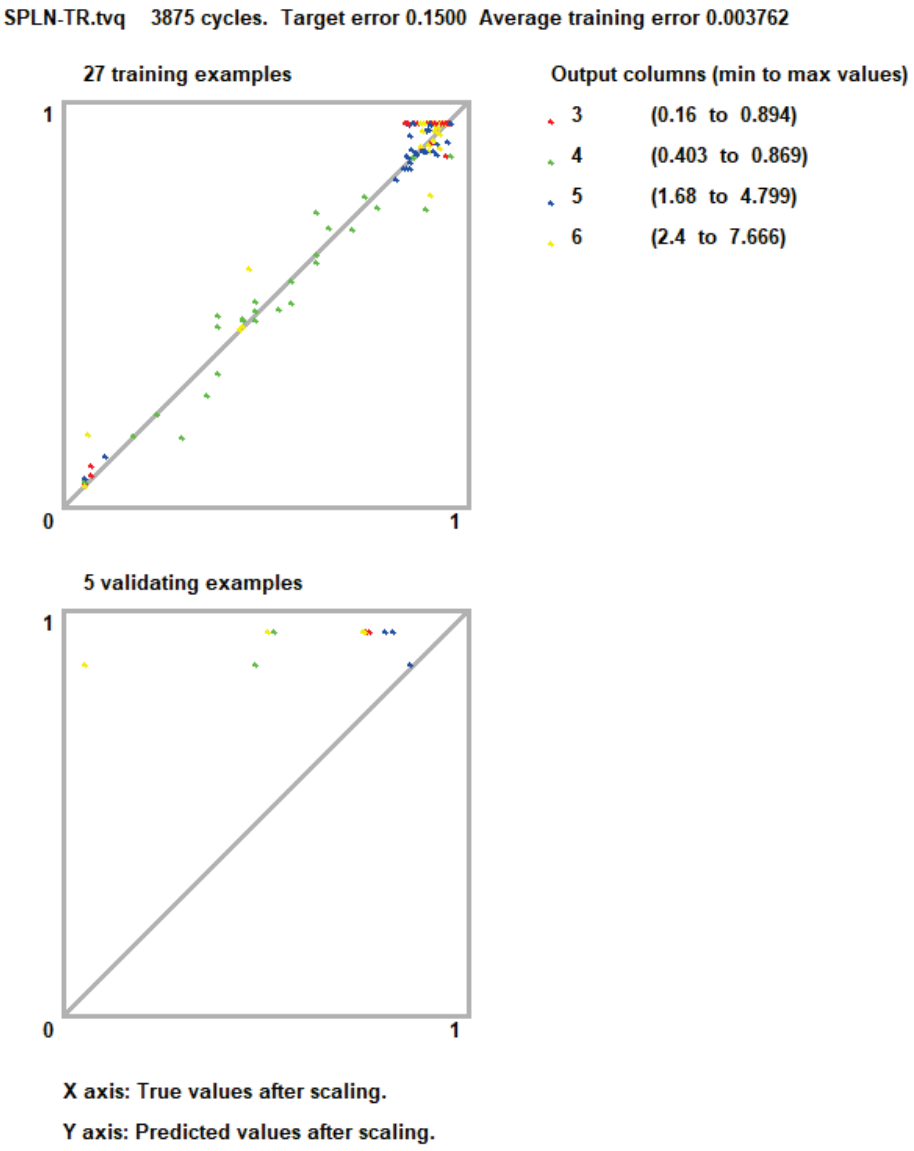


Figure 10. Value prediction chart.

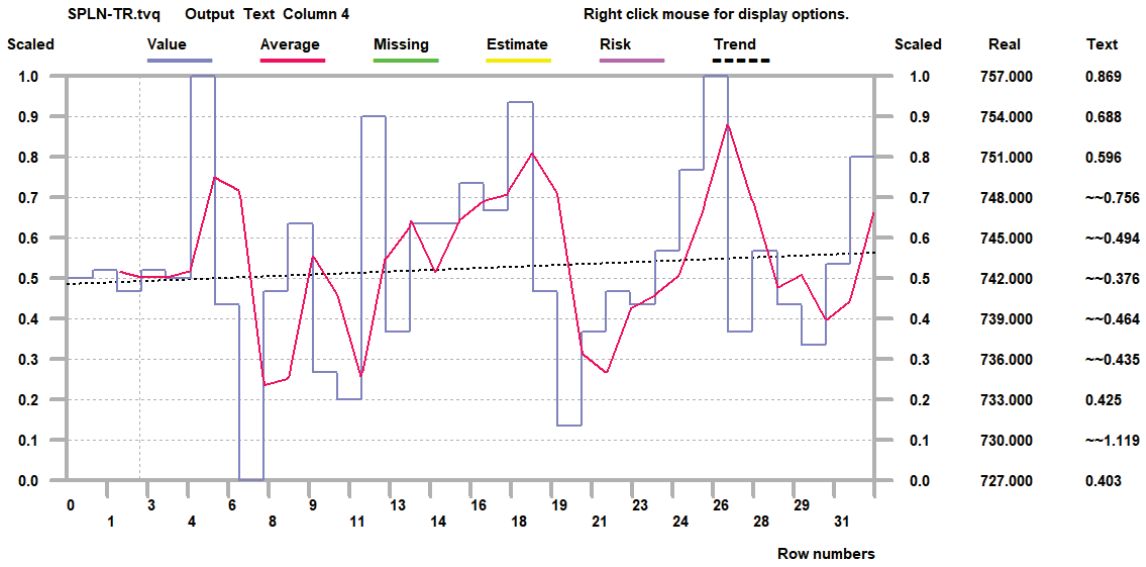


Figure 11. The value of the column representing the roughness Ra measured parallel.

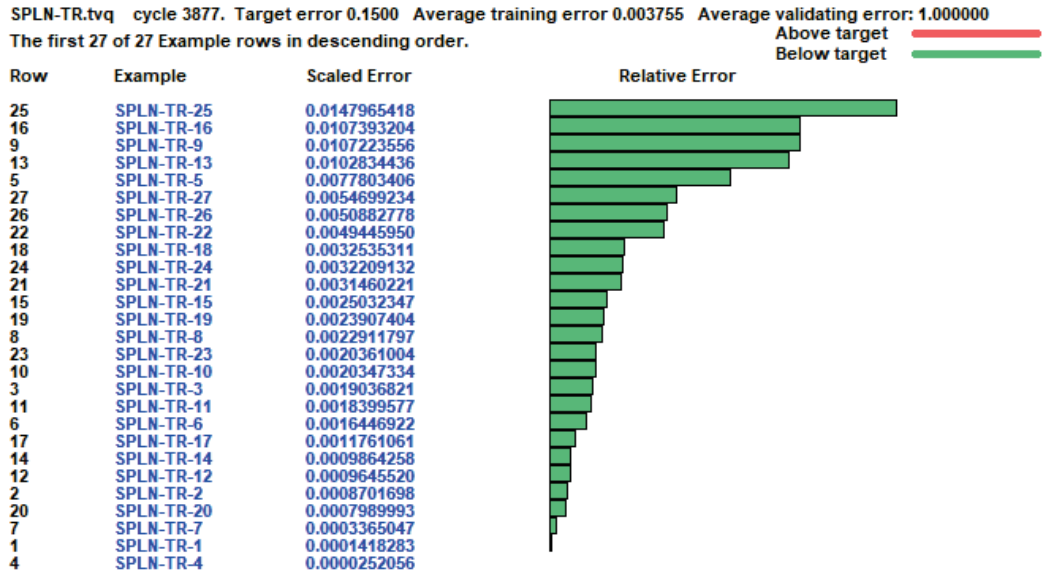


Figure 12. Information on neural network error values.

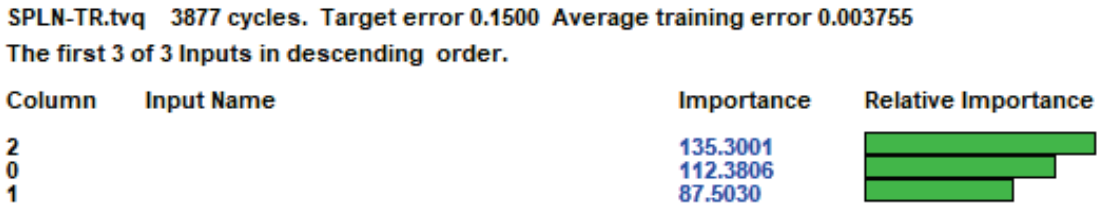


Figure 13. Importance values of the input columns of the neural network.

The sensitivity visualization illustrates the extent to which output values vary in response to changes in input values. The inputs are configured to median values, after which each input is incrementally adjusted from the minimum value to the maximum value. The output variation is documented as each input is incrementally raised from the minimum to the maximum, thereby determining the sensitivity to the change. Inputs that are deemed insensitive will not be shown.

The inputs are presented in a descending sequence based on sensitivity, starting with the most sensitive input. Figure 14 indicates that the feed per tooth input column values exhibit the highest sensitivity, while the cutting speed values demonstrate the lowest sensitivity.

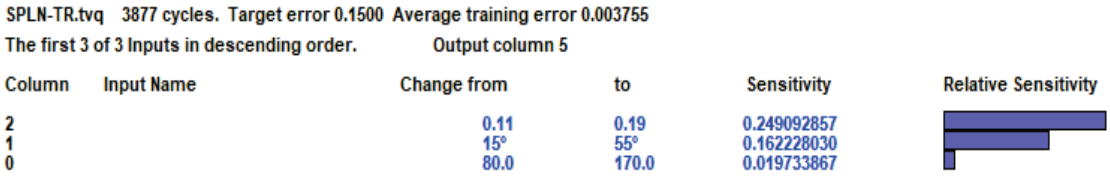


Figure 14. The sensitivity value of input columns on output values.

3.4. Analysis of Neural Network Predictions in MATLAB Alongside Statistical Evaluations in CAM

After the artificial neural network underwent training and learning, it produced time predictions for each cutting regime based on the new input values.

Table 3, presented below, displays the updated values of the cutting regimes and times, as determined by the neural network.

Table 3. The new input values and predicted times.

Experiment	Cutting Speed [m/min]	Feed Per Tooth [mm/tooth]	Estimated Time [s]
CICV-TR-1	80	0.11	964.18
CICV-TR-2	100	0.13	751.83
CICV-TR-3	120	0.15	528.36
CICV-TR-4	140	0.17	233.85
CICV-TR-5	160	0.19	305.43
CICV-TR-6	180	0.17	514.38
CICV-TR-7	200	0.15	404.33
CICV-TR-8	220	0.13	375.40
CICV-TR-9	240	0.11	365.49

To assess the accuracy percentage of the neural network, a simulation of the processing was conducted in the CAM software for each trial, as detailed in Appendix B.

Table 4 presents comparisons indicating an average prediction accuracy of 91.98%, accompanied by an average difference of −32 s. This accuracy and average could potentially improve if the predicted time value for the CICV-TR-4 surface were more precise. Despite this error, it can be concluded that the neural network is effectively applicable for predicting processing time.

Table 4. Analysis of the time estimates generated by RNA in relation to the time derived from CAM simulations.

Experiment	Running Time [s]		Prediction Error [s]	Prediction Accuracy Percentage [%]
	CAM Simulation Values	RNA Predicted Values		
CICV-TR-1	964	964	0	100
CICV-TR-2	745	751	−6	99.19
CICV-TR-3	523	528	−5	99.04
CICV-TR-4	407	233	174	57.24
CICV-TR-5	345	305	40	88.4
CICV-TR-6	481	514	−33	93.13
CICV-TR-7	393	404	−11	97.2
CICV-TR-8	367	375	−8	97.82
CICV-TR-9	381	365	16	95.8

The process variables represent the input values utilized for the creation of the artificial neural network in the MATLAB software and indicate the cutting modes. It is important to note that only the cutting speed and the feed per tooth are subject to variation, while the remaining cutting modes are maintained at constant levels.

The output data is represented by the machining time associated with each surface. The exposed machining time indicates the point at which the tool initiates the machining process, while auxiliary times and preparation times are regarded as constant and excluded from consideration. The times listed are the results derived from simulating each machining process within the CAM software.

4. Conclusions

Artificial neural networks (ANNs) are being integrated into various aspects of daily life, improving processes across multiple industries. In manufacturing, particularly within machining processes, artificial neural networks (ANNs) have shown significant potential for enhancing surface quality prediction and optimizing processes. This study highlights that the performance of artificial neural networks (ANNs) can be markedly improved by engaging in continuous training using precise and meticulously curated datasets.

This study utilized experimental data gathered from machining operations to effectively train an artificial neural network (ANN) with EasyNN software. The ANN was designed to predict surface roughness parameters, specifically Ra and Rt, based on input variables including cutting speed, feed per tooth, and tool inclination angle. The network demonstrated minimal error margins in the predictions of minimum, maximum, and average roughness, with feed per tooth recognized as the most significant input variable. The tool inclination angle demonstrated a negligible effect on the output variables, a finding that was corroborated by sensitivity analysis.

In a similar manner, a second artificial neural network (ANN) was developed using MATLAB to forecast machining time, utilizing a constrained dataset derived from experimental trials focused on concave cylindrical surfaces. Despite utilizing only nine datasets for training, the model achieved a prediction accuracy of 92%. This indicates that increasing the volume of training data could enhance the prediction accuracy towards 100%. This predictive capability provides substantial benefits for industrial applications, including accurate cost estimation, workload scheduling, and deadline forecasting.

Our findings are consistent with and build upon recent studies published in *Machines*, contributing to the advancement of the field. For example, Saric et al. (2023) in the paper [23] is illustrated the viability of machining time prediction through neural computing, attaining accuracies near 90%, yet did not provide comprehensive sensitivity analyses on input parameters. In a similar vein, another study conducted by Reiff C et al. (2021) [28] utilized deep learning techniques for the prediction of surface quality, with a primary emphasis on turning operations rather than milling of complex geometries. Additionally, Kotha Amarnath S. et al. (2025) [29] utilized neural networks for the purpose of tool wear monitoring; however, they did not investigate the direct integration of these networks into machining time estimation [3].

The current study presents a unique approach by simultaneously predicting surface quality and machining time, utilizing experimentally validated data. Additionally, it identifies the relative importance and sensitivity of each input variable. This work offers a robust, experimentally grounded validation of ANN potential in intelligent machining, in contrast to some previous approaches that relied heavily on simulated or theoretical data.

In future developments, the expansion of the dataset and the integration of additional variables, including tool wear, vibration, and cutting forces, have the potential to enhance the ANN models, thereby advancing their capability for real-time intelligent decision-making in manufacturing systems.

Author Contributions: Conceptualization, A.R.O.; methodology, A.R.O.; software, A.R.O.; validation, R.F.D. and A.R.O.; formal analysis, A.R.O.; data curation, A.R.O.; writing—original draft preparation, A.R.O.; writing—review and editing, R.F.D.; visualization, R.F.D.; supervision, R.F.D.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All data supporting the findings of this study are included within the article.

Acknowledgments: During the preparation of this manuscript, the authors used QuillBot for English language corrections (especially grammar). After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial Neural Network
SPLN-TR-(1–27)	Flat surface machined with toroidal cutter number 1–27
CICV-TR-(1–9)	Concave cylindrical surface machined with toroidal cutter no. (1–9)
CAM	Computer-Aided Manufacturing
CNC	Computer Numerical Control
CAD	Computer-Aided Design
Ra	Arithmetic Average Roughness
Rt	Maximum Height of the Roughness Profile
ap	Axial Depth of Cut
ae	Radial Depth of Cut
EasyNN	Easy Neural Network software
MATLAB	Matrix Laboratory (software for numerical computing)
RNA	Artificial Neural Network (Romanian abbreviation)

Appendix A

Generating Neural Network and Predictions

The initial phase of neural network generation involves supplying the requisite input and output data essential for the training and learning processes of the artificial neural network.

Variables - input

input | output

2x9 double

	1	2	3	4	5	6	7	8	9
1	80	80	80	170	170	170	210	210	210
2	0.1100	0.1500	0.1900	0.1100	0.1500	0.1900	0.1100	0.1500	0.1900

Variables - output

input | output

1x9 double

	1	2	3	4	5	6	7	8	9
1	964	734	601	508	399	337	430	343	292

Figure A1. Input and output data of the artificial neural network.

Variables - input_for_prediction

input_for_prediction

2x9 double

	1	2	3	4	5	6	7	8	9
1	80	100	120	140	160	180	200	220	240
2	0.1100	0.1300	0.1500	0.1700	0.1900	0.1700	0.1500	0.1300	0.1100

Figure A2. Input data for generating predictions.

Data loading is performed in the Neural Network/Data Manager module where the input and output data shown above are loaded.

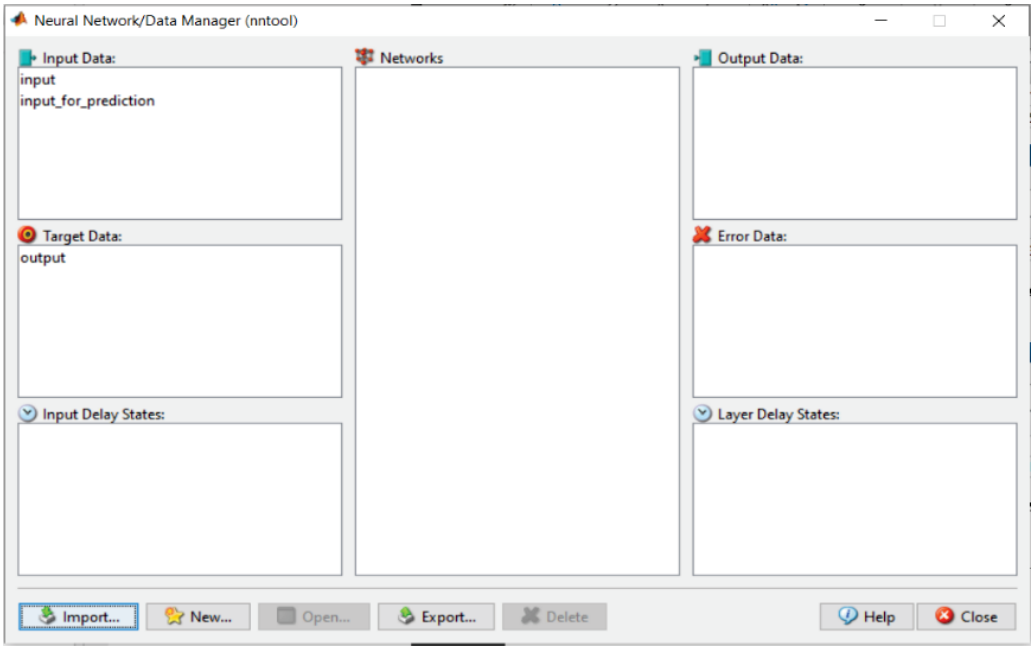


Figure A3. Data loading.

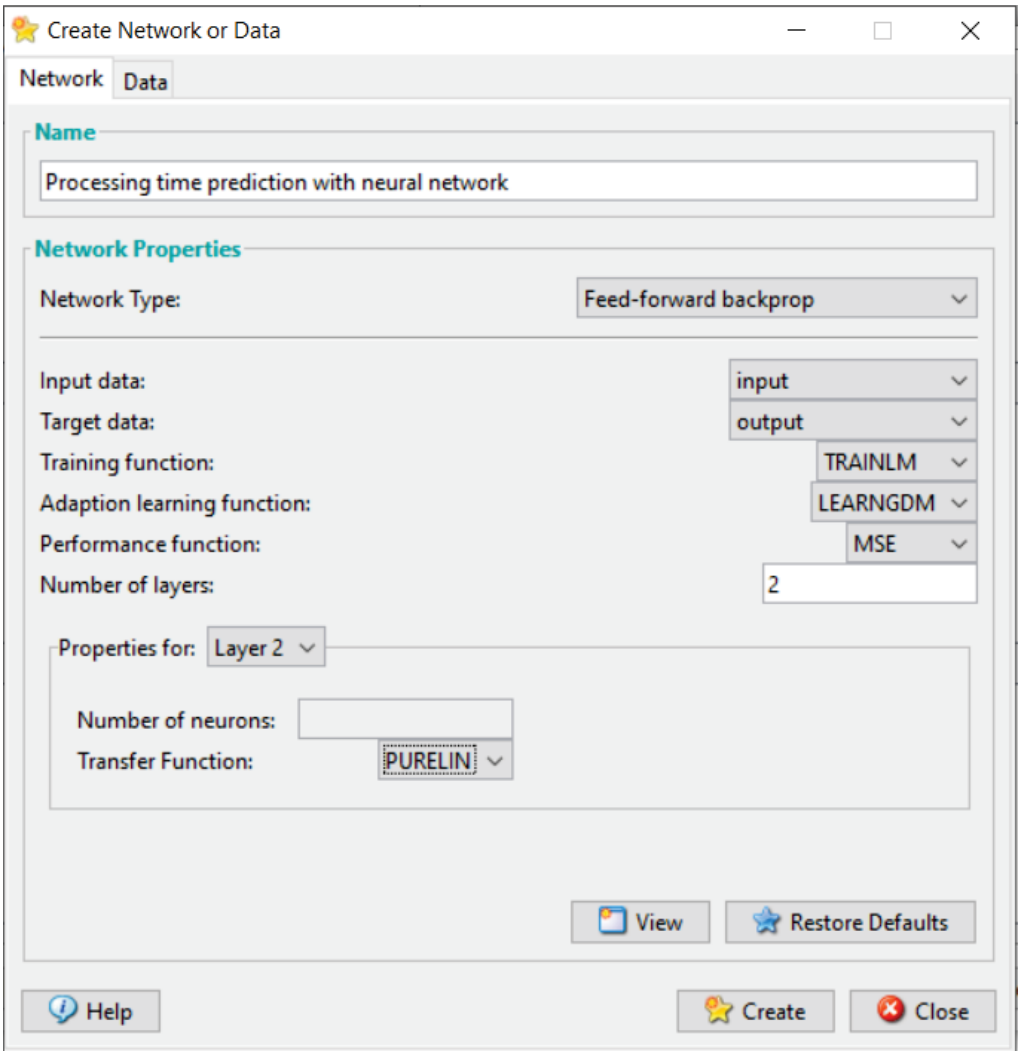


Figure A4. Creating the neural network for predicting working time.

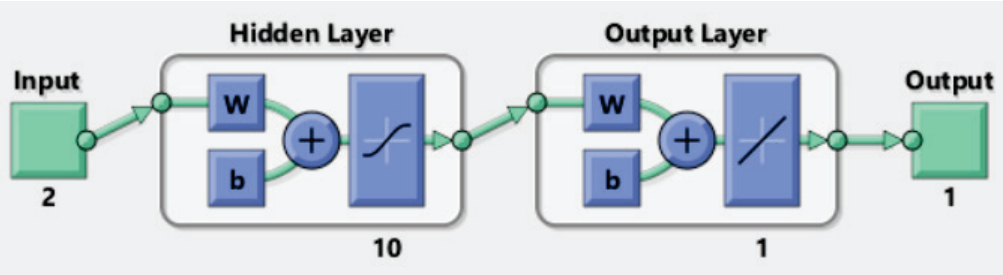


Figure A5. Architecture of the neural network for predicting working time.

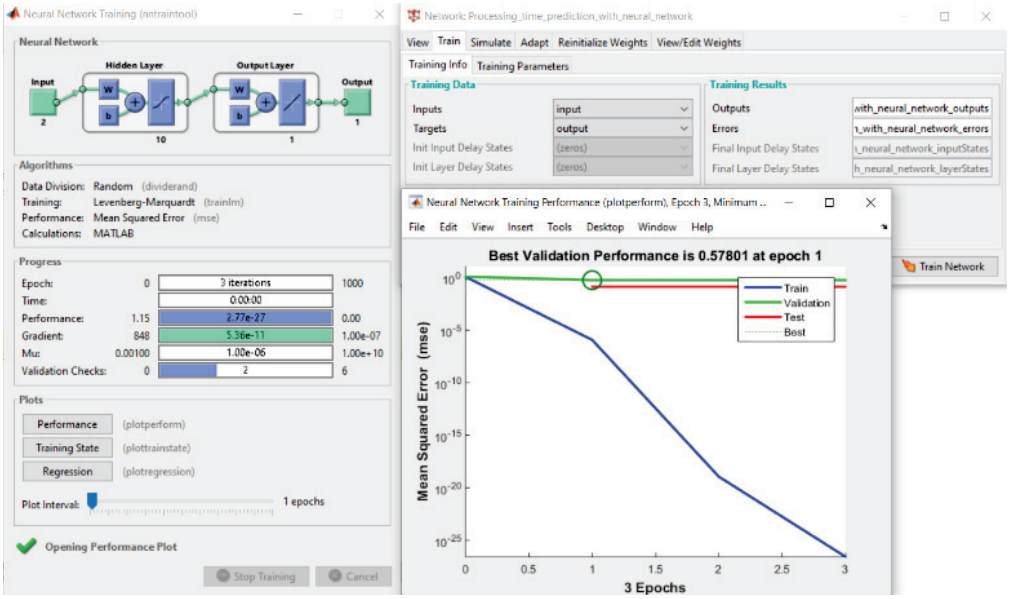


Figure A6. Training the neural network for predicting working time.

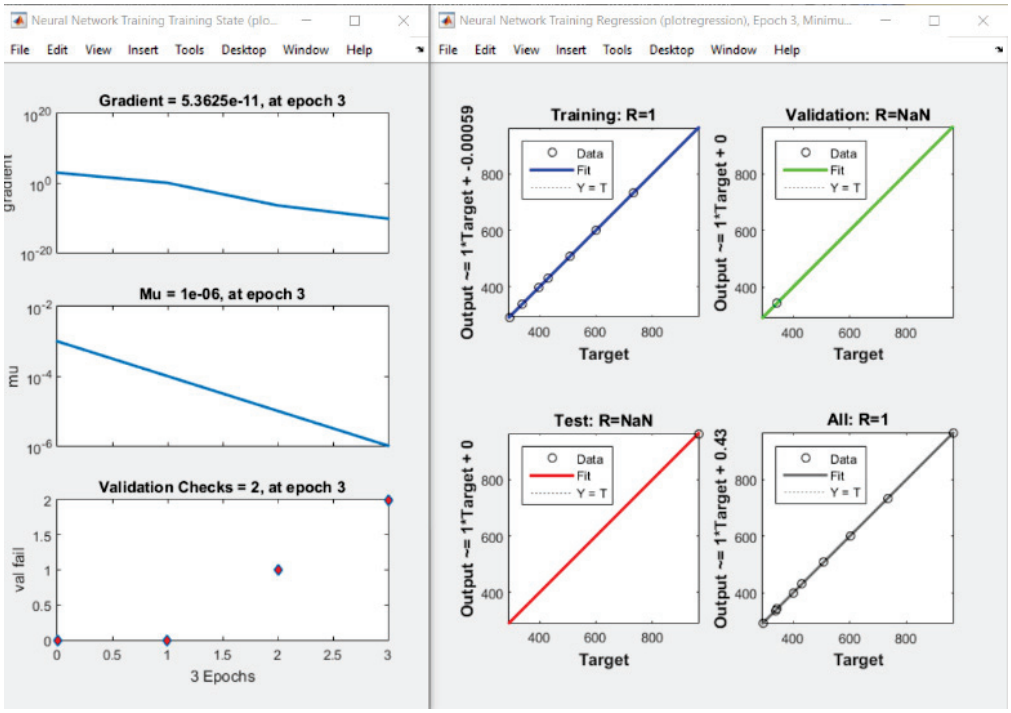


Figure A7. Teaching the neural network to predict working time.

Appendix B

Processing Time Simulations in CAM Software

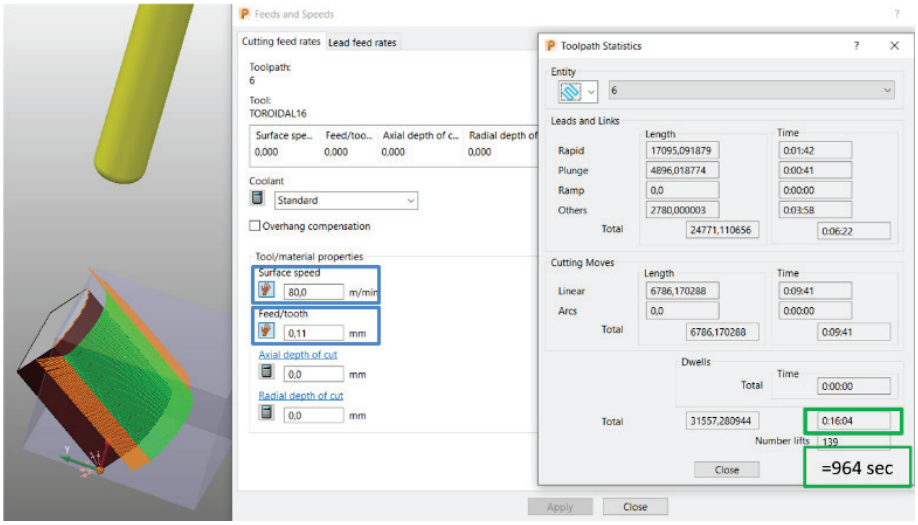


Figure A8. Machining time after CAM simulation of surface CICV-TR-1.

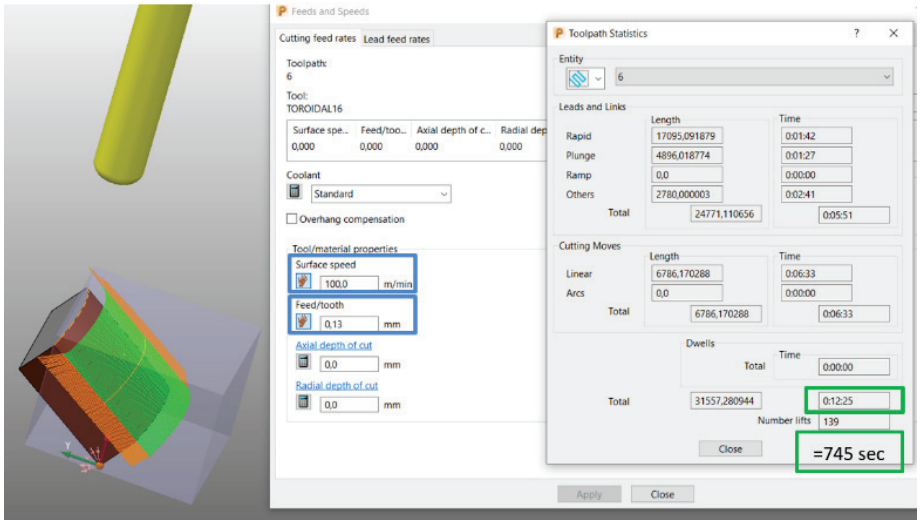


Figure A9. Machining time after CAM simulation of surface CICV-TR-2.

Table A1. New input values and machining times after CAM simulation.

Experiment	Cutting Speed [m/min]	Feed Per Tooth [mm/tooth]	CAM Simulation Time [s]
CICV-TR-1	80	0.11	964
CICV-TR-2	100	0.13	745
CICV-TR-3	120	0.15	523
CICV-TR-4	140	0.17	407
CICV-TR-5	160	0.19	345
CICV-TR-6	180	0.17	481
CICV-TR-7	200	0.15	393
CICV-TR-8	220	0.13	367
CICV-TR-9	240	0.11	381

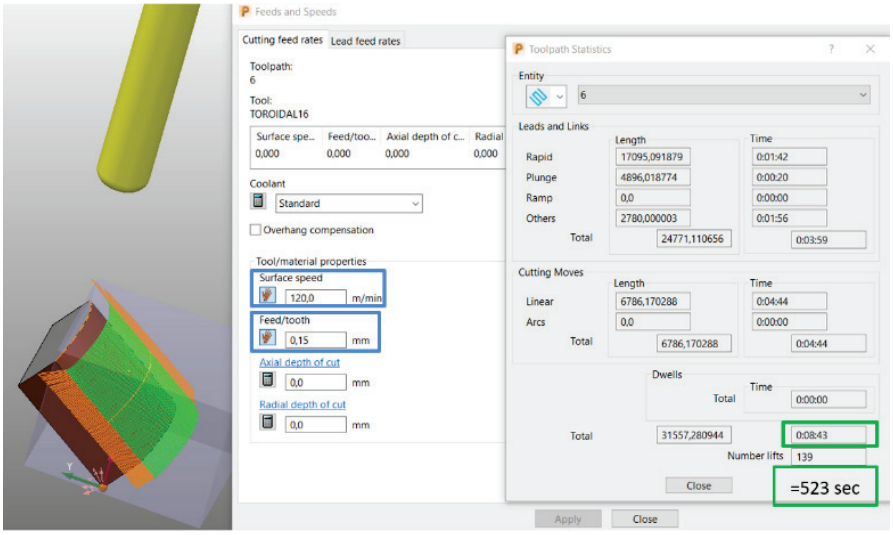


Figure A10. Machining time after CAM simulation of surface CICV-TR-3.

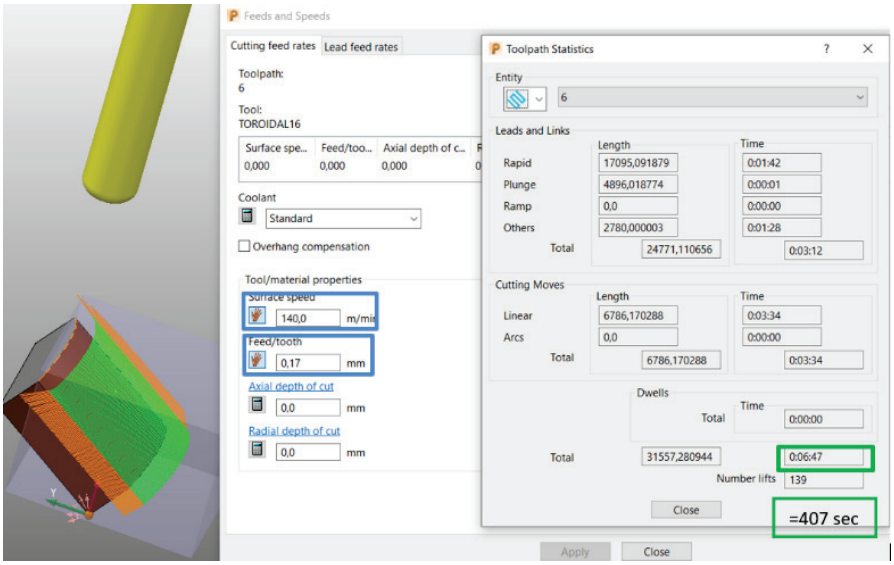


Figure A11. Machining time after CAM simulation of surface CICV-TR-4.

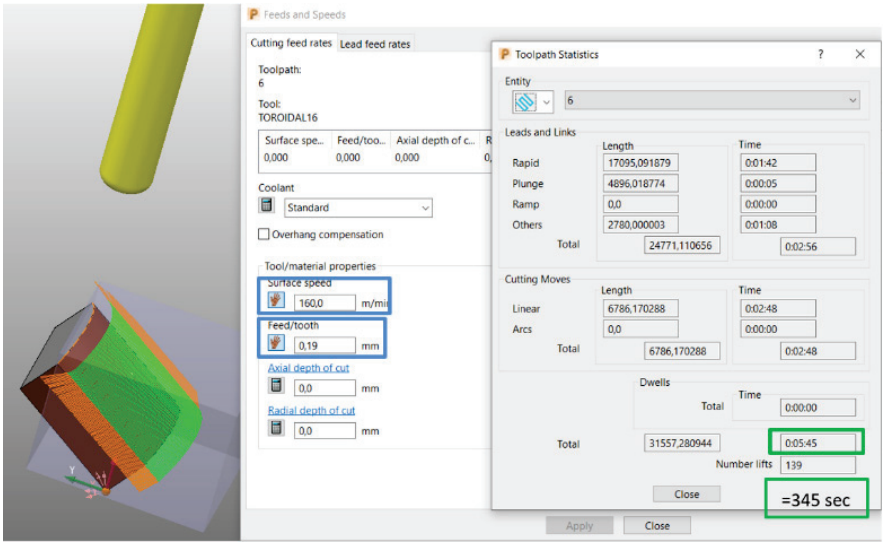


Figure A12. Machining time after CAM simulation of surface CICV-TR-5.

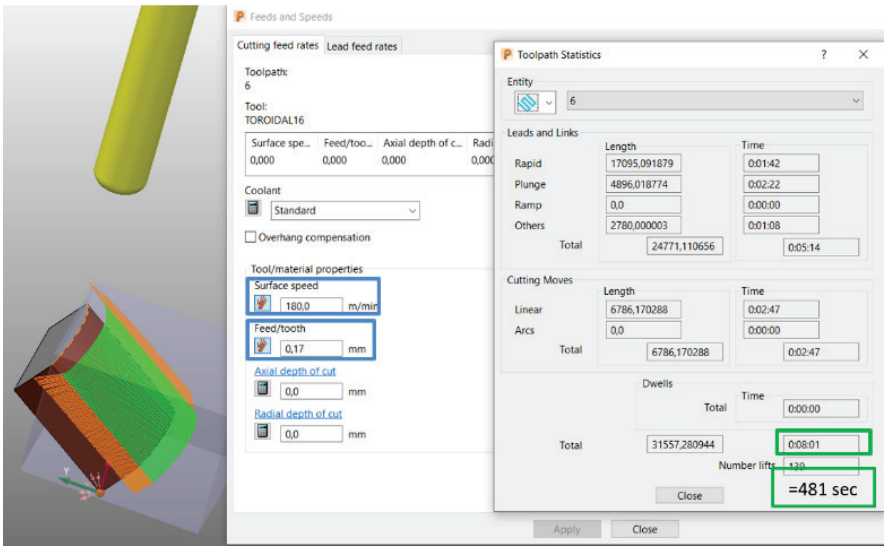


Figure A13. Machining time after CAM simulation of surface CICV-TR-6.

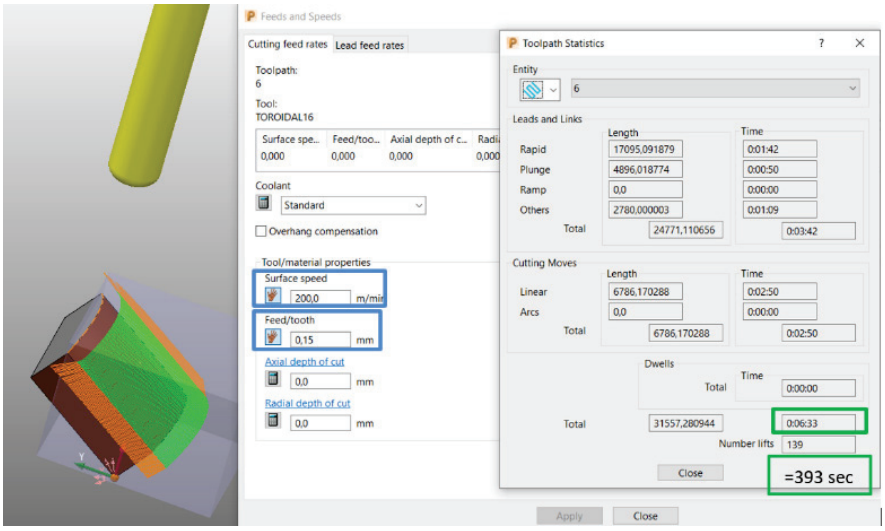


Figure A14. Machining time after CAM simulation of surface CICV-TR-7.

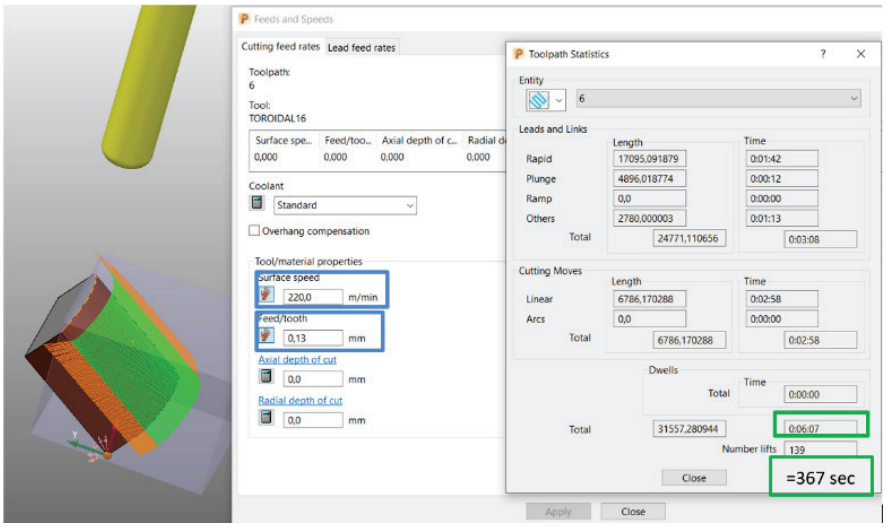


Figure A15. Machining time after CAM simulation of surface CICV-TR-8.

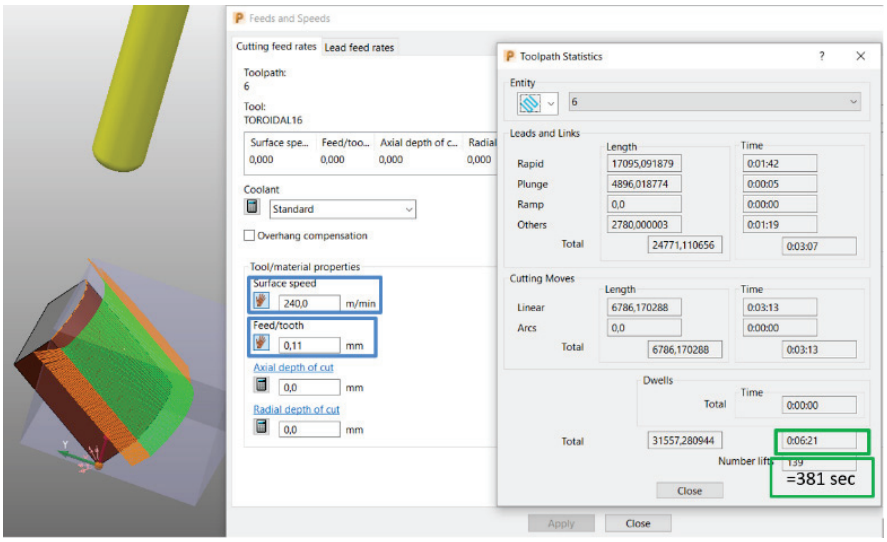


Figure A16. Machining time after CAM simulation of surface CICV-TR-9.

References

1. Pop-Suărășan, A.; Ungureanu, N. The implication imposed by prescriptive maintanance implementation into the industry 4.0. A Current state analysis. *Acta Teh. Napocesis Ser. Appled Math. Eng.* **2023**, *66*, 157–164.
2. Oșan, A.; Bănică, M.; Năsui, V. Processing of plane surfaces with toroidal milling versus ball nose end mill. In *Acta Technica Napocensis*; Series Applied Mathematics, Mechanics, and Engineering; Technical University of Cluj-Napoca: Cluj-Napoca, Romania, 2020.
3. Tansel, I.; Ozcelik, B.; Bao, W.; Chen, P.; Yenilmez, A. Selection of optimal cutting conditions by using GONNS. *Int. J. Mach. Tools Manuf.* **2006**, *46*, 26–35. [CrossRef]
4. De Filippis, L.; Serio, L.M.; Facchini, F.; Mummolo, F.F.A.G. ANN Modelling to Optimize Manufacturing Process. In *Advanced Applications for Artificial Neural Networks*; El-Shahat, A., Ed.; Intechopen: London, UK, 2018.
5. Mundada, V.; Narala, S.K.R. Optimization of Milling Operations Using Artificial Neural Networks (ANN) and Simulated Annealing Algorithm (SAA). *Mater. Today Proc.* **2018**, *5*, 4971–4985. [CrossRef]
6. Chen, Y.; Sun, R.; Gao, Y.; Leopold, J.A. ANN prediction model for surface roughness considering the effects of cutting forces and tool vibrations. *Measurement* **2017**, *98*, 25–34. [CrossRef]
7. Hesser, D.F.; Markert, B. Tool wear monitoring of a retrofitted CNC milling machine using artificial neural networks. *Manuf. Lett.* **2019**, *19*, 1–4. [CrossRef]
8. El-Mounayri, H.; Briceno, J.F.; Gadallah, M. A new artificial neural network approach to modeling ball-end milling. *Int. J. Adv. Manuf. Technol.* **2010**, *47*, 527–534. [CrossRef]
9. Al-Abdullah, K.I.A.-L.; Abdi, H.; Lim, C.P.; Yassin, W.A. Force and temperature modelling of bone milling using artificial neural networks. *Measurement* **2018**, *116*, 25–37. [CrossRef]
10. Kunto, M.; Aslan, A. A Review of Indirect Tool Condition Monitoring Systems and Decision-Making Methods in Turning: Critical Analysis and Trends. *Sensors* **2021**, *21*, 108.
11. Pimenov, D.Y. Artificial intelligence systems for tool condition monitoring in machining: Analysis and critical review. *J. Intell. Manuf.* **2023**, *34*, 2079–2121. [CrossRef]
12. Boujelbene, M.; Moisan, A.; Tounsi, N.; Brenier, B. Productivity enhancement in dies and molds manufacturing by the use of C1 continuous tool path. *Int. J. Mach. Tools Manuf.* **2004**, *44*, 101–107. [CrossRef]
13. Altintas, Y. *Manufacturing Automation Metal Cutting Mechanics. Machine Tool Vibrations and CNC Design*; Cambridge University: New York, NY, USA, 2007.
14. Yan, X.; Shirase, K.; Hirao, M.; Yasui, T. NC program evaluator for higher machining productivity. *Int. J. Mach. Tools Manuf.* **1999**, *39*, 1563–1573. [CrossRef]
15. Kataraki, P.S.; Abu Mansor, M.S. Automatic designation of feature faces to recognize interacting and compound volumetric features for prismatic parts. *Eng. Comput.* **2020**, *36*, 1499–1515. [CrossRef]
16. Ning, F.; Shi, Y.; Cai, M.; Xu, W.; Zhang, X. Manufacturing cost estimation based on the machining process and deep-learning method. *J. Manuf. Syst.* **2020**, *56*, 11–22. [CrossRef]
17. Atia, M.R.; Khalil, J.; Mokhtar, M. A Cost estimation model for machining operations.; an ann parametric approach. *J. Al-Azhar Univ. Eng. Sect.* **2017**, *12*, 878–885. [CrossRef]

18. Florjanič, B.; Kuzman, K. Estimation of time for manufacturing of injection moulds using artificial neural networks-based model. *Polim. Časopis Plast. Gumu.* **2012**, *33*, 12–21.
19. Yoo, S.; Kang, N. Explainable artificial intelligence for manufacturing cost estimation and machining feature visualization. *Expert Syst. Appl.* **2021**, *183*, 115430. [CrossRef]
20. Tsai, M.-H.; Lee, J.-N.; Tsai, H.-D.; Shie, M.-J.; Hsu, T.-L.; Chen, H.-S. Applying a Neural Network to Predict Surface Roughness and Machining Accuracy in the Milling of SUS304. *Electronics* **2023**, *12*, 981. [CrossRef]
21. Djurović, S.; Lazarević, D.; Ćirković, B.; Mišić, M.; Ivković, M.; Stojčetović, B.; Petković, M.; Ašonja, A. Modeling and Prediction of Surface Roughness in Hybrid Manufacturing–Milling after FDM Using Artificial Neural Networks. *Appl. Sci.* **2024**, *14*, 5980. [CrossRef]
22. Qu, D.; Liang, W.; Zhang, Y.; Gu, C.; Zhan, Y. Research on Machining Quality Prediction Method Based on Machining Error Transfer Network and Grey Neural Network. *J. Manuf. Mater. Process.* **2024**, *8*, 203. [CrossRef]
23. Saric, T.; Simunovic, G.; Simunovic, K. Estimation of Machining Time for CNC Manufacturing Using Neural Computing. *Int. J. Simul. Model.* **2016**, *15*, 663–675. [CrossRef]
24. Sable, Y.S.; Dharmadhikari, H.M.; More, S.A.; Sarris, I.E. Exploring Artificial Neural Network Techniques for Modeling Surface Roughness in Wire Electrical Discharge Machining of Aluminum/Silicon Carbide Composites. *J. Compos. Sci.* **2025**, *9*, 259. [CrossRef]
25. Mane, S.; Patil, R.B.; Al-Dahidi, S. Predictive Modeling of Surface Roughness and Cutting Temperature Using Response Surface Methodology and Artificial Neural Network in Hard Turning of AISI 52100 Steel with Minimal Cutting Fluid Application. *Machines* **2025**, *13*, 266. [CrossRef]
26. Abaza, B.F.; Gheorghita, V. Artificial Neural Network Framework for Hybrid Control and Monitoring in Turning Operations. *Appl. Sci.* **2025**, *15*, 3499. [CrossRef]
27. Turan, İ.; Özlü, B.; Ulaş, H.B.; Demir, H. Prediction and Modelling with Taguchi, ANN and ANFIS of Optimum Machining Parameters in Drilling of Al 6082-T6 Alloy. *J. Manuf. Mater. Process.* **2025**, *9*, 92. [CrossRef]
28. Reiff, C.; Buser, M.; Betten, T.; Onuseit, V.; Hoßfeld, M.; Wehner, D.; Riedel, O. A Process-Planning Framework for Sustainable Manufacturing. *Energies* **2021**, *14*, 5811. [CrossRef]
29. Kotha Amarnath, S.; Inturi, V.; Rajasekharan, S.G.; Priyadarshini, A. Combining Sensor Fusion and a Machine Learning Framework for Accurate Tool Wear Prediction During Machining. *Machines* **2025**, *13*, 132. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Type-3 Fuzzy Dynamic Adaptation of the Crossover Parameter in Differential Evolution for the Optimal Design of Type-3 Fuzzy Controllers for the Inverted Pendulum

Patricia Ochoa [1], Cinthia Peraza [1], Patricia Melin [1], Oscar Castillo [1] and Zong Woo Geem [2,*]

[1] Tijuana Institute of Technology, TecNM, Tijuana 22379, Mexico; martha.ochoa18@tectijuana.edu.mx (P.O.); cinthia.peraza@tectijuana.edu.mx (C.P.); pmelin@tectijuana.mx (P.M.); ocastillo@tectijuana.mx (O.C.)
[2] College of IT Convergence, Gachon University, Seongnam 13120, Republic of Korea
* Correspondence: geem@gachon.ac.kr

Abstract: This paper proposes a dynamic adaptation of the crossover parameter (CR) in Differential Evolution (DE) using Type-3 fuzzy logic. The strategy involves adjusting the CR value according to the progress of the algorithm by employing a Type-3 fuzzy system (T3FS) to handle the uncertainty and variability inherent in this parameter. To assess the effectiveness of the proposed approach, an inverted pendulum system controlled by a T3FS is employed. The controller parameters are optimized using an enhanced DE algorithm. The results demonstrate that dynamically adapting the crossover rate parameter significantly enhances both the efficiency and accuracy of the DE algorithm, compared to traditional static adjustments. Specifically, the proposed approach achieved an RMSE of 0.0127, outperforming traditional static adjustment methods, which resulted in an RMSE of 0.732. In addition, the results exhibit an improvement in the consistency of the outcomes, with a significantly lower standard deviation (0.0111) compared to conventional methods. These findings underscore the potential of Type-3 fuzzy systems (T3FSs) in enhancing evolutionary algorithms for optimizing fuzzy controllers in highly uncertain nonlinear systems. Furthermore, this emphasizes the capability of T3FSs to enhance evolutionary algorithms in optimizing fuzzy controllers for nonlinear systems subject to significant uncertainty.

Keywords: Type-3 fuzzy differential evolution; Type-3 fuzzy logic; Type-3 fuzzy controller

1. Introduction

The optimization of complex and nonlinear problems is a field of great interest in automatic control and computational intelligence. In this area, the Differential Evolution (DE) algorithm has established itself as a powerful tool due to its simplicity and effectiveness in exploring and exploiting the solution space [1]. However, the performance of DE critically depends on the proper configuration of its parameters, such as the scaling factor (F), population size, and particularly the crossover parameter (CR), which controls the exchange of information between candidate solutions. An inappropriate selection of these parameters can lead to premature convergence or a limited exploration of the search space.

To address this limitation, various adaptive strategies have been proposed to dynamically adjust the parameters during the evolutionary process. Among these strategies, the use of fuzzy logic-based systems has proven promising, especially for managing uncertainty and the nonlinear nature of problems [2–5]. Type-1 fuzzy logic allows modeling uncertainties through traditional membership functions [6], in control [7–22], in real-time

prediction [23], and in hybrid fuzzy-expert system control [24,25]. Type-2 fuzzy logic extends this capability by handling uncertainties within the membership functions themselves [26]. Some of the most recent applications with Type-2 fuzzy logic are mentioned below: energy-efficient routing [27], hydropower system management [28], and tracking for photovoltaic systems [29–31] and control systems [32,33]. Recently, Type-3 fuzzy logic has emerged as an innovative approach that incorporates an additional level of flexibility to handle more complex uncertainties through second-order interval fuzzy sets, enhancing representation and adjustment capabilities [34]. Some recent applications with Type-3 fuzzy logic (T3FL) have been in areas such as deep learning [35], diagnosis of mental disorders [36], control systems [37,38], and clustering [39].

In this context, this work proposes a strategy for the dynamic adaptation of CR in DE using a Type-3 fuzzy system (T3FS). The strategy consists of adjusting the CR value based on the algorithm's progress, allowing for a more precise adaptation to the problem characteristics and the evolutionary process. To evaluate the effectiveness of the proposed approach, an inverted pendulum plant is used as a classic nonlinear system in automatic control, which is controlled through an interval Type-3 Sugeno fuzzy system. This plant offers significant challenges due to its inherent instability and the need for fast and precise responses.

Despite advances in fuzzy control, significant limitations remain in the ability of traditional approaches to handle complex and dynamic uncertainties in nonlinear systems. Controllers based on Type-1 and Type-2 fuzzy logic exhibit constraints in terms of precision and robustness under scenarios with high levels of noise or uncertainty. This research gap highlights the need for more adaptive and efficient strategies, such as those offered by Type-3 fuzzy logic. In this context, the main contributions of our work are as follows:

- Proposal of a Type-3 fuzzy logic framework: We introduce an innovative approach that enables dynamic adaptation of the CR parameter, enhancing the system's response to complex uncertainties.
- Evaluation in a classical nonlinear system: We validate the performance of the approach in a well-known system, the inverted pendulum, demonstrating its effectiveness and robustness.
- Comprehensive comparison: We conduct a comparative analysis of the proposed method's performance against controllers based on Type-1 and Type-2 fuzzy logic, considering both noisy and noise-free scenarios.

The structure of this paper is as follows: Section 2 offers an overview of interval Type-3 fuzzy sets and their theoretical foundation. Section 3 describes the Differential Evolution algorithm, highlighting its relevance for optimization problems. Section 4 outlines the proposed methodology, integrating an IT3FS with DE to design a novel control strategy. Section 5 outlines the experimental setup and discusses the results obtained from various simulations. Section 6 offers a discussion of results. Finally, Section 7 summarizes the key findings.

2. Interval Type-3 Fuzzy Sets

Interval Type-3 fuzzy logic (IT3FL) is an extension of Type-2 fuzzy logic that provides an additional level of flexibility for modeling complex uncertainties. In general terms, fuzzy logic allows representing imprecise or uncertain information through fuzzy sets.

Evolution of Fuzzy Logic

Type-1: This uses well-defined membership functions to model uncertainty.

Type-2: This introduces second-order fuzzy sets, where membership functions can also be uncertain (defined by a range of values instead of a single value).

Type-3: This further extends the concept by incorporating third-order fuzzy sets, enabling the handling of more complex uncertainties in knowledge representation and decision-making.

The main characteristics of T3FL are as follows:

- Greater capacity for representing uncertainty, as it manages not only variability in data but also in rules and membership functions;
- Increased flexibility and robustness in decision-making compared to Type-1 and Type-2 approaches;
- Applications in complex problems, such as adaptive control, artificial intelligence, and optimization in highly uncertain environments.

The mathematical structure of fuzzy logic is presented in detail in various existing works in the literature [35]; for this work, some important mathematical aspects of Type-3 fuzzy logic are presented: a set $A^{(3)}$ is represented by the plot of a trivariate function, given by the membership function (MF) in the Cartesian product, where the universe is defined in the primary variable x. The MF is expressed by a function, and it is called a Type-3 membership function (T3MF).

$$\mu_{A^{(3)}} : X \times [0,1] \times [0,1] \rightarrow [0,1] \tag{1}$$

$$A^{(3)} = \{x \in X, u \in U \subseteq [0,1], v \in V \subseteq [0,1]\} \tag{2}$$

Here, U represents the universe for the secondary variable u, and V denotes the universe for the tertiary variable v. A Type-3 fuzzy set $A^{(3)}$ can also be defined as follows:

$$A^{(3)} = \int_{x \in X} \int_{u \in [0,1]} \int_{v \in [0,1]} \mu_{A^{(3)}}(x,u,v)/(x,u,v) \tag{3}$$

$$A^{(3)} = \int_{x \in X} \left[\int_{u \in [0,1]} \left[\int_{v \in [0,1]} \mu_{A^{(3)}}(x,u,v)/v \right] /u \right] /x \tag{4}$$

Figure 1 illustrates a graphical representation of a membership function for a T3FS, with its corresponding mathematical expression provided as follows:

$$A = \int_{x \in X} \int_{u \in [0,1]} \widetilde{\mu}_A(x,u)/(x,u) = \int_{x \in X} \left[\int_{u \in [0,1]} \widetilde{f}_x(u)/u \right] /x \tag{5}$$

where $\widetilde{\mu}_A(x,u) \in \left[\underline{\mu}_A(x,u) \underline{\mu}_A(x,u) \right]$.

$$A = \int_{x \in X} \mu_{A(x)}(u)/x = \int_{x \in X} \left[\int_{u \in [0,1]} \widetilde{f}_x(u)/u \right] /x \tag{6}$$

where $\mu_{A(x)}(u)$ is a T2FS.

$$\mu_{A(x)}(u) = \int_{u \in J_x} \widetilde{f}_x(u)/u = \int_{u \in [\underline{f}_x(u), \underline{f}_x(u)} 1/u \tag{7}$$

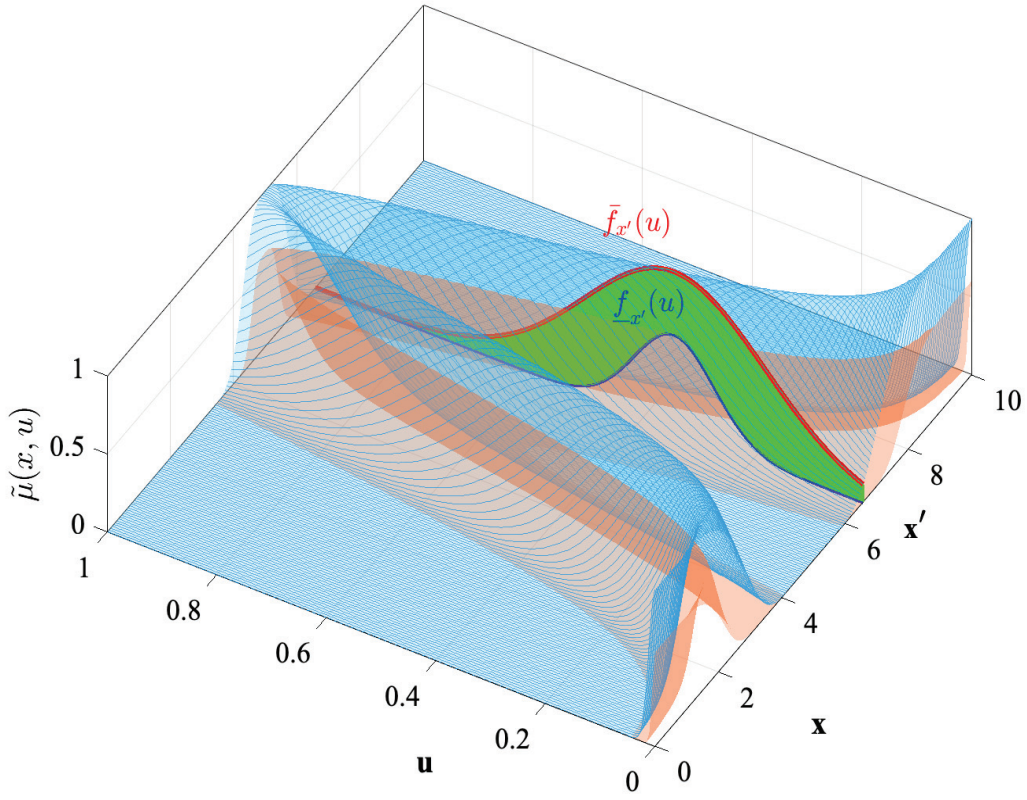


Figure 1. Graphical view of a T3FS membership function with a mathematical expression.

3. Differential Evolution

The DE algorithm is a population-based method that belongs to the family of evolutionary algorithms. Its goal is to find the best solution to a problem by minimizing (or maximizing) an objective function through operations inspired by biological evolution, such as mutation, recombination, and selection [1].

3.1. Initialization

An initial population of N individuals (solution vectors) of dimension D is generated, where each individual is represented as

$$x_{j,i,0} = rand_j(0,1)\cdot\left(b_{j,U} - b_{j,L}\right) + b_{j,L} \tag{8}$$

3.2. Mutation

For each individual x_i, a mutant vector v_i is generated from three randomly selected individuals, x_{r0}, x_{r1}, and x_{r2} (different from x_i):

$$v_{i,g} = x_{r_0,g} + \mathrm{F}\cdot\left(x_{r_1,g} - x_{r_2,g}\right) \tag{9}$$

where F is a scaling factor (typically in the range [0, 1]) that controls the magnitude of the mutation.

3.3. Crossover

The mutant vector v_i is crossed with the individual x_i to produce an offspring vector u_i, using a crossover parameter CR (in the range [0, 1]):

$$u_{i,g} = u_{j,i,g}\begin{cases} v_{j,i,g}\ if\ \left(rand_j(0,1) \le Cr\ or\ j = j_{rand}\right) \\ x_{j,i,g\ otherwise} \end{cases} \tag{10}$$

3.4. Selection

The new vector u_i is evaluated using the objective function $f(u_i)$. If the offspring has a better value than the parent, it replaces the individual in the next generation:

$$x_{i,g+1} = \begin{cases} u_{i,g} \; if \; f(u_{i,g}) \leq f(x_{i,g}) \\ x_{i,g} \; otherwise \end{cases} \tag{11}$$

The Differential Evolution algorithm offers several advantages, including its simplicity, ease of implementation, and the need for only a few parameters to tune. It is particularly efficient for continuous optimization problems as it does not require derivative information, making it suitable for a wide range of applications. DE is widely used in function optimization, automatic control, neural network training, and parameter tuning in dynamic systems, among other fields where robust and adaptive optimization techniques are needed.

4. Proposed Methodology

The inverted pendulum controlled with a Sugeno-type fuzzy system is a classic approach in automatic control that employs fuzzy logic to stabilize an inherently unstable system. In some related works, a Takagi–Sugeno model has been used [40–45]. In this method, the Sugeno fuzzy controller receives system variables as inputs, such as the pendulum's angle and angular velocity, and generates a control output that adjusts the force applied to the cart to maintain the pendulum's balance. The Sugeno Type-1 fuzzy system has proven effective in controlling the inverted pendulum since it can generate a fast and stable response without requiring an exact model of the system. However, its performance can be affected by parameter uncertainties or variations in system conditions. Figure 2 provides a visual representation of the controller, while Figure 3 illustrates the structure of the original Type-1 fuzzy system. As shown, the system has four inputs, follows a Sugeno-type approach, and includes a single output corresponding to the balance.

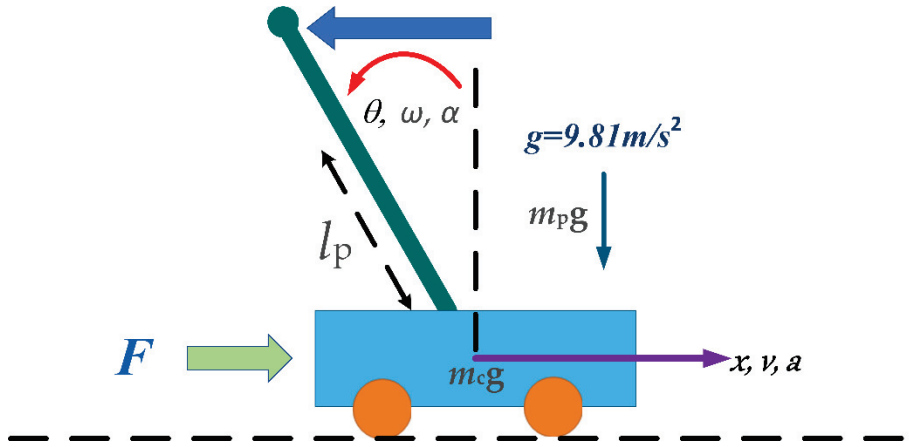


Figure 2. Inverted pendulum controller.

The goal of this problem is to maintain the pendulum's balance without it falling from the cart. The system consists of a straight-line rail, a cart, a pendulum, and a driving unit. The cart is free to move left or right along the rail, while the pendulum is attached at the center of the cart's top surface and can rotate around its pivot within the same vertical plane as the rail. Assuming the system is frictionless, Equations (12) and (13) describe the mathematical model of the inverted pendulum.

$$\propto = (m_c + m_p)g \; sig\theta - \left\{F + m_p l_p \omega^2 sin\theta\right\} cos\theta \left\{4/3(m_c + m_p) - m_p(cos\theta)^2\right\} l_p \tag{12}$$

$$a = 4/3\left\{F + m_p l_p \omega^2 sin\theta\right\} - m_p g\ sin\theta\ cos\theta\left\{4/3(m_c + m_p) - m_p(cos\theta)^2\right\} \quad (13)$$

The parameters m_c and m_p correspond to the mass of the cart and the pendulum, respectively, measured in kilograms [kg]. The gravitational acceleration is given as $g = 9.8\ \mathrm{m/s^2}$ The parameter l_p represents the distance from the center of the pendulum to the pivot, measured in meters [m], and is equal to half the length of the pendulum. The variable F denotes the driving force applied horizontally to the cart, measured in newtons [N]. The variables θ, ω, and α indicate, respectively, the pendulum's angle from the upright position, its angular velocity, and its angular acceleration, where the clockwise direction is considered positive. Similarly, the variables x, v, and a represent the cart's position relative to the rail's origin, its velocity, and its acceleration, with the rightward direction taken as positive. Figure 4 presents the rule set for this controller, which models the vehicle's behavior on the cart to keep it balanced.

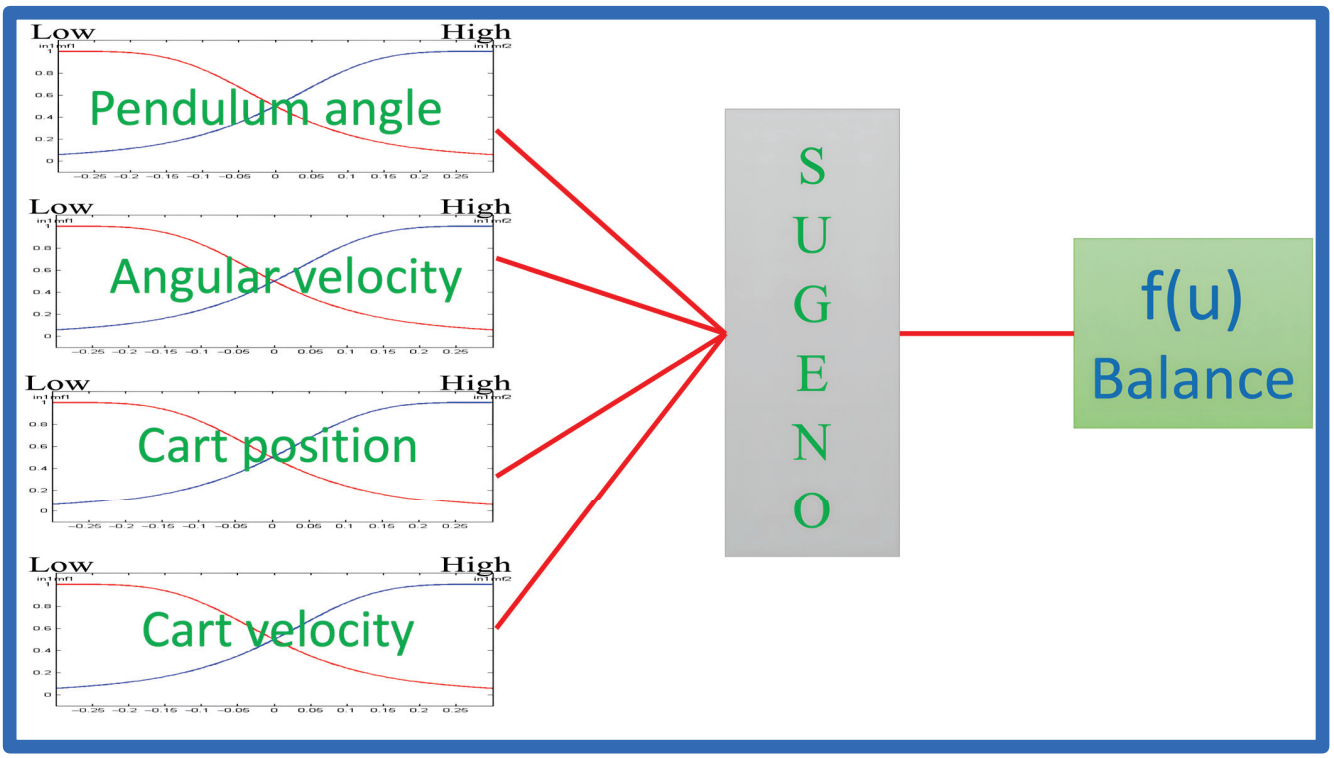


Figure 3. Structure of the original Type-1 fuzzy system with a Sugeno approach.

1. If (in1 is in1mf1) and (in2 is in2mf1) and (in3 is in3mf1) and (in4 is in4mf1) then (out is outmf1)
2. If (in1 is in1mf1) and (in2 is in2mf1) and (in3 is in3mf1) and (in4 is in4mf2) then (out is outmf2)
3. If (in1 is in1mf1) and (in2 is in2mf1) and (in3 is in3mf2) and (in4 is in4mf1) then (out is outmf3)
4. If (in1 is in1mf1) and (in2 is in2mf1) and (in3 is in3mf2) and (in4 is in4mf2) then (out is outmf4)
5. If (in1 is in1mf1) and (in2 is in2mf2) and (in3 is in3mf1) and (in4 is in4mf1) then (out is outmf5)
6. If (in1 is in1mf1) and (in2 is in2mf2) and (in3 is in3mf1) and (in4 is in4mf2) then (out is outmf6)
7. If (in1 is in1mf1) and (in2 is in2mf2) and (in3 is in3mf2) and (in4 is in4mf1) then (out is outmf7)
8. If (in1 is in1mf1) and (in2 is in2mf2) and (in3 is in3mf2) and (in4 is in4mf2) then (out is outmf8)
9. If (in1 is in1mf2) and (in2 is in2mf1) and (in3 is in3mf1) and (in4 is in4mf1) then (out is outmf9)
10. If (in1 is in1mf2) and (in2 is in2mf1) and (in3 is in3mf1) and (in4 is in4mf2) then (out is outmf10)
11. If (in1 is in1mf2) and (in2 is in2mf1) and (in3 is in3mf2) and (in4 is in4mf1) then (out is outmf11)
12. If (in1 is in1mf2) and (in2 is in2mf1) and (in3 is in3mf2) and (in4 is in4mf2) then (out is outmf12)
13. If (in1 is in1mf2) and (in2 is in2mf2) and (in3 is in3mf1) and (in4 is in4mf1) then (out is outmf13)
14. If (in1 is in1mf2) and (in2 is in2mf2) and (in3 is in3mf1) and (in4 is in4mf2) then (out is outmf14)
15. If (in1 is in1mf2) and (in2 is in2mf2) and (in3 is in3mf2) and (in4 is in4mf1) then (out is outmf15)
16. If (in1 is in1mf2) and (in2 is in2mf2) and (in3 is in3mf2) and (in4 is in4mf2) then (out is outmf16)

Figure 4. Rule set for the inverted pendulum controller modeling vehicle behavior.

To enhance the controller's robustness against more complex uncertainties, this work develops three Type-3 fuzzy systems, varying the Lower Scale parameter with values of 0.1, 0.4, and 0.8. Each of these configurations is represented through images that illustrate the structures of the T3FSs, allowing visualization of how the variation of the Lower Scale parameter affects uncertainty representation and, consequently, the controller's performance.

Three fuzzy systems have been developed for the control of the inverted pendulum, all sharing the same structure but differing in the configuration of the LowerScale parameter. This parameter influences the representation of uncertainty within the system, and in this study, three different values have been considered: 0.1, 0.4, and 0.8.

Figures 5–7 illustrate the membership functions of the inputs for each of these systems, showing how the variation in Lower Scale affects the dispersion and overlap of the fuzzy sets. This analysis allows for the evaluation of the impact of uncertainty on the controller's performance.

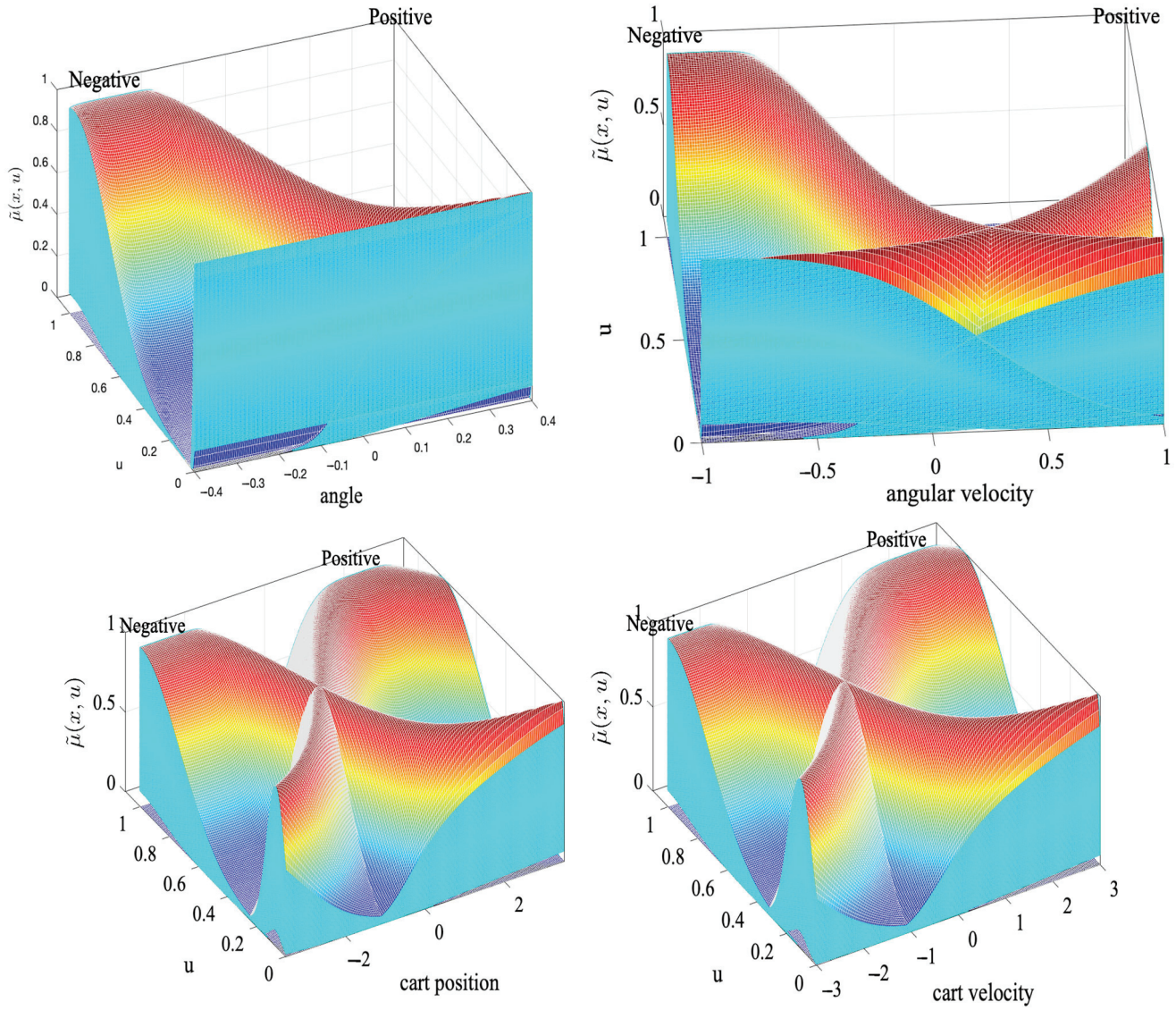


Figure 5. MFs of the inputs with Lower Scale of 0.1.

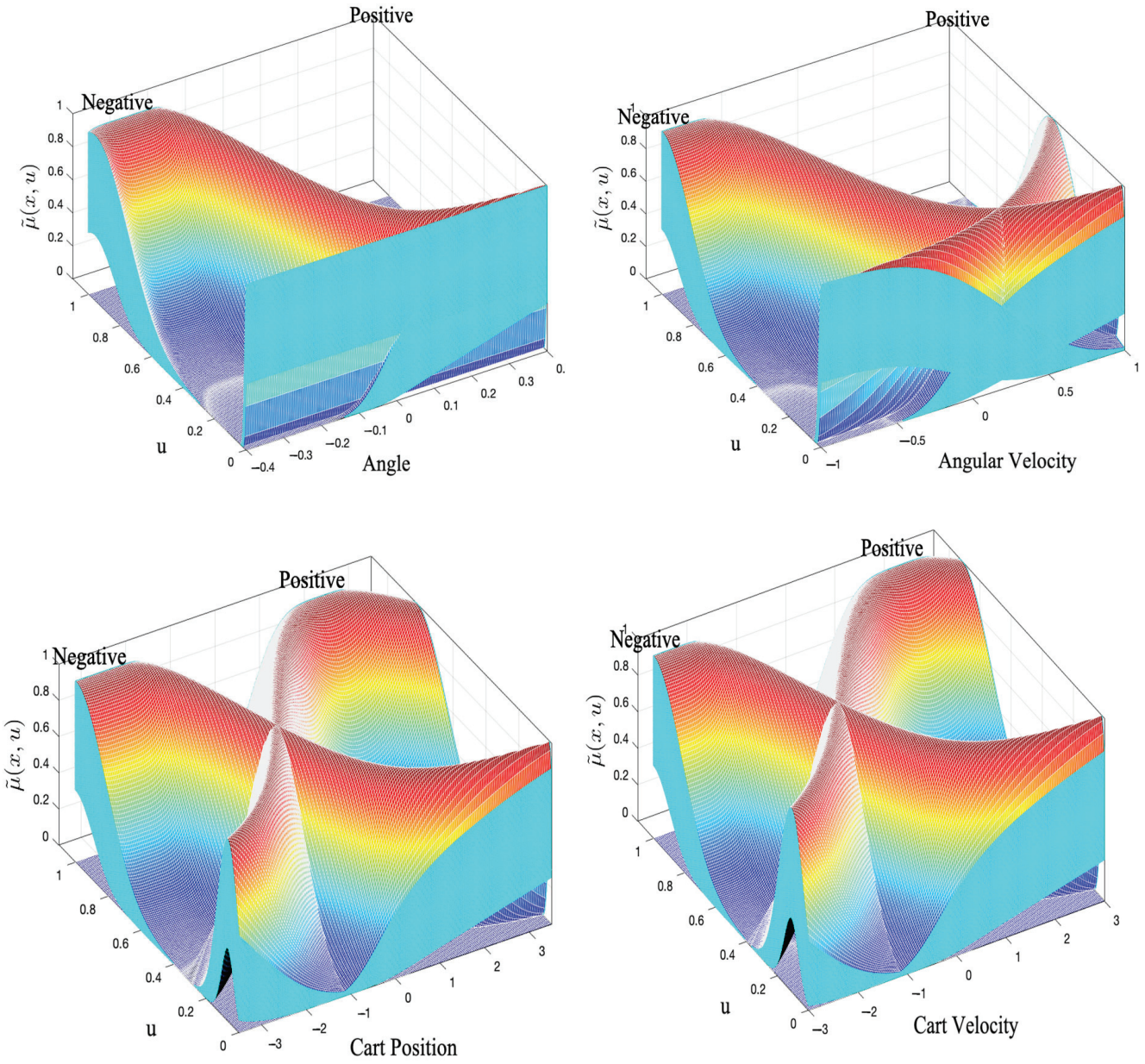


Figure 6. MFs of the inputs with Lower Scale of 0.4.

Figure 8 depicts the response surface of the fuzzy systems used for controlling the inverted pendulum. Since both the inputs and the rule set remain constant across the three evaluated systems, the shape of the surface does not change. This indicates that despite the variation in the Lower Scale parameter, the relationship between the input variables and the system's output remains the same, ensuring consistency in the controller's behavior.

In the context of enhancing the adaptability of the DE algorithm, an interval T3FS was designed to dynamically adjust the crossover rate (CR) during the optimization process. This system aims to improve the convergence behavior and performance of DE by responding to changes in the number of generations. Figure 9 illustrates the internal structure of the fuzzy system, where the input variable represents the current generation count and the output corresponds to the dynamically adjusted CR value. The idea of employing fuzzy logic to tune the DE parameters has been explored in several works [46–48], and this allows the algorithm to adapt its exploration and exploitation balance throughout the optimization.

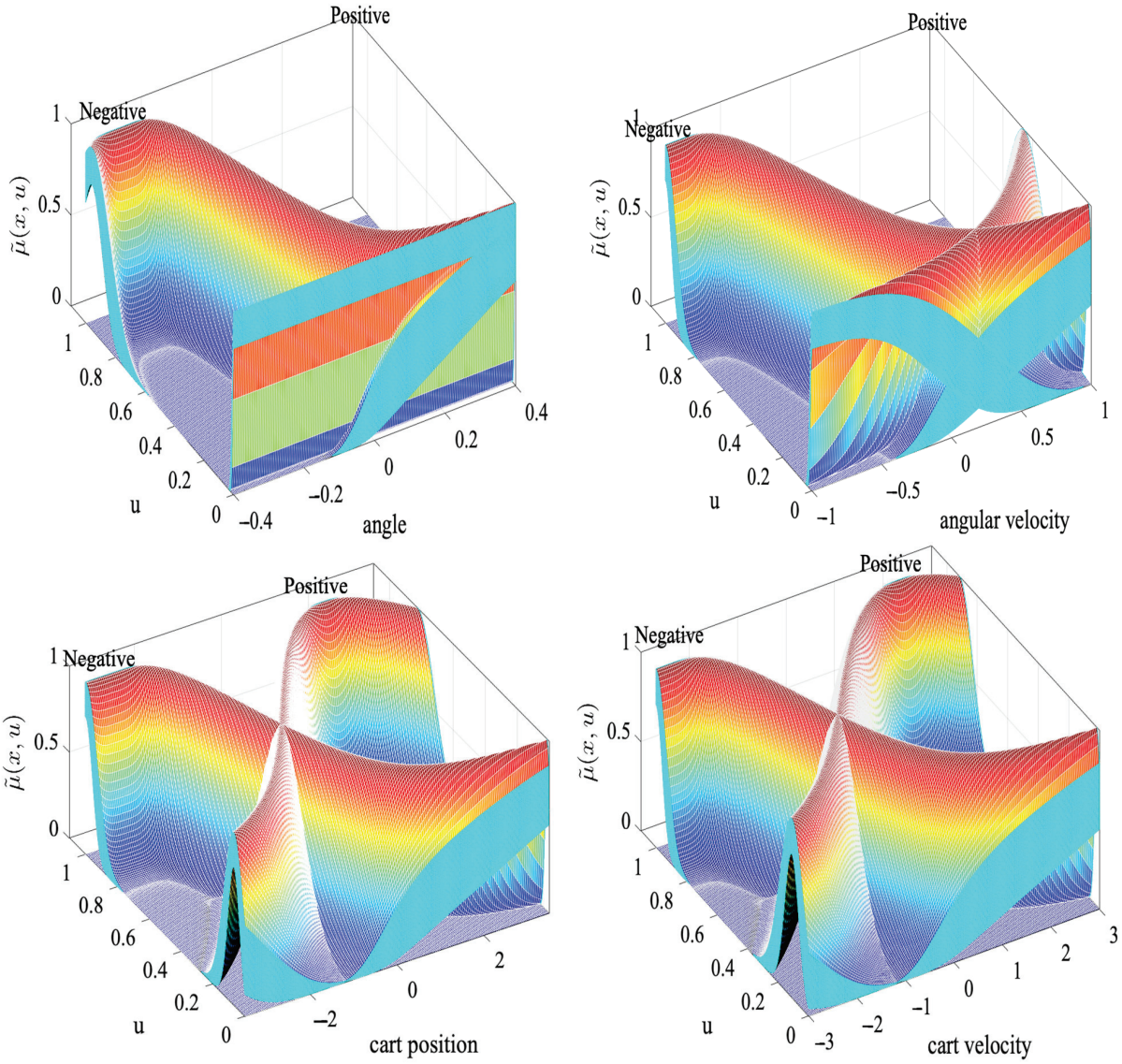


Figure 7. MFs of the inputs with Lower Scale of 0.8.

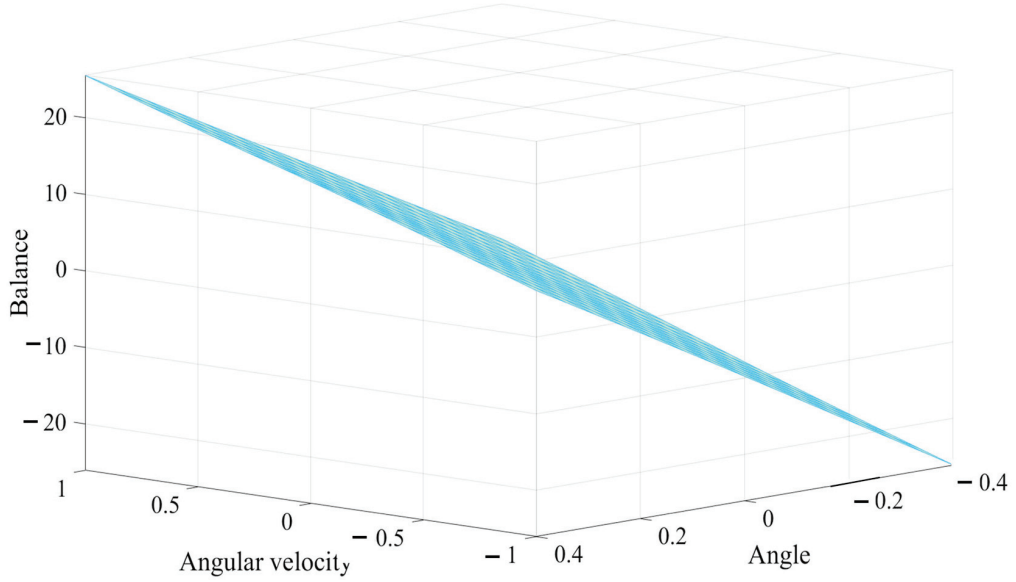


Figure 8. Response surface of the fuzzy control systems.

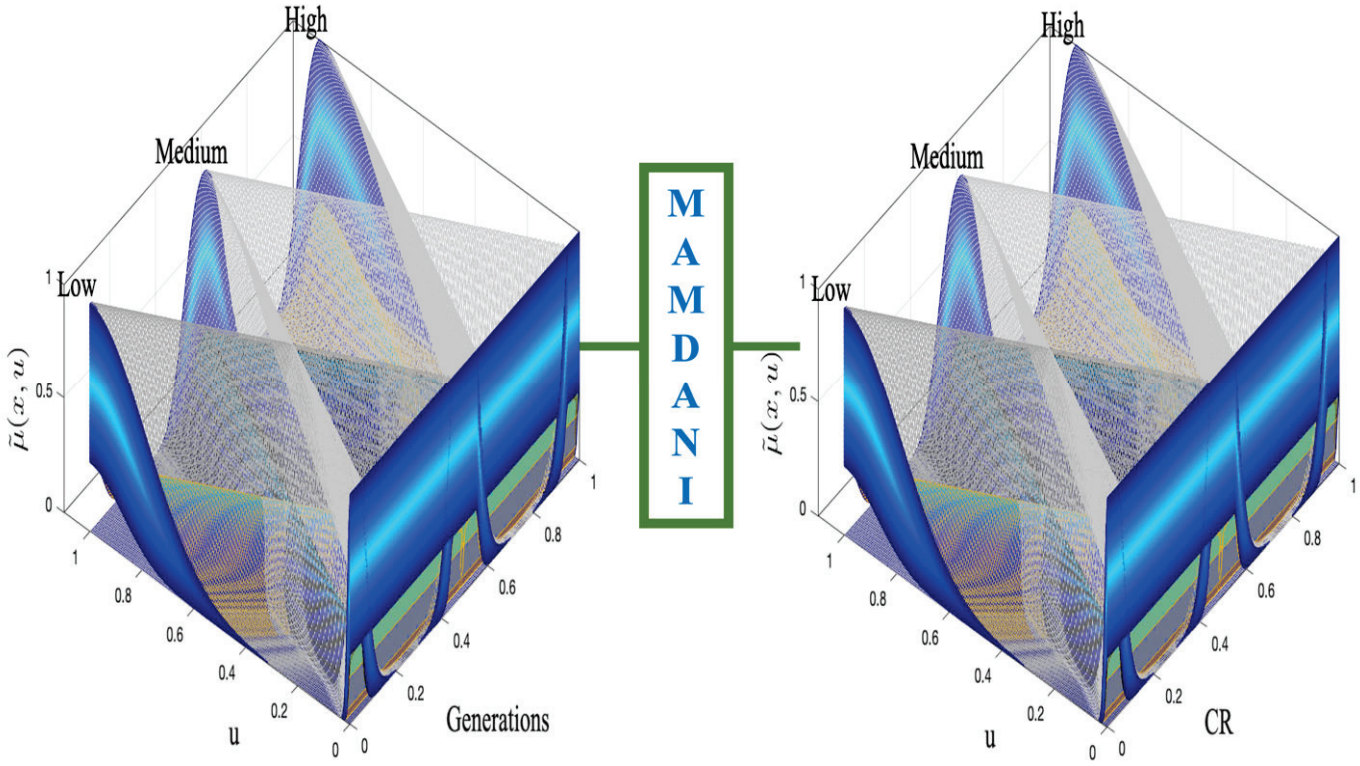


Figure 9. Interval T3FS for CR adjustment in Differential Evolution.

The fuzzy rules of the system were designed based on expert knowledge and previous research in the field. These rules are of a decreasing nature, meaning that as the number of generations increases, the value of the crossover rate (CR) is gradually reduced. This approach aims to encourage exploration in the early stages of the evolutionary process and promote exploitation in later stages.

Equation (14) shows how the CR value is computed through a weighted average of the consequent values, using the membership degrees μ_i^{CR} of each rule. The input variable (generations) is normalized to the interval [0, 1] to maintain consistency across different runs.

$$\mathrm{CR} = \frac{\sum_{i=1}^{r_{CR}} \mu_i^{CR}(\mathrm{CR}_{1i})}{\sum_{i=1}^{r_{CR}} \mu_i^{CR}} \tag{14}$$

The fuzzy rule base is composed of three rules that define the relationship between the number of generations and the crossover rate (CR). These rules are structured as follows:

Rule 1: If the generation is low, then the CR is high;

Rule 2: If the generation is medium, then the CR is medium;

Rule 3: If the generation is high, then the CR is low.

This rule configuration reflects a decreasing strategy, where the crossover rate is initially high to encourage exploration and is gradually reduced as generations increase to promote exploitation. The following figure, Figure 10, illustrates the surface generated by the previously described fuzzy system. This surface graphically represents the relationship between the input (number of generations) and the output (CR parameter value), showing how the crossover rate evolves as the algorithm progresses.

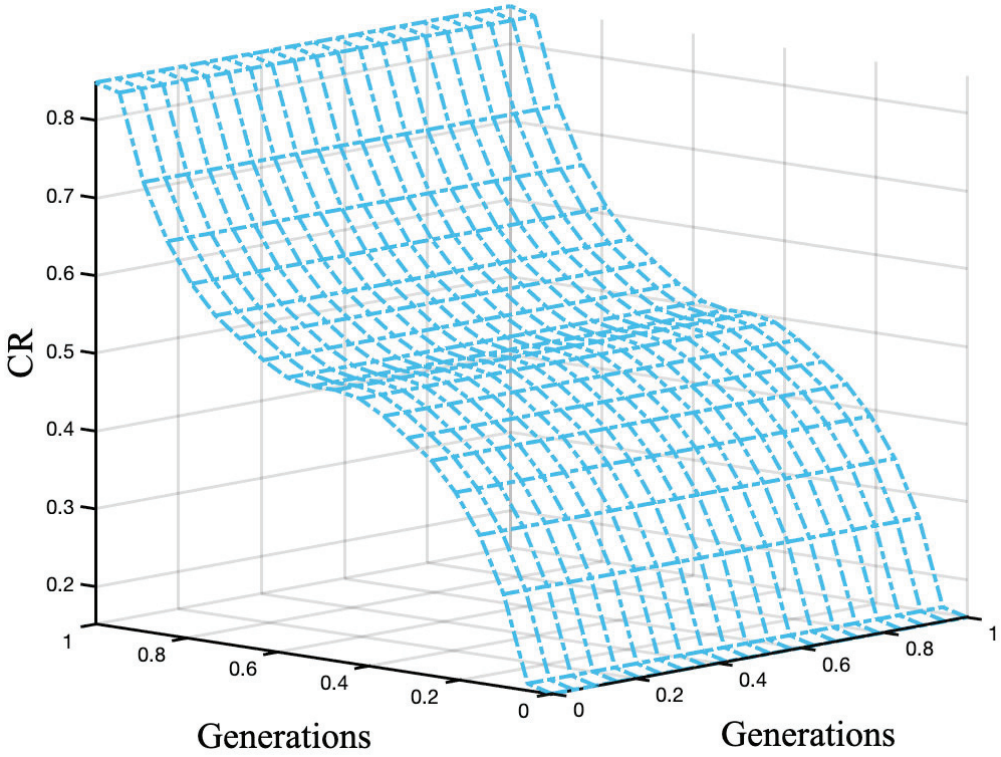


Figure 10. Surface of the fuzzy system for dynamic CR adjustment based on generations.

5. Experimental Setup and Results

For the experimentation of the fuzzy controller applied to the inverted pendulum system, the following metrics were used along with their respective performance equations to evaluate the system's performance:

$$\text{ITAE} = \sum_{t=1}^{N} t(x(t) - \hat{x}(t)) \tag{15}$$

$$\text{ITSE} = \sum_{t=1}^{N} t\left(x(t) - \hat{x}(t)^2\right) \tag{16}$$

$$\text{IAE} = \sum_{t=1}^{N} |x(t) - \hat{x}(t)| \tag{17}$$

$$\text{ISE} = \sum_{t=1}^{N} (x(t) - \hat{x}(t))^2 \tag{18}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N} (x_t - \hat{x}_t)^2} \tag{19}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (\underline{Y}_i - Y_i)^2 \tag{20}$$

Among these metrics, the root-mean-square error (RMSE) was chosen as the objective function for the experimentation, as it provides a direct measure of the model's accuracy in terms of mean squared error.

Moreover, to assess the practical feasibility of the proposed method, it is relevant to analyze its computational complexity from both the inference system perspective and the execution environment.

Type-3 Fuzzy Inference System:

Type-3 fuzzy logic introduces a higher computational load compared to Type-1 and Type-2 systems due to the increased dimensionality of the footprint of uncertainty. However, to mitigate this load, a simplified inference structure based on α-plane representation is employed, which significantly reduces the number of operations required during the type reduction process. Furthermore, by using a Sugeno-type model, the inference stage is further simplified, as the outputs are linear functions, which enhances computational efficiency.

Simulation Environment:

The controller has been evaluated exclusively in a simulation environment, which allows for the computational load associated with both the Type-3 fuzzy logic and the optimization algorithm to be handled without difficulty. Since the system is not executed in real time, the type reduction and rule evaluation processes do not pose a practical limitation. Therefore, the system's complexity does not compromise the viability of the proposed approach within the experimental framework used.

The first step in the experimental process was to validate three different fuzzy systems. In this phase, the LowerScale parameter in the inverted pendulum controller was adjusted to analyze its impact on performance. Each of the fuzzy systems was implemented, and the results obtained for each system are presented in Table 1.

Table 1. Experimental results of the fuzzy controller with Lower Scale adjustments.

Error	Lower Scale 0.1	Error	Lower Scale 0.4	Error	Lower Scale 0.8
IAEV	224	IAEV	1.4	IAEV	225
ISEV	628	ISEV	224	ISEV	633
ITAEV	11,200	ITAEV	11,300	ITAEV	11,300
ITSEV	31,700	ITSEV	31,800	ITSEV	31,900
MEE	1.96	MEE	1.96	MEE	1.96
RMSE	1.4	RMSE	1.41	RMSE	1.42

Based on the results presented in Table 1, the performance of the three evaluated fuzzy systems was analyzed. Considering RMSE as the main comparison criterion, it was observed that the fuzzy system with a Lower Scale value of 0.1 achieved the lowest error. Therefore, this configuration was selected for further development and experimentation, as it demonstrated superior performance in terms of control accuracy.

Once the T3FS for the inverted pendulum controller was defined—referred to as T3IP (Type-3 inverted pendulum)—the experimentation phase was carried out with the objective of evaluating its performance under different operating conditions. To this end, the original DE algorithm was used and applied to the simulation of the T3IP controller.

The experimentation was conducted in two main phases. In the first phase, simulations were performed without the presence of noise (DE + T3IP), allowing the system's behavior to be analyzed under ideal conditions. In the second phase, noise was introduced into the input signal—specifically a uniform random number—resulting in the DE + T3IP + N configuration. The noise level was set to 0.50 to assess the robustness and stability of the controller under external disturbances.

The experimental setup parameters are provided in Table 2, and the results of the experimentation are shown in Table 3, where the performance of the DE+T3IP and DE + T3IP + N algorithms is compared. The table includes the best, worst, mean, and standard deviation values obtained over 30 independent experiments, allowing a comprehensive evaluation of consistency and performance under both ideal and noisy conditions.

Table 2. Parameter configuration.

Parameter Configuration	
NP	20
D	25
GEN	20
F	0.5
CR	0.5
Number of experiments	30

Table 3. Evaluation metrics for the Type-3 fuzzy controller with and without noise.

RMSE	DE + T3IP	DE + T3IP + N
Best	0.732	0.302
Worst	5.68	4.5
Mean	2.38	0.546
Std	0.226	0.277

Overall Accuracy (Mean RMSE): The controller configuration DE+T3IP+N achieves a significantly lower average RMSE compared to DE+T3IP, indicating better general performance even in the presence of noise. This suggests that the Type-3 fuzzy system responds more flexibly and adaptively to noisy inputs—a key advantage of higher-order fuzzy systems.

Best Case (Lowest RMSE): The best RMSE was also achieved with the noisy configuration (0.302 vs. 0.732), reinforcing the idea that introducing noise might help the system avoid local minima and reach more optimal solutions.

Worst Case (Highest RMSE): Interestingly, the worst-case performance with noise (4.500) is slightly better than that without noise (5.680), suggesting that the controller maintains acceptable behavior even under perturbations.

Consistency (Standard Deviation): The standard deviation is slightly higher with noise (0.277 compared to 0.226), indicating slightly more variability across the 30 runs. However, this increased variability is moderate and acceptable considering the significant reduction in average error.

The next phase of the experimentation involves applying the DE algorithm with T3FL, where the CR parameter is dynamically adapted. The parameters used for this stage are presented in Table 4.

Table 4. Parameters used for the experimentation with adaptive DE.

Parameter Configuration	
NP	20
D	25
GEN	20
F	0.5
CR	Dynamic
Number of experiments	30

Table 5 shows the results obtained from this approach applied to the Type-3 inverted pendulum system, under two different configurations:

Table 5. Performance results of the T3DE+T3IP controller with and without noise.

RMSE	T3DE + T3IP	T3DE + T3IP + N
Best	0.133	0.0127
Worst	0.397	0.235
Mean	0.122	0.0245
Std	0.136	0.0111

T3DE+T3IP: Differential Evolution with Type-3 fuzzy logic adapting CR, without noise;

T3DE + T3IP + N: The same configuration, but with the addition of uniform random noise with a noise level of 0.5 applied to the input signal.

The table reports the best, worst, mean, and standard deviation values obtained from 30 experimental runs, to assess the controller's accuracy, robustness, and stability under both ideal and noisy conditions.

To further evaluate the performance of the proposed controller, Table 5 presents the results obtained from 30 experimental runs under both noise-free (T3DE + T3IP) and noisy (T3DE + T3IP + N) conditions. The analysis includes the best, worst, mean, and standard deviation of the RMSE metric. These results allow us to assess the controller's accuracy, robustness, and consistency when facing external disturbances.

Best Case Performance: The T3DE + T3IP + N configuration achieved the best individual result (0.0127), indicating that introducing noise significantly enhanced the controller's performance in the best-case scenario.

Worst Case Performance: While both configurations show performance degradation in the worst case, the worst result of T3DE + T3IP + N (0.235) is still substantially better than that of T3DE + T3IP (0.397).

Average Performance (Mean): The average RMSE drops from 0.122 to 0.0245, demonstrating a considerable overall improvement in performance when noise is introduced.

Standard Deviation (Std): The standard deviation also decreases from 0.136 to 0.0111, indicating that the system with noise is more consistent and less variable.

To compare the performance of the Type-3 inverted pendulum controller under different experimental configurations, the results presented in Tables 3 and 5 were analyzed. These tables show the system's performance using the Differential Evolution algorithm in both its original form and in the variant with dynamic adaptation of the CR parameter, evaluated under both noise-free and noisy conditions. To facilitate comparison, Table 6 summarizes the best RMSE values obtained in each configuration. Additionally, Figure 11 provides a graphical representation of these results, allowing for a clear and direct visualization of the controller's performance differences under the evaluated conditions.

Table 6. Comparison of the best RMSE results of the T3IP controller.

T3IP Controller	RMSE
DE + T3IP	0.732
DE + T3IP + N	0.302
T3DE + T3IP	0.133
T3DE + T3IP + N	0.0127

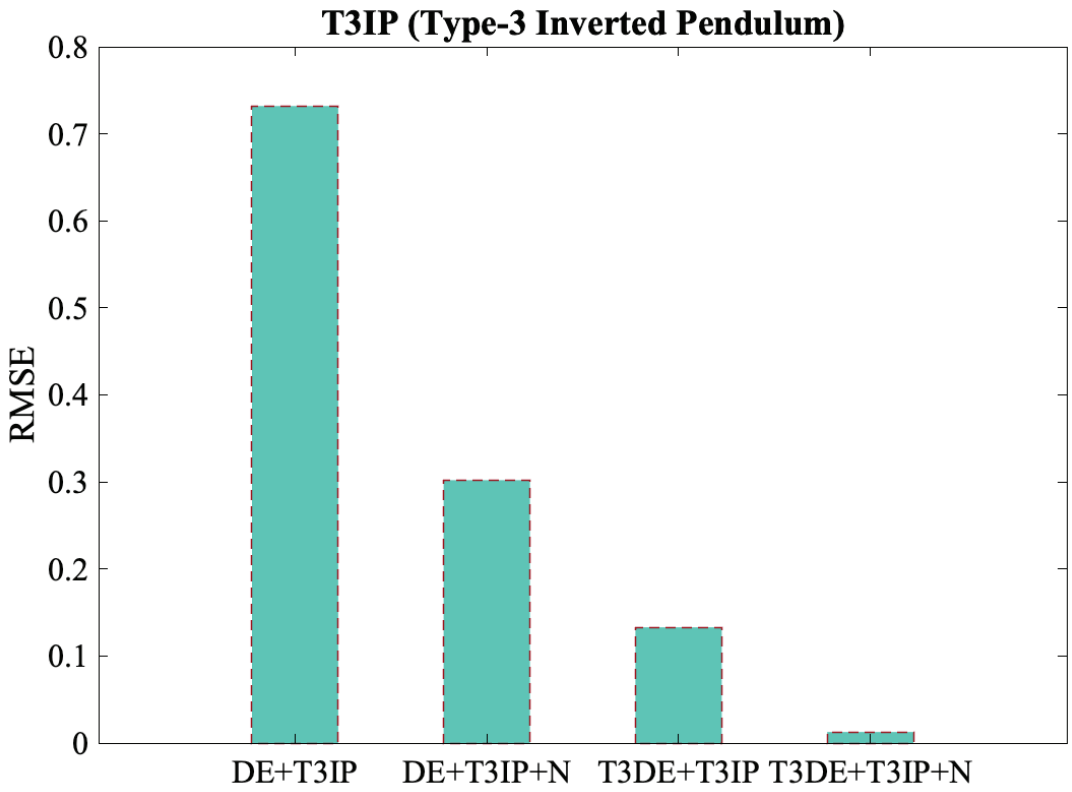


Figure 11. Comparison of best RMSE results for DE with different variables on the Inverted Pendulum.

The RMSE results indicate that the T3IP controller significantly improves its performance when using the adaptive Type-3 Fuzzy Differential Evolution algorithm (T3DE). This improvement is evident in both ideal and noisy conditions, where the lowest error (0.0127) is achieved, demonstrating greater precision and robustness of the system.

To validate the results obtained, two independent statistical tests were conducted to compare the performance of the controller under different configurations:

First Statistical Test: This test compares DE + T3IP + N ($\mu 1$) against DE + T3IP ($\mu 2$), based on the hypothesis that introducing noise improves the controller's performance, meaning that DE + T3IP + N yields a lower error than DE + T3IP.

Second Statistical Test: This comparison evaluates T3DE + T3IP + N ($\mu 1$) versus T3DE + T3IP ($\mu 2$), under the hypothesis that the Type-3 fuzzy controller with dynamic CR adaptation performs better when noise is introduced, implying that T3DE + T3IP + N achieves a lower error than T3DE + T3IP.

For both comparisons, a one-tailed (left-tailed) *Z*-test was used, since the goal is to determine whether the configuration with noise significantly reduces the error. The parameters used for these statistical tests are presented in Table 7.

Table 7. Specifications of the test parameters.

Parameter	Value
Significance	95%
Alpha	5%
H0	$\mu 1 \geq \mu 2$
Ha	$\mu 1 < \mu 2$ (claim)
Critical value	-1.645

Table 8 shows the results of the statistical tests conducted to compare the controller's performance under different configurations. Two independent one-tailed (left-sided) Z-tests were applied to determine whether the inclusion of noise significantly reduces the RMSE.

Table 8. Z-test results for the T3IP system.

Statistical Tests				
Case study	μ_1	μ_2	Z Value	Evidence
T3IP (Type-3 Inverted Pendulum)	DE+T3IP+N	DE+T3IP	−28.0987	Significant
	T3DE+T3IP+N	T3DE+T3IP	−3.9137	Significant

In the first case, DE+T3IP+N was compared with DE+T3IP, and in the second case, T3DE+T3IP+N was compared with T3DE+T3IP. In both tests, the alternative hypothesis assumes that the configuration with noise (μ_1) produces a lower error than the configuration without noise (μ_2). The calculated Z values were −28.0987 and −3.9137, respectively, both indicating statistically significant differences in favor of the noisy configurations.

Finally, a comparison was made with a previous study in which the inverted pendulum controller was implemented using the Differential Evolution (DE) algorithm under various experimental configurations [49]. Table 9 presents the RMSE values obtained for each configuration, including the best, worst, mean, and standard deviation.

Table 9. Performance comparison of the controller with previous and current configurations.

Inverted Pendulum RMSE							
	DE [49]	**DE+T1FS [49]**	**DE+T1FS+N [49]**	**DE+IT2FS [49]**	**DE +IT2FS+N [49]**	**T3DE+T3IP**	**T3DE+T3IP+N**
Best	0.58	0.1	0.0433	0.0183	0.014	0.133	0.0127
Worst	2.4	1.4	1.26	1.4	0.433	0.397	0.235
Mean	1.4	0.43	0.33	0.215	0.241	0.122	0.0245
Std	0.52	0.36	0.269	0.275	0.165	0.136	0.0111

To contextualize the performance of the approach, a direct comparison is presented with configurations previously reported in the literature. The table includes various combinations of DE with Type-1, Type-2, and Type-3 fuzzy systems, both with and without noise. The following section analyzes the performance of each configuration based on the obtained RMSE values.

Performance evolution with higher-order fuzzy systems: There is a clear performance improvement when moving from pure DE (mean RMSE of 1.4) to more sophisticated fuzzy systems such as the T1FS, IT2FS, and ultimately T3IP.

Effect of noise (N): The inclusion of noise consistently improves performance across all configurations. For instance, DE+T1FS has a mean RMSE of 0.43, which drops to 0.33 with DE+T1FS+N. A similar pattern is observed with the IT2FS and T3IP.

Proposed controller (T3DE+T3IP+N): This configuration shows the best overall performance among all evaluated approaches:

Best RMSE: 0.0127;

Lowest mean RMSE: 0.0245;

Smallest standard deviation: 0.0111.

These results indicate that the controller based on T3FL with dynamic CR adaptation and added noise not only enhances accuracy but also improves stability and consistency.

Comparison with IT2FS: Although IT2FS-based configurations perform better than the T1FS, they still do not reach the performance levels achieved by the proposed controller. This highlights the advantages of T3FSs for nonlinear control tasks.

Based on the best RMSE values reported in Table 9 for each of the evaluated configurations, Figure 12 was generated to provide a comparative graphical representation of these results. This visualization clearly highlights the impact of different control approaches and the role of noise on the system's performance.

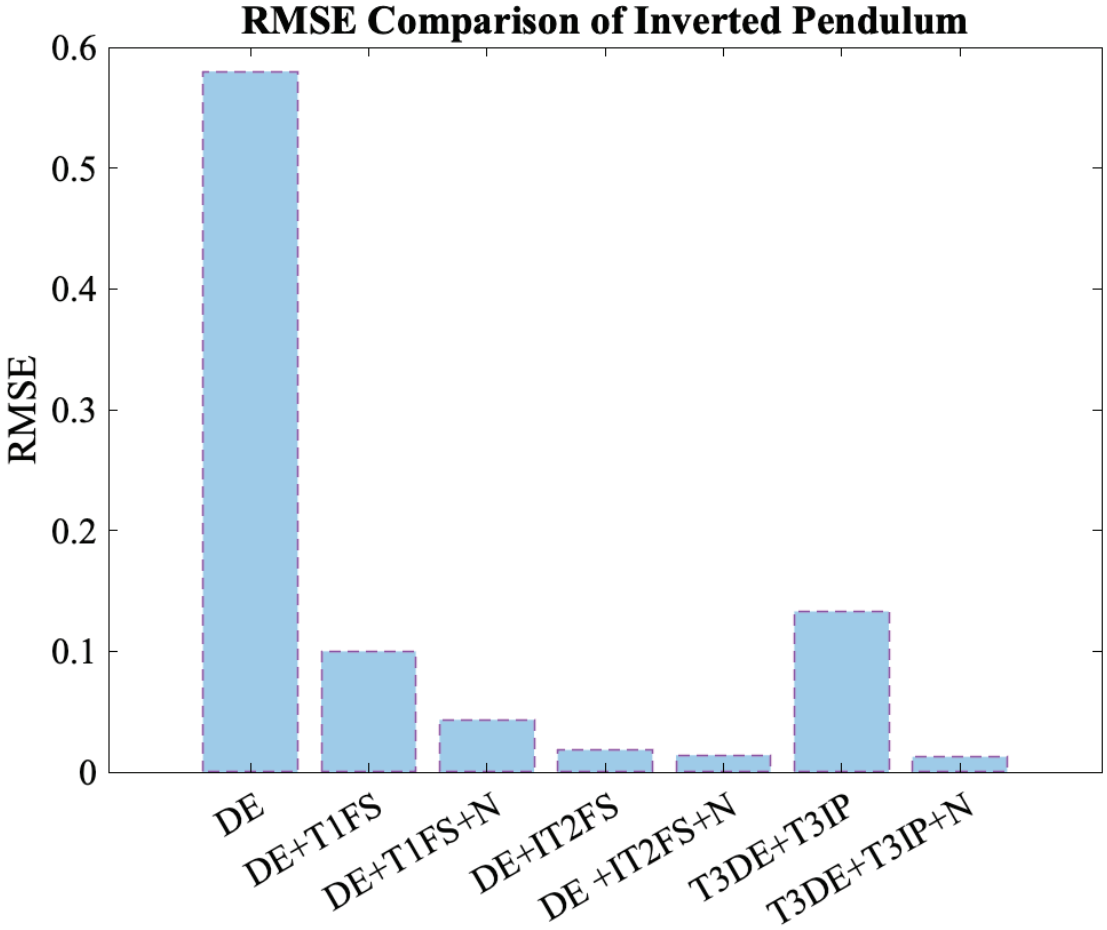


Figure 12. Comparison of the best RMSE results.

6. Discussion of Results

The results presented in this study demonstrate the effectiveness of the Type-3 fuzzy logic framework in addressing the limitations of traditional Type-1 and Type-2 fuzzy control systems. The proposed approach exhibits superior performance in terms of accuracy (measured by RMSE) and robustness, particularly under noisy conditions. These findings align with recent advances in adaptive control methodologies, further validating the potential of Type-3 fuzzy logic as a viable solution for complex nonlinear systems.

Compared to related studies, our framework provides a more flexible mechanism to handle dynamic uncertainties. Unlike previous approaches, which rely heavily on fixed or less adaptable parameters, the dynamic adaptation of the CR parameter in our method ensures consistent performance across a wide range of operational scenarios. Additionally, the results presented in Table 9 highlight the comparative advantages of the proposed approach under both noisy and noise-free conditions, reinforcing its robustness and adaptability. Another aspect to consider is the complexity involved in selecting and calibrating the additional parameters inherent in the Type-3 framework. Although the parameter selection process described in Section 4 provides a systematic approach, future research could explore automated or hybrid optimization techniques to simplify this process.

This work presented the design and evaluation of a control strategy based on an interval Type-3 fuzzy system (IT3FS) for the inverted pendulum system, with the main objective of minimizing the RMSE. Initially, the influence of the LowerScale parameter was analyzed, and the configuration with LowerScale = 0.1 yielded the best RMSE (1.4), making it the selected setup for the controller. Subsequently, experiments were conducted using the inverted pendulum system in two phases:

1. Using a standard Differential Evolution (DE) algorithm with the Type-3 fuzzy controller (T3IP), both with and without noise.
2. Using an enhanced Differential Evolution (T3DE) version, where the crossover rate (CR) is dynamically adjusted using a Type-3 fuzzy logic system.

The following RMSE results were obtained:

DE + T3IP: 0.732;
DE + T3IP + N: 0.302;
T3DE + T3IP: 0.133;
T3DE + T3IP + N: 0.0127.

These results show a substantial improvement in performance, particularly in the presence of noise. The T3DE+T3IP+N configuration outperformed all others, confirming that integrating T3FL both in the controller and in the parameter adaptation of the evolutionary algorithm enhances robustness and accuracy. To further assess the effectiveness of the proposal, a comparative analysis was conducted against previous fuzzy control strategies, including the following:

DE;
DE with Type-1 fuzzy systems (T1FSs);
DE with interval Type-2 fuzzy systems (IT2FSs).

The comparison, summarized in terms of the best, worst, mean, and standard deviation RMSE values, showed that the proposed T3DE + T3IP approach achieved the following:

The lowest mean RMSE: 0.0245;
The best performance in noisy environments: 0.0127;
The lowest variability: standard deviation of 0.0111.

Compared to the previous best (DE + IT2FS + N), which had a best-case RMSE of 0.014, the proposed method improved both the average performance and the consistency of the results. This demonstrates the advantage of using Type-3 fuzzy systems not only for control, but also for dynamic parameter tuning in optimization algorithms.

7. Conclusions

In conclusion, the dual application of T3FL, both within the controller and in the evolutionary optimization process, proved to be a robust and effective strategy for controlling nonlinear systems under uncertain and noisy conditions. The comparative results clearly validate the superiority of the proposal over traditional DE and lower-order fuzzy logic-based approaches.

As a main line of future work, we propose to further explore the Differential Evolution algorithm enhanced with Type-3 fuzzy logic, investigating alternative strategies for dynamic parameter adaptation and hybridization with other optimization techniques. This could lead to improved performance in complex and highly uncertain environments.

Additionally, it is considered promising to extend Type-3 fuzzy logic to other types of controllers, such as adaptive PID, predictive, or neuro-fuzzy controllers, aiming to enhance their robustness against disturbances and noise. This would allow for the development of more flexible and accurate control schemes.

Finally, a key goal is to apply these developments to real-world problems, such as the control of robotic systems, autonomous vehicles, industrial processes, or energy applications. This would not only validate the practical utility of the proposed approach but also open new opportunities for designing intelligent control systems based on T3FL.

This study has presented a Type-3 fuzzy logic framework that demonstrates significant advantages over traditional Type-1 and Type-2 approaches in terms of accuracy and robustness, especially under noisy conditions. However, like any approach, the proposed method has limitations that must be considered:

Limitations:

- Computational cost: The implementation of the Type-3 framework involves higher computational costs compared to Type-1 and Type-2 approaches. This could limit its application in systems with restricted computational resources or those requiring real-time responses.
- Parameter selection complexity: The need to select and calibrate several additional parameters can pose a challenge, particularly in systems with high dimensionality or changing conditions.

Future directions:

- Computational optimization: Future research could focus on improving the computational efficiency of the method, developing simplified versions, or leveraging specialized hardware such as GPUs.
- Automation of parameter selection: The use of automatic optimization techniques, such as evolutionary algorithms or machine learning, could simplify the parameter tuning process and enhance the framework's usability.

Author Contributions: P.O., C.P. and O.C. conducted the state-of-the-art and literature reviews, conducted the formal analysis, and developed the main idea. P.M. contributed to the formal analysis and methodology. O.C. and Z.W.G. provided supervision. Z.W.G. acquired funding. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & Energy, Republic of Korea (RS-2024-00441420; RS-2024-00442817).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We thank the Tijuana Institute of Technology and TECNM for their support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Price, K.V.; Storn, R.M.; Lampinen, J.A. *Differential Evolution*; Springer: Berlin/Heidelberg, Germany, 2005.
2. Pytel, K. Fuzzy logic applied to tuning mutation size in evolutionary algorithms. *Sci. Rep.* **2025**, *15*, 1937. [CrossRef] [PubMed]
3. Bingol, M.C. A Safe Navigation Algorithm for Differential-Drive Mobile Robots by Using Fuzzy Logic Reward Function-Based Deep Reinforcement Learning. *Electronics* **2025**, *14*, 1593. [CrossRef]
4. Li, K.; Liu, X. Fuzzy Adaptive Cooperative Prescribed Time Control for Nonlinear Vehicular Platoon Systems. *IEEE Trans. Fuzzy Syst.* **2025**, *33*, 1568–1577. [CrossRef]
5. Yilmaz, B.M.; Tatlicioglu, E.; Selim, E. Velocity observer design for a class of uncertain nonlinear mechanical systems: A self-adaptive fuzzy logic-based approach. *Int. J. Syst. Sci.* **2025**, *56*, 423–433. [CrossRef]
6. Zadeh, L.A. Fuzzy sets. *Inf. Control* **1965**, *8*, 338–353. [CrossRef]
7. Grasa, P.; Polyakhov, N. The direct adaptive trajectory control of robot manipulators via fuzzy logic algorithm. *WIT Trans. Inf. Commun. Technol.* **2025**, *1*, 9.

8. Zhao, X.; Zhao, Q.; Peng, J. Research on improved dynamic window approach based on fuzzy logic for adaptive obstacle avoidance control in unmanned surface vehicles. In Proceedings of the Fifth International Conference on Signal Processing and Computer Science (SPCS 2024) SPIE, Kaifeng, China, 23–25 August 2024; Volume 13442, pp. 643–651.
9. Cavus, M.; Dissanayake, D.; Bell, M. Deep-Fuzzy Logic Control for Optimal Energy Management: A Predictive and Adaptive Framework for Grid-Connected Microgrids. *Energies* **2025**, *18*, 995. [CrossRef]
10. Wang, Y.; Liu, J.; Yu, J. Lyapunov-derived fuzzy temperature control for thermoelectric cooling system using dynamic updating law. *Appl. Therm. Eng.* **2025**, *258*, 124600. [CrossRef]
11. Li, H.; Chou, W.; Meng, L. Finite-Time Adaptive Fuzzy Prescribed Performance Control for Teleoperation With Asymmetric Time Delays and Uncertainties. *Int. J. Robust. Nonlinear Control* **2025**, *35*, 2982–2995. [CrossRef]
12. Elnaghi, B.E.; Ismaiel, A.M.; El Sayed Abdel-Kader, F.; Abelwhab, M.N.; Mohammed, R.H. Validation of energy valley optimization for adaptive fuzzy logic controller of DFIG-based wind turbines. *Sci. Rep.* **2025**, *15*, 711. [CrossRef]
13. Madebo, N.W. Enhancing Intelligent Control Strategies for UAVs: A Comparative Analysis of Fuzzy Logic, Fuzzy PID, and GA-Optimized Fuzzy PID Controllers. *IEEE Access* **2025**, *13*, 16548–16563. [CrossRef]
14. Mohammad, S.I.; Yogeesh, N.; Raja, N.; Chetana, R.; Ramesha, M.S.; Vasudevan, A. Optimizing MIMO Antenna Performance Using Fuzzy Logic Algorithms. *Appl. Math.* **2025**, *19*, 349–364.
15. Jiang, Q.; Wang, P. A fuzzy logic-based coordinated adaptive control method for megawatt novel nuclear power systems. *Ann. Nucl. Energy* **2025**, *215*, 111245. [CrossRef]
16. Hosny, M.; Marei, M.I.; Mohamad, A.M. Adaptive hybrid virtual inertia controller for PMSG-based wind turbine based on fuzzy logic control. *Sci. Rep.* **2025**, *15*, 3757. [CrossRef]
17. Li, Z.; Dewantoro, G.; Xiao, T.; Swain, A. A Comparative Analysis of Fuzzy Logic Control and Model Predictive Control in Photovoltaic Maximum Power Point Tracking. *Electronics* **2025**, *14*, 1009. [CrossRef]
18. Zhang, W. Predefined-Time Fuzzy Control for Robotic Manipulators Driven by Compliant Actuators. *Int. J. Fuzzy Syst.* **2025**, *27*, 1–10. [CrossRef]
19. Wang, B.; Zhang, J.; Wu, D. Force–vision fusion fuzzy control for robotic batch precision assembly of flexibly absorbed pegs. *Robot. Comput.-Integr. Manuf.* **2025**, *92*, 102861. [CrossRef]
20. Pérez-Juárez, J.G.; García-Martínez, J.R.; Medina Santiago, A.; Cruz-Miguel, E.E.; Olmedo-García, L.F.; Barra-Vázquez, O.A.; Rojas-Hernández, M.A. Kinematic Fuzzy Logic-Based Controller for Trajectory Tracking of Wheeled Mobile Robots in Virtual Environments. *Symmetry* **2025**, *17*, 301. [CrossRef]
21. Kumar, V.; Mistri, A. Fuzzy logic-based synchronization of trajectory planning and obstacle avoidance for RRP SCARA robot. *Int. J. Interact. Des. Manuf.* **2025**, *19*, 3039–3052. [CrossRef]
22. Shufian, A.; Shib, S.K.; Dipto, D.R.; Rahman, M.T.; Hannan, N.; Fattah, S.A. Fuzzy logic-controlled three-phase dynamic voltage restorer for enhancing voltage stabilization and power quality. *Int. J. Electr. Power Energy Syst.* **2025**, *166*, 110517. [CrossRef]
23. Zheng, Y.; Na, Z.; Ji, W.; Lu, Y. An Adaptive Fuzzy SIR Model for Real-Time Malware Spread Prediction in Industrial Internet of Things Networks. *IEEE Internet Things J.* 2025; *in press*. [CrossRef]
24. Chotikunnan, P.; Roongprasert, K.; Chotikunnan, R.; Pititheeraphab, Y.; Puttasakul, T.; Wongkamhang, A.; Thongpance, N. Hybrid Fuzzy-Expert System Control for Robotic Manipulator Applications. *J. Robot. Control (JRC)* **2025**, *6*, 155–165. [CrossRef]
25. Nasir, N.M.; Ghani NM, A.; Nasir AN, K.; Ahmad, M.A.; Tokhi, M.O. Neuro-modelling and fuzzy logic control of a two-wheeled wheelchair system. *J. Low. Freq. Noise Vib. Act. Control* **2025**, *44*, 588–602. [CrossRef]
26. Mendel, J.M.; John, R.I.; Liu, F. Interval Type-2 Fuzzy Logic Systems Made Simple. *IEEE Trans. Fuzzy Syst.* **2006**, *14*, 808–821. [CrossRef]
27. Ambareesh, S.; Chavan, P.; Supreeth, S.; Nandalike, R.; Dayananda, P.; Rohith, S. A secure and energy-efficient routing using coupled ensemble selection approach and optimal type-2 fuzzy logic in WSN. *Sci. Rep.* **2025**, *15*, 38. [CrossRef] [PubMed]
28. Zeng, Y.; Hussein, Z.A.; Chyad, M.H.; Farhadi, A.; Yu, J.; Rahbarimagham, H. Integrating type-2 fuzzy logic controllers with digital twin and neural networks for advanced hydropower system management. *Sci. Rep.* **2025**, *15*, 5140. [CrossRef]
29. Boudia, A.; Messalti, S.; Zeghlache, S.; Harrag, A. Type-2 fuzzy logic controller-based maximum power point tracking for photovoltaic system. *Electr. Eng. Electromech.* **2025**, *1*, 16–22. [CrossRef]
30. Rinesh, S.; Arun, M.; Kumar, S.N.; Prajitha, C.; Kumar, A.S. Evaluating the type 2 fuzzy logic controller with multilayer perceptrons for optimal tracking of solar photovoltaic systems. *Int. J. Low-Carbon. Technol.* **2025**, *20*, 394–403. [CrossRef]
31. Belhadj, S.M.; Meliani, B.; Benbouhenni, H.; Zaidi, S.; Elbarbary ZM, S.; Alammer, M.M. Control of multi-level quadratic DC-DC boost converter for photovoltaic systems using type-2 fuzzy logic technique-based MPPT approaches. *Heliyon* **2025**, *11*, e42181. [CrossRef]
32. Hashim AA, A.; Ghani NM, A.; Ahmad, S.; Hashim, M.R.; Naharuddin NZ, A.; Irawan, A. Optimized Type-2 Fuzzy Logic Control for Low-Speed Vehicle Pedal Pressing Automation Using Hybrid Spiral Sine Cosine Algorithm. *IIUM Eng. J.* **2025**, *26*, 563–584. [CrossRef]

33. Benbrahim, M.; Bouhentala, M.; Ghanai, M.; Chafaa, K.; Essounbouli, N. Generalization to Type 2 of PSO-Optimized Type 1 PD Fuzzy Controller and its Application to a Quadrotor UAV. *Eng. Technol. Appl. Sci. Res.* **2025**, *15*, 21658–21664. [CrossRef]
34. Castillo, O.; Castro, J.R.; Melin, P. Interval type-3 fuzzy logic systems (IT3FLS). In *Interval Type-3 Fuzzy Systems: Theory and Design*; Springer International Publishing: Cham, Switzerland, 2022; pp. 45–49.
35. Sabahi, K.; Panahi, S.; Shokri Kalandaragh, Y.; Mohammadzadeh, A. Deep reinforcement learning tuned type-3 fuzzy PID controller: AC microgrid case study. *Electr. Eng.* **2005**, *107*, 1–14. [CrossRef]
36. Melin, P.; Castillo, O. A Type-3 Fuzzy-Fractal Approach for Diagnosis of Mental Disorders. In *Type-3 Fuzzy Logic and Fractal Theory for Medical Diagnosis*; Springer Nature Switzerland: Cham, Switzerland, 2025; pp. 47–55.
37. Sabetahd, R.; Mohammadzadeh, A.; Jafarzadeh, O.; Taghavifar, H.; Sakthivel, R.; Zhang, C. *Applied Type-3 Fuzzy Logic Systems and Controllers*; Studies in Fuzziness and Soft Computing; Springer: Cham, Switzerland, 2025.
38. Li, S.; Mohammadzadeh, A.; Zhang, C. An Applied Type-3 Fuzzy Controller for Gyroscopes. *Int. J. Fuzzy Syst.* **2025**, *27*, 1–18. [CrossRef]
39. Mozaffari, M.; Mazinani, S.M.; Khazaei, A.A. An energy efficient grid-based clustering algorithm using type-3 fuzzy system in wireless sensor networks. *Wirel. Netw.* **2025**, *31*, 109–125. [CrossRef]
40. Chen, F.; Qiu, X.; Alattas, K.A.; Mohammadzadeh, A.; Ghaderpour, E. A new fuzzy robust control for linear parameter-varying systems. *Mathematics* **2022**, *10*, 3319. [CrossRef]
41. Maraslidis, G.S.; Kottas, T.L.; Tsipouras, M.G.; Fragulis, G.F. Design of a Fuzzy Logic Controller for the Double Pendulum Inverted on a Cart. *Information* **2022**, *13*, 379. [CrossRef]
42. Shirgir, S.; Farahmand-Tabar, S. An enhanced optimum design of a Takagi-Sugeno-Kang fuzzy inference system for seismic response prediction of bridges. *Expert. Syst. Appl.* **2025**, *266*, 126096. [CrossRef]
43. Zaman, M.Q.; Wu, H.M. Genetic Algorithm Based Takagi-Sugeno Fuzzy Modelling of a Mecanum Wheeled Mobile Robot. *Int. J. Control Autom. Syst.* **2025**, *23*, 907–919. [CrossRef]
44. Gu, S.; Wang, Y.; Chou, Y.; Cong, J.; Lu, M.; Jiao, Z. Interpretable Style Takagi-Sugeno-Kang Fuzzy Clustering. *arXiv* **2025**, arXiv:2504.05125.
45. Pérez-del-Notario, I.; González, X.; Horanska, L.; Oria, P.; Fernandez, J.; Dimuro, G.; Bustince, H. Flexible fusion models for improving the early warning of river floodings, based on a generalization of Takagi-Sugeno-Kang inference system. *Expert. Syst. Appl.* **2025**, *240*, 121539.
46. Zhang, S.X.; Liu, Y.H.; Hu, X.R.; Zheng, L.M.; Zheng, S.Y. Fully Informed Fuzzy Logic System Assisted Adaptive Differential Evolution Algorithm for Noisy Optimization. *IEEE Trans. Fuzzy Syst.* 2025; *in press*.
47. Mehmood, A.; Raja MA, Z. Fuzzy-weighted differential evolution computing paradigm for fractional order nonlinear wiener systems. *Chaos Solitons Fractals* **2022**, *159*, 112160. [CrossRef]
48. Shukla, S.; Pandit, M. An optimum multi-objective dynamic scheduling strategy for a hybrid microgrid system using fuzzy rank-based modified differential evolution algorithm. In *Artificial Intelligence and Sustainable Computing: Proceedings of ICSISCET 2020*; Springer: Singapore, 2021; pp. 175–188.
49. Ochoa, P.; Castillo, O.; Soria, J. Optimization of fuzzy controller design using a differential evolution algorithm with dynamic parameter adaptation based on type-1 and interval type-2 fuzzy systems. *Soft Comput.* **2020**, *24*, 193–214. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Supporting Human–Robot Interaction in Manufacturing with Augmented Reality and Effective Human–Computer Interaction: A Review and Framework

Karthik Subramanian [1,*], Liya Thomas [2], Melis Sahin [3] and Ferat Sahin [1]

[1] Department of Electrical and Microelectronic Engineering, Rochester Institute of Technology, Rochester, NY 14623, USA; feseee@rit.edu
[2] School of Information, Rochester Institute of Technology, Rochester, NY 14623, USA; lst6940@g.rit.edu
[3] Departments of Biomedical and Electrical Engineering, Case Western Reserve University, Cleveland, OH 44195, USA; mxs1801@case.edu
* Correspondence: kxs8997@g.rit.edu

Abstract: The integration of Augmented Reality (AR) into Human–Robot Interaction (HRI) represents a significant advancement in collaborative technologies. This paper provides a comprehensive review of AR applications within HRI with a focus on manufacturing, emphasizing their role in enhancing collaboration, trust, and safety. By aggregating findings from numerous studies, this research highlights key challenges, including the need for improved Situational Awareness, enhanced safety, and more effective communication between humans and robots. A framework developed from the literature is presented, detailing the critical elements of AR necessary for advancing HRI. The framework outlines effective methods for continuously evaluating AR systems for HRI. The framework is supported with the help of two case studies and another ongoing research endeavor presented in this paper. This structured approach focuses on enhancing collaboration and safety, with a strong emphasis on integrating best practices from Human–Computer Interaction (HCI) centered around user experience and design.

Keywords: Situational Awareness; user experience; trust in automation; collaborative robotics; AR interface design; user-centered design

1. Introduction

The field of Human–Robot Interaction (HRI) has been rapidly evolving, with recent years seeing a significant focus on integrating Augmented Reality (AR). HRI, which combines the efforts of humans and robots, offers considerable flexibility and benefits, such as reduced production time [1,2]. Although HRI is used in various industries, such as medical and automotive, it is predominantly applied in collaborative industrial settings, such as assembly lines and manufacturing [3].

Recent research has used AR to address these challenges in HRI. AR has been shown to improve accuracy and efficiency, reduce cognitive workload, and minimize errors [4,5]. In HRI, AR is implemented through devices such as headsets, smartwatches, and projections, each tailored to tackle different aspects of the aforementioned issues [6–8]. However, the rise in AR usage has introduced a new challenge related to information visualization. Current research on AR in HRI focuses more on information architecture rather than the presentation of information to the user [9]. It is crucial to emphasize the user interface (UI) in AR applications, as it directly affects how users perceive, comprehend, and interact with the instructions and information provided [10]. This is especially important in high-speed, high-stress environments where humans work in close proximity to robots. Integrating Human–Computer Interaction (HCI) principles into AR interfaces can enhance both visualization and usability.

Most AR in robotics research focuses primarily on AR in manufacturing, rather than HRI as a whole [4,11–13]. These papers solely address the applications and benefits of AR in aiding the work between humans and robots in manufacturing settings. Regarding HRI as a whole, prior research also includes a survey paper on the use of Augmented Reality (AR) in HRI [5]. This paper is highly technical and introduces specialized terminology to categorize different AR hardware, overlays, and interaction points. However, none of the papers mentioned above address HCI, user interface design, or the importance of how visual information should be displayed. An AR survey paper briefly touches on HCI but is not focused on the field of HRI [9]. Instead, it looks at AR used for Situational Awareness across many fields of research.

HRI and HCI are closely related fields, yet they focus on different aspects of interaction. HCI primarily deals with the design and use of computer technologies, emphasizing how humans interact with digital interfaces such as software applications and websites. The goal of HCI is to improve the user experience, usability, and accessibility of these technologies. In contrast, HRI focuses on interactions between humans and robots, which involves not only digital interfaces but also physical, spatial, and social dynamics. HRI studies tend to focus on information architecture and ensuring that AR systems operate correctly, while HCI studies concentrate on information visualization and user research, prioritizing how effectively users understand the information. There is a gap in research at the intersection of HRI and HCI. Specifically, there is a lack of HRI papers that incorporate HCI principles to enhance AR visualizations, thereby optimizing efficiency, understanding, and task load. The principles of good design in AR from HCI can be used to improve the visualization of information within AR systems in HRI. This survey paper aims to provide a comprehensive literature review of key factors in HRI, AR visualizations that enhance HRI, and potential applications of HCI within AR in HRI. The paper will begin by discussing the key factors in HRI and research that has improved collaboration, trust, and safety over the years. It will then examine common challenges within HRI and prior work that uses AR to address these challenges. Finally, it will discuss the implementation of UI design practices in AR for HRI, outline good design practices based on HCI principles, and provide a framework for future work in AR with respect to HRI. The scope of this study is specifically focused on industrial applications, particularly assembly line tasks in manufacturing where humans and robots collaborate in shared workspaces.

2. Materials and Methods

In this study, four researchers collaborated to analyze and compile 164 papers to explore the integration of Augmented Reality (AR) into HRI and the incorporation of HCI principles. The updated PRISMA methodology was utilized to guide and justify the literature collection process used in this paper [14]. A total of 164 papers were gathered by three researchers from 8 databases and 11 registers. The databases included ACM Digital Library, Elsevier, IEEE Xplore, MDPI, PLOS, SAGE Journals, Springer, and Taylor & Francis. The registers used were ArXiv Preprints, Royal Society Publishing, Frontiers, Core Reader, OpenReview, Inderscience Online, Now Publishers, Nielsen Norman Group (NNG), Scribd, ISO Standards, and Thesai Publications, with some identified through Google Scholar.

The literature search was divided among three researchers, each focusing on different thematic areas:

- The first researcher focused on AR- and HRI-related content, using the following search strings: “HRI Key Principles”, “HRI Foundations”, “AR in HRI”, “Mixed Reality in HRI”, and “Hololens in HRI”.
- The second researcher concentrated on HCI materials, using the following search strings: “UI in HRI”, “UX in HRI”, “HCI in HRI”, “AR UI in HRI”, “HCI Foundations”, and “UX/UI Principles”.
- The third researcher specialized in Situational Awareness, using the following search strings: “Situational Awareness in HCI”, “Situational Awareness in HRI”, “Situational Awareness for AR”, and “Hololens Situational Awareness”.

No duplicate papers were found as each researcher explored different areas of the literature. In the initial screening process, 26 papers were excluded for not meeting the inclusion criteria, as they were non-peer-reviewed, irrelevant to the research topic, or outdated (published before 1990). The remaining 138 papers were retrieved, reviewed, and assessed for eligibility by all four researchers. Papers were required to contribute to at least one of the following areas: key HRI factors, AR use in HRI, HCI principles, integration of HCI in AR for HRI, or Situational Awareness in AR-HRI contexts. Eighteen additional papers were removed after this eligibility assessment. Of these, 4 papers gathered by the second researcher were excluded due to missing HCI integration within HRI. Fourteen papers gathered by the first researcher were excluded, with 9 focusing on contexts outside the scope of manufacturing and production in HRI, and 6 being VR-centric rather than AR-centric. In total, 119 papers were included in this review, including two studies cited in the final discussion on the proposed framework. Figure 1, generated from [15], visualizes the flow of identification, screening, retrieval, and eligibility of the literature gathered for this paper. Two additional citations were used in our paper to reference PRISMA.

The structured process of gathering and categorizing papers facilitated the development of a robust framework outlining critical AR elements necessary for advancing HRI, with a focus on improving collaboration, safety, and effective communication between humans and robots. With this comprehensive review of the methodology, the following sections will explore key factors that drive successful HRI, particularly focusing on collaboration, trust, and safety.

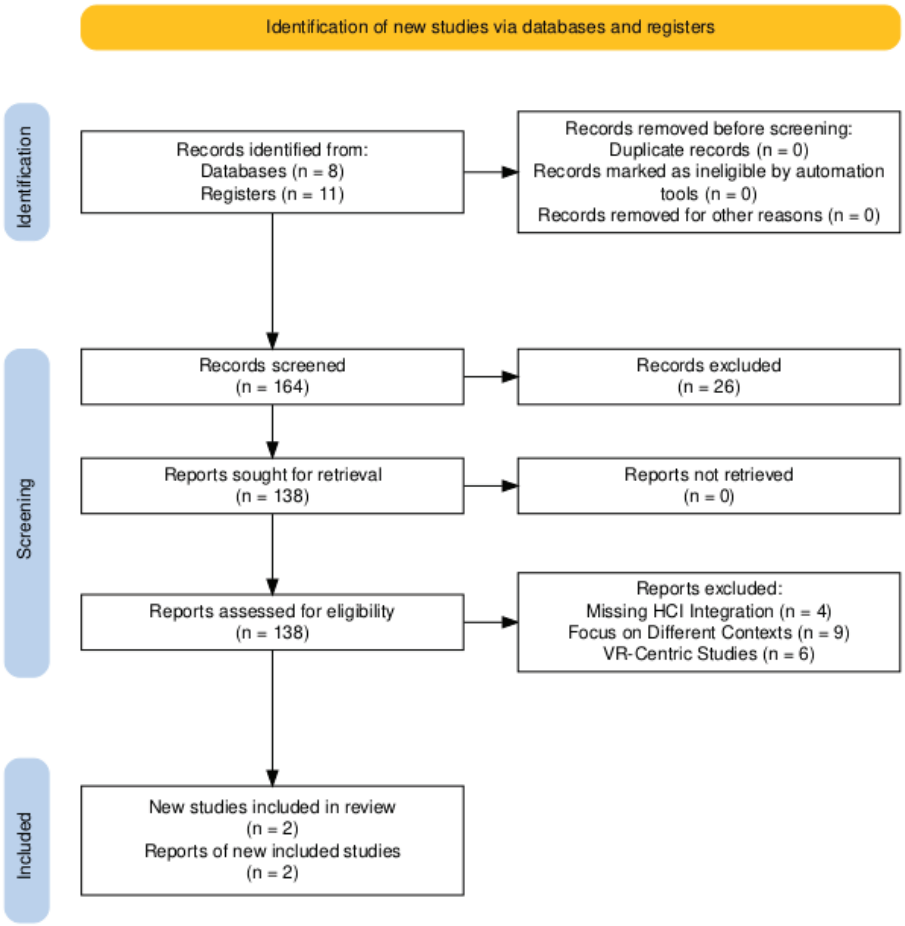


Figure 1. PRISMA diagram of the literature gathering and screening process, generated using [15].

3. Key Factors in HRI

Human–Robot Interaction (HRI) is the combination of both robots and humans working together. It combines the accuracy and speed of automated machines with the ease of adjustability and change in human labor [3,16]. HRI requires robots to be complex; they must be flexible enough to adapt to changing work environments and situations while also being autonomous [2,17]. All robots have control systems and software controlled by their human counterparts; some require full direction, such as teleoperated robots, while others operate alone, but are previously programmed to complete certain tasks [1]. Human–Robot Collaboration (HRC) refers to when humans and robots work together towards a common goal. HRC is a subsection of HRI and often utilizes cognitive robots that can plan their own actions by gathering information from the environment, understanding the human counterpart's intent and determining what steps need to be completed in order to finish the task [18]. These types of cognitive robots are advancing more in the field of HCI and require modern machine learning solutions that can build both the cognitive model and the behavioral block [19]. Typically, HRC is seen in industrial work settings including manufacturing and fabrication [20].

This paper mostly looks at HRC research in industrial manufacturing settings but also includes HRI and HRC research from other fields when relevant. Therefore, the umbrella term HRI will be used unless it is necessary to specify that the research was carried out for HRC. HRI can be beneficial for a number of reasons including low production costs, flexible adaptation to changing production layouts and tasks, and ergonomic benefits for humans [21]. Research in HRI tends to focus on improving fluency and efficiency between humans and robots through 2 key factors: collaboration on the one hand and trust and safety on the other. The framework for this section has been visualized in Figure 2.

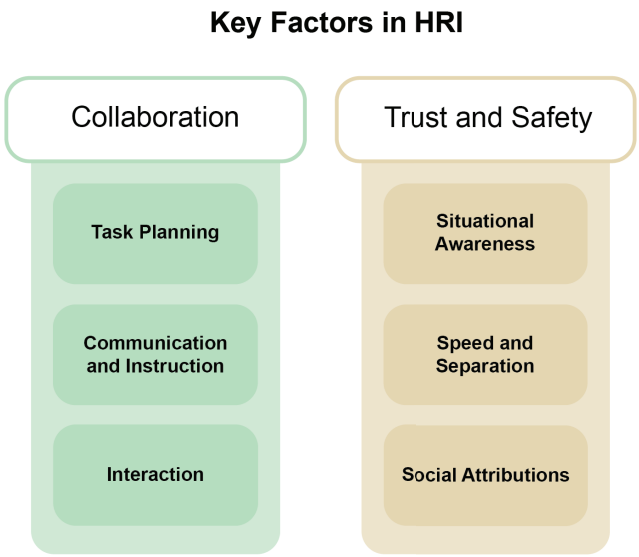


Figure 2. Framework for key factors in HRI.

3.1. Collaboration

Collaboration between humans and robots refers to how they work together. Bauer et al. categorize various types of HRI collaboration into five categories: cell, coexistence, sequential collaboration, cooperation, and responsive collaboration [18]. Each of these levels increases exponentially in terms of how closely robots and humans work and how much they interact with each other.

Efficient collaboration requires planning, communication and instruction, and interaction. Planning entails the allocation of tasks and deciding the pathways of the robots [18]. Communication and instruction pertain to how humans and robots deliver and send each other feedback or instructions. With proper planning and communication, successful inter-

action between humans and robots can be achieved. Interaction can also proceed smoothly and quickly with the integration of action planning. Action planning involves robots anticipating a human's movement and then planning their own actions accordingly. This subsection explores current research that aims to improve the planning, communication and interaction in HRI.

3.1.1. Task Planning

When robots and humans work together, proper planning is required to ensure safe interaction and a productive environment. One aspect of planning is task allocation or assigning proper tasks to both human individuals and robot entities. This ensures a smooth workflow and prevents overloading for any single participant. Lamon et al. [22] look at allocating tasks in manufacturing scenarios based on a robot's physical characteristics. The experiment looks at the physical capabilities of robot agents and human workers and then allocates tasks accordingly. Figure 3, taken from the paper, explains how tasks are broken down into a series of actions. These actions are assessed by the algorithm to determine whether they should be completed by a human or robot agent and are then allocated accordingly. The solution proposes which tasks are better suited for robot agents to remove easily physically fatiguing tasks from human workers. The results proved that tasks allocated by physical characteristics are suitable for fast-reconfigurable manufacturing environments. Another way to allocate tasks is by looking at trust [22]. Rahman et al. [23] consider the two-way trust of human trust in robots and human trust when creating a system that allocates and reallocates subtasks based on trust levels. Real-time trust methods and computational models of trust are used to calculate trust levels in both humans and robots. Subtasks are allocated using these calculations and reallocated in real time if trust levels drop. The results show that the allocation of subtasks and re-allocation based on trust lead to improved assembly performance compared to considering human trust alone or no trust at all.

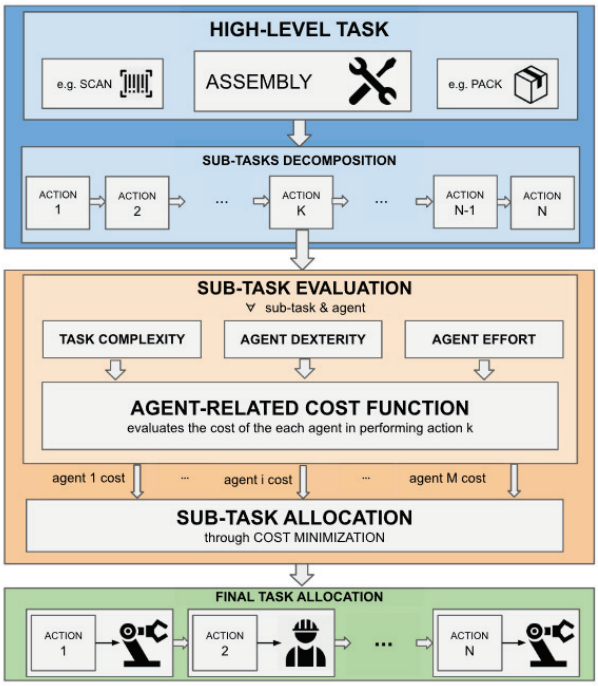


Figure 3. Task allocation graphic from Lamon et al. The diagram illustrates the process from high-level tasks to the designated sequence of actions. Each high-level task is broken down into a series of actions needed to accomplish it. The algorithm then assesses the suitability of each agent for performing these actions based on specific metrics. Finally, the actions are allocated to the agents to minimize the overall performance cost [22].

3.1.2. Communication and Instruction

Instruction and communication are crucial in any industrial task, but they are especially important when involving robot entities that collaborate with humans. In typical HRI settings, most communication occurs with a human operator who controls the robots in the environment and instructs all human workers. When working with teleoperated robots, the case is different. Teleoperated robots are controlled by a human operator from a different location. Oftentimes, teleoperated robots enter locations that are not suitable for humans like underground caves or sewage systems. These robots are fully operated from a machine, and communication can be difficult when the robot's environment is dark and cannot be properly viewed through its camera [24]. Glassmire et al. [25] add force feedback to teleoperated humanoid robot astronauts. The addition of haptic feedback was proven to improve the performance of the operator and decrease completion times.

3.1.3. Interaction

When humans and robots work together, it often involves taking turns or working on separate tasks, resulting in rigid actions and limited flow. In contrast, when humans work with other humans, they adapt to each other and begin to predict each other's actions, creating a more seamless and efficient workflow [26]. This type of flow is often attempted to be replicated by making robots autonomous using various systems that involve prediction and adaptation. Hoffman et al. [26] introduce an adaptive action selection system to robots that allows for anticipatory decision-making. The robots adapt to the human workers and use predictions to automate their next steps. The researchers found an increase in fluency in Human–Robot Interaction when the robots used anticipatory actions through a higher increase in concurrent motion and a decrease in time between human and robot action. Additionally, participants who interacted with robots that did not use anticipatory actions were often annoyed by the robots' lack of predictive behavior. Lasota et al. [27] conducted a similar study in which participants worked with an adaptive robot that used human-aware motion planning. Fluency between humans and robots was evaluated along with human satisfaction and perceived safety. The results concluded that team fluency increased with the adaptive robot, leading to faster task completion times, decreased idle times, and increased concurrent motion. Additionally, participants reported feeling comfortable and satisfied working with the adaptive robot. Yao et al. [28] used a different approach to increasing autonomy in robots: they used a Cyber-Physical Production System (CPPS) and IEC 6199, International Electrotechnical Commission (IEC), Geneva, Switzerland, 2012. [29] function blocks. The system was designed to allow for robot autonomy during various Human–Robot Interaction scenarios. Figure 4 shows eight categories of Human–Robot teamwork that are addressed by this system. The feasibility of this proposed system was tested in an assembly line case and proven to have potential capabilities to improve HRC fluency and increase the flexibility of manufacturing systems in future work.

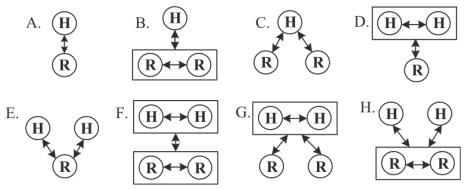


Figure 4. Different categories of human–robot teamwork [28]. In each case, a human is depicted by a circle labeled with an "H," whereas a robot is illustrated by a circle containing an "R." Double headed arrows indicate command flows between the humans and robot

Another process to increase autonomy in robots is through the use of digital twins. Digital twins are the digital representations of physical objects. In HRI, digital twins are typically used to model robotic mechanisms and simulate changes in tasks and manufacturing scenarios before implementing them in real life [30]. This virtual visualization enables workers to devise optimal solutions and saves time, money, and effort, as not

all changes need to be physically tested [31]. Liu et al. [31] creates a digital-twin-based design platform using a quad-play CMCO (configuration design–motion planning–xontrol development–optimization decoupling) model. The platform is used to design flow-type smart manufacturing systems that can be easily adjusted. A completed case study demonstrated that the digital-twin-based platform is both efficient and feasible. Rosen et al. [32] explored the topic of digital twins increasing the autonomy of robots in manufacturing settings. Information gathered by sensors or generated from the system is stored within a digital twin, allowing it to understand and represent the full environment and process state. The researchers propose that this information, along with the models of the digital twin, can be used to create simulations that anticipate the consequences of actions. This capability can enable the development of action-planning autonomous robots.

3.2. Trust and Safety

Trust and safety are important aspects of HRI [20,33,34]. When working with automated machines, it is necessary to ensure the safety of humans. Safety leads to human trust, which in turn increases fluency and productivity [20]. Neglecting the importance of safety in HRI can lead to serious injuries. Collision mitigation or avoidance is a necessary and critical system in HRI, especially Human–Robot Collaboration. One way to avoid injuries is through Situational Awareness (SA), which ensures that both humans and robots are aware of each other's locations at all times. Another key approach is Speed and Separation Monitoring (SSM), a fundamental safety strategy that dynamically adjusts robot operations based on real-time proximity and the relative velocity of humans in a shared workspace. This method is guided by standards such as ISO 10218 [35] and ISO/TS 15066:2016 [36], which outline the requirements for safe SSM. Additionally, trust in robots can be enhanced by introducing social attributes like facial expressions, which help create a sense of bonding between humans and robots. The presence of robust safety measures like SSM and SA, combined with the introduction of social attributes in robots, can significantly boost human workers' trust and acceptance of robotic systems. When workers feel confident that their safety is prioritized, they are more likely to engage positively with robots, leading to better collaboration and higher productivity.

Haddadin et al. [37] systematically evaluated the safety in HRI looking at real-world threats, dangers and injuries. The paper also included a summary and classification of possible human injuries when working with robots, including soft-tissue injuries and blunt trauma. Topics like trust and safety can be very broad and difficult to define, as they encompass human emotions, behavior, and psychology. Lee et al. [38] provides a detailed paper defining trust, its various aspects, and strategies for designing automation devices, such as robots, to gain trust.

A key challenge in HRI is measuring or evaluating trust and safety. Qualitative data, often collected through questionnaires, is typically used to assess perceived safety or trust. For example, Sahin et al. [33] conducted an experiment on perceived safety when working with robots, gathering participant data through questionnaires during and after trials. Similarly, Mauruta et al. [39] also used qualitative measures in the form of questionnaires to measure trust in their series of HRI experiments. On the other hand, Hancock et al. [40] explored various quantitative methods for measuring perceived trust in HRI. Their survey paper analyzed numerous papers and categorized quantitative measures into three sections: human-related, robot-related, and environmental. They concluded that robot-related measurements were most commonly associated with determining trust in HRI. There is no single correct method for measurement; qualitative and quantitative measures can be combined. For instance, Soh et al. [34] used both quantitative and qualitative methods to measure perceived safety, cross-checking quantitative data with qualitative questionnaire responses to ensure consistency.

3.2.1. Situational Awareness

A significant area of research concerning safety and trust in HRI is Situational Awareness (SA). In HRI, SA involves both the human being aware of the robot's position relative to themselves and the robot being aware of the human's position relative to itself [41]. SA is crucial for establishing trust, safety, and effective collaboration between humans and robots. Additionally, SA directly corresponds with staying "in the loop". Low Situational Awareness (SA) can lead to an "out-of-the-loop" problem, where operators are unable to effectively take control when automation and robots fail. This lack of awareness hinders their ability to respond promptly and accurately in critical situations, compromising safety and efficiency [42]. Situational Awareness can improve collaboration between humans and robots, but the way it is displayed is crucial to its benefits. For instance, in an experiment by Unhelkar et al. [43] comparing human and robotic assistants, a red light was used to indicate when the robot was getting close. This visual cue frightened the workers and put them on edge. As a result, they preferred the human assistant over the robot due to the unsettling nature of the red light. When it comes to measuring SA, the most prominent technique is the Situational Awareness Global Assessment Technique (SAGAT). It was developed by Mica R. Endsley [44,45] back in 1988 and is still used today. The SAGAT technique involves random pauses of the machine so the user cannot predict the robot's whereabouts or next moves. This allows for the comparison of the human's perception of the robot's location to the robot's actual location. This method is often paired with questions to the user about the robot's location or action. Unhelkar et al. [43], show a series of awareness questions asked during their user testing to measure the human's Situational Awareness. Newer techniques for measuring SA include the Situation Present Assessment Technique (SPAM) and the Tactical Situational Awareness Test (TSAT) for the small-unit tactical level [46,47].

3.2.2. Speed and Separation Monitoring

Speed and Separation Monitoring (SSM) is a safety mechanism in Human–Robot Collaboration that continuously tracks the distance between humans and robots. When a human enters a designated safety zone, the system prompts the robot to reduce speed or halt, thereby preventing collisions and ensuring a safe working environment. The framework for SSM, as established by Marvel and Norcross, underscores the importance of calculating the minimum protective distance at which a robot needs to stop, using parameters such as human speed, robot speed, and the robot's stopping time (TS). This is further enhanced by implementing external observer systems that continuously monitor the workspace, ensuring the robot can initiate a controlled stop if the separation distance is breached [48]. Complementing this approach, Kumar et al. [49] demonstrated the efficacy of using Time-of-Flight (ToF) sensor arrays mounted on robot links to dynamically adjust the robot's operational speed based on real-time distance measurements and relative velocities between the human and the robot. This methodology not only improves safety but also integrates self-occlusion detection to filter out false readings, thus enhancing system reliability. Additionally, Rosenstrauch et al. [50] highlighted the use of Microsoft Kinect V2 for skeleton tracking, which continuously detects and tracks human workers within the shared workspace. This allows for dynamic adjustments in the robot's speed, ensuring compliance with ISO/TS 15066 standards. Furthermore, Ganglbauer et al. [51] introduced a computer-aided safety assessment tool that incorporates the real-time digitization of human body parts and dynamic robot models to estimate potential contact forces, thereby providing immediate feedback on safety risks. Integrating these advanced SSM methodologies into HRI adheres to international safety standards and fosters safe and efficient working environments.

3.2.3. Social Attributions

A smaller but still important area of research within safety and trust in HRI is promoting social interactions between humans and robots. This includes exploring how humans

socially work with robots, adding emotional cues and social body language to robots, and understanding mental stress when working with robots. Through a Wizard of Oz and video ethnographic study at an elementary school, Oh and Kim [52] discovered that over time, children develop an emotional connection with robots. The character resembling robots displayed a variety of facial expressions, as shown in Figure 5. There were a total of 30 emotions that included sorrow, anger, fear, delight, curiosity, indifference, and more. The emotional bond created with robots can lead to the long-term use of robots, thereby extending their typically short lifespans. A similar emotional connection was created between workers and a robot co-worker in a study conducted by Sauppe et al. [53]. This study concluded that the social attributes of the robot facilitated this emotional connection, which in turn made the workers feel safe when near the robot. They began perceiving the robot as they would a human co-worker. Additionally, Bruce et al. [54] explored how robots that express facial features and engage with humans socially, such as by turning towards them, are more compelling to interact with and thereby increase trust within humans. The mental stress of human workers is also a crucial factor to consider when aiming to ensure safety and trust around robots. This is important because mental stress directly correlates with productivity and comfort. Lu et al. [55] documented a number of measurements that can be used to gauge the mental stress of workers when collaborating with robots. The addition of both monitoring mental stress and giving robots social attributes can improve the interaction between humans and robots, fostering trust and creating safer workspaces.

Although understanding the essential factors that contribute to effective HRI is crucial, it is equally important to recognize the challenges that persist in this field, which the following section addresses.

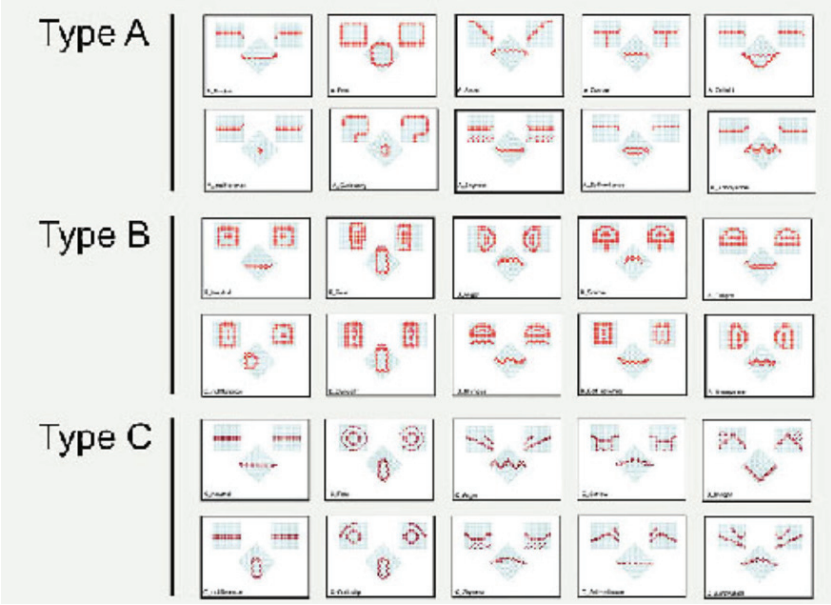


Figure 5. Facial expression range of the robot used by Oh and Kim [52].

4. Challenges in HRI

The field of HRI faces several key challenges that must be addressed to ensure effective and safe collaboration between humans and robots. Kumar et al. [20] identify the three main challenges in HRC as human safety, human trust in automation, and productivity. This paper identifies similar challenges, which include safety concerns, low levels of communication and collaboration, and the necessity for clear instruction and planning.

- **Safety Concerns**: Improving both human and robot awareness of each other's actions and intentions to prevent collisions and minimize risks of injuries.
- **Low Levels of Communication and Collaboration**: Building and maintaining trust in robotic systems through predictable, transparent actions.

- **Necessity for Clear Instruction and Planning**: Overcoming the limitations of pre-planned or rigid robot actions by enabling more dynamic, interactive, and responsive robot behaviors.

Although recent research offers innovative solutions and improvements, HRI still struggles with the need for constant feedback to enable natural interactions. Currently, most HRI scenarios involve pre-planned actions on computers or adaptive robots that lack real-time interactive capabilities.

Communication and collaboration between humans and robots also remain inadequate. Effective HRI requires seamless interaction, where both parties can understand and predict each other's actions. However, current systems often lack the capability to provide dynamic, ongoing communication. Rather, the communication is one-sided, with only the robot predicting, while the human cannot see its predictions or upcoming actions [26–28]. This deficiency hinders the fluidity of teamwork and reduces overall efficiency, leading to misunderstandings and delays. Figure 6 (left) depicts the robot's workspace setup where human interaction takes place. On the right, it illustrates how the robot from Lastoa et al.'s [27] experiment predicts human actions. While the most likely and frequently used paths are predicted, this method does not account for random actions or movements by humans. Additionally, the human cannot predict the robot's path, which could further assist in preventing collisions.

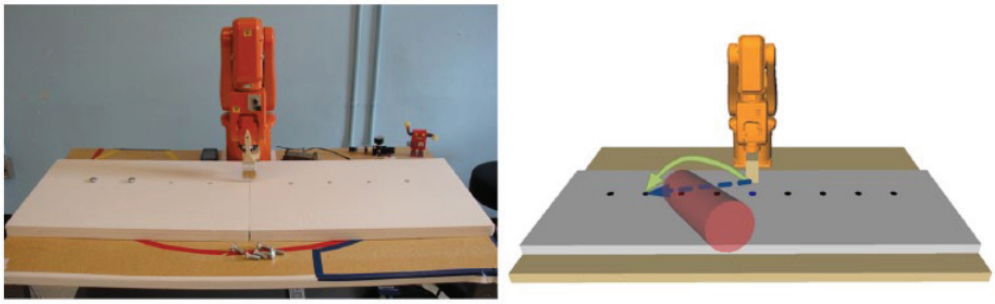


Figure 6. (**Left**) A collaboration workspace between a robot and a human. (**Right**) The robot's prediction model, displaying two types of robot movements: a conventional shortest-path motion (dashed arrow) and a human-aware motion (solid arrow). It also includes the predicted path of a human worker [27].

Furthermore, clear instruction and planning are crucial for optimizing HRI. Robots typically require precise instructions to perform tasks accurately. However, in dynamic and fast-paced environments, the ability to provide and update these instructions in real time is limited [56,57]. This limitation results in inefficiencies and increased cognitive workload for human operators, who must continuously monitor and adjust the robots' actions.

Lastly, safety concerns are important in HRI, as the close physical interaction between humans and robots in environments such as manufacturing and assembly lines introduces the risk of collision. Without real-time feedback and communication, it is challenging to ensure that both humans and robots can anticipate and respond to each other's movements, potentially leading to hazardous situations [58–61]. Although solutions such as flashing lights have been created to show the whereabouts of the robots, unless the light is always in constant view of the user, it cannot always be useful and reach its full potential for ensuring safety and Situational Awareness [43]. Figure 7 shows an example of lights near a collaborative robot that lights up to show the robot is in motion.

The next section demonstrates how Augmented Reality can enhance HRI and provide solutions to the challenges that currently exist in the field, showcasing relevant research that addresses these issues.

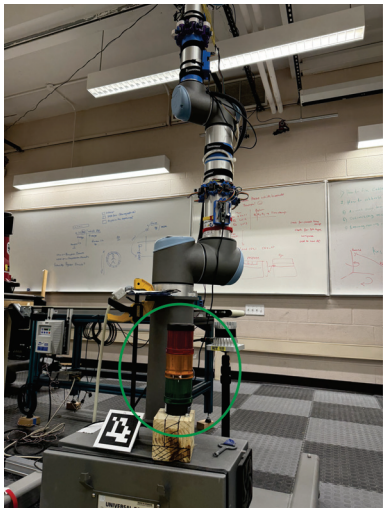


Figure 7. Operation lights on a collaborating robot circled in green. The lights represent the separation distance. Green is safe, red is unsafe, and amber is warning.

5. Augmented Visualizations Used to Improve HRI

The integration of Augmented Reality (AR) presents a promising solution to the challenges in HRI. AR combines virtual elements overlaid on a real-world environment. It is typically used to convey important information needed while performing a task or viewing an environment [4,11]. AR enables humans to receive immediate visual feedback about the status and intentions of robots. This capability allows for more intuitive and responsive interactions. Additionally, AR has been integrated into a variety of fields including medical, entertainment, military, product design, and manufacturing [4]. More recently, it has been used in HRI to improve collaboration and feelings of safety between robots and humans [62,63]. Within HRI, AR is typically presented in the form of wearable devices, such as head-mounted displays (HMDs) and smartwatches. It is also seen in the form of projections onto real-life surfaces [64]. This section will explore research demonstrating that AR can enhance planning and training, communication and instruction, and interaction within the field of HRI. Additionally, the exploration of AR to increase Situational Awareness and collision mitigation will be examined to understand how it improves trust and safety. Table 1 below categorizes all of the papers reviewed in this section.

Table 1. Summary of AR Applications in Human–Robot Interaction.

Reference	Category	Use of AR	Hardware
Fang et al. [65]	Collaboration	Path Plan	Display
Doil et al. [66]	Collaboration	Task Plan	HMD+
Wang et al. [67]	Collaboration	Task Plan	HMD
Andersson et al. [68]	Collaboration	Task Train	HMD
Saaski et al. [57]	Collaboration	Task Instr	HMD
Michalos et al. [69]	Collaboration	Task Instr	HMD+
Liu et al. [31]	Collaboration	Task Instr	Display
Kalpangam et al. [8]	Collaboration	Task Instr	Projection
Lunding et al. [70]	Collaboration	Communication	HMD
Tabrez et al. [71]	Collaboration	Communication	HMD
De Franco et al. [72]	Collaboration	Task Collab	HMD
Andronas et al. [73]	Collaboration	Interact. Cues	HMD
Gkournelos et al. [74]	Collaboration	Interact. Cues	Smartwatch
Qui et al. [75]	Collaboration	Interact. Cues	Web
Sonawwani et al. [76]	Collaboration	Interact. Cues	Projection
Tsamis et al. [59]	Trust and Safety	Visual. Actions	HMD
Palmarini et al. [58]	Trust and Safety	Visual. Actions	Tablet

Table 1. *Cont.*

Reference	Category	Use of AR	Hardware
Vogel et al. [60]	Trust and Safety	Safe Zone	Projection
Choi et al. [61]	Trust and Safety	Safe Zone	HMD
Hietanen et al. [77]	Trust and Safety	Safe Zone	HMD+
Lunding et al. [78]	Trust and Safety	Safe Zone	HMD
Eschen et al. [79]	Trust and Safety	Safe Maint	HMD
Papanastasiou et al. [80]	Trust and Safety	Safe OS	HMD+
Makris et al. [81]	Trust and Safety	Safe OS	HMD
Matsas et al. [82]	Trust and Safety	Collis. Avd	Projection

Abbreviations: **Path Plan**—Path Planning, **Task Plan**—Task Planning, **Task Train**—Task Training, **Task Instr**—Task Instruction, **Task Collab**—Task Collaboration, **Interact. Cues**—Interaction Cues, **Visual. Actions**—Visualizing Actions, **Safe Zone**—Safety Zone, **Safe Maint**—Safety Maintenance, **Safe OS**—Safety Operating System, **Collis. Avd**—Collision Avoidance, **HMD+**—Head-Mounted Displays and Screens.

5.1. Collaboration Utilizing AR

Similar to traditional HRI collaboration, collaboration in HRI using AR focuses on three main components: planning and training, communication and instruction, and interaction. The key difference is that AR is used to enhance the features and systems involved in these components, ensuring more efficient collaboration. Additionally, there are various ways in which AR can be integrated into the collaboration, expanding the interaction between humans and robots to include how AR interfaces with both. Suzuki et al. [5] provide a detailed categorization of the various approaches, interactions, and characteristics commonly found in AR applications within HRI. They categorize the three main approaches of how AR works with humans and robots as on-body, on-environment, and on-robot. This subsection will explore how AR is used in the field of HRI to improve planning and training, communication and instruction, and interaction, ultimately making collaboration more effective.

5.1.1. Planning and Training Using AR

Recent research has found that AR enriches path planning and training by offering a virtual approach rather than a physical one, making operations safer, easier, and more flexible [65]. AR can be used to understand a robot's status and planned actions, discuss and review plans with robots, and generate optimal plans before implementing them. Fang et al. [65] utilized AR to plan the path of an industrial robot. Despite encountering accuracy errors in their case study due to the application system, the concept of AR path planning was generally considered a success. Doil et al. [66] explored the use of AR in planning manufacturing tasks by placing virtual robots and machines within a real-world workspace to map out a manufacturing layout. This approach eliminates the need to model the work environment as required in a VR-based version, streamlining the planning process. Similarly, assembly workstation planning using AR is explored by Wang et al. [67]. The researchers developed an AR system that enables HRC assembly simulation based on current human worker mapping and motion data. This system allows operators to design optimal HRC assembly lines that mimic human assembly lines efficiently and cost-effectively. In HRI training, AR technology often provides virtual simulations of real-life interactions and situations, helping individuals better understand and become comfortable working with robots. Bischoff et al. [83] identified robot operation training as the most promising area of AR research in this field. When training with AR, human operators can safely practice complex tasks. Andersson et al. [68] developed a system that facilitates AR-based training, enabling this safe and effective learning environment. The system is divided into four main parts, interconnected via a Robot Operating System (ROS), to ensure flexibility and extensibility. The multimodal robot programming toolbox allows for various programming tools to interact with a data model through parameter requests, enabling the user to select desired tools during runtime, such as PbD or free drive mode,

to define robot poses or trajectories. This is presented visually in Figure 8. In contrast, Matsas et al. [84] present a similar training system that utilizes VR and is entirely virtual, without incorporating real-world elements.

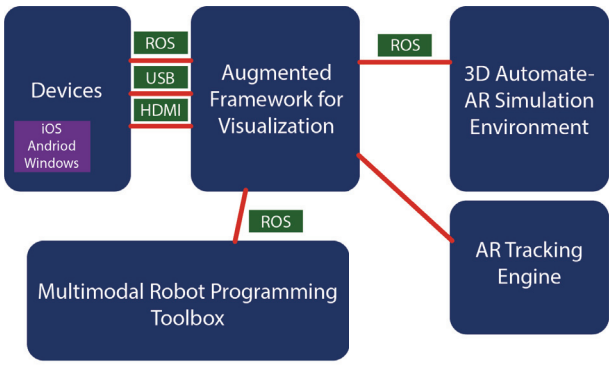


Figure 8. Visual representation of the training solution by Andersson et al. [68].

5.1.2. Employing AR for Communication and Instruction

In HRI, AR can be used both for communication and instruction. When giving instructions, AR is used for robots to give humans instructions proactively during work. When receiving instructions, AR can be used to view instruction plans during manufacturing and project instructions onto objects in assembly work. In an AR manufacturing case study by Saaski et al. [57], researchers found that participants completed assembly tasks faster when instructions were displayed virtually using AR and an HMD display rather than paper instructions. Participants identified the 3D animations of parts being assembled as the most useful AR feature. AR systems that provide instructions are not only helpful for workers but also for operators who oversee entire workplaces, including both human and robot workers. Michalos et al. [69] created an AR system that supports operators by allowing them to send AR instructions and maintain an overview of the workplace to ensure safety and display additional information through AR visualization. The results of the study demonstrated that the system significantly enhanced working conditions and the integration of the operator. Additionally, AR systems can support robots in giving instructions to human workers. Liu et al. [85] designed an AR system that virtually augments instructions intuitively to humans. These instructions are provided by robots through recognizing human motions and determining the next best course of action.

When using HMDs to display instructions and other important information, it is vital to ensure that virtual screen coordinates, eye level, and calibrations are properly adjusted to the wearer. Janin et al. [86] provide detailed procedures for finding the correct calibration parameters for displaying visuals on HMDs. Additionally, it is important to note that although utilizing HMDs for AR instructions can lead to positive results, they may not yet be ready for real-world applications. For example, an evaluation study by Evans et al. [56] concluded that while the Microsoft HoloLens has the potential to display AR instructions in an assembly line, it still requires improvements in accuracy before it can be applied in a real factory setting. In addition to just HMDs, Lunding et al. [70] used a hybrid approach to provide instruction and enhance communication. Instead of relying solely on HMD-based AR, they incorporated a web interface to achieve a more accurate visual representation when displaying assembly instructions. Additionally, this hybrid method establishes two-way communication between the operator and the robot. Aside from HMDs, projection mapping can also be used to display instructions and informational cues from the robot to a human worker. Kalpangam et al. [8] used vision-based object tracking to map out an environmental space, determine the location of various objects and then utilized projection mapping to overlay important cues and instructions onto tracked objects. Figure 9 illustrated the steps to aligning a car door projected over the physical door. A projected circle on the floor moved in real-time with the car door until it

reached the correct alignment. Once aligned, the circle provides feedback to let the user know that the task has been successfully completed.

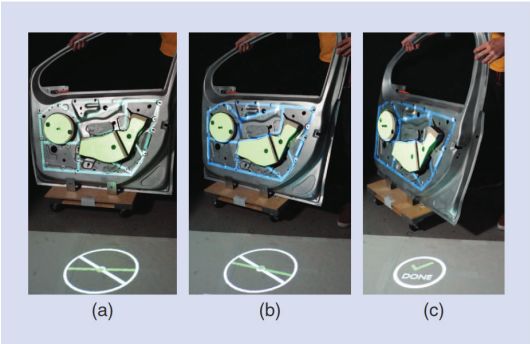


Figure 9. Projected instructions for car door alignment. (**a**) The starting projection for a car door alignment task. (**b**) The projection moving as the car door aligns. (**c**) Ending projection with the "Done" visualization [8].

Communication in HRI is vital for making informed decisions and adapting to unknown situations. AR-based communication and location guidance between robots and humans is explored by Tabrez et al. [71]. They use AR for both prescriptive and descriptive guidance techniques. For teleoperated robots, an AR interaction system can address the challenges of maintaining awareness due to a single ego-centric view. Green et al. [24] explore how AR enables operators to view a teleoperated robot through a 3D visual representation of the work environment and anticipate the robot's actions. This capability facilitates improved communication and collaboration between the operator and the teleoperated robot.

5.1.3. AR Promoting Interaction

Interaction in HRI often uses AR to promote fluency and productivity between humans and robots. This typically involves conveying the robots' upcoming plans and actions, allowing users to interact with robots intuitively. De Franco et al. [72] take a similar approach to promote natural interaction between robots and humans by creating an AR system that allowed humans to view the robot's status and future actions for collaborative tasks. Using an AR headset, the robot provided feedback to the human user, enabling them to understand the upcoming steps and be prepared. This improved their performance and was deemed helpful by 10 participants [72]. Andronas et al. [73] also promoted natural interactions using a Human–System and System–Human interaction framework. This framework allows robots and humans to communicate important information back and forth while completing their tasks. In general, the system increased operator trust and awareness and reduced human errors. Similar interaction systems can also be operated through smartwatches. Gkournelos et al. [74] promoted HRI interactions through a hybrid approach involving smartwatches. This method significantly aided the operator in integrating into an assembly line. Another approach to Human–Robot Interaction using AR is by allowing robots to manipulate virtual elements in a shared Augmented Reality workspace. In this setup, robots do not assist in the physical world but rather work proactively with humans by offering cues in the virtual augmented world. Qiu et al. [75] developed an experiment to test this setup and found that it worked effectively and could be useful for the future of HRC. Outside of the virtual world and elements, the AR interaction can also be projected onto a work surface to provide information about the robot's upcoming actions and intentions. Sonawami et al. [76] used projection mapping to display visual cues, highlighting the objects that human beings need to interact with. As seen in Figure 10, the robot's shadow is projected, highlighting its status. Additionally, the colored semicircles show users which blocks are safe to pick up. The results show that projection mapping improved safety and task efficiency.

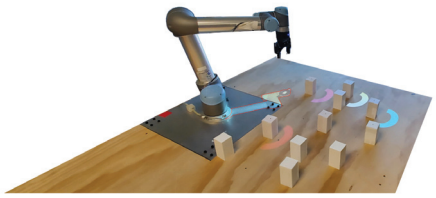


Figure 10. AR projections by Sonawani et al. that show a robot's status and interaction elements [76].

5.2. Applying AR to Improve Trust and Safety

AR can enhance trust and safety in HRI through two main approaches: prevention and correction. Prevention is achieved through Situational Awareness, where humans are actively shown the next steps and whereabouts of the robot to avoid collisions. This approach helps both robots and humans understand each other's locations, enabling safe interactions and reducing the likelihood of being taken by surprise if they are too close to each other. Collision mitigation and avoidance includes maintenance, which refers to the upkeep of a robotic system or environment to ensure it complies with safety protocols and functions appropriately. It also includes collision avoidance through the operator's overall view and management of an active HRI environment and the ability for robots to stop automatically if humans get too close to avoid collisions. In this subsection, we will examine research that demonstrates how AR systems can be integrated into HRI through Situational Awareness and collision mitigation to increase trust and safety for users in HRI environments.

5.2.1. Situational Awareness with AR

When it comes to AR, Situational Awareness is a widely researched topic in many different fields, including medical, military, and manufacturing [9,87]. In HRI, many researchers aim to promote safety by increasing user awareness of robots they are working with and their surroundings. These approaches may be similar to those discussed in the interaction section, but are grouped here because they are specifically designed to enhance Situational Awareness for safety and trust rather than interaction.

The first approach to improving Situational Awareness involves visualizing the intended path or motion of the robot. Both Tsamis et al. [59] and Palmarini et al. [58] use AR to show users the robot's next actions. Their goal is to help users become more context-aware and trust the robots more. Although Tsamis et al. [59] used an HMD and Palmarini et al. [58] used a handheld tablet, both experiments resulted in positive sentiments and increased perceived trust from participants. Figure 11 shows the AR seen through an HMD. The green tube indicates the robot's next path of motion, while the red bubble encloses the danger zone, where humans should be cautious of collisions.

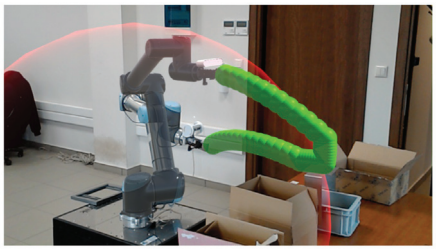


Figure 11. The AR visualization seen through an HMD that visualizes the robots' upcoming motions in green and the safety zone in red [59].

The second approach is similar to the red bubble; it is when AR is used to visualize safety zones. This allows users to see the area in which a robot operates and know when they are too close. Vogel et al. [60] displayed safety zones using multiple projectors around

the room, while Choi et al. [61] used an HMD and calculated the 3D locations of both the human and the robot using digital twins. Neither of these papers evaluated their system designs; instead, they focused on design architecture and making the system functional. Hietanen et al. [77] used both HMD and projection mapping to display safety zones. Both types of visualizations were tested in a user study. The AR in the HMD and the projection mapping both reduced task completion time and robot idle time. However, the projector was proven to show more noticeable improvements in safety.

Lunding et al. [78] combined both approaches by displaying robot safety zones and upcoming actions using an HMD. Their system, called RoboVisAR, also allows for users to easily customize the visualizations shown in the AR. The papers discussed employed one of three different methods to present visualizations of Situational Awareness. While each method was found to enhance SA, some approaches are proven to be more effective than others [12]. Generally, HMDs are found to improve efficiency. Furthermore, when compared to handheld displays and radio links, HMDs have been shown to offer better performance and reduce workload [6].

5.2.2. Using AR for Collision Mitigation and Avoidance

AR can make HRI conditions safer, but there are also mental and physical risks associated with the use of AR in environments with potentially dangerous robots. Bahaei and Gallina [88] created a framework called FRAAR (Framework for Risk Assessment in AR-equipped socio-technical systems) that can be used for the assessment of the risk of both social and technical issues that can arise when AR is introduced into any robot manufacturing system. Safe practices can also mean the maintenance and upkeep of a running system. Eschen et al. [79] utilized AR for the inspection and maintenance in the aviation industry. In a broader sense, Papanastiou et al. [80] and Makris et al. [81] both developed AR systems designed for operators to maintain safe and integrated environments in human–robot manufacturing settings. The accuracy of these safety measures is usually not precise, but instead, they are activated early to avoid human–robot collisions. This means that robots will be interrupted from their work and forced to stop moving very often, which can hurt the flow of a workspace. To minimize interruptions and maximize workflow, Matsas et al. [82] used proactive and adaptive techniques in a virtual immersive environment to test the performance metrics of a robot's safety enhancements. This approach aims to obtain precise measurements of how close a human can be before a robot is forced to stop. A virtual human body model is created and analyzed to determine the appropriate distances from the body that should trigger the robot's motion-stopping feature. Figure 12 shows the diagram of this body model.

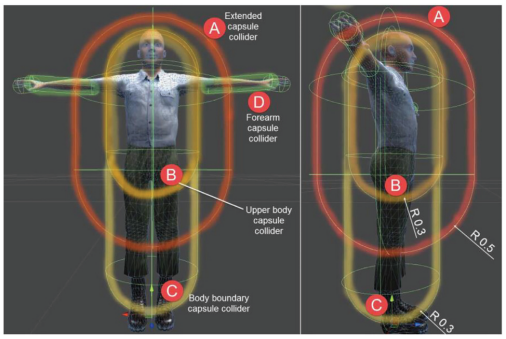


Figure 12. Diagram of the body model used by Matsas et al. to calculate the distance at which a robot should automatically shut down to avoid human collision. Different sizes of geometric primitives are used to create estimations of the regions of the human body, this can be seen in A through D [82].

The use of digital twins in Human–Robot Collaboration has shown significant promise in improving safety and mitigating collisions. According to Feddoul et al. [89], digital twins facilitate real-time monitoring, simulation, and prediction of potential safety hazards,

providing a comprehensive approach to preemptively address safety issues. Similarly, Maruyama et al. [90] proposed a twin-based digital safety control framework that focuses on accurate distance calculations and real-time safety monitoring, ultimately preventing collisions by maintaining a safe distance between humans and robots. Lee et al. [91] detailed the application of digital twins in collaborative robotic systems through the implementation of safety controllers and validation techniques, focusing on their role in offline testing, real-time monitoring, and predictive analysis. It contains a good overview of different types of simulation environments used for different types of collaborative applications. As we examine the impact of augmented visualizations on improving HRI, it becomes clear that good design principles are critical to maximizing their effectiveness, guiding the development of user-friendly AR systems.

6. Good Design Principles for AR

Although there is immense research and development of AR in HRI, there is a lack of focus on visual information, with more emphasis placed on information architecture [9]. The way information is displayed is crucial because it directly affects how users comprehend and act on it. To properly design how information is presented in AR, we must look to Human–Computer Interaction (HCI) design principles. HCI is a multidisciplinary study focused on understanding how humans use and interact with technology, and it plays a crucial role in design and research [92]. AR is a highly popular area in HCI research across various fields, including mental health, accessibility, automotive, education, and gaming. Many of these studies examine how AR interfaces can be improved for efficiency, user understanding, and enjoyment. Positive results from HCI work in AR demonstrate its potential to enhance user experience and engagement [93–95]. Integrating HCI principles into HRI is essential, as it can lead to more intuitive and effective interactions between humans and robots, ultimately improving user satisfaction and trust.

6.1. UI and UX Design

HCI differs from HRI particularly in terms of visual design, specifically looking at the user interface (UI) and user experience (UX) of AR features implemented in Human–Robot collaboration. From Norman [96] and Rogers [10] we can summarize the top principles for both UI and UX design. UI design refers to the visual elements users see and interact with, which include both static elements such as icons, colors, typography, and layout and interactive elements like buttons, scroll bars, and animations. The goal of UI design is to create an intuitive, easily learned, and easily navigated product. To create user-friendly interfaces, the top four principles to follow include the following:

1. **Simplicity**: Designs should be simple, clutter-free, and easy to navigate. Avoid unnecessary elements and keep the interface as straightforward as possible.
2. **Consistency**: Maintain internal and external consistency. The design elements should be consistent throughout the system and align with universal design standards. For example, a floppy disk icon universally signifies "save".
3. **Feedback**: Provide feedback through visual, auditory, or haptic means to indicate that an action has been successfully completed. This helps users understand the outcome of their interactions.
4. **Visibility**: Ensure that the cues are easy to see and understand. Visibility also refers to making the current state of the system visible to users. This helps users stay informed about what is happening within the system.

UX (User experience) design focuses on the flow of an app and how users interact with the system, rather than its appearance. It involves understanding users' behaviors, pain points, and motivations. Good UX design improves engagement and usability. The top four principles for good UX design as follows.

1. **Usability**: Ensure the user flow of the system is easy to navigate, learn, and use. The design should be intuitive and provide a seamless experience for users.

2. **Clarity**: The design should be straightforward and provide users with only essential information, avoiding unnecessary elements. Clarity ensures that users understand the interface and its functions easily.
3. **Efficiency**: The system should be designed for performance optimization, enabling users to accomplish their goals quickly. Efficient design minimizes the time and effort required to complete tasks.
4. **Accessibility**: Ensure that the system is accessible to all users, regardless of their abilities. This includes considering different devices, screen sizes, and assistive technologies to create an inclusive user experience.

When it comes to evaluating the interface and user experience of systems, including AR systems, two common subjective measures are the System Usability Scale (SUS) and the User Experience Questionnaire (UEQ) [97,98]. Both measures are well established with benchmarks and prior work involving user studies. The SUS asks participants to rate the usability of an interface using a five-point Likert scale. It includes 10 questions, each focusing on different aspects of UI usability [98]. The UEQ is a 26-question survey that uses a 7-point Likert scale. Participants rate the user experience of a system based on six attributes, including dependability and attractiveness [97].

6.2. User-Centered Design

A good practice in design is always to consider the user. This is known as User-Centered Design (UCD). It refers to designing with the user in mind, catering to their needs and considering their technology level and skill set. This approach ensures that the interface is accessible and user-friendly for the intended audience. UCD involves continuous user participation and feedback throughout the design process, ensuring that the final product effectively meets user requirements and provides a positive user experience [99,100].

Additionally, in HCI, there is a focus on designing for inclusion, accessibility and sustainability. These principles can be considered when designing AR visual elements for HRI to ensure that all workers can properly use and interact with the technology [101]. Furthermore, designing for diversity, equality, and inclusion can ensure unintended bias does not occur within the AR system [102]. This paper will not go into detail on the design guidelines needed for accessibility, sustainability, and design, but there are numerous resources that can be referenced, including design guidelines and previous work [103–106].

6.3. Current Design Principles Used for AR in HRI

Within HRI research, there are limited papers that acknowledge or explore HCI principles such as UI/UX design. Only four papers considered UX and UI design strategies when developing the AR systems used to improve HRI. Fang et al. [107] used interface design strategies when creating an AR system to aid in the seamless interaction between humans and robots. Their paper includes a section labeled "Visualization", which discusses each visual element in the AR and the purpose of its design. The user experience was evaluated based on how well the users understood the visual elements, their cues, and the information they were displaying. Alt et al. [108] directly examine principles for UI design, such as adaptability and workflow when creating a user interface to explain an AI-based robot program. User experience was assessed based on users' perceptions of the interface's usefulness, assistance, readability, and understanding. Similarly, when developing an AR head-mounted display for robot team collaboration, Chan et al. [7] designed the system with user workflow in mind. The system's user experience was evaluated based on task load and the increased efficiency of collaboration. Green et al. [24] is the only one of these papers to offer multiple design variations. They presented three different UI designs for their AR system to increase Situational Awareness, as shown in Figure 13. The user experience of these various UI designs was evaluated based on users' comfort and their ability to accurately read the robot's location at any given time. The papers above all evaluate user experience based on various qualitative and quantitative measurements that correctly correlate with what they are designing to improve. For example, Chan et al. [7]

used both the SUS and UEQ, mentioned above, to evaluave the usability and the user experince of the system design. There are also more universal methodologies for evaluating UX. Lindblom and Alenjung [109] created ANEMONE, an approach to evaluating UX in HRI. They conducted extensive research on HRI evaluation principles and UX evaluation principles, deriving from HCI, to develop this combined methodology that can be implemented into any AR user study in the field of HRI. Aside from incorporating HCI principles and evaluation methods, HRI research also includes a few papers regarding design guidelines, but they tend to be limited or highly specific. Jeffri and Rambli [110] offer guidelines for interface design in AR, but they focus specifically on the AR used in manual assemblies. Wewerka et al.'s [111] design guidelines are tailored for robotic process automation, and Zhao et al.'s [112] work is centered on the fabrication process. While these guidelines and tools can potentially be applied to design UI for AR in various HRI fields, they have only been tested in their respective domains. Consequently, their effectiveness in other fields remains uncertain. A better approach is to use HCI design guidelines for AR. These broader guidelines can be applied to any use of AR, including HRI. Since they are not specific to a particular industry, they offer flexibility and can be adapted to suit the specific needs and contexts of various designs, leading to more optimal outcomes [103].

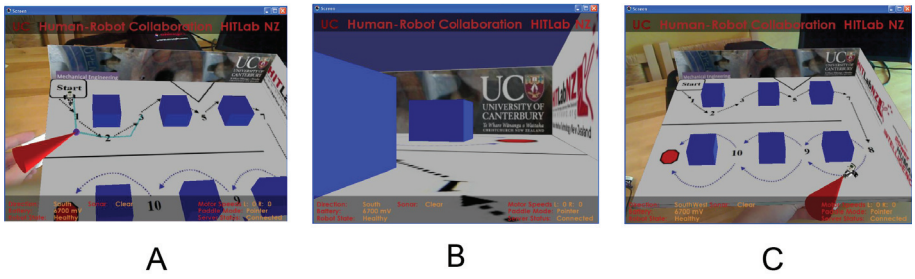


Figure 13. The user's view for three different UI screens created by Green et al. (**A**) Overview without planning, (**B**) immersive view, (**C**) overview with planning [24].

7. Evaluation and Continuous Improvement

Continuous evaluation and improvement are pivotal to ensuring that AR systems in HRI remain effective, user-friendly, and safe. This section elaborates on the methods and strategies for evaluating AR systems and outlines a process for continuous improvement.

7.1. Safety Metrics

Safety metrics are critical for evaluating the risk and potential hazards in HRI environments. They focus on ensuring that interactions between humans and robots are safe and free from accidents or injuries. These metrics assess how well the system maintains safe distances, prevents collisions, and handles emergency situations. By monitoring these metrics, it is possible to identify and mitigate risks, enhancing the overall safety and trust of human operators in robotic systems. The key safety and efficiency metrics for robot safety using speed and separation algorithms are defined in [113,114].

- **Collision Count:** The number of collisions or near misses between robots and humans. A lower collision rate indicates that the system effectively prevents accidents, ensuring the safety of human operators. Monitoring this metric helps quantify the direct impact of the AR-HRI system on safety. It is critical to track this to ensure that the implemented safety features are functioning as intended and to identify any areas needing improvement.
- **Safe Distance Maintenance:** The percentage of time the robot maintains a predefined safe distance from the human operator. This metric shows how consistently the system keeps humans out of harm's way. Maintaining a safe distance is crucial to prevent injuries and build trust between human operators and robots. Continuous

monitoring of this metric ensures that the system adapts to dynamic environments while prioritizing human safety.

- **Emergency Stops:** The frequency of emergency stops triggered by the system to prevent collisions. While emergency stops are necessary to prevent collisions, a high frequency may indicate overly conservative settings, which can disrupt workflow. Balancing safety and efficiency is key, and this metric helps in tuning the system. Tracking emergency stops helps in refining the sensitivity and response parameters of the system for optimal performance.

7.2. Efficiency Metrics

Efficiency metrics measure the productivity and effectiveness of the HRI system. They focus on how well tasks are completed. These metrics assess factors like task completion time, idle time, and path optimization. By evaluating efficiency, it is possible to determine if the system is not only safe but also effective in improving workflow and productivity in an HRI environment.

- **Task Completion Time:** The time taken to complete a task before and after implementing the AR-HRI system. Reducing task completion time while maintaining safety shows that the system is efficient [115]. This metric is vital for demonstrating that the system not only keeps humans safe but also enhances productivity. It provides insights into how well the AR integration and SSM are streamlining the workflow.
- **Idle Time:** The amount of time the robot is idle due to safety interventions. Minimizing idle time indicates that the system is effective without unnecessarily halting operations [115]. This balance is critical for maintaining a smooth and efficient workflow. By tracking idle time, one can assess the efficiency of the SSM algorithm in differentiating between real and false positives, thereby ensuring that the robot operates as smoothly as possible.
- **Path Optimization:** Changes in the robot's path efficiency, such as distance traveled and the smoothness of movements. Efficient path planning ensures that the robot performs its tasks optimally while avoiding humans [116]. Measuring path optimization helps in understanding the overall efficiency of the system. This metric is important for identifying any unnecessary detours or delays caused by the AR-HRI system and for optimizing the robot's navigation algorithms.

7.3. User Experience Metrics

User experience metrics evaluate the subjective perceptions and satisfaction of human operators interacting with robotic systems. These metrics focus on how safe, comfortable, and intuitive the interactions are for users. They often include assessments of perceived safety, workload, and emotional responses. Understanding user experience is crucial for ensuring that the system is user-friendly and meets the needs and expectations of its operators, leading to higher acceptance and better overall performance.

- **Perceived Safety:** Users' feedback on how safe they feel working alongside the robot. Perceived safety is crucial for user acceptance of robotic systems [117]. This qualitative metric helps to understand the psychological impact of the system on human operators. Gathering user feedback through surveys and interviews helps in assessing the effectiveness of the AR visualizations and the SSM in making users feel secure.
- **NASA TLX Scores:** Scores from the NASA Task Load Index, assessing perceived workload across six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration [44]. Assessing workload using NASA TLX provides insights into how the system affects the cognitive and physical demands on human operators [118]. Lower scores indicate a more user-friendly and less stressful interaction. This metric is essential for understanding how the AR-HRI system impacts overall user workload and for identifying areas that can be improved to reduce user strain.

- **SAM Scale Scores:** Measurements from the Self-Assessment Manikin (SAM) scale [119], assessing user arousal and valence. The SAM scale provides a subjective measure of users' emotional responses [120]. These scores can be used for understanding the feelings felt by the workers during interaction with the robots while using AR. Positive emotional experiences are essential for user satisfaction and acceptance. Correlating these scores with physiological data (described in the next section) can provide a holistic view of user experience and emotional state [121]. Figure 14 is a visual depiction of how subjects would convey their subjective state.
- **User Experience Questionnaire (UEQ):** The UEQ consists of 26 questions, using a 7-point Likert scale where participants evaluate the user experience of a system across six attributes, including dependability and attractiveness [97].

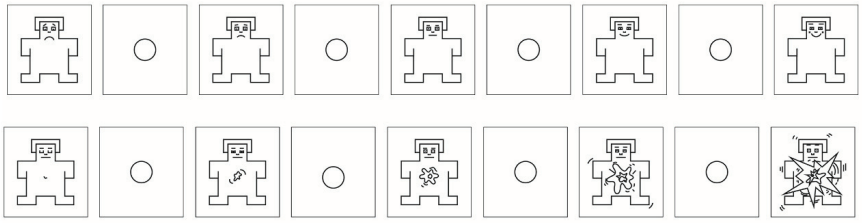


Figure 14. SAM used commonly to obtain subjective responses from human subjects [119].

7.4. Physiological Metrics

Physiological metrics involve measuring the physical and emotional responses of human operators through biometric data. These metrics can include heart rate, skin conductance, and other physiological signals that indicate stress, arousal, or emotional states. By monitoring these physiological responses, it is possible to gain insights into how the HRI system affects users' well-being and to make necessary adjustments to improve their comfort and performance. Savur and Sahin in [122] compile methods of physiological computing used in Human–Robot Collaboration.

- **ECG and Galvanic Skin Response (GSR):** Data from physiological sensors that relay electrocardiogram (ECG) and galvanic skin response (GSR) signals from the human operator can be used to train models that estimate arousal and valence from ECG and GSR signals [117]. These models can also provide data on the emotional states of users, which can be correlated with subjective measures from the SAM scale to validate the user experiences. Implementing these models helps in real-time monitoring and analysis of user states, enabling proactive adjustments to the AR-HRI system to improve user comfort and performance.

7.5. System Performance Metrics

System performance metrics evaluate the technical effectiveness and reliability of the HRI system. These metrics measure response time, accuracy, and the system's ability to detect and respond to issues. High system performance ensures reliable, efficient, and safe operations.

- **Response Time:** The time it takes for the system to detect a potential collision and respond. Faster response times are crucial for preventing accidents. This metric demonstrates the system's efficiency in real-time monitoring and intervention. Reducing response time is essential for ensuring that safety interventions are timely and effective, thereby minimizing the risk of accidents.
- **Positional Accuracy:** The accuracy of the system in detecting the positions of both the robot and the human. High positional accuracy is essential for the effective functioning of the system. This metric ensures that the system can reliably monitor and react to the positions of humans and robots. Accurate position tracking is vital for the SSM algorithm to function correctly and for providing precise AR visualizations.

Building upon AR in HRI, HCI design principles, and evaluation metrics, we propose a comprehensive framework for developing AR applications in HRI. This framework ensures that these systems are both effective and adaptable to the needs of human operators.

8. Framework for Developing Augmented Reality Applications in Human–Robot Collaboration

The framework for designing AR systems in HRI is meticulously structured to address the critical aspects of collaboration, trust, and safety. This framework is built upon the foundational principles of HCI and is supported by the comprehensive research findings outlined in the AR in HCI survey. The framework, visualized in Figure 15, is divided into three key components: Key Components of AR in HRI, Good Design Principles for AR, and Evaluation and Continuous Improvement.

Collaboration focuses on enhancing the synergy between humans and robots through effective planning, training, communication, and interaction. It emphasizes the need for AR systems to facilitate seamless communication and instruction, thereby improving task efficiency and reducing the likelihood of errors.

Trust and safety are paramount in HRI, and the framework highlights Situational Awareness and control as essential elements. By leveraging AR technologies, the framework aims to enhance users' Situational Awareness, enabling them to perceive, comprehend, and project the state of their environment accurately. The integration of control mechanisms, such as Speed and Separation Monitoring (SSM) algorithms, ensures that robots operate safely around human operators, dynamically adjusting their actions to maintain safe distances and prevent collisions.

Implementing HCI Principles for Good AR Design involves creating user interfaces (UI) and user experiences (UX) that are intuitive, user-centered, and designed to meet the specific needs of human operators. This includes utilizing AR to provide real-time feedback, improve usability, and enhance the overall user experience.

Evaluation and Continuous Improvement underscore the importance of using both quantitative and qualitative measures to assess the effectiveness of AR systems. By continuously collecting and analyzing data on system performance and user feedback, the framework advocates for an iterative design process that ensures that AR systems remain effective, user-friendly, and safe.

The framework outlined in this paper serves as a foundational guide for developing Augmented Reality (AR) applications in Human–Robot Interaction (HRI). Its practical relevance is demonstrated through specific experiments and the design of the elements that leverage different components of the framework. The following subsections describe implementations that focus on the key components of AR in HRI and good design principles of AR.

We showcase two case studies that are in accordance with the proposed framework. The first study highlights the use of AR for enhancing Situational Awareness, while the second study shows the utilization of digital twins with the help of AR to enable safety in HRI. Furthermore, ongoing studies that incorporate good practices from HCI are described in this section.

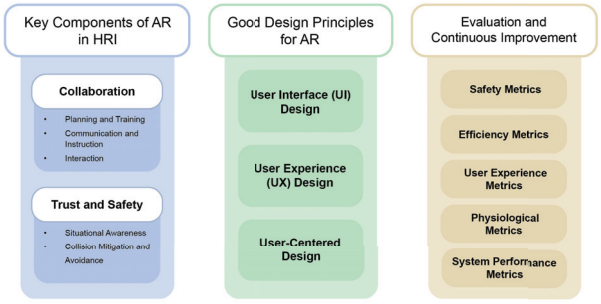


Figure 15. Framework for developing AR application in HRI.

8.1. Improving Collaboration, Safety, and Trust Using AR in HRI

Ensuring trust and safety in HRI is paramount, particularly in dynamic industrial environments. The presented framework for designing Augmented Reality (AR) systems in HRI underscores the importance of these components, as validated by recent experimental studies. The first study utilizes AR to enhance Situational Awareness, employing the Situational Awareness Global Assessment Technique (SAGAT) to measure improvements [123]. Wearing a head-mounted display, the user can see a green arrow that moves to point at the end effector of the robot. When the user is turned away from the robot, the arrow indicates where the robot is located in relation to the user. The green arrow is circled in Figure 16 below. Results indicate a significant reduction in perception errors (Average Perception Error decreased significantly) and increased operator confidence (Mean Confidence levels rose), reinforcing the framework's emphasis on Situational Awareness as a cornerstone of safety. The full results are available in the citation and not reported here for brevity. The study also reported an increase in the percentage of correct responses to accurately predict the direction of movement of the end effector of the robot and a trend towards faster response times, highlighting the effectiveness of AR displays in improving Situational Awareness. Figure 17 shows the results for the perception of error and the percentage of correct responses to the experiment, and Table 2 shows the statistical significance of the metrics. [123].



Figure 16. AR display shown to human workers through a HMD. Green arrow indicating robot's location is circled in red [123].

Table 2. Statistical significance of situational awareness metrics.

Metric	T-Statistic	*p*-Value
Average Perception Error	−3.5871	0.002106
Percent Correct	2.4179	0.02643
Mean Confidence	4.0171	0.008085
Mean Response Time	−1.7847	0.09115

The second study leverages mixed reality (MR) and digital twin technology to implement a Speed and Separation Monitoring (SSM) algorithm, which dynamically adjusts robot speed to prevent collisions [124]. This real-time control mechanism aligns perfectly with the framework's focus on maintaining safety through effective control strategies. Figure 18 shows the digital twins for the robot and the human; the human model uses geometric primitives to approximate the human body when calculating the minimum distance from the robot. The study's results, in Table 3, showed a maximum tracking error of 0.0743 m and a mean error of 0.0315 m, demonstrating high positional accuracy. Additionally, the SSM algorithm effectively adjusted the robot's speed in response to changes in separation distance, ensuring a safe distance between humans and robots. An example of this can be seen in Figure 19.

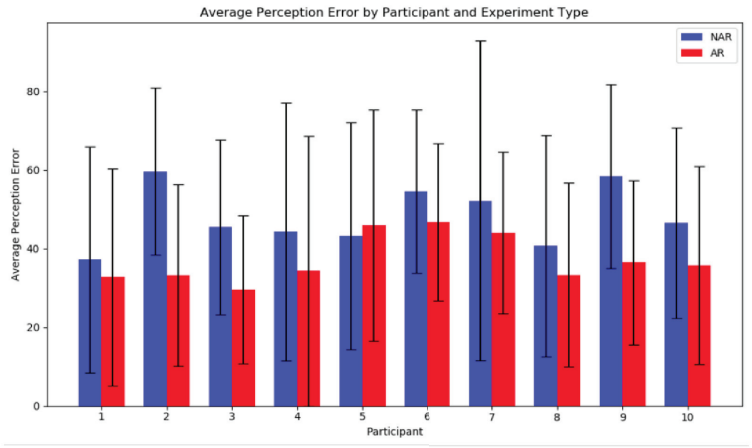


Figure 17. Results from Study 1: (**A**) Average Perception Error; (**B**) Mean Confidence. AR represents performance with the use of Augmented Reality, and NAR represents the lack of AR when performing the task [123].

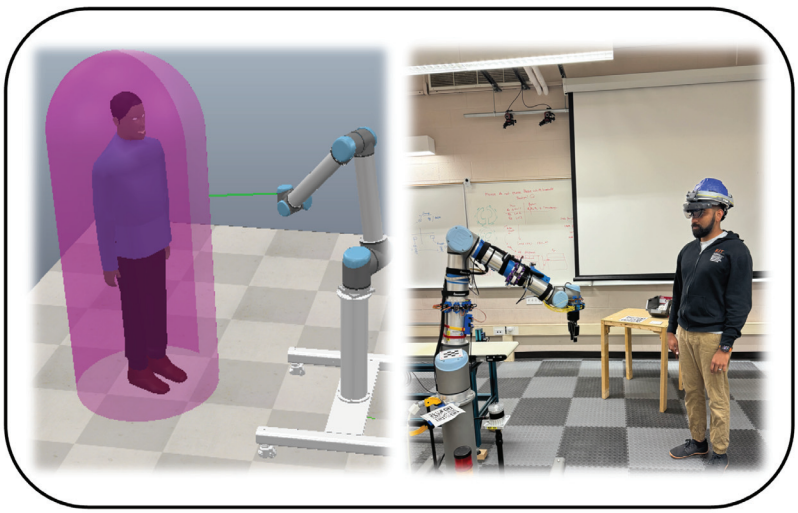


Figure 18. Geometric primitives used for approximating the human body for minimum distance measurements in a digital twin [124].

Table 3. Absolute error metrics: Hololens 2 tracking vs. Opti-Track Motion Capture System; values are in cms [124].

Error Metric	Value
Max	0.07435
Min	0.00607
Median	0.03087
Mean	0.03149
RMSE	0.03382
Std Dev	0.01234

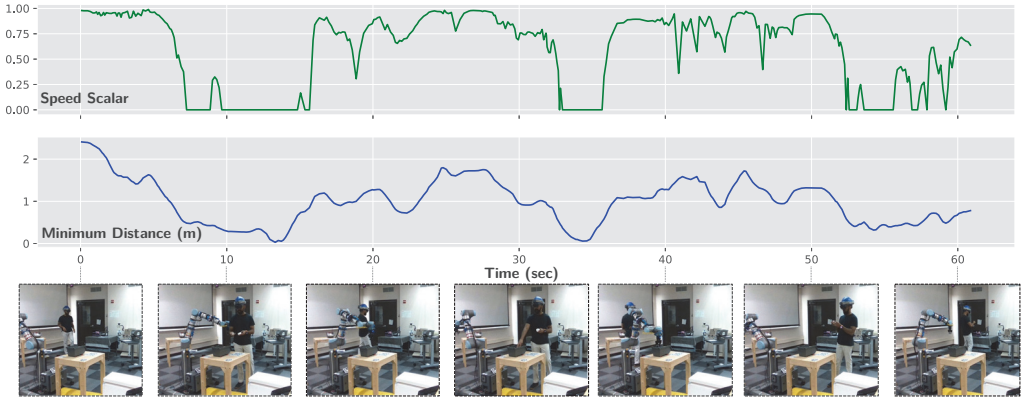


Figure 19. A stacked view of a 60 s experiment segment shows the system's performance using the Speed and Separation Monitoring algorithm. The top row indicates the speed scaling command to the robot, the middle row shows the minimum distance between human and robot, and the bottom row captures their positions every 10 s [124].

Both studies utilize quantitative measures such as tracking error, perception accuracy, and response times, alongside qualitative feedback on user confidence and system usability, providing robust empirical support for the framework. By prioritizing trust and safety, the framework ensures that AR systems are designed to enhance operational security and build user trust, making it an invaluable tool for advancing HRI in industrial applications. These findings show the practical relevance of the framework, demonstrating its ability to guide the development of AR systems that significantly improve safety and trust in HRI environments.

8.2. Collision Warnings Using Virtual Elements

When utilizing the Augmented Reality (AR) system, the wearer is immersed in a sophisticated, real-time environment that significantly enhances their Situational Awareness and safety during HRI. The wearer can expect to see a series of dynamic visual indicators, represented by spheres, which trace the path of the robot end effector's movements. These spheres provide an intuitive visual cue, changing color based on the proximity of the wearer to the robot's path. When the human operator is within a safe distance, the spheres remain blue (Figure 20A), indicating no immediate threat. However, if the wearer approaches a critical proximity threshold, the spheres turn red (Figure 20B), signaling a potential collision risk. This visual feedback allows the wearer to anticipate the robot's trajectory and adjust their movements accordingly, thereby preventing accidents. The system continuously updates these indicators, ensuring that the wearer has up-to-date information about the robot's position and path.

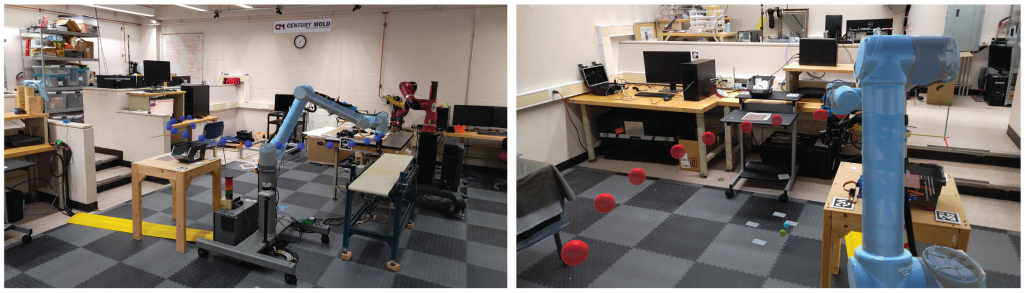


Figure 20. Spheres to show the robot end effector's movement. (**A**) Blue spheres when the human is in a safe proximity to the robot. (**B**) Red spheres when the human is in close proximity to the robot.

Implementing effective HCI practices should begin with user research and a thorough understanding of how the user will interact with the system. For the minimap design, it started with understanding user preferences and the optimal system design. Prior work by Green et al. [24] suggested that an overview of the robot's path was more successful in determining Situational Awareness than an immersive view. In the current lab, users have an immersive view. To provide them with an overview or bird's-eye view, a minimap was implemented to show the user, the robot, and the tables at all times. The mini-map overlay provides real-time updates on the positions of key elements within the workspace, including the base and tool of the robot, the main camera, and additional objects of interest. These elements are represented by distinct indicator objects on the mini-map, which are dynamically updated to reflect their relative positions scaled down to a manageable size for the wearer's view.

The mini-map serves as a crucial tool for spatial orientation, allowing the wearer to easily track the movements of both the robot and surrounding objects. Each indicator is color-coded for quick visual recognition: the base indicator is semi-transparent white, the tool indicator is cyan, the human indicator is green (Figure 21A) (changing to orange and then red based on proximity to the human operator) (Figure 21B,C), and other object indicators are grey. This color-coding not only helps in distinguishing between different elements but also provides immediate visual cues about the current operational status and potential hazards. The locations of the objects are linked to a fiducial and will be updated as the Hololens 2 is able to scan them.

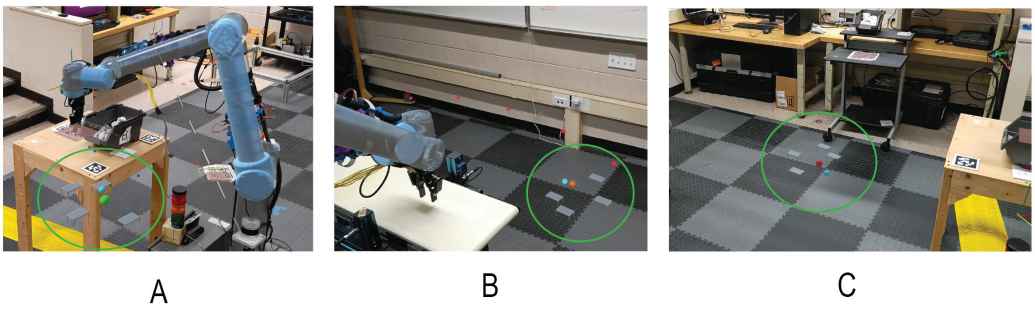


Figure 21. The minimap design is circled in green. (**A**) The human indicator is green in safe proximity. (**B**) The human indicator turns orange when in closer but still safe proximity. (**C**) The human indicator turns red when in close and dangerous proximity to the robot.

As the wearer navigates the workspace, the mini-map continuously updates, showing the real-time positions of the robot and other objects relative to the base position. This dynamic update mechanism ensures that the wearer always has the latest information about the robot's location, enhancing their ability to predict movements and avoid collisions.

Furthermore, the camera indicator changes color based on the distance between the robot and the human operator, turning red when within a critical threshold, thus providing an immediate warning of potential collision risks.

For the visualization of the design, UI principles such as simplicity, consistency, and visibility were all considered. The design of the minimap is kept simple, including only the essential elements: the robot, the human, and the tables. Additionally, the human and robot are both represented as spheres to avoid cluttering the minimap. The tables take the shape of actual tables rather than cubes or other 3D shapes to remain consistent with the real world. This way, users do not have to assume or remember the real-world item corresponding to the elements in the AR.

Visibility is ensured because the map moves with the person, ensuring they always see the map in the correct orientation. This means users do not have to calculate their direction relative to the map's direction to know where the robot is, enhancing ease of use and safety. Situational Awareness

9. Limitations

9.1. Scope Limitation

The scope of this study is focused on Human–Robot Collaboration in industrial settings, particularly in manufacturing settings like assembly lines. This targeted scope allows for in-depth exploration of relevant use cases and provides actionable insights applicable to these specific environments. While the current focus limits broader applicability, future work will continue to refine and expand the framework to address emerging challenges pertaining to other domains.

9.2. Technical Limitations and User Acceptance

While Augmented Reality (AR) offers significant potential for enhancing HRI, it is important to acknowledge several technical limitations and user acceptance challenges. AR systems often face constraints such as limited field of view, display resolution, and response time, which can impact their effectiveness in high-stakes environments like industrial settings. These technical issues can lead to missed cues or distractions, affecting continuous Situational Awareness and safety. Additionally, variations in AR device performance across different environments may reduce the consistency and reliability of critical safety alerts. When systems do not work as intended, they can lead to worker frustration.

User acceptance is another critical factor that influences the success of AR in HRI. Challenges include the comfort and ergonomics of AR headsets, the learning curve associated with new technologies, and potential resistance due to concerns about complexity or reliability. To address these barriers, we can look at tools like NASA-TLX to measure subjective cognitive load when using AR.

9.3. Plan for Long-Term Impact Assessment of AR in HRI

To evaluate the long-term impact of AR-enhanced HRI systems on trust and safety, it is crucial to establish robust evaluation metrics and methodologies. Key performance indicators should include not only immediate outcomes such as reduced safety incidents and improved task efficiency but also long-term measures like sustained user trust, system reliability, and continued user engagement. Methodologies for assessment could involve mixed-methods approaches that combine quantitative data with qualitative insights from user feedback and testing in both simulated and real-world environments. By collecting data across all the metrics listed in Section 7, it is possible to capture trends over time and provide valuable data on the long-term impact of AR-enhanced HRI systems in manufacturing settings.

10. Conclusions

This paper has explored the significant advancements in integrating Augmented Reality (AR) into Human–Robot Interaction (HRI) with a focus on manufacturing, emphasizing

the enhancement of collaboration, trust, and safety. By synthesizing findings from numerous studies, we have identified key challenges and proposed a comprehensive framework designed to address these issues effectively.

Our framework emphasizes the importance of Situational Awareness and control, leveraging AR technologies to enhance users' ability to perceive, comprehend, and project the state of their environment. The integration of Speed and Separation Monitoring (SSM) algorithms ensures that robots operate safely around human operators, dynamically adjusting their actions to maintain safe distances and prevent collisions.

Moreover, the implementation of Human–Computer Interaction (HCI) principles for AR design ensures that user interfaces (UI) and user experiences (UX) are intuitive, user-centered, and tailored to meet the specific needs of human operators. Continuous evaluation and improvement, through both quantitative and qualitative measures, are essential to maintain the effectiveness, usability, and safety of AR systems in HRI.

In conclusion, the proposed framework not only addresses the critical aspects of collaboration, trust, and safety but also provides a structured approach for the continuous development and refinement of AR applications in HRI for manufacturing applications. This research paves the way for safer and more efficient Human–Robot collaborations, ultimately contributing to the advancement of the field.

Author Contributions: K.S. studied various AR devices, such as headsets and projections, to understand their impact on HRI, while L.T. contributed towards HCI principles to enhance the usability and effectiveness of AR interfaces in HRI. M.S. studied existing methods of evaluating Situational Awareness in HRI and contributed towards integrating Situational Awareness in the proposed framework mentioned in this paper. Under the supervision of F.S., the team synthesized their findings to provide a comprehensive review of AR applications within HRI, highlighting key challenges and proposing solutions for enhancing collaboration, trust, and safety in HRI systems. All authors have read and agreed to the published version of the manuscript.

Funding: This material is based upon work supported by the National Science Foundation under Award No. DGE-2125362. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Thrun, S. Toward a Framework for Human-Robot Interaction. *Human–Computer Interact.* **2004**, *19*, 9–24. [CrossRef] [PubMed]
2. Scholtz, J. Theory and evaluation of human Robot Interactions. In Proceedings of the 36th Annual Hawaii International Conference on System Sciences, Big Island, HI, USA, 6–9 January 2003 ; p. 10. [CrossRef]
3. Sidobre, D.; Broquère, X.; Mainprice, J.; Burattini, E.; Finzi, A.; Rossi, S.; Staffa, M. Human–Robot Interaction. In *Advanced Bimanual Manipulation: Results from the DEXMART Project*; Siciliano, B., Ed.; Springer Tracts in Advanced Robotics; Springer: Berlin/Heidelberg, Germany, 2012; pp. 123–172. [CrossRef]
4. Ong, S.K.; Yuan, M.L.; Nee, A.Y.C. Augmented Reality applications in manufacturing: A survey. *Int. J. Prod. Res.* **2008**, *46*, 2707–2742. [CrossRef]
5. Suzuki, R.; Karim, A.; Xia, T.; Hedayati, H.; Marquardt, N. Augmented Reality and Robotics: A Survey and Taxonomy for AR-enhanced Human-Robot Interaction and Robotic Interfaces. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022–5 May 2022; CHI '22; pp. 1–33. [CrossRef]
6. Ruiz, J.; Escalera, M.; Viguria, A.; Ollero, A. A simulation framework to validate the use of head-mounted displays and tablets for information exchange with the UAV safety pilot. In Proceedings of the 2015 Workshop on Research, Education and Development of Unmanned Aerial Systems (RED-UAS), Cancun, Mexico, 23–25 November 2015; pp. 336–341. [CrossRef]
7. Chan, W.P.; Hanks, G.; Sakr, M.; Zhang, H.; Zuo, T.; van der Loos, H.F.M.; Croft, E. Design and Evaluation of an Augmented Reality Head-mounted Display Interface for Human Robot Teams Collaborating in Physically Shared Manufacturing Tasks. *ACM Trans. Hum.-Robot Interact.* **2022**, *11*, 31:1–31:19. [CrossRef]
8. Kalpagam Ganesan, R.; Rathore, Y.K.; Ross, H.M.; Ben Amor, H. Better Teaming Through Visual Cues: How Projecting Imagery in a Workspace Can Improve Human-Robot Collaboration. *IEEE Robot. Autom. Mag.* **2018**, *25*, 59–71. [CrossRef]

9. Woodward, J.; Ruiz, J. Analytic Review of Using Augmented Reality for Situational Awareness. *IEEE Trans. Vis. Comput. Graph.* **2023**, *29*, 2166–2183. [CrossRef] [PubMed]
10. Rogers, Y.; Sharp, H.; Preece, J. *Interaction Design: Beyond Human-Computer Interaction*, 6th ed.; John Wiley & Sons: Hoboken, NJ, USA, 2023. Available online: http://id-book.com (accessed on 9 September 2024).
11. Blaga, A.; Tamas, L. Augmented Reality for Digital Manufacturing. In Proceedings of the 2018 26th Mediterranean Conference on Control and Automation (MED), Zadar, Croatia, 19–22 June 2018; pp. 173–178. [CrossRef]
12. Caudell, T.; Mizell, D. Augmented Reality: An application of heads-up display technology to manual manufacturing processes. In Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences, Kauai, HI, USA, 7–10 January 1992; Volume 2, pp. 659–669. [CrossRef]
13. Novak-Marcincin, J.; Barna, J.; Janak, M.; Novakova-Marcincinova, L. Augmented Reality Aided Manufacturing. *Procedia Comput. Sci.* **2013**, *25*, 23–31. [CrossRef]
14. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Syst. Rev.* **2021**, *10*, 89. [CrossRef]
15. Haddaway, N.R.; Page, M.J.; Pritchard, C.C.; McGuinness, L.A. PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. *Campbell Syst. Rev.* **2022**, *18*, e1230. [CrossRef]
16. Ajoudani, A.; Zanchettin, A.M.; Ivaldi, S.; Albu-Schäffer, A.; Kosuge, K.; Khatib, O. Progress and prospects of the human–robot collaboration. *Auton. Robot.* **2018**, *42*, 957–975. [CrossRef]
17. Baratta, A.; Cimino, A.; Gnoni, M.G.; Longo, F. Human Robot Collaboration in Industry 4.0: A literature review. *Procedia Comput. Sci.* **2023**, *217*, 1887–1895. [CrossRef]
18. Bauer, A.; Wollherr, D.; Buss, M. Human-Robot Collaboration: A Survey. *I. J. Humanoid Robot.* **2008**, *5*, 47–66. [CrossRef]
19. Semeraro, F.; Griffiths, A.; Cangelosi, A. Human–robot collaboration and machine learning: A systematic review of recent research. *Robot. Comput.-Integr. Manuf.* **2023**, *79*, 102432. [CrossRef]
20. Kumar, S.; Savur, C.; Sahin, F. Survey of Human–Robot Collaboration in Industrial Settings: Awareness, Intelligence, and Compliance. *IEEE Trans. Syst. Man Cybern. Syst.* **2021**, *51*, 280–297. [CrossRef]
21. Matheson, E.; Minto, R.; Zampieri, E.G.G.; Faccio, M.; Rosati, G. Human–Robot Collaboration in Manufacturing Applications: A Review. *Robotics* **2019**, *8*, 100. [CrossRef]
22. Lamon, E.; De Franco, A.; Peternel, L.; Ajoudani, A. A Capability-Aware Role Allocation Approach to Industrial Assembly Tasks. *IEEE Robot. Autom. Lett.* **2019**, *4*, 3378–3385. [CrossRef]
23. Rahman, S.M.; Wang, Y. Mutual trust-based subtask allocation for human–robot collaboration in flexible lightweight assembly in manufacturing. *Mechatronics* **2018**, *54*, 94–109. [CrossRef]
24. Green, S.A.; Chase, J.G.; Chen, X.; Billinghurst, M. Evaluating the Augmented Reality Human–Robot Collaboration System. In Proceedings of the 2008 15th International Conference on Mechatronics and Machine Vision in Practice, Auckland, New Zealand, 2–4 December 2008.
25. Glassmire, J.; O'Malley, M.; Bluethmann, W.; Ambrose, R. Cooperative manipulation between humans and teleoperated agents. In Proceedings of the 12th International Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2004. HAPTICS '04. Proceedings, Chicago, IL, USA, 27–28 March 2004; pp. 114–120. [CrossRef]
26. Hoffman, G.; Breazeal, C. Effects of anticipatory action on Human–Robot teamwork: Efficiency, fluency, and perception of team. In Proceedings of the 2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI), Arlington, VA, USA, 10–12 March 2007; pp. 1–8. ISSN: 2167-2148. [CrossRef]
27. Lasota, P.A.; Shah, J.A. Analyzing the Effects of Human-Aware Motion Planning on Close-Proximity Human–Robot Collaboration. *Hum. Factors: J. Hum. Factors Ergon. Soc.* **2015**, *57*, 21–33. [CrossRef]
28. Yao, B.; Zhou, Z.; Wang, L.; Xu, W.; Yan, J.; Liu, Q. A function block based cyber-physical production system for physical human–Robot Interaction. *J. Manuf. Syst.* **2018**, *48*, 12–23. [CrossRef]
29. IEC 61499 Standard for Distributed Automation. Available online: https://iec61499.com/ (accessed on 9 September 2024).
30. Stark, R.; Fresemann, C.; Lindow, K. Development and operation of Digital Twins for technical systems and services. *CIRP Ann.* **2019**, *68*, 129–132. [CrossRef]
31. Liu, Q.; Leng, J.; Yan, D.; Zhang, D.; Wei, L.; Yu, A.; Zhao, R.; Zhang, H.; Chen, X. digital-twin-based designing of the configuration, motion, control, and optimization model of a flow-type smart manufacturing system. *J. Manuf. Syst.* **2021**, *58*, 52–64. [CrossRef]
32. Rosen, R.; Von Wichert, G.; Lo, G.; Bettenhausen, K.D. About The Importance of Autonomy and Digital Twins for the Future of Manufacturing. *IFAC-PapersOnLine* **2015**, *48*, 567–572. [CrossRef]
33. Sahin, M.; Savur, C. Evaluation of Human Perceived Safety during HRC Task using Multiple Data Collection Methods. In Proceedings of the 2022 17th Annual System of Systems Engineering Conference (SOSE), Rochester, NY, USA, 7–11 June 2022; pp. 465–470. [CrossRef]
34. Soh, H.; Xie, Y.; Chen, M.; Hsu, D. Multi-task trust transfer for human–Robot Interaction. *Int. J. Robot. Res.* **2020**, *39*, 233–249. [CrossRef]
35. *ISO 10218-1:2011*; Robots and Robotic Devices—Safety Requirements for Industrial Robots—Part 1: Robots. ISO: Geneva, Switzerland, 2011.

36. *ISO/TS 15066:2016*; Robots and Robotic Devices—Collaborative Robots. ISO: Geneva, Switzerland, 2016.
37. Haddadin, S.; Albu-Schäffer, A.; Hirzinger, G. Safe Physical Human-Robot Interaction: Measurements, Analysis and New Insights. In *Robotics Research*; Kaneko, M., Nakamura, Y., Eds.; Springer Tracts in Advanced Robotics; Springer: Berlin/Heidelberg, Germany, 2011; pp. 395–407. [CrossRef]
38. Lee, J.D.; See, K.A. Trust in automation: Designing for appropriate reliance. *Hum. Factors* **2004**, *46*, 50–80. [CrossRef] [PubMed]
39. Maurtua, I.; Ibarguren, A.; Kildal, J.; Susperregi, L.; Sierra, B. Human–robot collaboration in industrial applications: Safety, interaction and trust. *Int. J. Adv. Robot. Syst.* **2017**, *14*, 172988141771601. [CrossRef]
40. Hancock, P.A.; Billings, D.R.; Schaefer, K.E.; Chen, J.Y.C.; De Visser, E.J.; Parasuraman, R. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Hum. Factors: J. Hum. Factors Ergon. Soc.* **2011**, *53*, 517–527. [CrossRef] [PubMed]
41. Endsley, M. Theoretical underpinnings of situation awareness: A critical review. In *Situation Awareness Analysis and Measurement*; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2000; pp. 3–32.
42. Endsley, M.; Kiris, E. The Out-of-the-Loop Performance Problem and Level of Control in Automation. *Hum. Factors: J. Hum. Factors Ergon. Soc.* **1995**, *37*, 381–394. [CrossRef]
43. Unhelkar, V.V.; Siu, H.C.; Shah, J.A. Comparative performance of human and mobile robotic assistants in collaborative fetch-and-deliver tasks. In Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction, Bielefeld, Germany, 3–6 March 2014; pp. 82–89. [CrossRef]
44. Endsley, M. Direct Measurement of Situation Awareness: Validity and Use of SAGAT. In *Situation Awareness: Analysis and Measurement*; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2000.
45. Endsley, M. Situation awareness global assessment technique (SAGAT). In Proceedings of the IEEE 1988 National Aerospace and Electronics Conference, Dayton, OH, USA, 23–27 May 1988; Volume 3; pp. 789–795. [CrossRef]
46. Bew, G.; Baker, A.; Goodman, D.; Nardone, O.; Robinson, M. Measuring Situational Awareness at the small unit tactical level. In Proceedings of the 2015 Systems and Information Engineering Design Symposium, Charlottesville, VA, USA, 24 April 2015; pp. 51–56. [CrossRef]
47. Endsley, M.R. A Systematic Review and Meta-Analysis of Direct Objective Measures of Situation Awareness: A Comparison of SAGAT and SPAM. *Hum. Factors* **2021**, *63*, 124–150. [CrossRef]
48. Marvel, J.A.; Norcross, R. Implementing Speed and Separation Monitoring in collaborative robot workcells. *Robot. Comput.-Integr. Manuf.* **2017**, *44*, 144–155. [CrossRef]
49. Kumar, S.; Arora, S.; Sahin, F. Speed and Separation Monitoring using On-Robot Time-of-Flight Laser-ranging Sensor Arrays. In Proceedings of the 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), Vancouver, BC, Canada, 22–26 August 2019; pp. 1684–1691. [CrossRef]
50. Rosenstrauch, M.J.; Pannen, T.J.; Krüger, J. Human robot collaboration—using kinect v2 for ISO/TS 15066 Speed and Separation Monitoring. *Procedia CIRP* **2018**, *76*, 183–186. [CrossRef]
51. Ganglbauer, M.; Ikeda, M.; Plasch, M.; Pichler, A. Human in the loop online estimation of robotic speed limits for safe human robot collaboration. *Procedia Manuf.* **2020**, *51*, 88–94. [CrossRef]
52. Oh, K.; Kim, M. Social Attributes of Robotic Products: Observations of Child-Robot Interactions in a School Environment. *Int. J. Design* **2010**, *4*, 45–55.
53. Sauppé, A.; Mutlu, B. The Social Impact of a Robot Co-Worker in Industrial Settings. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Republic of Korea, 18–23 April 2015; pp. 3613–3622. [CrossRef]
54. Bruce, A.; Nourbakhsh, I.; Simmons, R. The role of expressiveness and attention in Human–Robot Interaction. In Proceedings of the 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292), Washington, DC, USA, 11–15 May 2002; Volume 4, pp. 4138–4142. [CrossRef]
55. Lu, L.; Xie, Z.; Wang, H.; Li, L.; Xu, X. Mental stress and safety awareness during Human–Robot collaboration—Review. *Appl. Ergon.* **2022**, *105*, 103832. [CrossRef] [PubMed]
56. Evans, G.; Miller, J.; Pena, M.; MacAllister, A.; Winer, E. Evaluating the Microsoft HoloLens through an Augmented Reality assembly application. In Proceedings of the SPIE Defense + Security, Anaheim, CA, USA, 9–13 April 2017; p. 101970V. [CrossRef]
57. Sääski, J.; Salonen, T.; Liinasuo, M.; Pakkanen, J.; Vanhatalo, M.; Riitahuhta, A. Augmented Reality Efficiency in Manufacturing Industry: A Case Study. In Proceedings of the DS 50: Proceedings of NordDesign 2008 Conference, Tallinn, Estonia, 21–23 August 2008.
58. Palmarini, R.; del Amo, I.F.; Bertolino, G.; Dini, G.; Erkoyuncu, J.A.; Roy, R.; Farnsworth, M. Designing an AR interface to improve trust in Human-Robots collaboration. *Procedia CIRP* **2018**, *70*, 350–355. [CrossRef]
59. Tsamis, G.; Chantziaras, G.; Giakoumis, D.; Kostavelis, I.; Kargakos, A.; Tsakiris, A.; Tzovaras, D. Intuitive and Safe Interaction in Multi-User Human Robot Collaboration Environments through Augmented Reality Displays. In Proceedings of the 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN), Vancouver, BC, Canada, 8–12 August 2021; pp. 520–526. ISSN: 1944-9437. [CrossRef]
60. Vogel, C.; Schulenburg, E.; Elkmann, N. Projective- AR Assistance System for shared Human-Robot Workplaces in Industrial Applications. In Proceedings of the 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), Vienna, Austria, 8–11 September 2020; Volume 1, pp. 1259–1262. ISSN: 1946-0759. [CrossRef]

61. Choi, S.H.; Park, K.B.; Roh, D.H.; Lee, J.Y.; Mohammed, M.; Ghasemi, Y.; Jeong, H. An integrated mixed reality system for safety-aware Human–Robot collaboration using deep learning and digital twin generation. *Robot. Comput.-Integr. Manuf.* **2022**, *73*, 102258. [CrossRef]
62. Bassyouni, Z.; Elhajj, I.H. Augmented Reality Meets Artificial Intelligence in Robotics: A Systematic Review. *Front. Robot. AI* **2021**, *8*, 724798. [CrossRef]
63. Costa, G.D.M.; Petry, M.R.; Moreira, A.P. Augmented Reality for Human–Robot Collaboration and Cooperation in Industrial Applications: A Systematic Literature Review. *Sensors* **2022**, *22*, 2725. [CrossRef]
64. Franze, A.P.; Caldwell, G.A.; Teixeira, M.F.L.A.; Rittenbruch, M. Employing AR/MR Mockups to Imagine Future Custom Manufacturing Practices. In Proceedings of the 34th Australian Conference on Human-Computer Interaction, New York, NY, USA, 6 April 2023; OzCHI '22; pp. 206–215. [CrossRef]
65. Fang, H.; Ong, S.; Nee, A. Robot Path and End-Effector Orientation Planning Using Augmented Reality. *Procedia CIRP* **2012**, *3*, 191–196. [CrossRef]
66. Doil, F.; Schreiber, W.; Alt, T.; Patron, C. Augmented Reality for manufacturing planning. In Proceedings of the Workshop on Virtual Environments 2003, New York, NY, USA, 22–23 May 2003; EGVE '03; pp. 71–76. [CrossRef]
67. Wang, Q.; Fan, X.; Luo, M.; Yin, X.; Zhu, W. Construction of Human-Robot Cooperation Assembly Simulation System Based on Augmented Reality. In *Virtual, Augmented and Mixed Reality. Design and Interaction*; Chen, J.Y.C., Fragomeni, G., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020; pp. 629–642. [CrossRef]
68. Andersson, N.; Argyrou, A.; Nägele, F.; Ubis, F.; Campos, U.E.; Zarate, M.O.d.; Wilterdink, R. AR-Enhanced Human-Robot-Interaction—Methodologies, Algorithms, Tools. *Procedia CIRP* **2016**, *44*, 193–198. [CrossRef]
69. Michalos, G.; Karagiannis, P.; Makris, S.; Tokçalar, Ö.; Chryssolouris, G. Augmented Reality (AR) Applications for Supporting Human-robot Interactive Cooperation. *Procedia CIRP* **2016**, *41*, 370–375. [CrossRef]
70. Lunding, R.; Hubenschmid, S.; Feuchtner, T. Proposing a Hybrid Authoring Interface for AR-Supported Human–Robot Collaboration. 2024. Available online: https://openreview.net/forum?id=2w2ynC3yrM&noteId=ritvr8VKmu (accessed on 6 July 2024).
71. Tabrez, A.; Luebbers, M.B.; Hayes, B. Descriptive and Prescriptive Visual Guidance to Improve Shared Situational Awareness in Human–Robot Teaming. In Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, Online, 9–13 May 2022.
72. De Franco, A.; Lamon, E.; Balatti, P.; De Momi, E.; Ajoudani, A. An Intuitive Augmented Reality Interface for Task Scheduling, Monitoring, and Work Performance Improvement in Human-Robot Collaboration. In Proceedings of the 2019 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), Budapest, Hungary, 3–5 July 2019; pp. 75–80. [CrossRef]
73. Andronas, D.; Apostolopoulos, G.; Fourtakas, N.; Makris, S. Multi-modal interfaces for natural Human-Robot Interaction. *Procedia Manuf.* **2021**, *54*, 197–202. [CrossRef]
74. Gkournelos, C.; Karagiannis, P.; Kousi, N.; Michalos, G.; Koukas, S.; Makris, S. Application of Wearable Devices for Supporting Operators in Human-Robot Cooperative Assembly Tasks. *Procedia CIRP* **2018**, *76*, 177–182. [CrossRef]
75. Qiu, S.; Liu, H.; Zhang, Z.; Zhu, Y.; Zhu, S.C. Human-Robot Interaction in a Shared Augmented Reality Workspace. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October 2020–24 January 2021; pp. 11413–11418. ISSN: 2153-0866, [CrossRef]
76. Sonawani, S.; Amor, H.B. When and Where Are You Going? A Mixed-Reality Framework for Human Robot Collaboration. 2022. Available online: https://openreview.net/forum?id=BSrx_Q2-Akq (accessed on 21 June 2024).
77. Hietanen, A.; Latokartano, J.; Pieters, R.; Lanz, M.; Kämäräinen, J.K. AR-based interaction for safe Human–Robot collaborative manufacturing. *arXiv* **2019**, arXiv:1909.02933.
78. Lunding, R.S.; Lunding, M.S.; Feuchtner, T.; Petersen, M.G.; Grønbæk, K.; Suzuki, R. RoboVisAR: Immersive Authoring of Condition-based AR Robot Visualisations. In Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, New York, NY, USA, 11–15 March 2024; HRI '24; pp. 462–471. [CrossRef]
79. Eschen, H.; Kötter, T.; Rodeck, R.; Harnisch, M.; Schüppstuhl, T. Augmented and Virtual Reality for Inspection and Maintenance Processes in the Aviation Industry. *Procedia Manuf.* **2018**, *19*, 156–163. [CrossRef]
80. Papanastasiou, S.; Kousi, N.; Karagiannis, P.; Gkournelos, C.; Papavasileiou, A.; Dimoulas, K.; Baris, K.; Koukas, S.; Michalos, G.; Makris, S. Towards seamless human robot collaboration: Integrating multimodal interaction. *Int. J. Adv. Manuf. Technol.* **2019**, *105*, 3881–3897. [CrossRef]
81. Makris, S.; Karagiannis, P.; Koukas, S.; Matthaiakis, A.S. Augmented Reality system for operator support in human–robot collaborative assembly. *CIRP Ann.* **2016**, *65*, 61–64. [CrossRef]
82. Matsas, E.; Vosniakos, G.C.; Batras, D. Prototyping proactive and adaptive techniques for Human–Robot collaboration in manufacturing using virtual reality. *Robot. Comput.-Integr. Manuf.* **2018**, *50*, 168–180. [CrossRef]
83. Bischoff, R.; Kazi, A. Perspectives on Augmented Reality based Human–Robot Interaction with industrial robots. In Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566), Sendai, Japan, 28 September–2 October 2004; Volume 4, pp. 3226–3231. [CrossRef]
84. Matsas, E.; Vosniakos, G.C. Design of a virtual reality training system for human–robot collaboration in manufacturing tasks. *Int. J. Interact. Des. Manuf. (IJIDeM)* **2017**, *11*, 139–153. [CrossRef]

85. Liu, H.; Wang, L. An AR-based Worker Support System for Human-Robot Collaboration. *Procedia Manuf.* **2017**, *11*, 22–30. [CrossRef]
86. Janin, A.; Mizell, D.; Caudell, T. Calibration of head-mounted displays for Augmented Reality applications. In Proceedings of the IEEE Virtual Reality Annual International Symposium, Seattle, WA, USA, 18–22 September 1993; pp. 246–255. [CrossRef]
87. Mitaritonna, A.; Abásolo, M.J.; Montero, F. An Augmented Reality-based Software Architecture to Support Military Situational Awareness. In Proceedings of the 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), Istanbul, Turkey, 12–13 June 2020; pp. 1–6. [CrossRef]
88. Sheikh Bahaei, S.; Gallina, B. Assessing risk of AR and organizational changes factors in socio-technical robotic manufacturing. *Robot. Comput.-Integr. Manuf.* **2024**, *88*, 102731. [CrossRef]
89. Feddoul, Y.; Ragot, N.; Duval, F.; Havard, V.; Baudry, D.; Assila, A. Exploring human-machine collaboration in industry: A systematic literature review of digital twin and robotics interfaced with extended reality technologies. *Int. J. Adv. Manuf. Technol.* **2023**, *129*, 1917–1932. [CrossRef]
90. Maruyama, T.; Ueshiba, T.; Tada, M.; Toda, H.; Endo, Y.; Domae, Y.; Nakabo, Y.; Mori, T.; Suita, K. Digital Twin-Driven Human Robot Collaboration Using a Digital Human. *Sensors* **2021**, *21*, 8266. [CrossRef] [PubMed]
91. Shaaban, M.; Carfì, A.; Mastrogiovanni, F. Digital Twins for Human-Robot Collaboration: A Future Perspective. In *Intelligent Autonomous Systems 18*; Lee, S.G., An, J., Chong, N.Y., Strand, M., Kim, J.H., Eds.; Lecture Notes in Networks and Systems; Springer: Cham, Switzerland, 2024; Volume 795, pp. 429–441. [CrossRef]
92. Carroll, J.M. *HCI Models, Theories, and Frameworks: Toward a Multidisciplinary Science*; Elsevier: Amsterdam, The Netherlands, 2003.
93. Nazari, A.; Alabood, L.; Feeley, K.B.; Jaswal, V.K.; Krishnamurthy, D. Personalizing an AR-based Communication System for Nonspeaking Autistic Users. In Proceedings of the 29th International Conference on Intelligent User Interfaces, New York, NY, USA, 18–21 March 2024; IUI '24'; pp. 731–741. [CrossRef]
94. von Sawitzky, T.; Wintersberger, P.; Riener, A.; Gabbard, J.L. Increasing trust in fully automated driving: Route indication on an Augmented Reality head-up display. In Proceedings of the 8th ACM International Symposium on Pervasive Displays, New York, NY, USA, 12–14 June 2019; PerDis '19. [CrossRef]
95. Chang, C.J.; Hsu, Y.L.; Tan, W.T.M.; Chang, Y.C.; Lu, P.C.; Chen, Y.; Wang, Y.H.; Chen, M.Y. Exploring Augmented Reality Interface Designs for Virtual Meetings in Real-world Walking Contexts. In Proceedings of the 2024 ACM Designing Interactive Systems Conference, New York, NY, USA, 1–5 July 2024; DIS '24; pp. 391–408. [CrossRef]
96. Norman, D.A. *The Design of Everyday Things*, Revised Edition; Basic Books: New York, NY, USA, 2013.
97. Laugwitz, B.; Held, T.; Schrepp, M. Construction and Evaluation of a User Experience Questionnaire. In *HCI and Usability for Education and Work*; Holzinger, A., Ed.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2008; Volume 5298, pp. 63–76. [CrossRef]
98. Brooke, J. SUS—A quick and dirty usability scale . In *Usability Evaluation in Industry*; CRC Press: Boca Raton, FL, USA, 1996.
99. Vredenburg, K.; Mao, J.Y.; Smith, P.W.; Carey, T. A survey of user-centered design practice. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA, 20–25 April 2002; CHI '02; pp. 471–478. [CrossRef]
100. Abras, C.; Maloney-Krichmar, D.; Preece, J. User-Centered Design. In *Bainbridge, W. Encyclopedia of Human-Computer Interaction*; Sage Publications: Thousand Oaks, CA, USA, 2004.
101. Geerts, D.; Vatavu, R.D.; Burova, A.; Vinayagamoorthy, V.; Mott, M.; Crabb, M.; Gerling, K. Challenges in Designing Inclusive Immersive Technologies. In Proceedings of the 20th International Conference on Mobile and Ubiquitous Multimedia, New York, NY, USA, 5–8 December 2021; MUM '21; pp. 182–185. [CrossRef]
102. Tanevska, A.; Chandra, S.; Barbareschi, G.; Eguchi, A.; Han, Z.; Korpan, R.; Ostrowski, A.K.; Perugia, G.; Ravindranath, S.; Seaborn, K.; et al. Inclusive HRI II: Equity and Diversity in Design, Application, Methods, and Community. In Proceedings of the Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, New York, NY, USA, 13–16 March 2023; HRI '23; pp. 956–958. [CrossRef]
103. Ejaz, A.; Syed, D.; Yasir, M.; Farhan, D. Graphic User Interface Design Principles for Designing Augmented Reality Applications. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 0100228. [CrossRef]
104. Knowles, B.; Clear, A.K.; Mann, S.; Blevis, E.; Håkansson, M. Design Patterns, Principles, and Strategies for Sustainable HCI. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, New York, NY, USA, 7–12 May 2016; CHI EA '16; pp. 3581–3588. [CrossRef]
105. Nebeling, M.; Oki, M.; Gelsomini, M.; Hayes, G.R.; Billinghurst, M.; Suzuki, K.; Graf, R. Designing Inclusive Future Augmented Realities. In Proceedings of the Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 11–16 May 2024; CHI EA '24; pp. 1–6. [CrossRef]
106. Samaradivakara, Y.; Ushan, T.; Pathirage, A.; Sasikumar, P.; Karunanayaka, K.; Keppitiyagama, C.; Nanayakkara, S. SeEar: Tailoring Real-time AR Caption Interfaces for Deaf and Hard-of-Hearing (DHH) Students in Specialized Educational Settings. In Proceedings of the Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 11–16 May 2024 2024; CHI EA '24; pp. 1–8. [CrossRef]
107. Fang, H.C.; Ong, S.K.; Nee, A.Y.C. Novel AR-based interface for Human–Robot Interaction and visualization. *Adv. Manuf.* **2014**, *2*, 275–288. [CrossRef]

108. Alt, B.; Zahn, J.; Kienle, C.; Dvorak, J.; May, M.; Katic, D.; Jäkel, R.; Kopp, T.; Beetz, M.; Lanza, G. Human-AI Interaction in Industrial Robotics: Design and Empirical Evaluation of a User Interface for Explainable AI-Based Robot Program Optimization. *arXiv* **2024**, arXiv:2404.19349.
109. Lindblom, J.; Alenljung, B. The ANEMONE: Theoretical Foundations for UX Evaluation of Action and Intention Recognition in Human-Robot Interaction. *Sensors* **2020**, *20*, 4284. [CrossRef]
110. Jeffri, N.F.S.; Rambli, D.R.A. Guidelines for the Interface Design of AR Systems for Manual Assembly. In Proceedings of the 2020 4th International Conference on Virtual and Augmented Reality Simulations, New York, NY, USA, 14–16 February 2020; ICVARS '20; pp. 70–77. [CrossRef]
111. Wewerka, J.; Micus, C.; Reichert, M. Seven Guidelines for Designing the User Interface in Robotic Process Automation. In Proceedings of the 2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW), Gold Coast, Australia, 25–29 October 2021; pp. 157–165. ISSN: 2325-6605. [CrossRef]
112. Zhao, Y.; Masuda, L.; Loke, L.; Reinhardt, D. Towards a Design Toolkit for Designing AR Interface with Head-Mounted Display for Close-Proximity Human–Robot Collaboration in Fabrication. In *Collaboration Technologies and Social Computing*; Takada, H., Marutschke, D.M., Alvarez, C., Inoue, T., Hayashi, Y., Hernandez-Leo, D., Eds.; Springer: Cham, Switzerland, 2023; pp. 135–143. [CrossRef]
113. Marvel, J.A. Performance Metrics of Speed and Separation Monitoring in Shared Workspaces. *IEEE Trans. Autom. Sci. Eng.* **2013**, *10*, 405–414. [CrossRef]
114. Kumar, S.P. Dynamic Speed and Separation Monitoring with On-Robot Ranging Sensor Arrays for Human and Industrial Robot Collaboration. Ph.D. Thesis, Rochester Institute of Technology, Rochester, NY, USA, 2020.
115. Scalera, L.; Giusti, A.; Vidoni, R.; Gasparetto, A. Enhancing fluency and productivity in Human–Robot collaboration through online scaling of dynamic safety zones. *Int. J. Adv. Manuf. Technol.* **2022**, *121*, 6783–6798. [CrossRef]
116. Zanchettin, A.M.; Lacevic, B. Safe and minimum-time path-following problem for collaborative industrial robots. *J. Manuf. Syst.* **2022**, *65*, 686–693. [CrossRef]
117. Savur, C. A Physiological Computing System to Improve Human–Robot Collaboration by Using Human Comfort Index. Ph.D. Thesis, Rochester Institute of Technology, Rochester, NY, USA, 2022.
118. Chacón, A.; Ponsa, P.; Angulo, C. Cognitive Interaction Analysis in Human–Robot Collaboration Using an Assembly Task. *Electronics* **2021**, *10*, 1317. [CrossRef]
119. Bradley, M.M.; Lang, P.J. Measuring emotion: The self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* **1994**, *25*, 49–59. [CrossRef]
120. Betella, A.; Verschure, P.F.M.J. The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions. *PLoS ONE* **2016**, *11*, e0148037. [CrossRef] [PubMed]
121. Legler, F.; Trezl, J.; Langer, D.; Bernhagen, M.; Dettmann, A.; Bullinger, A.C. Emotional Experience in Human–Robot Collaboration: Suitability of Virtual Reality Scenarios to Study Interactions beyond Safety Restrictions. *Robotics* **2023**, *12*, 168. [CrossRef]
122. Savur, C.; Sahin, F. Survey on Physiological Computing in Human–Robot Collaboration. *Machines* **2023**, *11*, 536. [CrossRef]
123. Sahin, M.; Subramanian, K.; Sahin, F. Using Augmented Reality to Enhance Worker Situational Awareness in Human-Robot Interaction. In Proceedings of the 2024 IEEE Conference on Telepresence , California Institute of Technology, Pasadena, California, USA, 16–17 November 2024; Accepted for Presentation. [CrossRef]
124. Subramanian, K.; Arora, S.; Adamides, O.; Sahin, F. Using Mixed Reality for Safe Physical Human–Robot Interaction. In Proceedings of the 2024 IEEE Conference on Telepresence , California Institute of Technology, Pasadena, California, USA, 16–17 November 2024; Accepted for Presentation. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Proposed Multi-ST Model for Collaborating Multiple Robots in Dynamic Environments

Hai Van Pham [1,*], Huy Quoc Do [1], Minh Nguyen Quang [2], Farzin Asadi [3] and Philip Moore [4]

[1] School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi 100000, Vietnam; huydq@soict.hust.edu.vn
[2] School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi 100000, Vietnam; minh.nguyenquang@hust.edu.vn
[3] Department of Computer Engineering, OSTIM Technical University, Ankara 06374, Turkey; farzin.asadi@ostimteknik.edu.tr
[4] School of Information Science and Engineering, Lanzhou University, Lanzhou 730030, China; ptmbcu@gmail.com
* Correspondence: haipv@soict.hust.edu.vn

Abstract: Coverage path planning describes the process of finding an effective path robots can take to traverse a defined dynamic operating environment where there are static (fixed) and dynamic (mobile) obstacles that must be located and avoided in coverage path planning. However, most coverage path planning methods are limited in their ability to effectively manage the coordination of multiple robots operating in concert. In this paper, we propose a novel coverage path planning model (termed *Multi-ST*) which utilizes the spiral-spanning tree coverage algorithm with intelligent reasoning and knowledge-based methods to achieve optimal coverage, obstacle avoidance, and robot coordination. In experimental testing, we have evaluated the proposed model with a comparative analysis of alternative current approaches under the same conditions. The reported results show that the proposed model enables the avoidance of static and moving obstacles by multiple robots operating in concert in a dynamic operating environment. Moreover, the results demonstrate that the proposed model outperforms existing coverage path planning methods in terms of coverage quality, robustness, scalability, and efficiency. In this paper, the assumptions, limitations, and constraints applicable to this study are set out along with related challenges, open research questions, and proposed directions for future research. We posit that our proposed approach can provide an effective basis upon which multiple robots can operate in concert in a range of 'real-world' domains and systems where coverage path planning and the avoidance of static and dynamic obstacles encountered in completing tasks is a systemic requirement.

Keywords: spiral spanning tree algorithm; coverage path planning; multi-ST algorithm; multiple robot operation; multiple robot coordination; knowledge-based systems

1. Introduction

Research has investigated the management of autonomous robots with studies addressing both single and multiple robot operation. Robots are required to navigate a specified operating space (hereafter termed: the defined operating environment (DOE)) where there are static (fixed) and dynamic (moving) obstacles. Such navigation requires the identification of an efficient path the robot will follow to complete the set task where the path will be defined using Coverage Path Planning (CPP) [1–4]. When considering dynamic obstacles where multiple mobile robots operate in concert (defined as arranging something by mutual agreement or coordination), the robots are, in effect, dynamic obstacles and CPP must accommodate all static and dynamic obstacles in 'real-time'.

Robots can take many manifestations which include robots operating at ground level [using two-dimensional CPP] to unmanned aerial vehicles (UAV) [using three-dimensional

CPP] where the goal is to optimize the operation of multiple UAVs while avoiding collisions [5,6]. We defer a discussion on related research to Section 2. In this study we are interested in the operation of multiple mobile robots operating in concert, the goal being to autonomously optimize the paths taken by the mobile robots within a DOE to achieve a pre-defined task [2] while implementing the avoidance of static and dynamic obstacles [4].

In this paper, we present our proposed *Multi-ST* model which combines intelligent robot reasoning with knowledge-based reasoning techniques to realize effective and flexible CPP in dynamic unknown environments. The *Multi-ST* model uses the Spiral Spanning Tree Coverage (SSTC) algorithm to enable a robot (or multiple robots operating in concert) to autonomously implement path planning decisions. The proposed model applies knowledge inference with the SSTC algorithm to enable optimal CPP with effective obstacle avoidance. In our research we have divided the CPP algorithms into two approaches using the spanning tree algorithm (STA) [7,8]: (i) *offline*, and (ii) *online*. We introduce these methods in Section 3.2 with the proposed *Multi-ST* model introduced in Section 4.

In experimental testing, we have evaluated the *Multi-ST* model in a simulation case study which implements physical robots in a cleaning task in a DOE including a comparative analysis of alternative current approaches under the same conditions. The results are reflected in performance metrics which include the *coverage percentage*, *path length*, and *execution time*. The reported results demonstrate the effectiveness of the *Multi-ST* model and its capability to implement the avoidance of static and moving obstacles by multiple robots operating in concert in a DOE. Moreover, the comparative analysis shows that our *Multi-ST* model outperforms existing models in coverage quality, robustness, and scalability.

Our contribution builds on earlier published studies where CPP with linguistic and semantic methods have been employed. Our contribution may be summarized as follows:

- The *Multi-ST* model enables single and collaborating multiple robots to achieve optimal CPP while implementing avoidance of both static and dynamic obstacles. Novel obstacle avoidance is implemented using a rule-based approach with updating of the rules in the knowledge base (KB) to optimize CPP.
- The proposal set out in this paper provides an effective basis upon which the complex problems resulting from CPP by multiple collaborating mobile robots can be solved (or at least mitigated). Where multiple robots operate in concert in a DOE, each robot can exchange the coverage and obstacle information with other collaborating robots.
- Multiple robots operating in a DOE can complete a defined task more effectively than is the case for a single robot which provides: (a) operational advantages in diverse domains and systems, and (b) significant savings in terms of time with improved effectiveness when employed in the exploration and coverage of unknown environments.

The limitations of this study are set out along with related challenges, open research questions, and proposed directions for future research. We posit that our proposed approach can provide an effective basis upon which multiple robots can operate in concert in a range of 'real-world' domains and systems where CPP and the avoidance of static and dynamic obstacles encountered in completing tasks is a systemic requirement.

The remainder of this paper is as follows: related research isconsidered in Section 2. An overview of robot sensor design and development is set out in Section 3.1. Preliminaries are introduced in Section 3 where algorithm design is considered. The proposed *Multi-ST* model is set out in Section 4 with the processing by robots discussed in Section 5.2. Testing and evaluation is described in Section 6 with the results provided in Section 7. Section 8 provides a discussion with consideration of open research questions and directions for future research. The paper closes with concluding observations in Section 9.

2. Related Research

This paper is not a systematic survey of the field as the literature contains a large volume of published studies investigating CPP in heterogeneous domains as discussed

in [1,2,9] where a multiplicity of published studies proposing methods to address CPP are introduced. For a survey on multi-robot CPP for model reconstruction and mapping see [10] with an overview of formulations and solution procedures for the multiple traveling salesman problem provided in [11]. Choset in [2] provides a survey of recent results related to CPP and "coverage for robotics". Galceran and Carreras in [1] address CPP with a survey of the "most successful" CPP methods with a focus on the achievements made in the past decade. Galceran and Carreras consider CPP methods with the field of CPP for successful works. Significant examples of studies related to CPP that are relevant to the research which forms the focus of this paper are considered in this section.

CPP algorithms that rely on the SSTC algorithm and/or CPP problem constraints can be classified into two basic approaches: (i) complete coverage algorithms, and (ii) optimal coverage algorithms. In summary: complete coverage algorithms cover the entire DOE. In some studies, the algorithm is required to optimize the DOE including consideration of factors and conditions such as power and the maximum elapsed time. When the work area is large, complete coverage in CPP cannot be guaranteed; this is a limitation not shared by the optimal coverage algorithms.

The online sensor-based CPP is addressed by Gabriely et al. using an SSTC algorithm to implement CPP using the Spiral-STC in [8]. The algorithm has been implemented for a single robot operating alone with reported results claiming to realize a "guaranteed return-time bound". For multi-robot operations, Hazon and Kaminka [12] propose the multi-robot SSTC algorithm termed MSTC where improvement in coverage time is achieved by backtracking optimization. However, the results for the MSTC algorithm are dependent on parameters to compute the edge cost in covering and spanning graphs. Agmon et al. have introduced an SSTC algorithm with two phases [7]: (i) a polynomial-time tree-construction algorithm for *offline* coverage, and (ii) an *online* coverage of a finite terrain based on a spanning-tree.

The multi-robot forest coverage (MFC) algorithm is a polynomial-time multi-robot coverage algorithm based on an algorithm introduced by Even et al. in [13]. The MFC algorithm is proposed in [14] where the approach utilizes multiple minimal spanning trees to cover the terrain generated by the *min-max* tree cover algorithm. In practical situations, there are many physical constraints that must be considered and addressed in CPP and analysis of these approaches identifies weaknesses in that they fail to consider constraints implicit in robot operation.

Zheng et al. in [15] consider CPP and the MFC algorithm providing simulation results that demonstrate that coverage times for the MFC method have been tested in scenarios of alternative multi-robot coverage algorithms. Zheng et al. show that solving several versions of multi-robot coverage problems with minimal coverage time is an NP-hard problem [16,17] and argue that this provides motivation for designing polynomial-time constant-factor approximation algorithms. Even et al. [13] consider constant factor approximation algorithms for covering the nodes of graphs using trees (*rooted* or *unrooted*) and argue that such problems are related to location routing and the traveling salesman problem [16]. In considering NP hard problems Cheeseman et al. in [18] consider "where the really hard problems are". Additional discussion on random graphs can be found in [19–21] with a guide to the theory of NP-completeness as it relates to computers and intractability provided in [22]. We consider NP hard problems as they relate to this study CPP in Section 8.

CPP algorithms incorporating constraint satisfaction have been addressed in recently published studies. For example in [7,12–14] [introduced previously in this section] CPP for a single robot is addressed with consideration of the measurement of robot energy limitation. Zheng et al. in [23] present a study where CPP relates to multi-robot forest coverage where the DOE can be covered more effectively and efficiently using multi-robot operation. Zheng et al. observe that coverage times are "at most eight times larger than optimal" and their reported experimental results show that the proposed method performs "significantly better than existing multi-robot coverage algorithms". Pengzhan Chen et

al. have investigated CPP using deep reinforcement learning to target a destination for dynamic obstacle avoidance [24]. Studies have applied a multi-strategy particle swarm approach using exponential noise in order to solve low altitude penetration in secure space [25].

An interesting study in [26] proposes a new method for the path planning of the robotic arm mounted on the free-floating unmanned spacecraft where the path of the arm is described in the joint space by a spline function and the coordinates of spline knots form the vector of decision variables. The proposed model generates a collision-free path of the arm that "results in the desired position and orientation of the gripper and the desired attitude of the spacecraft". The reported results show that the path obtained using the proposed model is shorter and smoother than the bi-directional Rapidly-exploring Random Trees (RRT) algorithm and the RRT path.

Complete CPP algorithms are represented in the decomposition strategies in both static and dynamic environments [1], a decomposition strategy being defined as sub-regions or cells that use space using simple trapezoidal or triangular shapes [27]. As noted in [28] a "Morse function" can be utilized for CPP in dynamic environments. For sensor-based robot systems, there is a range of graph-based methods [29] including 3D data [30,31], and landmarks [32] in CPP, with grid-based decomposition as proposed in [2,33]. Sipahioglu et al. in [34] propose a novel algorithm for sensor-based [robot] coverage of narrow environments by considering the energy capacities of the robots. In the model proposed in [34], path planning was "modeled by a Generalized *Voronoi diagram-based graph* to guarantee complete sensor-based coverage".

A number of algorithms can be used to 'section' a DOE based on a grid strategy including the 'energy-aware' algorithm neural networks [35,36], and spanning trees [8]. The studies reduce the complexity required in terms of computational overhead incurred in determining a coverage path with the majority of CPP algorithms being based on the grid decomposition strategy. In related works, robot path planning is used to apply geometric models [37]. CPP for a mobile robot has been achieved by visiting a map in a dynamic environment [38], for example, selected map locations have been determined in [39] with CPP in DOE without obstacles being addressed in [9].

Huang et al. in [40] have investigated multi-robot coverage path planning (MCPP) algorithms to solve the CPP problem of robots in specific areas including outdoor environments, different landscapes, and emergency search and rescue tasks. The inherent complexities are noted along with a method to address such complexities. In summary, Huang et al. describe the "visual fields" of the robots by constructing hierarchical and topological relationships among the cells in the same and different layers. Reported results claim that the CPP achieved is an "efficient MCPP algorithm".

Research has investigated intelligent robot CPP with a view to energy saving for both single and multiple robots in uncertain environments [41–44]. Hameed in [41] considers intelligent CPP for agricultural robots and autonomous machines in a 3D Terrain. Shnaps and Rimon in [42] address the online coverage of unknown planar environments by a battery-powered autonomous mobile robot operating with a limited energy capacity battery is considered. CPP in situations where energy constraint is the significant feature is discussed by Wei and Isler in [43,44] where energy-efficient CPP for general terrain by an autonomous vehicle is addressed. Applied energy-constrained multi-robot sensor network systems are designed to minimize energy cost which is a solution for aggregating multiple robots [34,40].

In a related study, Tang et al. in [45] note that for large-scale tasks, effective coverage with CPP can benefit greatly from the use of multiple robots operating in concert. Tang et al. introduce the M^* algorithm for multi-robot coverage path planning (mCPP) based on spiral spanning tree coverage (Spiral-STC) which includes physical constraints such as "terrain *traversability* and material load capacity". In a comparative analysis with the state-of-the-art in mCPP for regular grid maps and 'real-field-terrains' in simulation environments, the reported results support the conclusion that the M^* method outperforms current existing

SSTC-based mCPP methods. However, this algorithm cannot guarantee to find the optimal solution and importantly has not been verified and validated in experimental testing using actual physical robots.

In a paper entitled "On optimal coverage of a tree with multiple robots" Aldana-Galván et al. in [46] study the algorithmic problem of optimally covering a tree with k mobile robots where the tree is known to all robots. The goal here is to assign a walk to each robot in such a way that the union of these walks covers the whole tree. The study considers a variant in which the robots must rendezvous periodically at the same vertex in a certain number of moves and it is shown that for the *two cost functions* the problem formulation is different. For the "cover time minimization problem" the problem is shown to be NP-hard when k is part of the input irrespective of the requirement for periodic rendezvous. Aldana-Galván et al. posit that "the cover length minimization problem can be solved in polynomial time when periodic rendezvous are not required, and it is NP-hard otherwise".

The A^* algorithm is an approach to CPP employed in [47,48]. The A^* algorithm has been used to control autonomous mobile robots as shown in [49,50]. The Spanning Tree Coverage (STC) algorithm has been used in CPP [8] with knowledge reasoning applied in [4,51] where it is argued that robot CPP and the STC algorithm with knowledge reasoning represents the optimal approach for decision-making for single robot operation. While the methods identified have shown significant experimental results there are limitations when implemented in dynamic environments where coordinated multiple mobile robots operating with optimal CPP are the preferred solution to realize the designated task.

Some advanced techniques such as knowledge graphs (KG) and fuzzy knowledge graphs (FKG) with reasoning can be applied to the decision-making of robots for healthcare domains [52–55]. In [52] a "novel Q-learning-based FKG-Pairs approach for extreme cases in decision making" is presented with the conclusion drawn that the FKG-Pairs approach is considered one of the best models to solve classification problems based on "uncertain and amplitude input datasets" as it can infer and find the output labels of new samples that are not in the fuzzy rule base. Further work in this area is presented in [54] where a novel fuzzy KG pairs approach to decision making is proposed.

In [55] a proposed deep learning model integrated with a KG to enable the monitoring of human behavior in forest protection situations is shown. Pham et al. in [55] address the modeling of the relationship between human behavior and forest protection using a deep learning model integrated with a KG. A deep learning technique has been investigated as it relates to the use of a robot(s) in forest fire rescues [11]. Applied optimal coverage using a deep learning model integrated with a KG for monitoring human behavior in forest protection has demonstrated the capability to solve complex problems in decision making [56] and applied smart autonomous robots in forests can enhance the performance of robots using ChatGPT in finding the road under uncertain environments [57–59].

Other techniques, including reasoning and the crowd-drawing effect, can be applied in 'real-world' applications of single and multiple robots in multiple decision making [3,60]. There is an interesting correlation between the discussion in [60] and the social force model [61], crowd dynamics, and the selection of an available route(s) during evacuation in emergency situations [62]. A robot(s) with *a priori* knowledge of the architectural layout of an area may be a useful guide.

From this brief review of relevant related research, we may conclude that CPP has been shown to be successful in many domains using a range of methods and algorithms. In this paper, we present our proposed *Multi-ST* model which combines robot reasoning with knowledge reasoning techniques using the SSTC algorithm. The proposed model is designed to enable optimal CPP while avoiding both static and dynamic obstacles and experimental results show that the proposed approach enables the avoidance of obstacles by multiple robots operating in a dynamic DOE with improvements in CPP and efficiency.

3. Preliminaries

In Section 2 we have identified a number of significant relevant CPP methods and algorithms. In this section we briefly consider networked robot systems (NRS) and provide an overview of spanning Trees (ST) (see Section 3.2), the online Spiral STC algorithm (see Section 3.3), and the full spiral-STC algorithm in (see Section 3.4). The proposed *Multi-ST* model is introduced in Section 4.

3.1. Networked Robot Systems

Prior to considering our proposed *Multi-ST* model, we provide a brief overview of sensors and related technologies as they relate to this study, such technologies are generally termed NRS. In this study, it is assumed that robots are equipped with networked onboard sensors designed to provide the data required to enable the detection of all obstacles, boundaries, and directions of travel. Moreover, the sensors will allow robot(s) to identify and scan the surrounding mega-cells and also in the current mega-cell.

Sensor technologies represent a large and complex topic and a detailed consideration is beyond the scope of this paper; however, readers interested in advances in robotics and related sensor technologies may consider the research documented in [63–68] where NRS and advances in sensor technologies are addressed.

Kala et al. in [63] consider emerging methods and their application in modeling, identification and robot control with the "multi-heterogeneous sensor data fusion method via convolutional neural network for fault diagnosis of wheeled mobile robot" operation addressed in [64]. The "3D self-deployment of jumping robot sensor nodes for improving network performance in obstacle dense environment" is addressed by Zhang et al. in [65].

Network Robot Systems (NRS) are introduced with related definitions in [66] "as is understood in Europe, USA and Japan". Moreover, it describes some of the NRS projects in Europe and Japan and presents a summary of the papers published in a special issue on the topic. Sanfeliu et al. in [67] extend previous research in a study investigating "decentralized sensor fusion for ubiquitous networking robotics in urban areas". The paper considers the architecture for the environment and sensors that have been built for the European Ubiquitous Networking Robotics in Urban Sites project.

Borenstein et al. in [68] have studied sensors and related techniques for the positioning of mobile robots. Seven categories for positioning systems are identified: (1) *Odometry*, (2) *Inertial Navigation*, (3) *Magnetic Compasses*, (4) *Active Beacons*, (5) *Global Positioning Systems*, (6) *Landmark Navigation*, and (7) *Model Matching*. However, given the multiplicity of "ingenious approaches", Borenstein et al. caveat that "for reasons of brevity, not all could be cited in this article".

There are seminal research studies where advanced sensor design has been realised; see for example the work of Sheila Nirenberg et al. and colleagues in [69,70]. The research had a focus on human vision with research addressing the "classification of Retinal Ganglion Cells" [69]; results from numerous studies have shown that such cells exhibit an array of responses to visual stimuli resulting in the notion that "these cells can be sorted into distinct physiological classes, such as linear versus non-linear or on versus off". In [70] Retinal Ganglion Cells have been shown to "act largely as independent encoders" and "correlated firing among neurons is widespread in the visual system". Nirenberg et al. have, in subsequent research, applied their research into human vision to robotic sensor systems with robots having vision sensors that can enable autonomous robot control. This work has interesting potential applications for robot control given the developments in artificial intelligence (AI) and machine vision as discussed in [71,72].

The design of autonomous robotic systems must address the required hardware and software systems where such systems will incorporate the hardware [and software functionality] in the form of sensors, actuators, and communication [between robots] where multiple robots operating in concert is clearly required. The operating system of a robot requires a variety of sensors including *vision*, *sound*, *location*, and an *accelerometer*. Moreover,

depending on the operating environment (e.g., an external location) robots may need *cloud-based systems*, *GPS*, and possibly the *Internet-of-Things* (IoT) connectivity.

3.2. Spanning Trees

Considered in terms of graph theory, an ST T for an undirected graph G is a sub-graph that is a tree that includes all of the vertices of G. A graph G may have several STs but a graph that is not connected will not contain an ST. If all of the edges of graph G are also edges of a ST T of G, then graph G is an identical tree to ST T.

A deterministic ST algorithm is introduced by Pettie and Ramachandran in [73] where it is established that the algorithmic complexity of the minimum ST problem is equal to its decision-tree complexity. The ST algorithm introduced by Pettie and Ramachandran aims to identify an ST for a graph with n vertices and m edges that runs in time $\{O(T^*(m,\ n))\}$ where T^* is the minimum number of edge-weight comparisons needed to determine the solution. In operation, when viewed from the perspective of this study, the ST is implemented using two methods: (i) *offline* and (ii) *online*:

***off-line* ST:** Provides *a priori* knowledge of the DOE to the robot(s) relating to static and dynamic obstacles. The algorithm finds an optimal path in linear time $O(n)$ [73] where n is the number of cells in the DOE. Pettie and Ramachandran note that "…although our time bound is optimal, the exact function describing it is not known at present …", for a detailed discussion on the topic see [73].

***on-line* ST:** Provides *no a priori* knowledge of the dynamic DOE where the location of static obstacles will be *a priori* known. This method addresses the related mathematics using information such as the location of static [generally known] and/or dynamic [generally unknown] obstacles. Dynamic obstacles will include cooperating with multiple robots operating in concert in the same DOE. Using the knowledge database, the online method can manage both static and dynamic obstacles. In the *online* ST algorithm, the robot incrementally constructs a spanning tree for a grid representing the DOE. The process is as follows:

Step 1 The path of the robot is formed from the mega-cell containing the initial position of the robot to an empty neighbor mega-cell in a counter-clockwise direction.

Step 2 During the construction of the spanning tree, the robot subdivides every cell it encounters into four identical sub-cells of $[2 \times 2]$ size.

Step 3 Following step 2, the robot follows a sub-cell path that circumnavigates the incrementally constructed spanning tree until the defined operational environment is covered.

Step 4 The robot then returns to the initial (starting) position in the opposite direction, avoiding traversing any cell previously traversed twice (because the mega-cell size is of $[2 \times 2]$ size).

The ST algorithm is based on the following:

Assumption: the size of mega-cell is $[2 \times 2]$ where the smaller cell size is $[1]$.

Input: for the robot's initial position cell there is no *a priori* knowledge of the DOE.

Recursive ST $(w,\ x)$ **function:** where (x) is a cell containing the starting robot position and (w) is the parent cell.

Initial ST $(null,\ s)$ **function:** where (s) is the starting cell.

The algorithmic process is as follows:

1 Mark the current cell (x) as an old cell

2 Scan the neighboring cell of (x) in counter-clockwise order
While (x has free neighboring cell):
Cell (y) is free
Construct a spanning tree edge from $(x\ to\ y)$

Move to a sub-cell of (y) as described below
Execute function STC $(w,\ x)$
End While

3 If the current cell is not the initial cell $(x \# s)$ return to the parent -cell

4 Terminate function STC $(w,\ x)$

The path-finding procedure of the algorithm is not complex and the relative simplicity of the process is a result of a robot recognizing mega-cells with no obstacles. When assessing adjacent mega-cells, the robot prioritizes them in a counter-clockwise sequence (i.e., *up*, *left*, *down*, and finally *right*. For scenarios where no mega-cell is identified, the robot chooses to revert to a parent cell. The construction process for the STA edge is modeled in Figure 1.

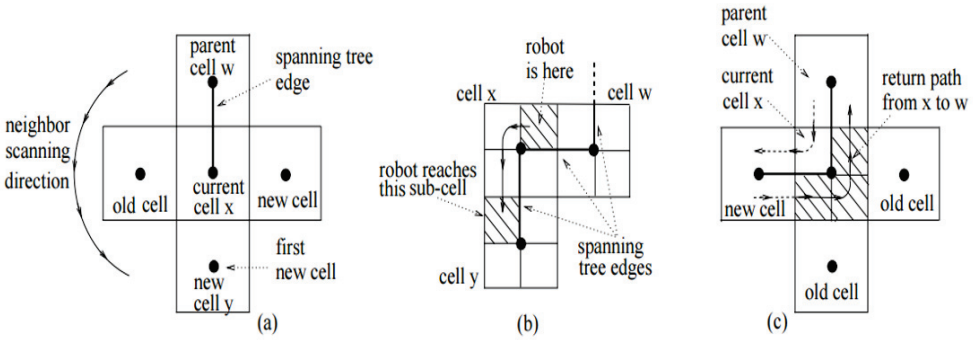


Figure 1. A model showing the path-finding procedure of the STA; (**a**) counterclockwise scanning of four neighbors. (**b**) a move from x to a new cell y. (**c**) a return from x to a parent cell w.

3.3. The Online Spiral Spanning Tree Algorithm

The ST algorithm has been shown to provide an effective basis upon which CPP can be achieved in unknown environments. However, it is not without limitations as it may not be capable of covering the whole of the DOE due to the potential for unidentified obstacles. This limitation has spurred the development of the *online* spiral spanning tree (OSST) algorithm which is an extension of the ST algorithm with the goal of ensuring complete coverage of the DOE.

In the CPP process path-finding introduces increased complexity and complication as a robot may try to traverse a mega-cell containing obstacles. Consequently, the assumption that a robot traverses the right side of spanning tree edges is not universally accurate. In certain instances, as depicted in Figure 2, robots must navigate in varying directions to guarantee complete coverage. When a robot encounters such a situation, as illustrated in Figure 2, it must pass through the edge of the spanning tree, as modeled in Figure 3.

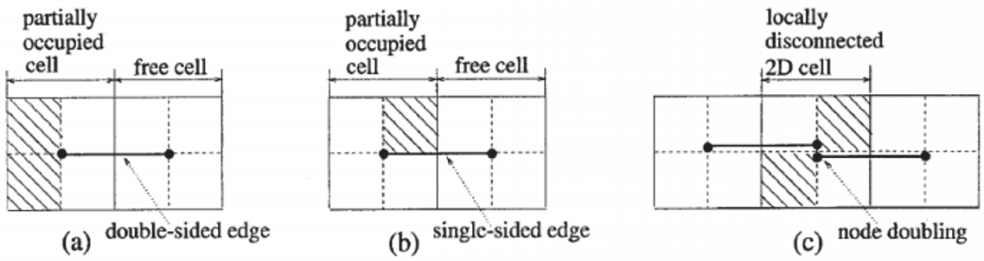


Figure 2. Special side-edges in STC consisting of (**a**) double-sided edge; (**b**) single-side edge; (**c**) node doubling.

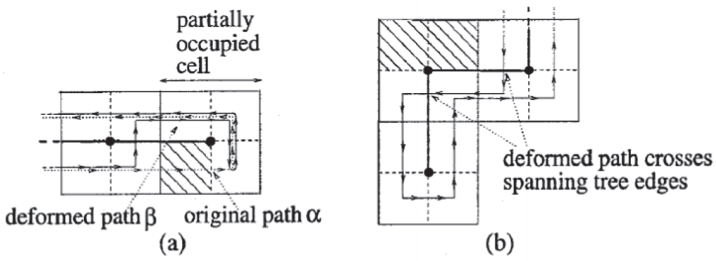


Figure 3. The approach to manage (i.e., avoid) special obstacles consisting of (**a**) partially occupied cell; (**b**) deformed path crosses spanning tree edges.

3.4. The Full Spiral Spanning Tree Algorithm

Here, we introduce the algorithmic process for the full spiral spanning tree (FSST) Algorithm:

Initialization: Call ST $(Null,\ S)$ where (S) S is the starting cell

FSST $(p,\ c)$

1 Mark the current cell (x) as an old cell

2 Scan the neighbor cell of (c) in counter-clockwise order.
    While ((c) has new obstacles-free neighbor):
        If ((n) find obstacle):
            If there is a cell available return (n)
            Else continue to scan other neighbors
            If (n) is suitable
                Construct a spanning tree edge from $(c\ to\ n)$
                Move to cell (n)
                Call recursive STC-full $(c,\ n)$
    End While

3 If the current cell is not the initial cell, return to the parent cell

4 Terminate FSST $(p,\ c)$

The pseudo-code is as follows:

```
FSST() {
        current.setStatus = false;
        if (not scan neighbor cell) {
                neighbor = findNewNeighbor(current);
        } else {
                robotWork (x, y);
        }
}
```

4. The Proposed *Multi-ST* Model

In Section 3 we have provided an overview of a number of significant relevant CPP methods and algorithms including the ST algorithm (see Section 3.2), and the FSST algorithm (see Section 3.4). In this section we introduce our proposed *Multi-ST* Model to collaborate multiple robots in a dynamic DOE as shown in Figure 4.

The proposed model consists of seven steps as follows: Step 1, identify multiple objectives for collaborative multiple decision-making in support multiple robots; Steps 2–3, consider an objective function as updated decision variables by dealing with multiple decision-making of Robots; Step 4, consider constraints in multiple decision-making for aggregation of multiple robots' tasks; Step 5, a primary objective of the FSST algorithm is to ensure that each robot constructs a distinct spanning tree during the navigation of the DOE; Step 6, consider rule and update its weight to the Robotic Knowledge Base (KB);

Step 7, find the appropriate rules considered by experts and applied these rules' weights in the KB. In the processing steps, multiple robot operations can result in reductions in the time taken to cover the DOE given the parallel robot processing implemented in the FSST algorithm. In robot simulations in the experiments, the robot applies existing rules with good behaviors in the Inference Engine for coverage path planning.

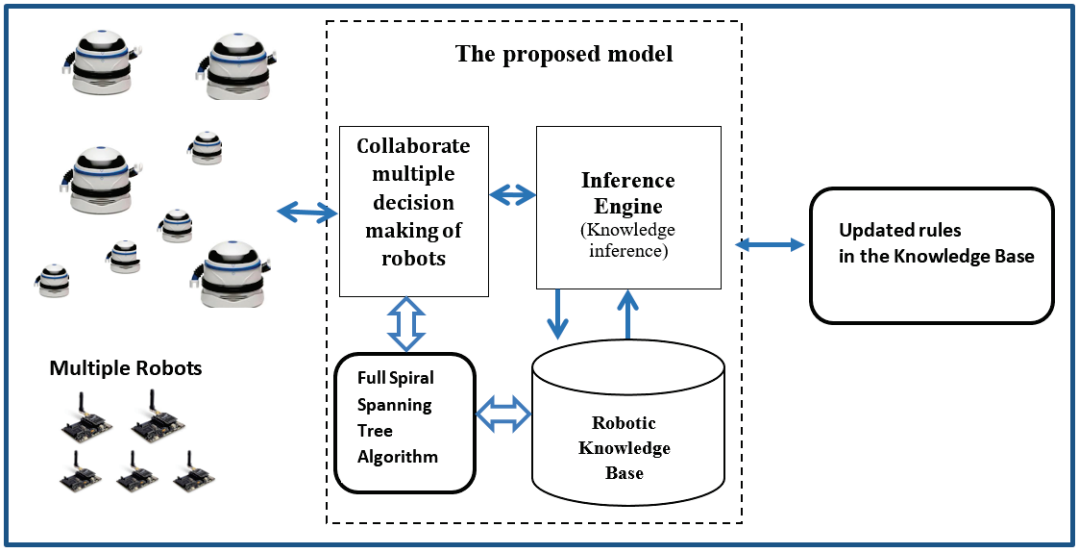


Figure 4. The proposed Multi-ST Model.

4.1. Conceptual Terms and Definitions

We can formulate a conceptual term and definition as follows:

1. Let CR be an objective function designed to minimize the optimal paths in priority selection of coverage robots in decision-making when dealing with multiple robots;
2. Let $(D_j(Y))$ be presented as a linguistic value in the variables and these linguistic values in the fuzzy weights $[0, 1]$;
3. Assume that a collaboration of multiple decision-making robots is present in domain S;
4. Let E be the decision makers' input for the proposed Multi-ST Model in domain S;
5. Let $RB^S = \{RB_1^S, RB_2^S, \ldots, RB_m^S\}$ be a set of Robots S, where m is the number of Robots.
6. Let the quantified semantics work in the mapping of (CR), presented in linguistic values for (w_j) fall in the range [0,1].
7. Calculate an objective function value for $(D_j(Y))$ in domain S;

4.2. Collaborate Multiple Decision Making of Robots

The proposed algorithm is described as follows:

Step 1: Implement collaborative multiple decision-making to support multiple robots. The proposed model identified multiple objectives for the optimal paths in the priority selection of coverage robots as shown in Equation (1):

$$CR = \frac{Min}{x} \sum w_j.D_j(Y) \quad (1)$$

where the constraints are as follows:

- $\sum_k Y_{ik} = 1 \forall i \in N$
- $\sum_{i \in P_j UQ_i} x_{i1} \geq 1$
- $x_{ik} = \{ \frac{1, If\ option\ i\ is\ decided\ aRobot\ to\ visit\ (k)}{0, other}$

Step 2: Consider an objective function by dealing with multiple decision-making of Robots;

Step 3: The decision variables $\left(Y_{ij}\right)$ can be considered objectives in the multiple criteria decision-making of robots.

Step 4: Consider constraints in multiple decision-making for the aggregation of multiple robots' tasks $\left(P_i^S\right)$ where $\left(\left(P_i^S \cap Q_j^S\right) = (\theta, \forall i, j \in \{1, 2, \cdots r\})\right)$.

Step 5: Identify all paths identified in monitoring multiple robots using the FSST algorithm.

Step 6: Consider the rule and update its weight to the Knowledge Base for each node of $\left(V_i^S\right)$ visited, consider Rule(j) represented by $\left(R_j\right)$ and apply $\left(V_i^S\right)$ as follows. In fuzzy rule-based reasoning, the fuzzy rules are generally expressed for two input variables and one output as a decision variable, as given by Equation (2):

$$\begin{cases} R^i : \text{if } x_1 \text{ is } ConditionA_{i1} \text{ and } x_2 \text{ is } A_{i2} \text{ then } y \text{ is } V_i \\ \vdots \\ R^n : \text{if } x_1 \text{ is } ConditionA_{n1} \text{ and } x_2 \text{ is } A_{n2} \text{ then } y \text{ is } V_n \end{cases} \tag{2}$$

where x_1 and x_2 represent input-quantitative variables ($x_1 \in X_1, x_2 \in X_2$) to a fuzzy system and y represents the output-decision variable ($y \in Y$). A_{i1}, A_{i2} and B_i are fuzzy subsets of X_1, X_2 and Y, respectively. When the non-fuzzy input data "x_1 is x_1^* and x_2 is x_2^*" are given, the matched context degree $\left(w_j\right)$ is calculated by the *weight* $\left(w_j\right)$: representative of $\left(V_i^S\right)$ and $\left(w_j\right)$.

Step 7: Find the appropriate rules considered by decision-makers and apply these weights and rules in the KB. *IF* upon checking the KB, the rule currently being considered exists in the KB, *THEN* action the considered rule *ELSE* a new rule matched in the proposed model that can be added to the KB.

5. Experimental Results

In this section we present a multi-robot CPP case study in Section 5.1 with robot processing in CPP addressed in Section 5.2 and an analysis of the *Multi-ST* model implemented with the KB in Section 5.3.

5.1. Case Study of Multi-Robot Coverage Model

In operation, mobile robots must be capable of working independently and/or collaboratively within the DOE. However, mobile robots operating in concert present several potential issues, including:

- Effective collaboration: consider two mobile robots A and B. For robot A robot B is a potential dynamic obstacle and vice versa. Avoiding each other while completing the designated tasks requires managed collaboration.
- Division of the operating environment: how can robots cooperate in concert to assist with another robot's errors while working to complete the designated task.

The design of the Multi-ST (MSTC) algorithm is predicated on the following assumptions related to the coverage of the DOE which is decomposed into cells that are used by a robot(s) to navigate the DOE (introduced in Section 3.2):

1. A robot can move continuously among two sub-cells without obstacles in a vertical or horizontal direction.
2. There will be an initial position for each robot and the robot terminates its run at the initial starting point.
3. A complete coverage path is applied through a robot that has scanned the entire work area of a dynamic environment.

4. The robot's coordination works in each cell to avoid blocking another collaborating robot.

The implementation process is as shown below; the stepwise process implemented in the OMST algorithm is modeled in Figure 5 where the steps are as follows:

1 Robot R_1 moves to cell $[1,\ 1]$ and Robot R_2 moves to cell $[0,\ 2]$
2 In step 1 mega-cell $[0,\ 0]$ is marked as an old cell previously entered into the visited array for robot R_1 and this is repeated for the mega-cell $[2,\ 0]$ for the robot R_2
3 In step 2 while searching for new adjacent cells, robot R_2 failed to locate or identify new cells above or on the left because they belonged to the spanning tree of robot R_1. Therefore,robot R_2 moves to the cell below which was mega-cell $[2,\ 2]$
4 Robot R_1 failed to locate a new cell, and therefore, would return to the previous cell at step 3
5 The FSST algorithm terminates when the whole of the DOE has been covered and the robot, e.g., robot R_1, has returned to its original initializing cell

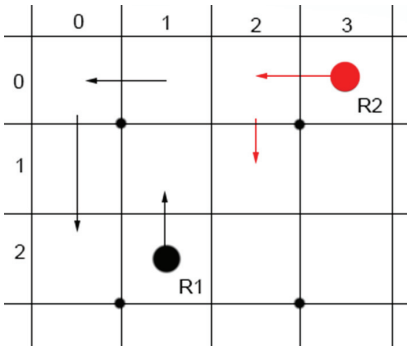


Figure 5. The stepwise process implemented in the OMST algorithm. Shown is a scenario where two robots share visited cell information.

As shown in Figure 6 the STC algorithm prioritizes a robot visit to a cell not previously visited followed by a cell that is reachable from an obstacle with the parent cell being given the lowest priority. For instance, when robot R_2 visits cell $[1,\ 1]$, its mega-cell is marked as an old mega-cell; however, robot R_2 would still enter this mega-cell as it is reachable, despite the under mega-cell of R_2 being unreachable. This scenario results in a high probability of collision between two robots when they enter the same mega-cell.

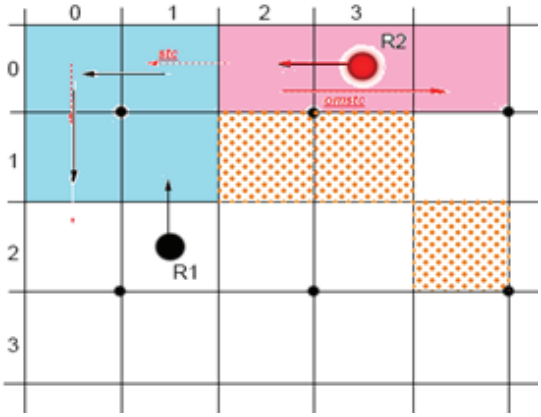


Figure 6. Two robots operating in the FSST algorithm where robot R_1 operates in the blue cells with robot R_2 operating in the red cells. The three hatched cells represent cells containing obstacles.

To address this potential collision scenario, the proposed solution to this issue is the FSST algorithm which has been conceived to ensure that a robot locates all unvisited cells

before returning to its parent cell. In the scenario modeled in Figure 6, when the mega-cell containing cell $[0,\ 0]$ is marked as an old cell by robot R_1, robot R_2 is not allowed to enter cell $[0,\ 0]$ because that mega-cell has been added to the visited array. Consequently, robot R_2 would return to its previous cell.

The primary objective of the FSST algorithm is to ensure that each robot constructs a distinct spanning tree during the navigation of the DOE. Moreover, multiple robot operations can result in reductions in the time taken to cover the DOE given the parallel robot processing implemented in the FSST algorithm.

5.2. Robot Processing in Coverage Path Planning Collaborations

Each robot runs its own algorithm to create its own convex hull during the exploration process. The process adopted to traverse a DOE may be summarized as follows:

- The initial position of the robot is given as input to the algorithm and the output is the convex hull of that robot. The robot can scan neighboring cells clockwise (up-left-down-right) starting from its parent and create the convex hull clockwise.
- Following the scanning of the neighboring cells, the unvisited cells will be marked as visited. The robot moves to the right side of the convex hull and eventually reaches the position of the starting sub-cell. It is assumed that in practice the robot is capable of leaving traces in the sub-cells it covers.
- To track the traces created in the terrain we use a visited array so that robots can determine whether that sub-cell has been covered or not. It is also assumed that there is a detection device in the actual robot capable of checking the traces of sub-cells in the current cell and its four neighbors immediately. It is assumed that we will use a stack to store information about nodes after each robot visit. The steps and the state of the stack are represented as follows:
    - Push the current node onto the stack, find the first neighboring node (n) counter-clockwise from the current node.
    - If node (n) is not empty the robot moves from the current node to node (n). Mark the current node as the pre-node of node (n). Mark node (n) as visited.
    - If node (n) is empty pop the current node from the stack.
    - The robot moves from the current node to the node at the top of the stack. Mark the pre-node of the node at the top of the stack as the current node.
    - Repeat the process until the stack is empty.

The convex hull forms a closed curve that visits all sub-cells exactly once. The combination of areas covered by the convex hulls generated by different robots constitutes the entire DOE; therefore, choosing the starting positions for the robots will greatly affect the creation of each robot's convex hull as well as the overall optimal working time of the collaborating robots. Collision is also a common issue for multiple robots; however, in this algorithm, as each robot builds its own separate convex hulls, there will be no convex hulls overlapping with each other or with other robots, thus avoiding collisions with other robots in the same working area. Table 1 shows in tabulated form a description of the steps and the related state(s) of the stack(s).

Table 1. Description of steps and state of the stack.

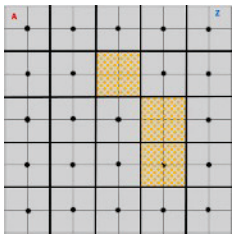
Stack R1	Stack R2	
A (Initial node of Robot 1)	Z (Initial node of Robot 2)	

Table 1. *Cont.*

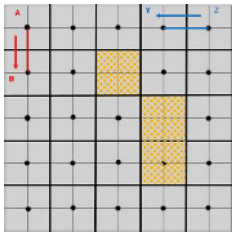
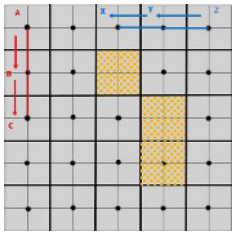
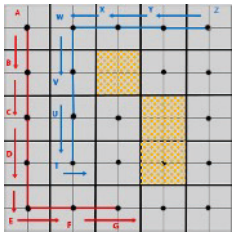
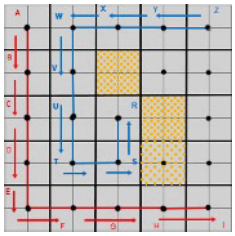
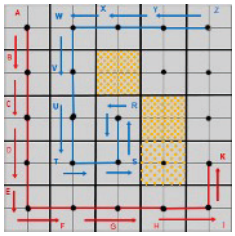
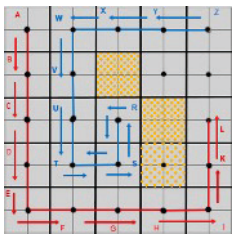
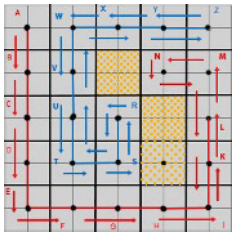
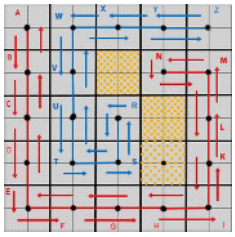
Stack R1	Stack R2	
A-B	Z-Y	
A-B-C	Z-Y-X	
…	…	…
A-B-C-D-E-F-G	Z-Y-X-W-V-U-T	
Push I	Push R	…
A-B-C-D-E-F-G-H-I	Z-Y-X-W-V-U-T-S-R	
Push K	Pop R	…
A-B-C-D-E-F-G-H-I-K	Z-Y-X-W-V-U-T-S	
Push L	Pop S	…
A-B-C-D-E-F-G-H-I-K-L	Z-Y-X-W-V-U-T	
…	…	…

Table 1. *Cont.*

Stack R1	Stack R2	
A-B-C-D-E-F-G-H-I	Pop Z	
…	null	…
Pop A	null	
null		…
	Finish	

In summary, the algorithm ensures that the convex hulls generated are separate for each robot. Moreover, a key motivating factor for using multiple robots is the ability to reduce coverage time by parallel processing and to this end the proposed algorithm is implemented in parallel by each robot to ensure optimal efficiency.

5.3. Analysis of the Multi-ST with the Knowledge Base

A major drawback of the ST algorithm is that it fails to accommodate and handle dynamic obstacles in the DOE. For example, in a scenario where a robot encounters a dynamic obstacle [within its current cell] it will stop all motion until the obstacle leaves the cell, this causes delays and inefficiencies in the robot's coverage and the defined task. To address this limitation, our proposed *Multi-ST* model applies the robot KB to enable the avoidance of obstacles, make subsequent action decisions, and enhance the performance and robustness of CPP in uncertain environments. When using the KB robots can locate and apply an appropriate rule from the KB. All the steps can be repeated until the robot completes the coverage of the DOE and its actions in multiple decision-making objectives. The events and rules that apply are in Tables 2 and 3.

Table 2. Robot navigation event rules and event codes.

Robot Event Rules and Event Code			
Event Code	**Rule Name**	**Event Code**	**Rule Name**
a1	Mobile obstacle ahead	a10	Wall on the left
a2	Mobile obstacle on the left	a11	Wall on the right
a3	Mobile obstacle on the right	a12	Wall behind
a4	Mobile obstacle behind	b1	Go ahead
a5	Another robot ahead	b2	Turn left
a6	Another robot on the left	b3	Turn right
a7	Another robot on the right	b4	Backward
a8	Another robot behind	b5	Stand
a9	Wall ahead		

Table 3. Examples of defined rules in the knowledge-base. The rules are essentially event-condition-action rules [74]. For the rules shown in this table, the assumption (IF) represents a condition which results in an action that, in turn, triggers a particular rule (code), e.g., for code 1: $IF\ (a1)\ THEN\ (b2)$.

Defined Rules in the Knowledge-Base					
Code	**IF**	**THEN**	**Code**	**IF**	**THEN**
1	a1	b2	44	a1&&a7&&a8	b2
2	a2	b1	45	a2&&a3&&a4	b1
3	a3	b1	46	a2&&a3&&a5	b4
4	a4	b1	47	a2&&a3&&a8	b1
5	a5	b2	48	a2&&a4&&a5	b3
6	a6	b1	49	a2&&a4&&a7	b1
7	a7	b1	50	a2&&a5&&a7	b4
8	a8	b1	51	a2&&a5&&a8	b3
9	a1&&a2	b3	52	a2&&a7&&a8	b1
10	a1&&a3	b2	53	a3&&a4&&a5	b2
11	a1&&a4	b2	54	a3&&a4&&a6	b1
12	a1&&a6	b3	55	a3&&a5&&a6	b4
13	a1&&a7	b2	56	a3&&a5&&a8	b2
14	a1&&a8	b2	57	a3&&a6&&a8	b1
15	a1&&a8	b1	58	a4&&a5&&a6	b3
16	a2&&a4	b1	59	a4&&a5&&a7	b2
17	a2&&a5	b3	60	a4&&a6&&a7	b1
18	a2&&a7	b1	61	a5&&a6&&a7	b4
19	a2&&a8	b1	62	a5&&a6&&a8	b3
20	a3&&a4	b1	63	a6&&a7&&a8	b1
21	a3&&a5	b2	64	a1&&a2&&a3&&a4	b5
22	a3&&a6	b1	65	a1&&a2&&a3&&a8	b5
23	a3&&a8	b1	66	a1&&a2&&a7&&a4	b5
24	a4&&a5	b2	67	a1&&a2&&a7&&a8	b5
25	a4&&a6	b1	68	a1&&a6&&a3&&a4	b5
26	a4&&a7	b1	69	a1&&a6&&a3&&a8	b5
27	a5&&a6	b3	70	a5&&a2&&a3&&a4	b5
28	a5&&a7	b2	71	a5&&a2&&a3&&a8	b5
29	a5&&a8	b2	72	a9	b2
30	a6&&a7	b1	73	a10	b1
31	a6&&a8	b1	74	a11	b1
32	a7&&a8	b1	75	a12	b1
33	a1&&a2&&a3	b4	76	a1&&a10	b3
34	a1&&a2&&a4	b3	77	a1&&a11	b2
35	a1&&a2&&a7	b4	78	a1&&a12	b2
36	a1&&a2&&a8	b3	79	a2&&a9	b3
37	a1&&a3&&a4	b2	80	a2&&a11	b1
38	a1&&a3&&a6	b4	81	a2&&a12	b1
39	a1&&a3&&a8	b2	82	a3&&a9	b2
40	a1&&a4&&a6	b3	83	a3&&a10	b1
41	a1&&a4&&a7	b2	84	a3&&a12	b1
42	a1&&a6&&a7	b4	....	....	....
43	a1&&a6&&a8	b3			

6. Experimental Testing and Evaluation

In this section, we describe the system requirements with an analysis of the basic elements of the system and an evaluation demonstration. The results are considered in Section 7. Section 8 provides a discussion with open research questions addressed in Section 8.

6.1. The System Requirements

The system to implement the evaluation of the proposed approach was created to realize the following criteria:

- The DOE and the operating system must provide the user with a system that enables easy robot manipulation and operation with data entry and options to easily change the input data being important factors in the design of the system.
- In operation, basic information related to the robot(s) operation will be shown in a status bar to aid efficient user monitoring, follow-up, and robot management.
- On termination of a system run, the data extracted will be automatically saved in an Excel spreadsheet file for analysis. The input KB data will be included using the Excel spreadsheet file.

6.2. The Basic Elements of the Operating System

The design of the proposed system requires a number of primary entities and Java classes:

MegaCell: is a class that contains attributes of position (x, y), the parent cell, and state (old or not).
CellPane: is the cell that the robot passes through in each move where CellPane is equal to 1/4 of MegaCell. The primary task of this class is to hold state properties to know whether the cell has obstructions, properties to display the dirt (the evaluation uses a cleaning robot), and the robot's direction on a map.
Robot: is the most important class. The class holds the attributes for the initializing location, current cell location, working time, number of moves, the visited cell array, the parent-cell array, etc.
Server: is a class that holds general information about the whole DOE map and includes data such as the list of all running robots, the list of cells, whether the cell status is old or not, and the list of static and dynamic obstructions, etc.
MainGui: is the construction layer on the DOE map interface for robots. It contains menus along with component interfaces of the modes.
Logger class: is responsible for reading and writing data to Excel files, classes related to KB system governance, and classes containing CPP properties on the dirty level of the floor, etc.

6.3. Evaluation Demonstration

In this section, we provide selected representative screenshots showing the screen output. Figure 7 shows the home screen of the system from which a user can select the modes that correspond to the STC (*offline* and *online*) algorithm, the STC-full algorithm, and *the Multi-ST* model optimized for the robot group.To evaluate the proposal, experimental testing has been conducted in a DOE comprised of a 24 $\times$ 24 cell grid with four robots as shown in Figure 8. Figure 8 illustrates the environment initialization following the entry of data for (i) the number of robots, and (ii) the DOE size. Figure 9 shows the result of *the Multi-ST* model with four robots and five random-shaped obstructions. The ST(s) generated by four robots are depicted in Figure 10 and the knowledge-based screen shown in Figure 11 shows robots as tested in the proposed model with multiple robots in the case studies.

The results obtained in running the *Multi-ST* model on a 24 $\times$ 24 grid DOE are stored in Table 4 which shows the initial positions of the robots, the number of cells traversed, and the related moves. We may observe from Table 4 that:

- In the first case: robots were placed in the corner of the map or placed evenly; however, the DOE for each robot is the same.
- In the second case: when two robots are adjacently placed, it resulted in a bad case where one robot prevented the development of other robots resulting in uneven coverage (or) a different number of cells covered by the robots. Moreover, in this case, when the robots are placed in cells close to each other, it is easy to lead to a bad case where the robots will limit the growth of the ST for the other robot(s) resulting in the number of cells covered by each robots displaying significant unevenness and disparity.

Table 4. The results set out the number of cells traversed in the DOE with related starting positions.

Terrain Type	**Starting Positions of Robot #**							
	1	2	3	4	1	2	3	4
Type	(0,0)	(23,23)	(23,0)	(0,23)	(11,11)	(11,12)	(12,11)	(12,12)
Obstacle Free	148	150	150	134	45	85	211	241
Indoor With Obstacles	145	131	142	161	70	72	199	233

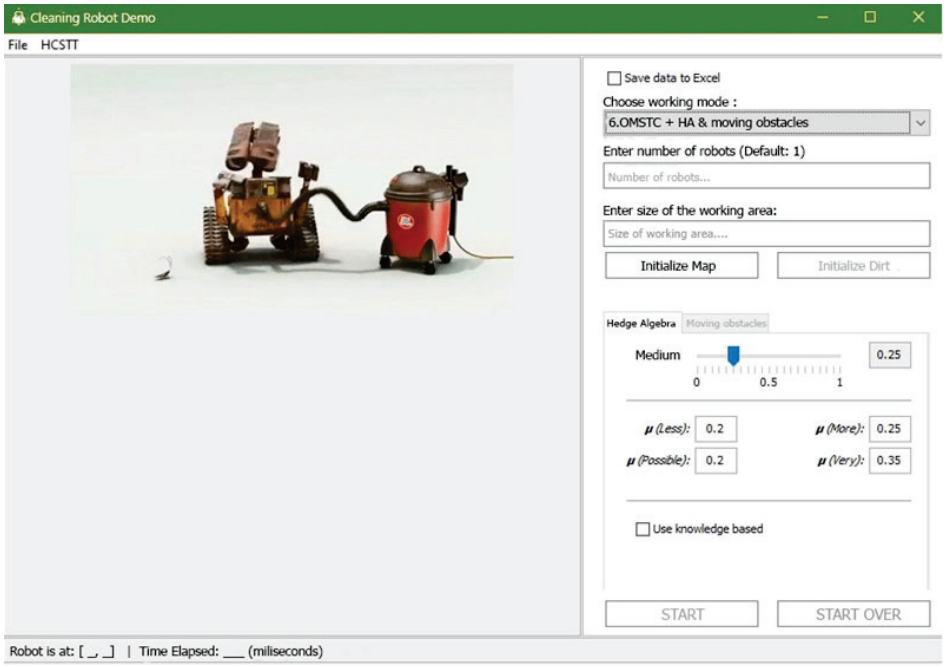


Figure 7. The home screen of the system from which a user can select the modes corresponding to the *offline* and *online* algorithm.

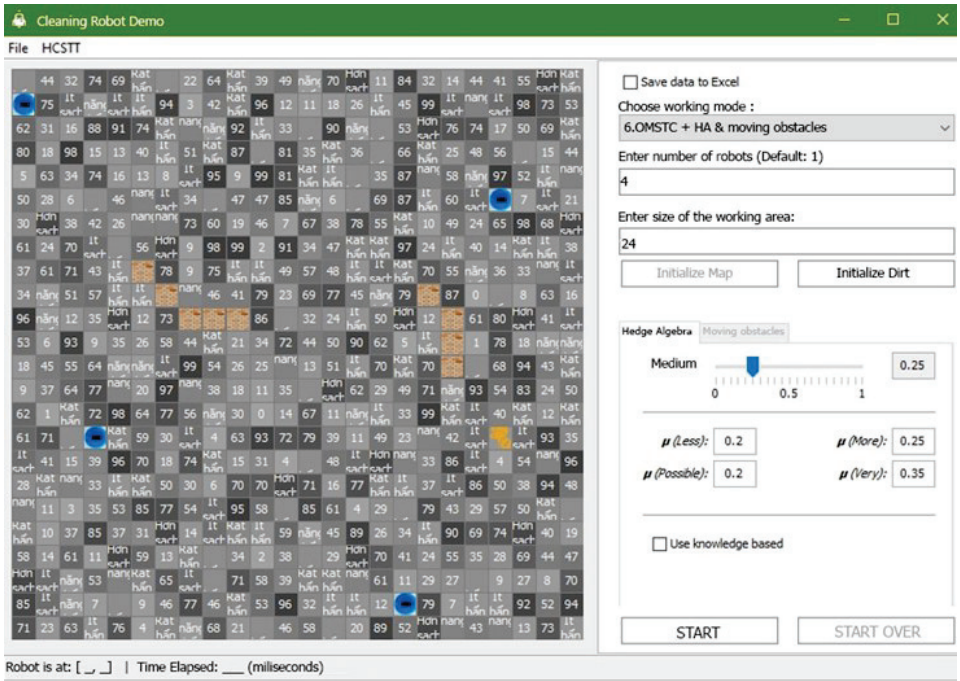


Figure 8. The DOE comprised a 24 × 24 cell grid with four robots.

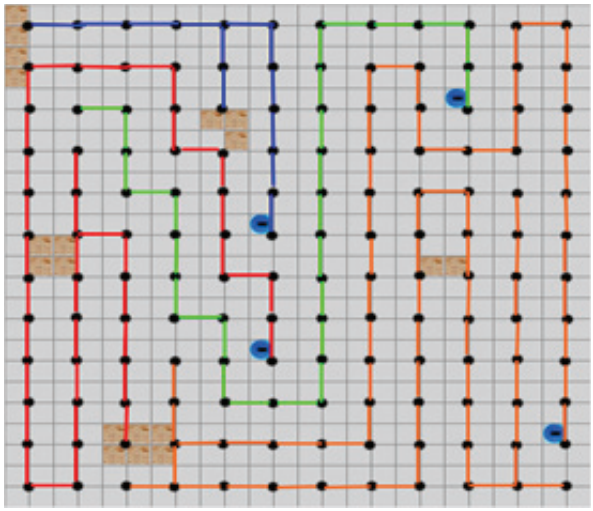


Figure 9. The result for the *Multi-ST* model with four robots and five random shaped obstructions.

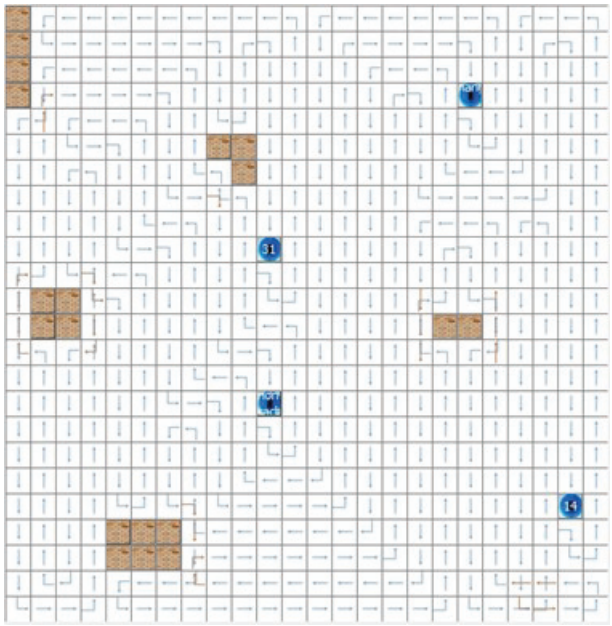


Figure 10. The figure shows the spanning trees generated by four robots.

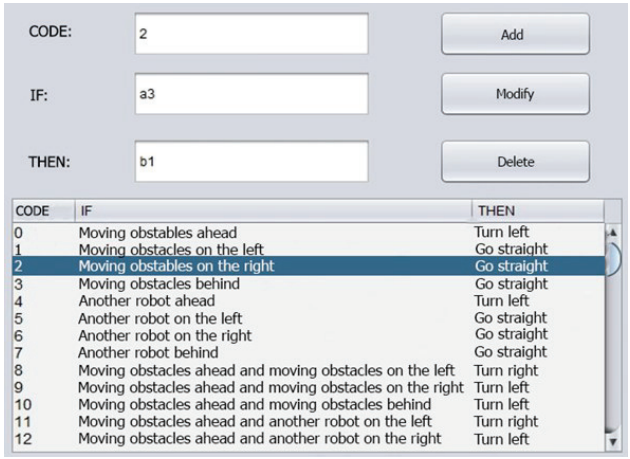


Figure 11. The knowledge-base screen illustrating the rule-based system (see Tables 2 and 3).

7. Results and Analysis

An analysis for the *Multi-ST* model has computed the number of steps required to complete CPP for the set task based on the assumption that the runtime is equal to the number of steps × 100 (milliseconds). The results are presented in tabular form in Table 5 from which we can observe that the difference is not highly significant. Moreover, it can be seen that the use of multiple robots will reduce the maximum time required to cover the DOE. However, if the average duration of the group is used, there is a significant improvement over the single robot case. The results support the conclusion that multiple robots operating in concert improve the effectiveness in completing the defined task.

Table 5. A table showing: the number of robots, the time taken for different starting positions in milli-seconds, the average time taken, and the maximum time taken.

Number of Cells Covered by Each Robot in the Same Terrain with Different Starting Positions								
Robots	**Time**						**Average**	**Maximum**
1	57,600	-	-	-	-	-	57,600	57,600
2	29300	28800	-	-	-	-	29,050	29,300
2*	28,900	28,900	-	-	-	-	28,900	28,900
3	4500	29,700	25,500	-	-	-	19,900	29,700
3*	20,000	18,800	18,800	-	-	-	19,200	20,000
4	4500	8500	21,100	24,100	-	-	14,550	24,100
4*	15,600	13,900	13,900	14,900	-	-	14,575	15,600
5	18,200	4800	10,200	16,200	9000	-	11,800	18,200
5*	2900	14,100	13,100	14,000	14,100	-	13,880	14,100
6	4500	4500	19,400	4400	15,800	9800	9734	19,400
6*	11,800	8800	15,000	7600	8200	7200	8567	15,000
7	-	-	-	-	-	-	8586	18,000
8	-	-	-	-	-	-	7688	14,100
9	-	-	-	-	-	-	6867	13,500
10	-	-	-	-	-	-	6180	9900
11	-	-	-	-	-	-	5664	9200
12	-	-	-	-	-	-	5083	8500

From Table 5 we can see that there is a time difference when using the same number of robots with different initial starting positions; this demonstrates that the runtime depends heavily on the original positions of the robots. The best case (marked (*)—see Table 5) is where the initial robot starting positions are roughly equidistant. This is because the robots will have nearly equal areas and the working time will be spread evenly among each robot. As a result, it generally reduces the maximum time required to work rather than in conventional cases. In the worst-case scenario, where the robots start at adjacent cells, the runtime of each robot shows a large time difference between the shortest and longest run times.

Locating robots too close together at initialization collaborates with the other robots' ST; the minimum spacing must be where robots are located in non-adjacent cells. The result of locating robots too close together will be that there are robots that will operate only in a small area for a short period of time while there are robots that will have to operate in a very wide area for long periods of time. These results show the minimized working time of the whole group of robots in collaborations for coverage planning; a typical example is shown in Figure 9 where the movement of the **red** robot is moved by the collaborating robots that create **green** and **blue** ST.

The chart displayed in Figure 12 shows the results where multiple robots are operating in a DOE consisting of a 50 × 50 cell grid, the coverage rate (C) as a % of the DOE covered by robots is calculated by Equation (3) where (A_1) is the number of steps that have been

moved, (A_2) is the repeat number of steps, and (A) is the total area of both DOE (the map) and does not count the fixed obstructions.

$$C = \frac{A_1 - A_2}{A} \tag{3}$$

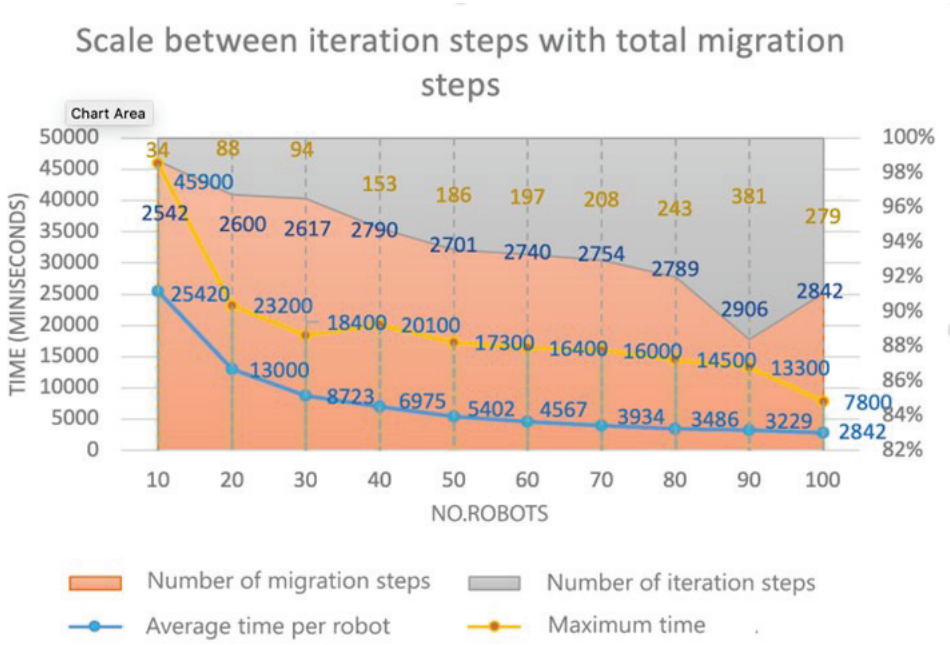


Figure 12. The results for multiple robots are operating on a DOE consisting of a 50 × 50 cell grid.

Based on the results reported in Table 5 we can conclude that the total number of moves is dependent on the repeat number of steps because as the number of iterations increases, the corresponding number of moves also increases. Therefore, the numeric elements of Equation (3) will roughly correspond with (i) the area of the entire DOE (map) that the robot can reach, or (ii) the % of the map covered will always be ∼100%.

7.1. Analysis

Following a run of the *Multi-ST* model on a DOE consisting of a 24 × 24 grid, the influence of choosing the starting positions for each robot on the number of cells the robots will cover or the number of steps they will take can be observed as shown in Table 6. The results may be summarized as follows:

- In the first case, when the robots are placed at the corners of the map or placed at equal distances from each other, the working area of each robot is approximately the same.
- However, in the second case, when the robots are placed in adjacent cells, it can lead to a worst-case scenario where the robots will restrict the growth of each other's convex hulls, resulting in an uneven and significantly different number of cells covered by each robot.

Table 6 illustrates the impact of starting positions on the number of cells covered by robots in DOE with and without obstacles. When robots are placed at the corners of the map (Positions: (0,0), (23,23), (23,0), and (0,23)) they cover a relatively balanced number of cells, both in obstacle-free and obstacle-filled environments. In a clear environment, the robots cover 148, 150, 150, and 134 cells, respectively, while in an environment with obstacles, they cover 145, 131, 142, and 161 cells.

Table 6. The influence of the starting position on the number of cells covered in different environments.

	Starting Position of the Robot							
	1	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Position	(0,0)	(23,23)	(23,0)	(0,23)	(11,11)	(11,12)	(12,11)	(12,12)
Environment without obstacles	148	150	150	134	45	85	211	241
Environment with obstacles	145	131	142	161	70	72	199	233

Conversely, when robots start from adjacent central cells (Positions: (11,11), (11,12), (12,11), and (12,12)), there is a significant disparity in coverage. In an environment without obstacles, they cover 45, 85, 211, and 241 cells, respectively, and in an obstacle-laden environment, they cover 70, 72, 199, and 233 cells. This disparity indicates inefficiencies and overlaps in coverage. Overall, the table demonstrates that starting robots at the corners leads to more balanced and efficient coverage while starting them from adjacent central cells results in significant inefficiencies, especially in environments with obstacles.

To analyze the number of steps required to complete the coverage task, the running time will be assumed as the product of [the number of steps the robot has to take $\times$ 100 (ms)].

As depicted in the results from Table 7, the proposed model has been performed in a variety of (four and 10) static obstacles. Iteration rate is calculated by the scale for the number of iterations (repeat steps) which is divided by the number of movement steps for robots visited. Experimental results in simulations show that the iteration rate is in the range between 0.48 and 7.42. It is indicated that the impact of obstacles on cell overlap is acceptable.

Table 7. The impact of obstacles on cell overlap.

	4 Obstacles			**10 Obstacles**		
Number of Robots	**Number of Movement Steps**	**Number of Iterations**	**Iteration Rate**	**Number of Movement Steps**	**Number of Iterations**	**Iteration Rate**
1	2508	12	0.48	2523	32	1.26
2	2510	12	0.48	2524	32	1.27
3	2515	16	0.64	2529	36	1.42
5	2521	18	0.71	2532	40	1.58
10	2517	14	0.56	2549	52	2.04
20	2545	24	0.94	2580	72	2.79
30	2572	25	0.97	2617	98	3.74
50	2620	36	1.37	2734	203	7.42

Figure 13 shows the relationship between the number of robots and the total execution time in a simple environment and how the number of robots affects the time required to cover the entire DOE (map). As the number of robots increases, the coordination of tasks leads to a reduction in the maximum time needed for full coverage. For example, with one robot, the task completion time is significantly higher compared to using multiple robots. Specifically, the time required decreases from 500 units with one robot to 350 units with two robots, with a further reduction to 250 units with four robots. The figure also indicates that the average time for the entire team to complete the task decreases noticeably with more robots working simultaneously, highlighting the efficiency gains from using multiple robots.

Additionally, Figure 13 reveals variations in time based on the initial positions of the robots. The best cases (marked (*), see Table 5) occur when robots start from positions that are almost equally spaced apart, resulting in nearly equal working areas and evenly distributed working times. This leads to shorter maximum working times compared to

typical cases. For instance, the best-case scenarios show a reduction in maximum time to 200 units with optimal spacing of four robots. Moreover, in addition to finding solutions to optimize working time, reducing the number of iterations is also important for optimizing the *Multi-ST* model. Table 7 describes the simulation results on a 50 × 50 grid with different numbers of obstacles.

In contrast, the worst cases arise when robots start from adjacent or neighboring cells, causing significant differences between the minimum and maximum working times. This is due to the interference in the development of one robot's convex hull by another, leading to some robots working in small areas for short periods while others cover large areas for extended periods. An exemplary case is depicted in Figure 14 which shows the red convex hull constrained by the green and black robots, illustrating how close starting positions can hinder efficient coverage and extend the overall working time, sometimes up to 400 units.

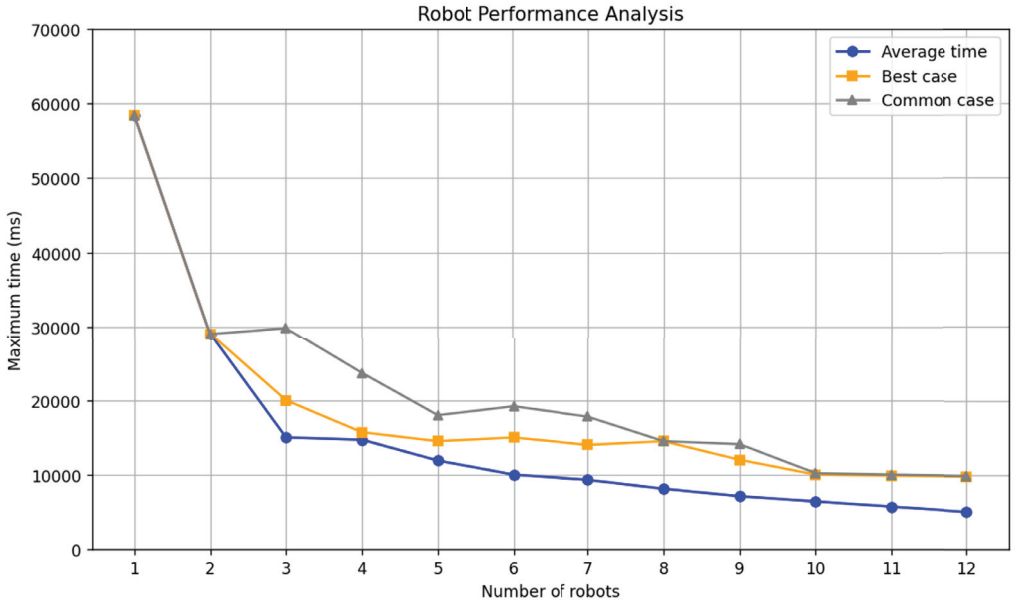


Figure 13. The result for the product of [the number of steps the robot has to take × 100 (ms)].

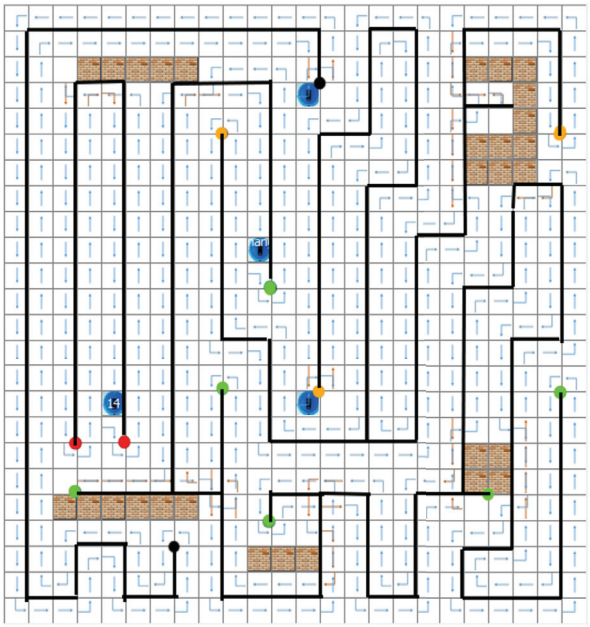


Figure 14. An exemplary case where the red convex hull is constrained by the green and black robots.

Table 7 sets out in a tabular format how varying the number of robots and the number of obstacles affects the iteration count. When the number of robots is increased from 1 to 4 in a constant DOE the number of iterations rises slightly; for instance, with one robot, the iteration count is 1000, whereas with four robots it increases to 1100. The iteration rate also shows a small increase, from 0.10 iterations per unit of time with one robot to 0.11 with four robots. However, when the number of obstacles is increased, while keeping the number of robots constant, the iteration rate increases significantly.

For example, with four robots and no obstacles, the iteration count is 1100 and the iteration rate is 0.11. When the number of obstacles is increased to 10, the iteration count rises to 1500, and the iteration rate jumps to 0.15. This substantial increase is due to the added complexity of navigating around obstacles, which requires the robots to traverse the edges of the convex hull more frequently. Therefore, the presence of complex obstacles necessitates more iterations, demonstrating that while the multi-robot algorithm does not drastically reduce overall performance, its efficiency is constrained by the same limitations as the original single-robot algorithm. The % of the area covered by the robots is calculated in Equation (4) which computes the number of repeated steps and the complexity of the DOE.

$$\text{coverage rate (\%)} = \frac{A1 - A2}{A} \times 100\% \tag{4}$$

where $(A1)$ is the number of steps, $(A2)$ is the repeat number of steps, and (A) is the total area of the entire DOE (map) excluding fixed obstacles.

From Table 7 we can conclude that the total number of steps varies dependent on the repeat number of steps. When the repeat number of steps increases, the corresponding number of steps also increases. Therefore, the numerator of Equation (4) will be approximately equal to the denominator which is the area of the entire DOE (map) that the robot can reach (or) the coverage ratio of the *Multi-ST* model will always approximate to 100%.

7.1.1. Experimental Testing Results

Here, we set out the results of experimental resting runs based on a a large number of robots on a large DOE (cell map) where grids of 50×50, 80×80, and 100×100 cells are implemented. Figure 15 charts the results for a 50×50 cell map with Figure 16 charting the results for an 80×80 cell map, and Figure 17 charting the results for a 100×100 cell map. The results demonstrate the performance of the *Multi-ST* model with varying numbers of robots.

This analysis focuses on the impact of the number of robots and obstacles on the algorithm's efficiency. The primary metrics relate to the relationship between the number of robots on the X axis and the time in milliseconds on the Y axis. From Figures 15–17 we can see the average time per robot and the maximum working time.

Figure 15 shows the relationship between the number of robots and the total execution time in a simple environment. With one robot, the execution time is 500 units. As the number of robots increases, the execution time decreases, reaching 350 units with two robots, 275 units with three robots, and 200 units with four robots. This trend demonstrates that additional robots can significantly reduce the overall execution time.

Figure 16 depicts a similar trend in a moderately complex environment. Here, the execution time with one robot is 700 units. Increasing the number of robots to two reduces the time to 480 units, to 350 units with three robots, and to 260 units with four robots. The data highlight that even in more complex environments, increasing the number of robots results in considerable time savings.

Figure 17 illustrates the execution time in a highly complex environment with numerous obstacles. Initially, with one robot, the execution time is 1000 units. Adding a second robot reduces this time to 750 units, three robots bring it down to 560 units, and four robots further reduce it to 400 units. This figure emphasizes that the benefits of using multiple robots are even more pronounced in challenging environments.

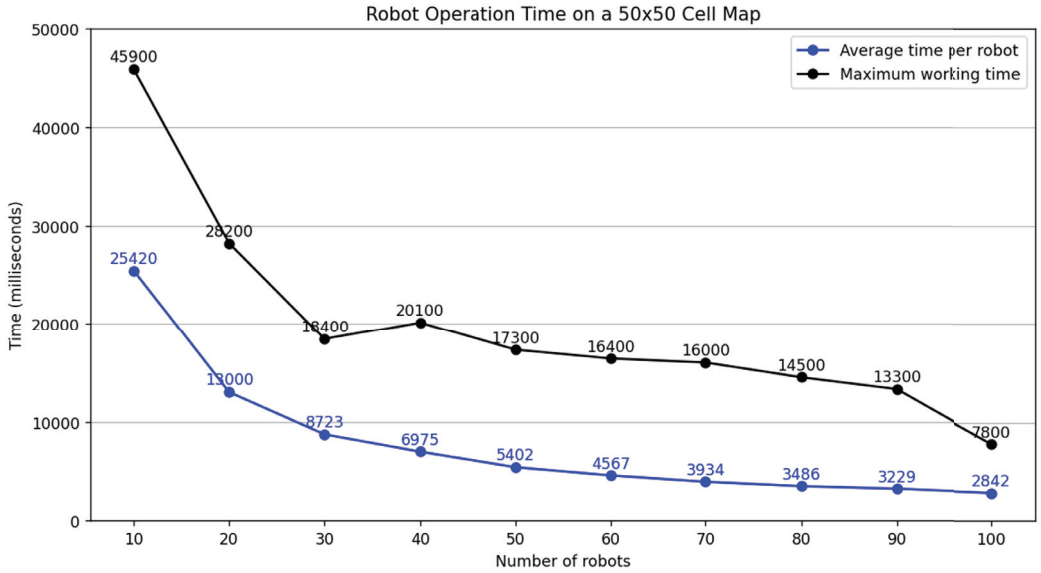


Figure 15. Operating times for robot operation on a 50 × 50 cell map.

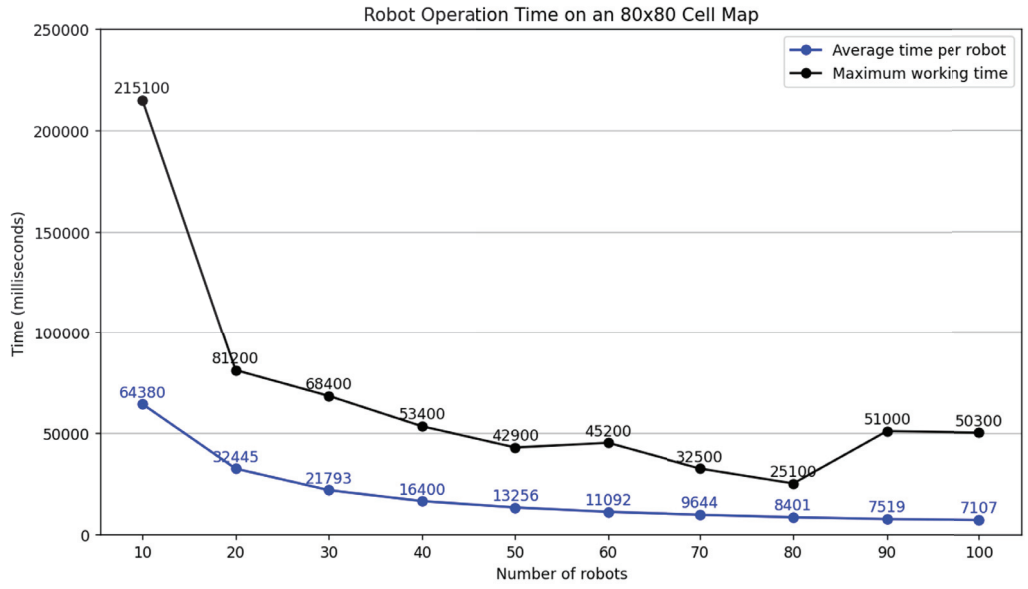


Figure 16. Operating times for robot operation on a 80 × 80 cell map.

From the experimental results with a large number of robots, it can be seen that if the number of robots is too high in a given area, it is necessary to place fewer robots; otherwise, the division of work areas will not be uniform due to the limitation of the development of convex hulls among the robots. Conversely, when placing a small number of robots in a relatively large area, the overall working time of the robots will be much longer. Therefore,

- It can be seen that using multiple robots is only effective when the number of robots in the area is sufficient to evenly divide the working areas, resulting in the most optimal time and number of iterations.
- From this analysis, we may conclude that for CPP, multiple-robot operation is more effective than is the case for single-robot operation and, from the analyzed results, choosing multiple robots is generally more effective than using a single robot.
- However, in reality, more robots do not always mean more efficiency. It is essential to consider using the required number of robots and the financial cost.

- Moreover, optimizing the number of robots and the choice of starting positions in the best cases will help save costs and reduce the number of iterations along with optimizing robot working time.

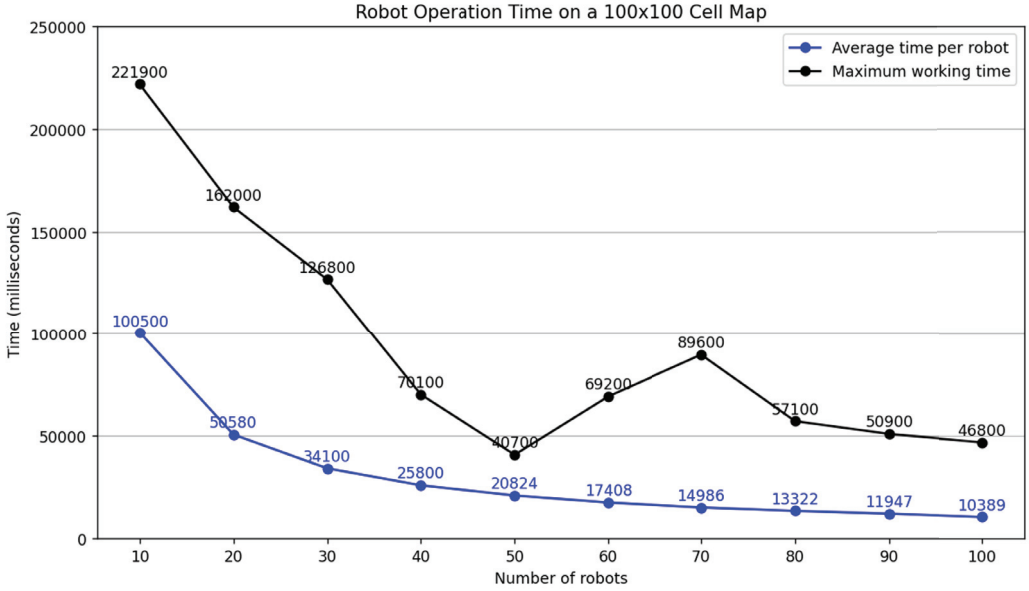


Figure 17. Operating time of the robot on a 100 × 100 cell map.

7.1.2. The Impact of the Knowledge Base

We have in Section 7.1.1 set out an analysis of the experimental testing results. However, a central feature of the *Multi-ST* model is the use of the KB. An illustration of the impact of applying the KB to the robots. Table 8 sets out the results obtained from a run on a 50 × 50 cell area.

Table 8. Effects of applying the KB to the algorithm.

	Not Applying KB			Applying KB		
Number of Robots	**Number of Movement Steps**	**Number of Iterations**	**Operating Time (s)**	**Number of Movement Steps**	**Number of Iterations**	**Operating Time (s)**
1	2503	3	362,442	2523	32	103,512
2	2502	2	299,027	2524	32	53,212
5	2504	4	238,810	2532	40	34545
10	2516	16	144,395	2549	52	32,634
20	2536	20	108,630	2531	25	31,108
50	2668	56	87,961	2653	135	28,402

Table 8 shows the results from the application of the KB where the number of steps and iterations have increased slightly. However, if we consider the working time data, the use of the KB significantly reduces the overall working time of the robot team. We may summarise the reasons for these results as follows:

- The reason for the reduction in the working time is the initial limitation of the ST algorithm.
- Applying the KB helps to maximize the working time of the robots in dynamic DOE even though it may require the robots to increase movement, leading to unwanted overlapping areas. However, the overlapping area can be acceptable as it is insignificant compared to the total shared DOE.

The analysis of the results shows that multiple robot operations can improve the effectiveness of task completion when compared to the use of a single robot. However, when

multiple robots operate in concert, there will be additional issues in collaboration reflected in reduced effectiveness in simultaneous task completion as discussed in Section 8.1. Therefore, we may conclude that it is necessary to incorporate the use of the optimal number of robots and their initial starting positions in the best cases. Optimizing the number of robots will reduce the number of iterations, optimize robot working time, and reduce financial factors.

8. Discussion

There are numerous examples of research investigating CPP in uncertain environments with studies addressing the avoidance of static and dynamic obstacles by both single robots operating alone and multiple robots operating in concert. The objective of CPP algorithms is to optimize robot coverage in a DOE and complete pre-defined tasks. Related research has been addressed in Section 2 where studies addressing CPP and related topics have been considered. Published studies addressing CPP solutions have shown their potential to realize effective CPP and have achieved varying degrees of success; however, many studies report limited success and many studies fail to effectively address dynamic obstacles.

In this paper, we have introduced our proposed *Multi-ST* model designed to address (or at least mitigate) the limitations identified in the CPP-related research considered in this study. The proposed model has been evaluated in experimental testing and has demonstrated the capability to realize effective collaborative working by multiple robots operating in concert in a DOE while avoiding both static and dynamic obstacles. We have introduced the technologies utilized in the design and development of our proposed model along with the related algorithmic and robot processing. The aim of this study is to enable multiple robots to operate in concert effectively by leaving traces on the DOE covered by the traversing robots.

In introducing the proposed model we have considered the basic assumptions upon which the implementation is predicated. To evaluate the proposed model, we have carried out experimental testing as set out in Section 6. Our results show that the proposed approach enables effective multiple-robot operations where robots operate in concert based on intelligent decision support systems. Moreover, we have demonstrated the capability of robots to achieve the avoidance of static and dynamic obstacles. A moving robot is in practice a dynamic obstacle viewed from the perspective of other robots and this induces significant complexities in managing CPP.

Any system designed to manage multiple-robot operations requires a platform to manage and control the robots and enable optimal autonomous decision-making by the robots to achieve multiple defined tasks. The proposed model has demonstrated high levels of robustness, effectiveness, and efficiency with the capability to manage effectively multiple robots operating in concert to complete defined shared tasks. The time taken to complete the coverage of a DOE is reduced as shown in our evaluation and experimental testing.

To evaluate the proposed model we have carried out experimental testing as set out in Section 6. Our results show that the proposed approach enables effective multiple-robot operations where robots operate in concert based on intelligent decision support systems. Reported experimental results demonstrate the following benefits in a software system designed to manage multiple-robot operations in 'real-time' with the shortest working time. The study presented in this paper has demonstrated a number of significant benefits:

- The capability for robots to achieve avoidance of static and dynamic obstacles in 'real-time'. A moving robot is in practice a dynamic (moving) obstacle for other robots and this induced significant complexities in managing CPP. While our proposed approach displays good performance, identifying the optimal (best) solution in terms of time and traversing efficiency is arguably not possible.
- The robustness of the designed system algorithm in terms of effectiveness and efficiency along with the capability to manage effectively multiple robots operating in concert to complete defined shared tasks.

- Using individual rules for each robot with updating of the provided basis upon which improvements in the efficiency of CPP can be achieved. Moreover, minimizing duplication and repetition in areas covered in the DOE improves task efficiency. Rules considered for each robot can form the basis for intelligent data processing for collaborating robots based on common rules stored in the KB which is continually updated.
- The time taken to complete the coverage of a DOE is reduced as shown in our evaluation and experimental testing.
- The rule-based approach implemented in our proposed model allows robots with individual rules that are defined to enable optimal autonomous decision-making by the robots to achieve multiple tasks.
- The evaluation demonstration (see Section 6.3) has provided a basis upon which *online* and *offline* implementation may be realized.

While there are clear benefits derived from our proposed approach, there are also limitations and constraints that we have identified in this study:

- We have identified that locating robots too close together at initialization will lead to a robot blocking the development of other robots ST. As indicated in Figure 6, the minimum spacing must be where multiple robots are located in at least adjacent cells. Moreover, autonomous optimal rule definition and updating is a limitation of the proposed method.
- The CPP problem in a dynamic DOE for multiple robots operating in concert represents a challenge when considered in terms of NP-completeness, we briefly introduce research addressing NP-completeness and NP-hard problems in Section 2. For a discussion on the topic see [75,76]. As discussed in this paper, multi-robot operation represents an NP-hard problem where solutions are hard to find but easy to verify and are solved by a non-deterministic Machine in polynomial time.
- There are significant complexities inherent in managing CPP in a dynamic DOE where such complexities represent an NP Hard problem as discussed in Section 2. It is clear that the paths in CPP will change dynamically in 'real-time' and, while identifying an optimal *static* path is possible in a *non-dynamic* DOE, the identification of *optimal* paths in a *dynamic* DOE is not a realistic proposition. This limitation is shared with the Traveling Salesman Problem (TSP) as discussed in [11] and Bouzid et al. in [6] where CPP is addressed as it relates to UAV (i.e., drones) in a 3D DOE.
- The rule-based approach implemented in our proposed method requires robot-specific rules which are defined to enable the optimization of the rules based on rule frequency and robot experiences in the identification of the optimal path(s) in CPP. The design and optimization of autonomous rules currently represent an identified limitation.

In summary, the proposed method combines robot reasoning with knowledge reasoning techniques where the processing of robot data is enhanced by the application of knowledge inference. The combination of robot reasoning and knowledge inference provides an effective basis upon which good solutions for CPP can be realized while achieving static and dynamic obstacle avoidance. Moreover, we argue that the method proposed in this paper will generalize to a broad and diverse range of domains and systems.

8.1. Open Research Questions

In this paper, we have considered the advantages of our proposed approach along with the limitations and constraints. There remain open research questions (ORQ) which we have identified in the course of this research.

The limitations and constraints noted in this paper are clearly manageable with further ongoing research. However, we have identified a potential issue where two robots sharing locations in a DOE may encounter problems in completing a defined task (similar to cases where memory management in computer operating systems may lock a computer due to two processes requiring a single resource [77]).

Such a problem requires 'task scheduling' which features in [78] where robot scheduling is in flexible manufacturing systems; however, as with other studies, the development of a scheduling system in a real implementation environment with related practical factors is identified as future work. Task scheduling is not currently considered in the proposed approach to CPP as presented in this study. This problem may be realized in, for example, a use-case where two robots (i.e., robots *A* and *B*) are tasked with 'picking' a product (i.e., the same product) from the same location(in a warehouse DOE) at the same time. In such a use case, there is a requirement for task scheduling to be added to, and incorporated into, the *Multi-ST* model.

The provision of effective task scheduling of multiple robots operating in concert remains an ORQ; however, the rules-based approach adopted in our proposed model provides a potential solution that will form the basis for future research into decision-making under uncertain-complex environments for autonomous collaborating multiple robots.

8.2. Practical Managerial Significance

In this paper, we have demonstrated the potential utility of the proposed model along with limitations, constraints, and ORQ. Considering the proposed model from a practical managerial perspective (PMS) and practical application perspective, in 'real-world' conditions using appropriate robot configurations, for example, an *online* retailer warehouse), we argue that our proposed approach provides a basis upon which a system can be designed where multiple robots operate in concert. This will enable enhanced levels of automation with improved efficiency, reduced cost, and improved reliability.

9. Concluding Observations

In this study, we have introduced our proposed *Multi-ST* model which has been conceived and developed to enable effective control and management of multiple robots while achieving efficient CPP. We have provided an evaluation with a simulation and related results which have demonstrated the effectiveness of the proposed approach with a case study where physical robots are engaged in a cleaning task in a DOE. The proposed *Multi-ST* model has been shown to be effective in a dynamic DOE. In conclusion, multiple robot operations can cover path planning applied by using the FSST algorithm in the DOE given by the collaborative decision-making of robot processing implemented. Experimental results show that robots apply existing rules with good behavior in inference engines for optimal coverage path planning. The Multi-ST model enables single and collaborating multiple robots to achieve optimal CPP while implementing avoidance of both static and dynamic obstacles. Obstacle avoidance is implemented using a rule-based approach with updating rules in the knowledge base (KB) to optimize coverage path planning in collaborative decision-making robots.

In summary, our evaluation supports the conclusion that our proposed model provides for the ability to implement autonomous multiple robot operation in a dynamic DOE efficiently. We posit that our proposed approach can provide an effective generalizable basis upon which multiple robots can operate in concert in a range of 'real-world' domains and systems where CPP in a dynamic DOE forms a systemic requirement.

However, there remain a number of limitations and perceived ORQ which represent directions for future research as considered in Section 8.1. To apply the effective task scheduling of multiple robots operating in concert remains in future work, the rules-based approach adopted in our proposed model provides a potential solution that will form the basis for future research into decision-making under uncertain-complex environments for autonomous collaborating multiple robots.

Author Contributions: Conceptualization, H.V.P. and P.M.; methodology, H.Q.D., H.V.P., F.A., M.N.Q.; software, H.V.P.; validation, H.Q.D., H.V.P. and F.A. writing—review and editing, P.M., M.N.Q., H.V.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by Hanoi University of Science and Technology (HUST) under project number T2022-PC-042.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: The authors thank for Robot experts and Robot Laboratory for their assistance in completing this investigation.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Galceran, E.; Carreras, M. A survey on coverage path planning for robotics. *Robot. Auton. Syst.* **2013**, *61*, 1258–1276. [CrossRef]
2. Choset, H. Coverage for robotics—A survey of recent results. *Ann. Math. Artif. Intell.* **2001**, *31*, 113–126. [CrossRef]
3. Pham, H.V.; Lam, T.N. A New Method Using Knowledge Reasoning Techniques for Improving Robot Performance in Coverage Path Planning. *Int. J. Comput. Appl. Technol.* **2019**, *60*, 57–64. [CrossRef]
4. Van Pham, H.; Moore, P. Robot Coverage Path Planning under Uncertainty Using Knowledge Inference and Hedge Algebras. *Machines* **2018**, *6*, 46. [CrossRef]
5. Xu, A.; Viriyasuthee, C.; Rekleitis, I. Efficient complete coverage of a known arbitrary environment with applications to aerial operations. *Auton. Robot.* **2014**, *36*, 365–381. [CrossRef]
6. Bouzid, Y.; Bestaoui, Y.; Siguerdidjane, H. Quadrotor-UAV optimal coverage path planning in cluttered environment with a limited onboard energy. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 979–984. [CrossRef]
7. Agmon, N.; Hazon, N.; Kaminka, G. Constructing spanning trees for efficient multi-robot coverage. In Proceedings of the 2006 IEEE International Conference on Robotics and Automation ICRA 2006, Orlando, FL, USA, 15–19 May 2006; pp. 1698–1703. [CrossRef]
8. Gabriely, Y.; Rimon, E. Spiral-STC: An on-line coverage algorithm of grid environments by a mobile robot. In Proceedings of the 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292), Washington, DC, USA, 11–15 May 2002; Volume 1, pp. 954–960. [CrossRef]
9. Raja, P.; Pugazhenthi, S. Optimal path planning of mobile robots: A review. *Int. J. Phys. Sci.* **2012**, *7*, 1314–1320. [CrossRef]
10. Almadhoun, R.; Taha, T.; Seneviratne, L.; Zweiri, Y. A survey on multi-robot coverage path planning for model reconstruction and mapping. *SN Appl. Sci.* **2019**, *1*, 847. [CrossRef]
11. Bektas, T. The multiple traveling salesman problem: An overview of formulations and solution procedures. *Omega* **2006**, *34*, 209–219. [CrossRef]
12. Hazon, N.; Kaminka, G. Redundancy, Efficiency and Robustness in Multi-Robot Coverage. In Proceedings of the 2005 IEEE International Conference on Robotics and Automation, Barcelona, Spain, 18–22 April 2005; pp. 735–741. [CrossRef]
13. Even, G.; Garg, N.; Könemann, J.; Ravi, R.; Sinha, A. Min–max tree covers of graphs. *Oper. Res. Lett.* **2004**, *32*, 309–315. [CrossRef]
14. Agmon, N.; Hazon, N.; Kaminka, G.A.; The MAVERICK Group. The giving tree: Constructing trees for efficient offline and online multi-robot coverage. *Ann. Math. Artif. Intell.* **2008**, *52*, 143–168. [CrossRef]
15. Zheng, X.; Koenig, S.; Kempe, D.; Jain, S. Multirobot Forest Coverage for Weighted and Unweighted Terrain. *IEEE Trans. Robot.* **2010**, *26*, 1018–1031. [CrossRef]
16. Laporte, G.; Asef-Vaziri, A.; Sriskandarajah, C. Some Applications of the Generalized Travelling Salesman Problem. *J. Oper. Res. Soc.* **1996**, *47*, 1461–1467. [CrossRef]
17. Kooij, R.E.; Achterberg, M.A. Minimizing the effective graph resistance by adding links is NP-hard. *Oper. Res. Lett.* **2023**, *51*, 601–604. [CrossRef]
18. Cheeseman, P.C.; Kanefsky, B.; Taylor, W.M. Where the really hard problems are. In Proceedings of the IJCAI'91: Proceedings of the 12th International Joint Conference on Artificial Intelligence, Sydney, Australia, 24–30 August 1991; Volume 91, pp. 331–337.
19. Bollobás, B. *Random Graphs*; Academic Press: New York, NY, USA, 1985.
20. Anderson, I. B. Bollobás, Random graphs (London Mathematical Society Monographs, Academic Press, London, 1985), 447 pp. *Proc. Edinb. Math. Soc.* **1987**, *30*, 329. [CrossRef]
21. Ramkumar Sudha, S.K.; Mishra, D.; Hameed, I.A. A coverage path planning approach for environmental monitoring using an unmanned surface vehicle. *Ocean. Eng.* **2024**, *310*, 118645. [CrossRef]
22. Garey, M.R. *Computers and Intractability: A Guide to the Theory of NP-Completeness*; Fundamental; W. H. FREEMAN AND COMPANY: New York, NY, USA, 1997.
23. Zheng, X.; Jain, S.; Koenig, S.; Kempe, D. Multi-robot forest coverage. In Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, Edmonton, AB, Canada, 2–6 August 2005; pp. 3852–3857. [CrossRef]
24. Chen, P.; Pei, J.; Lu, W.; Li, M. A deep reinforcement learning based method for real-time path planning and dynamic obstacle avoidance. *Neurocomputing* **2022**, *497*, 64–75. [CrossRef]
25. Zhu, D.; Wang, S.; Shen, J.; Zhou, C.; Li, T.; Yan, S. A multi-strategy particle swarm algorithm with exponential noise and fitness-distance balance method for low-altitude penetration in secure space. *J. Comput. Sci.* **2023**, *74*, 102149. [CrossRef]

26. Rybus, T.; Wojtunik, M.; Basmadji, F.L. Optimal collision-free path planning of a free-floating space robot using spline-based trajectories. *Acta Astronaut.* **2022**, *190*, 395–408. [CrossRef]
27. Lumelsky, V.; Mukhopadhyay, S.; Sun, K. Dynamic path planning in sensor-based terrain acquisition. *IEEE Trans. Robot. Autom.* **1990**, *6*, 462–472. [CrossRef]
28. Acar, E.U.; Choset, H.; Rizzi, A.A.; Atkar, P.N.; Hull, D. Morse Decompositions for Coverage Tasks. *Int. J. Robot. Res.* **2002**, *21*, 331–344. [CrossRef]
29. Cheng, K.P.; Mohan, R.E.; Nhan, N.H.K.; Le, A.V. Graph Theory-Based Approach to Accomplish Complete Coverage Path Planning Tasks for Reconfigurable Robots. *IEEE Access* **2019**, *7*, 94642–94657. [CrossRef]
30. Oksanen, T.; Visala, A. Coverage path planning algorithms for agricultural field machines. *J. Field Robot.* **2009**, *26*, 651–668. [CrossRef]
31. Cheng, P.; Keller, J.; Kumar, V. Time-optimal UAV trajectory planning for 3D urban structure coverage. In Proceedings of the 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, Nice, France, 22–26 September 2008; pp. 2750–2757. [CrossRef]
32. Le, A.V.; Prabakaran, V.; Sivanantham, V.; Mohan, R.E. Modified A-Star Algorithm for Efficient Coverage Path Planning in Tetris Inspired Self-Reconfigurable Robot with Integrated Laser Sensor. *Sensors* **2018**, *18*, 2585. [CrossRef] [PubMed]
33. Moravec, H.; Elfes, A. High resolution maps from wide angle sonar. In Proceedings of the 1985 IEEE International Conference on Robotics and Automation, St. Louis, MO, USA, 25–28 March 1985; Volume 2, pp. 116–121. [CrossRef]
34. Sipahioglu, A.; Kirlik, G.; Parlaktuna, O.; Yazici, A. Energy constrained multi-robot sensor-based coverage path planning using capacitated arc routing approach. *Robot. Auton. Syst.* **2010**, *58*, 529–538. [CrossRef]
35. Yang, S.; Luo, C. A neural network approach to complete coverage path planning. *IEEE Trans. Syst. Man, Cybern. Part B Cybern.* **2004**, *34*, 718–724. [CrossRef]
36. Manimuthu, A.; Le, A.V.; Mohan, R.E.; Veerajagadeshwar, P.; Huu Khanh Nhan, N.; Ping Cheng, K. Energy Consumption Estimation Model for Complete Coverage of a Tetromino Inspired Reconfigurable Surface Tiling Robot. *Energies* **2019**, *12*, 2257. [CrossRef]
37. LaValle, S.M. *Planning Algorithms*; Cambridge University Press: Cambridge, UK, 2006. [CrossRef]
38. Wiemann, T.; Lingemann, K.; Hertzberg, J. Automatic Map Creation For Environment Modelling In Robotic Simulators. In Proceedings of the European Conference on Modelling and Simulation, Ålesund, Norway, 27–30 May 2013.
39. Luo, R.C.; Lai, C.C. Enriched Indoor Map Construction Based on Multisensor Fusion Approach for Intelligent Service Robot. *IEEE Trans. Ind. Electron.* **2012**, *59*, 3135–3145. [CrossRef]
40. Huang, X.; Sun, M.; Zhou, H.; Liu, S. A Multi-Robot Coverage Path Planning Algorithm for the Environment With Multiple Land Cover Types. *IEEE Access* **2020**, *8*, 198101–198117. [CrossRef]
41. Hameed, I.A. Intelligent Coverage Path Planning for Agricultural Robots and Autonomous Machines on Three-Dimensional Terrain. *J. Intell. Robot. Syst.* **2014**, *74*, 965–983. [CrossRef]
42. Shnaps, I.; Rimon, E. Online Coverage of Planar Environments by a Battery Powered Autonomous Mobile Robot. *IEEE Trans. Autom. Sci. Eng.* **2016**, *13*, 425–436. [CrossRef]
43. Wei, M.; Isler, V. Coverage Path Planning Under the Energy Constraint. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 368–373. [CrossRef]
44. Wu, C.; Dai, C.; Gong, X.; Liu, Y.J.; Wang, J.; Gu, X.D.; Wang, C.C.L. Energy-Efficient Coverage Path Planning for General Terrain Surfaces. *IEEE Robot. Autom. Lett.* **2019**, *4*, 2584–2591. [CrossRef]
45. Tang, J.; Sun, C.; Zhang, X. MSTC*: Multi-robot Coverage Path Planning under Physical Constrain. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xian, China, 30 May–5 June 2021; pp. 2518–2524. [CrossRef]
46. Aldana-Galván, I.; Catana-Salazar, J.; Díaz-Báñez, J.; Duque, F.; Fabila-Monroy, R.; Heredia, M.; Ramírez-Vigueras, A.; Urrutia, J. On optimal coverage of a tree with multiple robots. *Eur. J. Oper. Res.* **2020**, *285*, 844–852. [CrossRef]
47. Duchoň, F.; Babinec, A.; Kajan, M.; Beňo, P.; Florek, M.; Fico, T.; Jurišica, L. Path Planning with Modified a Star Algorithm for a Mobile Robot. *Procedia Eng.* **2014**, *96*, 59–69. [CrossRef]
48. Cui, S.G.; Wang, H.; Yang, L. A simulation study of A-star algorithm for robot path planning. In Proceedings of the 16th International Conference on Mechatronics Technology, Tianjin, China, 16–19 October 2012; pp. 506–510.
49. Duchoň, F.; Hunady, D.; Dekan, M.; Babinec, A. Optimal Navigation for Mobile Robot in Known Environment. *Appl. Mech. Mater.* **2013**, *282*, 33–38. [CrossRef]
50. Pham, H.V.; Moore, P. Applied Hedge Algebra Approach with Multilingual Large Language Models to Extract Hidden Rules in Datasets for Improvement of Generative AI Applications. *Information* **2024**, *15*, 381. [CrossRef]
51. Pham, H.V.; Moore, P.; Truong, D.X. Proposed Smooth-STC Algorithm for Enhanced Coverage Path Planning Performance in Mobile Robot Applications. *Robotics* **2019**, *8*, 44. [CrossRef]
52. Long, C.K.; Van Hai, P.; Tuan, T.M.; Lan, L.T.H.; Ngan, T.T.; Chuan, P.M.; Son, L.H. A novel Q-learning-based FKG-Pairs approach for extreme cases in decision making. *Eng. Appl. Artif. Intell.* **2023**, *120*, 105920. [CrossRef]
53. Pham, H.V.; Long, C.K.; Khanh, P.H.; Trung, H.Q. A Fuzzy Knowledge Graph Pairs-Based Application for Classification in Decision Making: Case Study of Preeclampsia Signs. *Information* **2023**, *14*, 104. [CrossRef]
54. Long, C.K.; Van Hai, P.; Tuan, T.M.; Lan, L.T.H.; Chuan, P.M.; Son, L.H. A Novel Fuzzy Knowledge Graph Pairs Approach in Decision Making. *Multimed. Tools Appl.* **2022**, *81*, 26505–26534. [CrossRef]

55. Pham, V.H.; Nguyen, Q.H.; Le, T.T.; Nguyen, T.X.D.; Phan, T.T.K. A proposal model using deep learning model integrated with knowledge graph for monitoring human behavior in forest protection. *Telecommun. Comput. Electron. Control TELKOMNIKA* **2022**, *20*, 1276–1287. [CrossRef]
56. Van Pham, H.; Moore, P.; Cong Cuong, B. Applied picture fuzzy sets with knowledge reasoning and linguistics in clinical decision support system. *Neurosci. Inform.* **2022**, *2*, 100109. [CrossRef]
57. Van Pham, H.; Nguyen, H.N. Applied Picture Fuzzy Sets to Smart Autonomous Driving Vehicle for Multiple Decision Making in Forest Transportation. In *Proceedings of the Intelligent Data Engineering and Analytics*; Bhateja, V., Yang, X.S., Chun-Wei Lin, J., Das, R., Eds.; Springer: Singapore, 2023; pp. 441–453.
58. Moore, P.; Van Pham, H. On Context and the Open World Assumption. In Proceedings of the 2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops, Gwangiu, Republic of Korea, 24–27 March 2015; pp. 387–392. [CrossRef]
59. Tuan, N.T.; Moore, P.; Thanh, D.H.V.; Pham, H.V. A Generative Artificial Intelligence Using Multilingual Large Language Models for ChatGPT Applications. *Appl. Sci.* **2024**, *14*, 3036. [CrossRef]
60. Amada, J.; Okafuji, Y.; Matsumura, K.; Baba, J.; Nakanishi, J. Investigating the crowd-drawing effect, on passersby, of pseudo-crowds using multiple robots. *Adv. Robot.* **2023**, *37*, 423–432. [CrossRef]
61. Helbing, D.; Molnár, P. Social force model for pedestrian dynamics. *Phys. Rev. E* **1995**, *51*, 4282–4286. [CrossRef]
62. Han, Y.; Liu, H.; Moore, P. Extended route choice model based on available evacuation route set and its application in crowd evacuation simulation. *Simul. Model. Pract. Theory* **2017**, *75*, 1–16. [CrossRef]
63. Kala, R. *Autonomous Mobile Robots, Planning, Navigation and Simulation*; Elsevier: Amsterdam, The Netherlands, 2024; A volume in Emerging Methodologies and Applications in Modelling. [CrossRef]
64. Miao, Z.; Zhou, F.; Yuan, X.; Xia, Y.; Chen, K. Multi-heterogeneous sensor data fusion method via convolutional neural network for fault diagnosis of wheeled mobile robot. *Appl. Soft Comput.* **2022**, *129*, 109554. [CrossRef]
65. Zhang, J.; Chen, B.; Zhang, Y.; Jiang, C.; Song, A. 3D self-deployment of jumping robot sensor nodes for improving network performance in obstacle dense environment. *Measurement* **2023**, *207*, 112410. [CrossRef]
66. Sanfeliu, A.; Hagita, N.; Saffiotti, A. Network robot systems. *Robot. Auton. Syst.* **2008**, *56*, 793–797. [CrossRef]
67. Sanfeliu, A.; Andrade-Cetto, J.; Barbosa, M.; Bowden, R.; Capitán, J.; Corominas, A.; Gilbert, A.; Illingworth, J.; Merino, L.; Mirats, J.M.; et al. Decentralized Sensor Fusion for Ubiquitous Networking Robotics in Urban Areas. *Sensors* **2010**, *10*, 2274–2314. [CrossRef]
68. Borenstein, J.; Everett, H.R.; Feng, L.; Wehe, D. Mobile robot positioning: Sensors and techniques. *J. Robot. Syst.* **1997**, *14*, 231–249. [CrossRef]
69. Carcieri, S.M.; Jacobs, A.L.; Nirenberg, S. Classification of Retinal Ganglion Cells: A Statistical Approach. *J. Neurophysiol.* **2003**, *90*, 1704–1713. [CrossRef] [PubMed]
70. Nirenberg, S.; Carcieri, S.M.; Jacobs, A.L.; Latham, P.E. Retinal ganglion cells act largely as independent encoders. *Nature* **2001**, *411*, 698–701. [CrossRef] [PubMed]
71. Zaid, I.M.; Sajwani, H.; Halwani, M.; Hassanin, H.; Ayyad, A.; AbuAssi, L.; Almaskari, F.; Samad, Y.A.; Abusafieh, A.; Zweiri, Y. Virtual prototyping of vision-based tactile sensors design for robotic-assisted precision machining. *Sens. Actuators A Phys.* **2024**, *374*, 115469. [CrossRef]
72. Sumi, M. Simulation of artificial intelligence robots in dance action recognition and interaction process based on machine vision. *Entertain. Comput.* **2025**, *52*, 100773. [CrossRef]
73. Pettie, S.; Ramachandran, V. An Optimal Minimum Spanning Tree Algorithm. *J. ACM* **2002**, *49*, 16–34. [CrossRef]
74. Berndtsson, M.; Lings, B. *Logical Events and Eca Rules*; Technical Report HS-IDA-TR-95-004; University of Skovde: Skövde, Sweden, 1995.
75. Hochba, D.S. Approximation algorithms for NP-hard problems. *ACM Sigact News* **1997**, *28*, 40–52. [CrossRef]
76. Tindell, K.W.; Burns, A.; Wellings, A.J. Allocating hard real-time tasks: An NP-hard problem made easy. *Real-Time Syst.* **1992**, *4*, 145–165. [CrossRef]
77. Kim, H.; Rajkumar, R.R. Memory reservation and shared page management for real-time systems. *J. Syst. Archit.* **2014**, *60*, 165–178. [CrossRef]
78. Samsuria, E.; Mahmud, M.S.A.; Abdul Wahab, N.; Romdlony, M.Z.; Zainal Abidin, M.S.; Buyamin, S. Adaptive fuzzy-genetic algorithm operators for solving mobile robot scheduling problem in job-shop FMS environment. *Robot. Auton. Syst.* **2024**, *176*, 104683. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Development and Evaluation of a Tool for Blind Users Utilizing AI Object Detection and Haptic Feedback

Georgios Voutsakelis [1], Ioannis Dimkaros [2], Nikolaos Tzimos [1], George Kokkonis [2,*] and Sotirios Kontogiannis [3]

[1] Department of Business Administration, University of Western Macedonia, 51100 Greven, Greece; dba00012@uowm.gr (G.V.); dba00020@uowm.gr (N.T.)
[2] Department of Information and Electronic Engineering, International Hellenic University, 57400 Thessaloniki, Greece; giannisdimkaros@gmail.com
[3] Department of Mathematics, University of Ioannina, 45110 Ioannina, Greece; skontog@uoi.gr
* Correspondence: gkokkonis@ihu.gr

Abstract: This paper presents the development and evaluation of a smartphone application designed to improve accessibility for blind users. It uses the lightweight EfficientDet-lite2 model and the comprehensive COCO dataset in order to provide real-time object detection. The novelty of the application is in the integration of haptic feedback, which is activated when users touch objects that are detected on the screen, combined with audio notifications that announce the name of the detected object in multiple languages. This multimodal feedback mechanism helps blind users to recognize, explore, and move within their environment more effectively and safely. Extensive usability and user experience evaluation was conducted with blind and blindfolded users. The evaluation assessed the usability, effectiveness, accessibility, and user satisfaction and experience of the application. Additionally, a comparative analysis was performed between the use of haptic feedback and scenarios where haptic feedback was disabled. The results show a higher level of user satisfaction, greater ease of use, and significant potential for improving the independence of blind people when the haptic feedback is enabled. The findings also suggest that the inclusion of haptic feedback significantly enhances the user experience. This study underlines the importance of multimodal feedback systems in assistive technologies and the potential of mobile applications to provide accessible solutions for blind users.

Keywords: assistive technology; visual impairment; haptic interfaces; smartphone applications; artificial intelligence (AI); object detection

1. Introduction

Generally speaking, digital accessibility in technology is essential in ensuring that more people can use technologies and people with disabilities can more effectively use digital applications, web platforms, and devices. In combination with assistive technology (AT), which includes a wide variety of applications and devices, the functional abilities of people with disabilities and their daily lives can be greatly improved. For people that have lost their sight or people that have impaired vision, this technology includes Braille displays, screen readers, voice recognition software, and more. The development of AT has been crucial to providing blind and visually impaired users with the tools they need to effectively interact with digital content and navigate their physical environment more independently.

Recent advances in machine learning and artificial intelligence (AI) have further expanded the possibilities for AI and have improved multiple assistive technologies that

have long been in use. Those improvements include real-time object detection and contextual understanding through accessible interfaces. These novelties are vital to promoting independence and improving the quality of life of blind users [1,2].

Object detection technology has advanced significantly with the emergence of large datasets and deep learning models. One of the prominent models used for object detection is EfficientDet-lite2, which, due to its lightweight architecture, is designed for efficient performance on mobile devices. The COCO (Common Objects in Context) dataset is widely used for evaluating and training object detection models. It is ideal for developing applications that require real-time object detection in a variety of environments, as it includes a wide and diverse set of everyday objects.

Specifically for blind users, smartphone applications that have object detection capabilities can significantly improve the interaction of those users with their environment and can be a powerful tool for visually impaired users. The navigation and interaction of blind users with their physical environment can be more efficient and safe because the user can receive immediate feedback about their environment by using real-time object detection. This technology can be particularly beneficial in unfamiliar settings, by providing important information about obstacles and other objects in the user's path [3].

For all people, but especially for blind users, the sense of touch is very important because it can transmit information. In the context of assistive technology, haptic feedback can provide valuable signals to blind users, enhancing their ability to perceive and interact with digital content and physical environments. Moreover, when it is integrated with object detection technology, haptic feedback can provide immediate and intuitive feedback about the detected objects, improving the spatial awareness of the user and giving them more confidence in navigating their surroundings.

The primary goal of this study is to develop an application that uses object detection to assist blind users in locating objects within their environment. The use of haptic feedback in this application represents a significant innovation. By combining visual information from the camera, audio feedback, and haptic interaction, the application creates a multimodal feedback system that supports the diverse needs of blind users. This approach aligns with the principles of universal design, which support the creation of products and environments that can be used by all people, to the greatest extent possible, without the need for customization or specialized design.

In addition, this research evaluates the usability, the accessibility, and the user experience of the application through a structured questionnaire administered to both blind users and sighted people simulating blindness. A significant novelty of this application is the incorporation of haptic feedback to provide haptic cues when users interact with detected objects. This combination of technologies aims to enhance the user's ability to explore and understand their environment, offering a practical solution to some of the challenges faced by the visually impaired.

This study contributes to the field of assistive technology by demonstrating how advanced artificial intelligence models and innovative feedback mechanisms can be incorporated into accessible mobile applications. The findings of this study provide valuable insights into user preferences and the effectiveness of different types of feedback in improving usability and accessibility. By focusing on the needs and experiences of both blind and sighted users simulating blindness, this research highlights the importance of inclusive design in technology development.

The remainder of this paper is organized as follows. Section 2 reviews related work in the areas of object detection technology and assistive applications for visually impaired users. Section 3 describes the methodology used to develop and evaluate the application, including technical implementation and user interface design. Section 4 presents the results

of the usability evaluation, highlighting the important findings and the user feedback. In Section 5, a discussion is made about the novelty and implications of these findings, the usefulness of the haptic feedback feature, and the areas for future research. Finally, Section 6 concludes the paper, summarizing the contributions and potential impact of this study.

2. Related Work

The development of assistive technologies, and particularly those aimed at improving the independence and quality of life of visually impaired individuals, has gained significant importance in recent years. As technology continues to advance, the integration of artificial intelligence (AI) and machine learning into these technologies has opened new doors for innovative applications. One of the most important areas of focus has been the use of object detection technologies in order to assist individuals with visual impairments in recognizing their environment and moving within it.

Object detection, which involves recognizing and classifying objects within an image or video stream, has evolved through the implementation of deep learning models such as EfficientDet-lite2, which is known for its efficiency on mobile devices [4]. These technologies have been trained using large datasets such as the COCO (Common Objects in Context) dataset [5], which contains a wide variety of labeled objects commonly encountered in everyday life. By incorporating these object detection systems into mobile applications, developers have been able to create accessible tools that offer real-time object identification, allowing visually impaired users to interact better and more confidently with their surroundings.

At the same time, there has been an increasing emphasis on designing accessible interfaces and integrating multimodal feedback systems—such as haptic and auditory feedback—into assistive technologies. Haptic feedback, in particular, has proven to be an important feature for visually impaired users, providing immediate tactile confirmation when they interact with digital or physical elements [6]. This has become an increasingly popular approach in applications designed to assist visually impaired individuals, as it allows users to confirm their interactions without relying on visual feedback. Similarly, audio notifications offer an additional layer of accessibility, allowing users to receive verbal signals about their environment or actions. By combining both the haptic and auditory feedback, developers can create a more comprehensive and inclusive user experience, ensuring that visually impaired users receive complementary indications about their surroundings.

Several studies have demonstrated the value of combining AI-based object detection with accessible interface designs. For instance, wearable devices and mobile applications that integrate object detection and real-time feedback mechanisms have shown great potential in making it easier for users to navigate both familiar and unfamiliar environments. Research has also highlighted the importance of adapting these technologies to meet the specific needs of visually impaired individuals [6]. This includes designing applications that are not only accurate and efficient in their detection capabilities but also intuitive and easy to use, with clear and consistent feedback. Also, the use of mobile platforms for providing these solutions has proven particularly beneficial, as smartphones are widely accessible and have the necessary computing power in order to run sophisticated machine learning models.

Kuriakose et al. [7] present a review of multimodal navigation systems designed to assist visually impaired users. The paper examines various approaches that combine sensory modalities, such as auditory, haptic, and visual cues, to enhance navigation in indoor and outdoor environments. The authors analyze the strengths and limitations of these systems, and propose specific features that can improve multimodal navigation

systems in relation to their accessibility while simultaneously emphasizing the importance of AI to overcome the various challenges in multimodal systems.

Also, Kuriakose et al. [8] developed DeepNAVI, an assistive navigation system for the blind using a smartphone and a bone conduction headset, incorporating obstacle, distance, scene, position, and motion detection to enhance spatial awareness without losing situational awareness. This system ensures portability and convenience while providing navigation aids to visually impaired users.

Rakkolainen et al. [9] conducted a review on technologies enabling multimodal interaction in extended reality (XR), including virtual and augmented reality, with an emphasis on accessibility. The review highlights how multimodal interfaces, including haptic and auditory feedback, can be used to make XR environments more accessible to visually impaired individuals. The paper emphasizes the potential for using XR technologies as tools for enhanced perception for users with vision impairments.

Jiménez et al. [10] developed an assistive locomotion device that provides haptic feedback to guide visually impaired individuals. The device uses sensors to detect obstacles and offers real-time haptic feedback to help users navigate safely. The study highlights how haptic feedback can improve physical environment awareness and autonomy for visually impaired individuals by providing precise haptic guidance.

See et al. [11] developed a wearable haptic feedback system designed to help visually impaired and blind individuals detect obstacles in multiple regions around them. The system uses sensors to monitor the environment and provides targeted haptic feedback to alert users to obstacles in different directions, improving their spatial awareness and mobility.

Palani et al. [12] conducted a behavioral evaluation comparing map learning between touchscreen-based visual and haptic displays, targeting both blind and sighted users. The study assessed how effectively visually impaired individuals can learn and navigate maps using haptic feedback compared to traditional visual methods. The findings suggest that haptic displays are highly effective for conveying spatial information, offering an accessible alternative for map learning for users with visual impairments.

Raynal et al. [13] introduced FlexiBoard, a portable multimodal device that provides tactile and tangible feedback to assist visually impaired users in interpreting graphics and spatial information. The system supports touch-based interaction, allowing users to explore physical representations of graphics while receiving auditory feedback for additional context. The study demonstrates the effectiveness of tactile graphics in enhancing learning and accessibility for people with vision impairments.

Nikolaos Tzimos et al. [14] examined the integration of haptic feedback into virtual environments to enhance user interaction, particularly benefiting visually impaired individuals when exploring the tactile internet.

Accessibility features in mobile apps are critical to ensuring that visually impaired users can effectively interact with digital content. These features often include screen readers, haptic feedback, and voice recognition technologies. Recent developments have led to the integration of artificial intelligence and machine learning to further improve accessibility.

Billi et al. [15] proposed a methodology for evaluating the accessibility and usability of mobile applications, combining user-centered design principles with accessibility guidelines. Their study shows how a systematic approach to evaluation can improve both the usability and accessibility of mobile apps, ensuring they meet the needs of users with disabilities.

Ballantyne et al. [16] examined existing accessibility guidelines for mobile apps, analyzing how well these guidelines are applied to mobile app design. The study assesses compliance with accessibility standards in various applications, highlighting common

barriers faced by users with disabilities, including visual impairments. Their findings highlight the need for better enforcement and clearer guidelines to improve the inclusion of mobile apps.

Park et al. [17] developed an automated evaluation tool designed to evaluate the accessibility of Android applications. The tool provides real-time feedback on app accessibility compliance, identifies common issues, and offers recommendations for improvements. This research points out the importance of automating the accessibility assessment process to make it easier for developers to create more inclusive apps.

Acosta-Vargas et al. [18] focused on the accessibility assessment of Android mobile applications, using human factors and system interaction methodologies. The study identified key accessibility issues faced by users with disabilities and presented recommendations for improving mobile app accessibility through better adherence to design principles and guidelines.

Mateus et al. [19] conducted a user-centered evaluation of mobile applications, focusing on their accessibility for the visually impaired. The study compared the experiences of visually impaired users with the results of automated accessibility assessment tools. The research highlighted the discrepancies between automated assessments and actual user experiences, underscoring the importance of user involvement in the assessment process.

Acosta-Vargas et al. [20] explored the accessibility of native mobile applications for people with disabilities in their scope review. The review highlighted significant gaps in accessibility compliance and the need for better tools and methodologies to ensure that mobile apps are accessible to all users, particularly those with vision and physical impairments.

Object detection technologies have been significantly improved, particularly in noisy or dynamic environments. For instance, Zheng et al. [21] proposed an asymmetric trilinear attention net with noisy activation functions to enhance automatic modulation classification under signal distortions, demonstrating how attention mechanisms can stabilize performance in challenging conditions. Similarly, Liu et al. [22] changed the multi-scale feature hierarchy in detection transformers (DETRs), addressing scale-variant object detection.

3. Materials and Methods

The application in this study leverages artificial intelligence (AI) for object detection and incorporates haptic feedback to enhance user interaction. The primary goal of the application is to provide real-time object detection and feedback to users, enabling them to navigate inside their environment more effectively.

The application was developed for Android devices using Android Studio, the official Integrated Development Environment (IDE) for Android app development. The primary programming language used was Kotlin, which is fully supported by Android Studio and offers modern features that enhance productivity and code safety.

The user interface (UI) was designed to be simple and intuitive, with a focus on accessibility. The main screen displays a live camera feed with overlaid bounding boxes around detected objects. Users can interact with the application by touching the screen, which triggers haptic and voice feedback. The application was designed to be simple and easy to be used by users with vision problems. The application does not have any buttons, and upon opening it, it directly performs its basic function, which is to recognize objects through the camera, as shown in Figure 1. Of course, there is the possibility of adjustments for the type of vibration, the selection of AI object recognition libraries, and the number of processors, but these are intended for use by the research team.

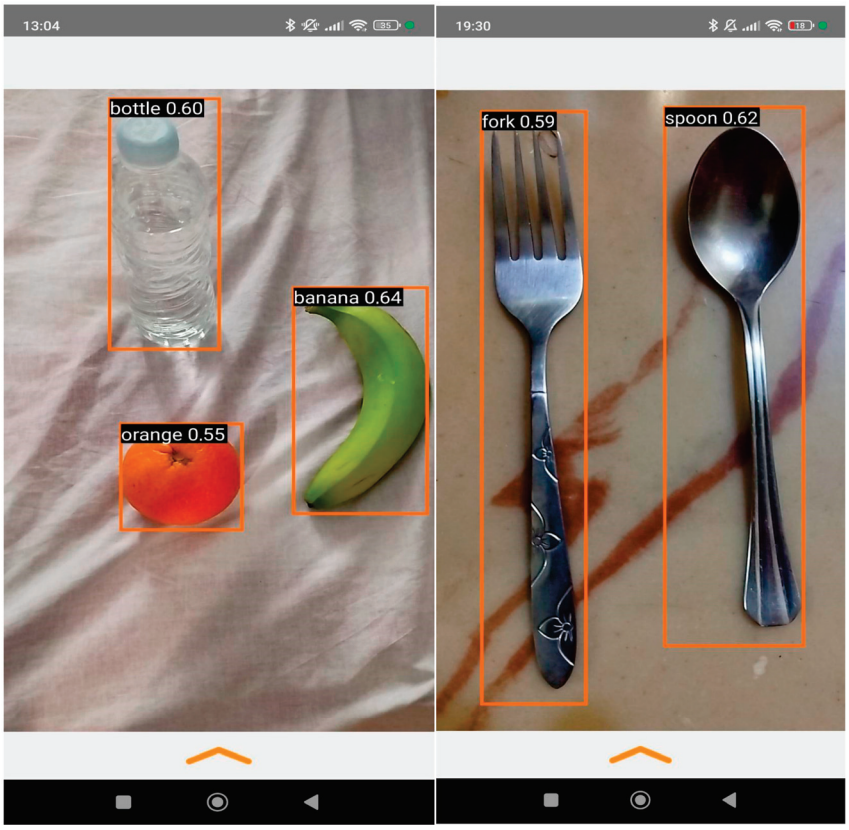


Figure 1. Object detection via the app.

The application utilizes the TensorFlow Lite framework for implementing the AI-based object detection model. TensorFlow Lite is a lightweight version of TensorFlow designed specifically for mobile and embedded devices, providing efficient inference with minimal latency. The model was integrated into the application using the TensorFlow Lite Object Detection API, which provides pre-built functions for loading the model, processing images, and interpreting the detection results. The API also supports customization of detection parameters such as confidence thresholds and the maximum number of objects to detect.

The core functionality of the application relies on the EfficientDet-lite2 model, a lightweight convolutional neural network (CNN) optimized for mobile devices. EfficientDet-lite2 is pre-trained on the COCO (Common Objects in Context) dataset, which includes a wide range of everyday objects.

In order to determine the object detection model that we will use in our application, we compared the EfficientDet-Lite models alongside SSD MobileNetV2 variants to select a model optimized for mobile assistive technology. EfficientDet-Lite and SSD MobileNet Models are the two categories of model that are commonly used in object detection of edge computing devices [23]. As shown in Table 1, for the COCO 2017 validation dataset, on a Pixel 4 smartphone using 4 threads on CPU, EfficientDet-Lite2 achieves 33.97% mAP at 69 ms latency. It combines accuracy and speed. While SSD MobileNetV2 320 $\times$ 320 is faster (24 ms), its low accuracy (20.2% mAP) risks missing critical objects, and SSD FPNLite 640 $\times$ 640 (28.2% mAP) has higher latency (191 ms) [24,25].

Table 1. Comparative analysis of lightweight object detection models [24,25].

Model	Size (MB)	Latency (ms)	mAP (%)
EfficientDet-Lite0	4.4	37	25.69
EfficientDet-Lite1	5.8	49	30.55
EfficientDet-Lite2	7.2	69	33.97
EfficientDet-Lite3	11.4	116	37.70
EfficientDet-Lite4	19.9	260	41.96
SSD MobileNetV2 320 × 320	6.7	24	20.2
SSD MobileNetV2 FPNLite 640 × 640	4.3	191	28.2

The application was developed by extending the open-source TensorFlow Lite Object Detection Android Demo [26]. The official TensorFlow Lite demo provides a well-tested tool for real-time object detection. It is optimized for mobile devices, ensuring low latency and efficient resource usage. Its modular design allowed the integration of new features like haptic feedback and multilingual voice support. We simplified the interface to prioritize voice and haptic feedback over visual elements.

We conducted performance testing using Android Profiler on a Xiaomi Redmi Note 11 (Xiaomi Corporation, Beijing, China; Snapdragon 680, 4GB RAM)under default application settings. The runtime metrics of the application are depicted in Table 2.

Table 2. Runtime metrics of the application.

Metric	Value
Inference Latency	70 ms
FPS	60 FPS
CPU Usage	27%
Memory Usage	150 MB (256 MB allocated to the Profiler)

While the base project provided the basic object detection functionality, the following modifications were implemented to make the app suitable for visually impaired users, enhancing accessibility through multimodal feedback mechanisms.

The application triggers vibrations when the user interacts with detected objects on the screen. The vibration patterns were designed to provide intuitive feedback.

The haptic feedback system was integrated with the object detection results, ensuring that users receive tactile feedback when they touch the screen near a detected object. This feature aims to enhance the user experience by providing an additional sensory cue that complements the visual and auditory feedback.

We added vibration patterns using Android's Vibrator class (for API version lower than 26) to provide tactile feedback when users interact with detected objects.

The vibration pattern that was used was a square pulse of vibration with an "on" duration of 10 ms. This effect is triggered continuously while the user is touching the object. The vibration starts immediately (0 ms delay), when the user touches the detected object and it vibrates repeatedly for 10 ms. The time gap between the pulses depends on the smartphone CPU and is near 10 ms.

The specific vibration pattern was chosen to provide pleasant, rapid, and intuitive feedback to the user [27]. For android phones with an API Level equal to or higher than 26,

we apply the VibratorManager and the VibrationEffect library, while for API levels lower than 26, we enforce the Vibrator API.

At the current research version, the application uses fixed vibration parameters during experimental evaluation. The user can change the vibration intensity through the android system settings based on their preferences.

To further assist visually impaired users, the application includes text-to-speech (TTS) functionality. The TTS engine reads out the names of detected objects in a selected language, providing an additional layer of information. Our implementation utilizes Android's native TextToSpeech API (version 9.0+) which offers comprehensive language support through several key features. The system currently supports eight major languages (English, Greek, Spanish, French, German, Italian, Japanese, and Mandarin). The application translates the object labels from English (as provided by the COCO dataset) to other supported language, such as Greek, using a custom HashMap that maps each object category to its corresponding Greek translation. This ensures that users receive feedback in their native language.

When the user touches the screen within the bounding box of a detected object, the speakText() method is triggered. This method retrieves the object's label (e.g., "cup," "chair") from the detection results. A predefined HashMap (e.g., "cup" → "φλιτζάνι") translates the label to Greek using android's TextToSpeech engine to vocalize the translated label in Greek. The system maintains synchronization between haptic and audio feedback, with vibration and speech output beginning immediately upon touch detection.

The flowchart diagram of the application is given in Figure 2.

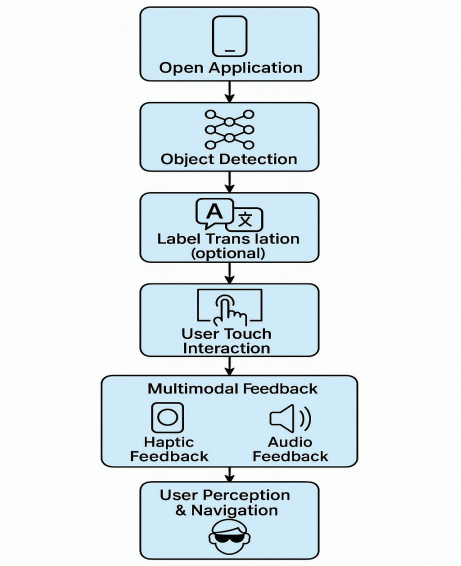


Figure 2. Flowchart diagram of the application.

The application also includes a settings menu that allows users to adjust various parameters, such as the detection sensitivity (where lower sensitivity detects more objects but may include false positives), the number of objects to detect simultaneously, and the type of feedback (haptic, voice, or both). The settings menu is accessible via a sliding panel, ensuring that it does not obstruct the main camera view and cannot be easily activated by mistake. The settings menu is depicted in Figure 3.

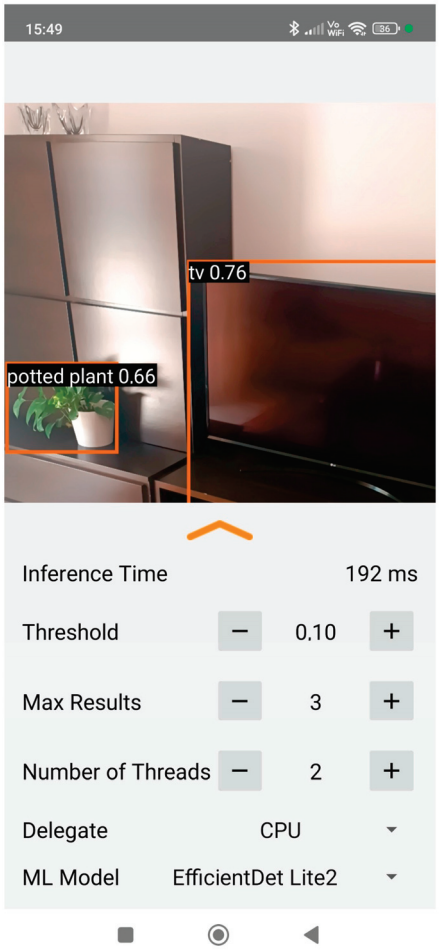


Figure 3. Settings menu of the application.

3.1. Experimental Design

An evaluation was conducted to compare user performance and satisfaction between the two versions of the application, with haptic feedback and with no haptic feedback, and to evaluate the usefulness and user experience of haptic feedback in aiding navigation and object recognition for visually impaired users.

The evaluation focused on several key metrics, including the following:

Usability: Measured through a questionnaire that assessed how easy the application was to use and how well it met the users' needs.

Effectiveness: Evaluated based on the accuracy of object detection and the responsiveness of the feedback system.

User Satisfaction: Assessed through questions about the overall user experience, including the quality of haptic and voice feedback.

Ease of Learning: Measured by how quickly users were able to learn and use the application effectively.

All participants were informed about the experiment's purpose and procedures and provided informed consent before participating. The study was conducted in accordance with ethical guidelines, while data confidentiality and anonymity were maintained. Participants were also given the option to withdraw from the study at any time without any consequences.

We categorized the research on the application's usability into two distinct studies. The first study evaluated the app's usability for people who have vision problems. The primary reason for including both blindfolded and blind participants in our study was

to understand, in general, the differences in perception in matters of usability between sighted and blind people.

We made sure that the physical environment of the experimental tests was as unchanged as possible. For this reason, we used specific rooms whose interiors remained unchanged despite the changes in the users participating in the tests. In all cases, the tests were conducted in a designated room on the premises of their training institution. We also made sure that there were no external factors that influenced the survey users. Inside those rooms, we placed three objects that the application had a high probability of recognizing.

To determine the optimal observation distance between the observer and the objects for detection, we must consider the human field of view (FOV) and the observation width, which is regarded as the horizontal span of the scene detectable by the observer. The relationship is expressed by Equation (1) and depicted in Figure 4.

$$\text{Observation distance} = \frac{\text{Observation Width}/2}{\tan\left(\frac{\text{FOV}}{2}\right)} \quad (1)$$

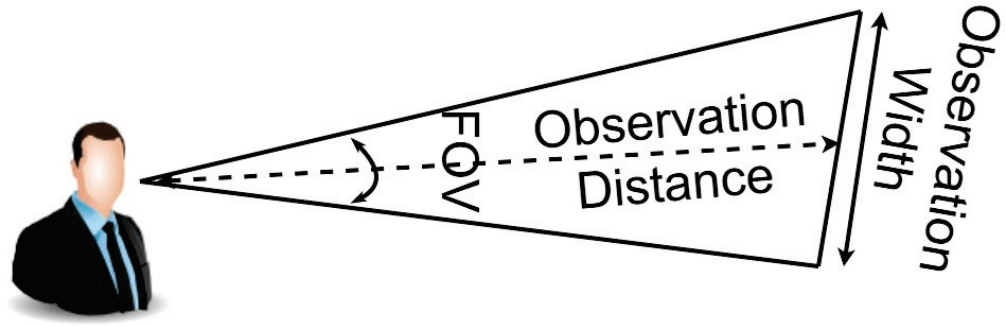


Figure 4. Human field of view vs. observation distance.

The human field of view (FOV) for object recognition varies depending on the level of visual attention required and the type of visual processing involved. For central vision where a high concentration of object recognition is required, the FOV is 5–10°. This is where the highest visual acuity occurs, allowing for detailed object recognition, reading, and fine motor coordination [28]. For object detection in the parafoveal vision, the FOV is between 10 and 30°. This region helps in recognizing objects without directly looking at them, while it lacks fine detail compared to the central vision [29].

The distance proposed for object detection with the help of the phone's camera to compensate for people with vision problems is equal to the distance resulting from Equation (1) for parafoveal object recognition with a maximum FOV of 30 degrees. If the maximum observation width is equal to 2.00 m, the maximum distance of the observer, according to Equation (1), is approximately 3.73 m.

Participants were seated in a comfortable, predetermined position. Each smartphone provided had both versions of the application pre-installed and readily accessible for comparison.

While the application settings were simplified for initial usability testing, changing this setting currently requires assistance from a support person. Future updates will prioritize voice-guided navigation and WCAG compliance to enable fully independent use by blind users.

Before the evaluation, the research team conducted a training session to familiarize participants with the app. They were given time to practice with both versions until they felt confident proceeding to the assessment phase.

Some blind participants were assisted by a support person during initial device use, while others received guidance from the research team. Not all the participants had prior

experience with similar applications. The team or support person helped position the phone correctly and directed their fingers to the appropriate areas of the screen.

The user each time had to locate the three objects in front of them at a distance of 3.5 m, with the help of the device. Every time the user changed, we changed the order of the versions that they were going to evaluate in order to have more objective results. Also, for the same purpose, we changed the location of the objects every time the evaluation of one version ended and the evaluation of the other version of the application was going to begin. We placed particular emphasis on user satisfaction and ease of use of the two versions of the application, and, in particular, we wanted to see if the version with vibration offered more satisfaction and was easier to use than the simple version, rather than focusing on the speed of locating the objects.

In the experimental design, we used a simplified environment with a limited number of objects (3–5 items). By minimizing the environmental variables, we could accurately measure the impact of vibration feedback on object detection performance. Standardized and stable conditions ensured comparable results across all participants (blind and blindfolded).

In the experimental scenario with no haptic feedback, the vibration feature was deactivated. The deactivation of the haptic feedback ensured that the users experienced only audio feedback, allowing for precise comparison of user experience between the two experimental conditions.

The study included 18 participants with normal or corrected-to-normal vision and 8 blind participants. Those with normal vision performed the assessment with their eyes closed (blindfolded), though it is important to note that blindfolding cannot fully replicate the experience of visual impairment. Consequently, simulations—no matter how carefully designed—cannot perfectly mirror real-world scenarios.

To address this limitation, our testing protocol follows a multi-stage process, involving method validation, preliminary testing, simulation trials, and final functionality evaluations, before progressing to user acceptance testing. These phases are vital for advancing assistive and inclusive technologies, as the insights gained—even from simulations—can enhance our understanding of disability-related needs and spur innovation. The findings and methodologies presented in this study aim to contribute to the scientific community, informing future research and development in the field.

Regarding the gender of the participants, 15 were male and 11 were female. The age range was 13 people from 17 to 25 years old, 5 people from 26 to 35 years old, 4 people from 36 to 45 years old, and finally 4 people who were 46 years old and over. All sighted participants were unfamiliar with such kind of applications prior to the study. The detailed distribution of age, gender, and vision status of the participants is depicted in Table 3.

Table 3. Detailed distribution of the 26 participants (8 blind, 18 blindfolded).

Age Group	Blind (Male/Female)	Blindfolded (Male/Female)
17–25 years	2M/2F	6M/3F
26–35 years	1M/0F	3M/1F
36–45 years	0M/2F	1M/1F
46+ years	1M/0F	2M/1F

3.2. Performance Measures

The users were asked to complete a 30-question questionnaire, based on a usability questionnaire (USE) [30] and the User Experience Questionnaire (UEQ) [31].

We used both the USE (Usefulness, Satisfaction, and Ease of Use) Questionnaire and the User Experience Questionnaire (UEQ) to provide a comprehensive evaluation of the system's usability and user experience from different perspectives. The USE Questionnaire focuses on usability aspects, measuring usefulness, ease of use, ease of learning, and satisfaction. This questionnaire is particularly valuable for assessing functional usability and efficiency. The UEQ extends beyond usability, evaluating attractiveness, efficiency, dependability, stimulation, and novelty. This allows us to measure users' subjective experiences and emotional engagement with the system. By combining these two questionnaires, we ensured a holistic assessment, addressing both practical usability concerns and user experience factors to gain a well-rounded understanding of the application's effectiveness and appeal.

The USE Questionnaire was divided into five sections: demographics, usefulness, ease of use, ease of learning, and satisfaction. Moreover, we evaluated the effect of the haptic feedback in both questionnaires in order to have a comparative analysis for blind and blindfolded users, as well as a comparative analysis for scenarios with haptic feedback activated and not.

The questionnaire responses were rated using a Likert scale, as presented in Table A1 in Appendix A. For the USE Questionnaire, we used a 5-point Likert scale (1 to 5), while the UEQ used a 7-point scale (1 to 7). To evaluate each question, we calculated the mean and standard deviation of participants' responses. Additionally, the mean values were normalized for better comparison between the four categories in the tables below.

Regarding the comparative analyses, the study deliberately prioritized qualitative indicators, usability, and user experience, as these dimensions are critical for evaluating assistive technologies. However, we acknowledge the value of integrating quantitative metrics such as task completion time or error rates, as it could provide a more comprehensive assessment of system performance in future studies.

4. Results

4.1. USE Questionnaire Evaluation

The data collected from the USE Questionnaire were analyzed using descriptive statistics. The results of the participants' responses are presented below in Tables 4 and 5. The analysis focused on identifying trends and patterns in the data, such as the overall satisfaction with the application and the perceived usefulness of the haptic feedback system.

The comparison results of the two versions of the application, with or without haptic feedback, for the two user categories, blind and blindfolded users, are presented in Table 4.

The next step was to calculate the mean values of the four main factors of the questionnaire, as shown in Table 5 for each category user.

The analysis of the USE Questionnaire responses, as shown in Tables 4 and 5, provides a clear indication of the impact of haptic (vibration) feedback on the usability of the application. Four key dimensions were assessed, usefulness, ease of use, ease of learning, and satisfaction, across two groups of participants, blind users and blindfolded users, and two application conditions, with and without vibration feedback.

From Tables 4 and 5, it is evident that blind users consistently rated the application higher when vibration feedback was enabled. Specifically, usefulness responses (questions 1–3) have an average increase of approximately 18.34% compared to the non-vibration feedback. A similar trend was observed in **ease of use**, where participants highlighted the intuitive and user-friendly nature of the interface when enhanced with haptic feedback. The most notable difference was in satisfaction (questions 10–12), where users expressed a greater desire to use the application regularly and recommend it to others when vibration

was present (90% vs. 71.67%). This suggests that the vibration feedback enhanced their sense of control, engagement, and trust in the system.

Table 4. Questionnaire results for usability evaluation.

Category of Users	Blind Users		Blindfolded Users	
Questions	With Vibration	No Vibration	With Vibration	No Vibration
Usefulness				
1. It helps me be more effective	80%	55%	62.22%	62.22%
2. It is useful	90%	70%	72.22%	68.88%
3. It gives me more control over the activities in my life	75%	65%	73.33%	68.88%
Ease of Use				
4. It is easy to use	95%	75%	91.11%	84.44%
5. It is user friendly	100%	80%	90%	81.11%
6. I can recover from mistakes quickly and easily	80%	70%	78.88%	73.33%
Ease of Learning				
7. I learned to use it quickly	100%	100%	100%	97.78%
8. I easily remember how to use it	100%	100%	100%	95.56%
9. It is easy to learn to use it	100%	100%	97.78%	93.33%
Satisfaction				
10. I feel I need to have it	85%	70%	44.44%	44.44%
11. I would recommend it to a friend	85%	70%	64.44%	64.44%
12. It is fun to use	100%	75%	78.89%	77.78%

Table 5. Cumulative results of usability evaluation.

Category of Users	Blind Users		Blindfolded Users	
Usability Evaluation	With Vibration	No Vibration	With Vibration	No Vibration
Usefulness	81.67%	63.33%	69.25%	66.67%
Ease of Use	91.67%	75%	86.67%	79.63%
Ease of Learning	100%	100%	99.26%	95.56%
Satisfaction	90%	71.67%	62.59%	62.22%

Blindfolded users also reported higher scores in most categories under the vibration-enabled condition, albeit with slightly smaller margins. For example, their reported usefulness increased from 66.67% without vibration to 69.25% with vibration. In the ease of use and ease of learning dimensions, scores remained high overall, indicating that the application was perceived as accessible and intuitive regardless of the user's prior visual experience. However, satisfaction again showed a noticeable difference, with users expressing slightly more positive sentiment (62.59% vs. 62.22%) when the vibration feature was active. Notably, all users, regardless of group, achieved very high scores in ease of learning category (near or at 100%), suggesting a short learning curve and quick familiarity with the interface. Overall, as summarized in Table 5, the results clearly support that vibration feedback improves user usability across all four evaluation factors.

4.2. UEQ Evaluation

The User Experience (UX) Questionnaire (UEQ) results are presented in Table 6. They describe how the participants perceived the quality and design of the application, particularly in relation to its efficiency, attractiveness, dependability, stimulation, and novelty. We used the short version of the UEQ, so as not to be too exhausting for the interviewers, in combination with the USE Questionnaire and the open-response interview. The evaluation of UX also involved two user categories, blind and blindfolded, and two interface conditions, with and without vibration feedback.

Table 6. Cumulative results of user experience evaluation (UEQ).

Category of Users		**Blind Users**		**Blindfolded Users**	
	Questions	**With Vibration**	**No Vibration**	**With Vibration**	**No Vibration**
1.	Supportive	78.57%	64.29%	77.71%	71.43%
2.	Easy	92.86%	92.86%	85.71%	87.29%
3.	Efficient	60.71%	57.14%	71.43%	65.14%
4.	Clear	71.43%	60.71%	84.14%	84.14%
5.	Exciting	89.29%	78.57%	69.86%	65.14%
6.	Interesting	89.29%	78.57%	76.14%	76.14%
7.	Inventive	96.43%	71.43%	71.43%	73.00%
8.	Leading edge	89.29%	64.29%	69.71%	68.29%

For blind users, the results show consistently higher ratings with vibration feedback enabled. The application was more supportive (5.5 vs. 4.5) and more inventive (6.75 vs. 5.0) with vibration feedback, suggesting a strong improvement in the user's experience. Particularly high scores were recorded for ease of use (6.5 for both conditions), indicating that the application was easy to use regardless of vibration. Additionally, the categories "Exciting" (6.25 vs. 5.5) and "Interesting" (6.25 vs. 5.5) also show that the haptic interaction offers more engaging and stimulating interaction.

Among blindfolded users, the pattern is similar, though with slightly less contrast between the two conditions. Notably, the "Clear" metric remained constant at a high value of 5.89 for both versions, indicating that users could interpret and understand the application effectively in either case. The vibration improved the categories "Supportive" (5.44 vs. 5.00) and "Efficient" (5.00 vs. 4.56), demonstrating that the haptic feedback provided an additional layer of usability and control. Interestingly, the categories "Exciting" and "Interesting" showed minor improvements or no difference.

In summary, the UEQ results reinforce the findings from the USE Questionnaire: vibration feedback positively influences user experience by enhancing both usability (e.g., efficiency, supportiveness) and experience (e.g., excitement, inventiveness). Blind users, in particular, showed a more significant appreciation of the system's enhanced feedback, underscoring the importance of multimodal interaction design in assistive technologies.

4.3. Test Reliability

The outcome of the Cronbach's Alpha reliability coefficient of the questionnaire for the application evaluation is depicted in Table 7.

Table 7. Haptic feedback consistency factor.

Cronbach's Alpha	With Haptic Feedback	No Haptic Feedback
α	0.89	0.84

Table 7 depicts that both conditions show strong internal consistency ($\alpha > 0.8$), indicating that the reliably measures of the overall user experience. The haptic feedback interaction improves the coherence of user perceptions across all dimensions (usefulness, ease, learning, and satisfaction). The application with haptic feedback enabled has a higher internal consistency reliability coefficient than without haptic feedback, which means that the respondents' answers remain more stable in the case with haptic feedback enabled. Consequently, the application with haptic feedback demonstrated broader user acceptance and satisfied a greater proportion of users.

4.4. Analysis of the Results

User feedback was collected from open questions, and the participants highlighted key areas for improving the usability and functionality of the assistive application. A suggestion was the need for enhanced object recognition, both in terms of the number of identifiable items and the reliability of the recognition system. Users demanded more detailed object descriptions, including the identification of object color. Additionally, spatial information was considered essential; participants requested functionality that provides not only the distance to detected objects but also their relative position (e.g., left, right, in front) in relation to the user. This feedback points to the importance of integrating more advanced computer vision capabilities and contextual awareness features in future development stages.

Furthermore, users positively highlighted the value of a simple and intuitive interface, with support for voice commands and natural-sounding audio output. The haptic feedback was suggested as a helpful aid for orientation and navigation. Device compatibility was also mentioned, with users recommending broader support for both Android and iOS platforms, including wearables like smartwatches. An additional proposed feature was a scanning function for reading product labels, prices, and expiration dates in environments like supermarkets. Overall, the feedback highlights the necessity of combining functional accuracy with user-centered design to create a more effective, accessible, and inclusive assistive application for visually impaired individuals. Nearly all participants reported satisfaction with the application, describing it as enjoyable and entertaining. Blindfolded participants, who simulated visual impairment during testing, describe the application as valuable tool for environmental navigation. Several blind participants expressed enthusiasm about their first experience with the application, noting they discovered a novel method for identifying nearby objects. They also noted that they would use this application as an additional tool to help them in their daily mobility.

From the processing of the USE Questionnaire, the comparisons of the two versions of the application are presented in Table 4. We see that the app with vibration presents better results than that without vibration in ten out of twelve questions of the questionnaire. This is also reflected in the calculation of the mean values of the four main factors of the questionnaire, as shown in Table 5.

By observing the users who participated in the research, some important conclusions were drawn regarding how they interact with the application. Most users initially showed interest and eagerness to try the application.

Younger ages, on average, became familiar with it more quickly and continued to use the application at a faster rate. Additionally, the different lengths of fingers and the long

nails of women made the precision of the movements more difficult. In general, users used different strategies to locate the detected items on the touch screen.

Another important finding was that haptic interaction helped users understand the exact position of objects and the size they occupied on the screen. This is due to the fact that when the user's finger touched the object, the device began to vibrate, while when the finger left the boundaries of the object, the vibration stopped; thus, the user understood both the exact position of the object on the screen and the size of this object.

The most important observation was that when the haptic feedback was enabled in the application, users identified the objects that had been detected with the help of technical intelligence more quickly and easily, while large variations were observed regarding the time and effort it took each user to identify all the objects that were detected by the application.

5. Discussion

This paper qualitatively evaluates the role of haptic feedback on object detection applications for visually impaired people that can be implemented on popular touchscreen devices such as tablets and mobile phones. To evaluate the effectiveness of haptics, a mobile phone Android application for visually impaired users was developed, supporting haptic feedback. Then, the application interface functionality was put to the test using the USE Questionnaire and the UEQ, handed out to the application testers via Google Forms. The usability evaluation of the application was carried out separately on two subsets of users. In one set, participants had normal vision but they were blindfolded for the purpose of the research, and in a second set, blind users participated. Haptic feedback scored higher marks on all four USE Questionnaire factors, so we can consider that the use of haptics increases usability in these kinds of applications.

The first contribution of the present study is the creation of an application for smartphones and tablets, which gives tactile and vocal feedback to the user. The second contribution is evaluating and comparing the use of haptic feedback on these kinds of applications. The third contribution of this study is the realization of the different perceptions of usability, in general, between sighted and blind users.

For practical reasons, we investigated the effect of haptic feedback in object detection applications. We can say with certainty that haptic feedback improves usefulness and use satisfaction. Moreover, we can also say that haptics enhance both accessibility and engagement, particularly for users with visual impairments. The performed evaluation also highlights the potential of combining intuitive and simple UI design with multimodal feedback to support inclusive technology adoption.

Unlike commercial applications that are generally based on visual substitution (e.g., Seeing AI [32] and EnvisionAI [33]) or human assistance (e.g., Be My Eyes [34]), this paper proposes haptic interaction to be used by people with visual impairments for environmental/spatial exploration. While current commercial solutions are useful in specific domains like OCR or live video support, they lack the tactile feedback mechanisms critical for spatial awareness. A qualitative comparison between commercial applications is depicted in Table 8.

Table 8. Comparison with commercial applications.

Feature	Proposed Application	Seeing AI	Be My Eyes	Envision AI
App type	Research purposed	Commercial/ free	Commercial/ free	Commercial/ free
Object detection	Real-time AI object detection	Object/scene recognition	No	Real-time AI object detection
Haptic feedback	Vibration patterns + touch interaction	No	No	No
Multimodal feedback	Haptic + audio (multimodal)	Audio-only	Live video (human-assisted)	Audio-only
User independence	High (no external assistance)	Moderate (many buttons)	No (Relies on volunteers)	High
Real-time interaction	Yes	Depends	Volunteer-dependent	Yes

5.1. Open Research Questions and Future Work

Several promising research directions emerge from this work. Future studies should investigate the development of more intuitive haptic interaction patterns to enhance usability and personalization mechanisms to adapt haptic feedback to individual user preferences and accessibility requirements. The updated application version should enhance the user experience without becoming overwhelming. Moreover, more research should be carried out on how this app should be designed to improve mobility for people with visual impairments.

As for learning and adaptation, there should be research on how users can learn and adapt the application more easily. What strategies can be employed to facilitate rapid learning of the app and how specific haptic feedback patterns influence users' psychological and emotional states? What types of haptic interactions are most effective for conveying emotions, such as urgency, excitement, or comfort? Exploring these questions could lead to valuable insights that contribute to developing more effective, engaging, and user-centric smartphone applications.

In future work, the authors plan to extend their investigation, using a larger visually impaired target group to ensure the results are more secure and representative. There will also be a quantitative evaluation of the application based on the number of errors made by users while trying to identify the location of the detected object. Future development will focus on preparing the application for official app marketplaces, with language customization features. Another future goal is to increase the dataset of this application so that it detects more objects than those of the COCO dataset. As artificial intelligence models become more efficient and lighter, we hope to use new models that are proposed by the scientific community. While the COCO dataset served as a practical baseline for evaluating haptic feedback's usability, a future version of the application will include an enlarged dataset based on the needs of visually impaired users. Transfer learning and data augmentation will be employed to enhance the detection of objects. Additionally, we will validate object size distributions against real usage scenarios to ensure better performance across varying distances.

To further enhance accessibility and user experience, we plan to develop an adaptive vibration system that changes haptic feedback. The users will adjust vibration strength (amplitude) via a slider in the app settings. Multiple vibration patterns (e.g., single pulses, rhythmic sequences) should be available for user selection. The user should also be able to change vibration frequency and pulse duration based on their preferences.

The generalization of our findings is limited due to practical constraints by the small blind participant sample size. While this was necessary for initial haptic feedback usability validation, follow-up studies should address this constraint by using a larger sample size of blind users.

The absence of quantitative performance metrics (e.g., task completion time and error rates) represents a limitation, as these could complement subjective usability data. Future work will adopt a hybrid evaluation framework to address this gap.

5.2. Innovations and Contributions

This study introduces several innovations in the field of assistive technologies for visually impaired users. First, we present a novel integration of real-time AI object detection with haptic feedback, creating a multimodal interface that combines tactile and auditory feedback to enhance spatial awareness and environmental interaction. Our application enables users to physically "explore" detected objects through touch-triggered vibrations, offering a more intuitive and immersive experience. Also, we conducted a comparative usability and user experience study evaluating the impact of combining haptic feedback with audio interfaces on blind users' environment exploration. The evaluation results demonstrate that multimodal interaction enriched with haptic feedback provides better results compared to audio-only interfaces. These contributions improve assistive technology by integrating haptic feedback in multimodal interaction for visually impaired users, especially for spatial awareness and user autonomy. Furthermore, in our evaluation, we combined two questionnaires—assessing usability and user experience—to conduct a comprehensive analysis not only of the application's performance but, more critically, of how haptic interaction improves human–computer interaction for visually impaired users.

6. Conclusions

The haptic application presented in this article can improve the mobility of visually impaired people. Traditionally, people with visual impairments explored their environment either with their hands or with a cane. As technology and artificial intelligence evolve, we are discovering new tools that can help guide people with visual impairments.

We created a haptic smartphone application, which allows users to detect and explore their surroundings through haptic and auditory feedback. We assessed and evaluated the use of haptic feedback for its usability. We found that the use of haptic feedback increased the efficiency of and users' satisfaction with the application.

Author Contributions: Conceptualization, G.V. and N.T.; methodology, G.K.; software, I.D.; validation, G.V.; formal analysis, G.V.; investigation, G.V. and I.D.; resources, G.V. and N.T.; data curation, G.V.; writing—original draft preparation, G.V.; writing—review and editing S.K.; visualization, G.K.; supervision, G.K.; project administration, G.K.; funding acquisition, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We appreciate the dedicated individuals who participated in the evaluation phase of the smartphone application, especially in the "Association of the Blind and Visually Impaired of Western Macedonia, Greece". Your commitment to helping us refine and improve our application has been invaluable. The time and effort you invested in exploring and evaluating the application are deeply appreciated.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Questionnaire used for evaluation.

Question	Available Answers				
	Demographics				
1. Age	Under 18	18–30		30–50	50+
2. Gender	Male	Female			Other
3. Do you often use similar applications?	1 (Not Often)				5 (Very Often)
	USE Questionnaire				
	Usefulness				
4. It helps me be more effective	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)
5. It is useful	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)
6. It gives me more control over the activities in my life	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)
	Ease of Use				
7. It is easy to use	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)
8. It is user friendly	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)
9. I can recover from mistakes quickly and easily	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)
	Ease of Learning				
10. I learned to use it quickly	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)
11. I easily remember how to use it	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)
12. It is easy to learn to use it	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)
	Satisfaction				
13. I feel I need to have it	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)
14. I would recommend it to a friend	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)
15. It is fun to use	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)
	UEQ (short version)				
	Is the Application:				
16. Obstructive	1 2 3 4 5 6 7				Supportive
17. Complicated	1 2 3 4 5 6 7				Easy
18. Inefficient	1 2 3 4 5 6 7				Efficient
19. Clear	1 2 3 4 5 6 7				Confusing
20. Boring	1 2 3 4 5 6 7				Exciting
21. Not interesting	1 2 3 4 5 6 7				Interesting
22. Conventional	1 2 3 4 5 6 7				Inventive
23. Usual	1 2 3 4 5 6 7				Leading edge
	Comparison Questionnaire Between The Two Versions				
24. The Vibration Pattern was pleasant	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)
25. Which version did you find more pleasant: with or without Haptic Feedback?	With Haptic Feedback	Without Haptic Feedback		None of them	Both of them
26. Which version do you consider more effective: with or without Haptic Feedback?	With Haptic Feedback	Without Haptic Feedback		None of them	Both of them
27. Which version do you consider easier to use: with or without Haptic Feedback?	With Haptic Feedback	Without Haptic Feedback		None of them	Both of them
28. Which version do you consider easier to learn: with or without Haptic Feedback?	With Haptic Feedback	Without Haptic Feedback		None of them	Both of them
29. Which version would you recommend to a friend: with or without Haptic Feedback?	With Haptic Feedback	Without Haptic Feedback		None of them	Both of them
	Open Response				
30. What improvements would you suggest to make the application more useful?	[Open text box]				

References

1. Wang, J.; Wang, S.; Zhang, Y. Artificial intelligence for visually impaired. *Displays* **2023**, *77*, 102391. [CrossRef]
2. Okolo, G.; Althobaiti, T.; Ramzan, N. Assistive Systems for Visually Impaired Persons: Challenges and Opportunities for Navigation Assistance. *Sensors* **2024**, *24*, 3572. [CrossRef] [PubMed]
3. Surougi, H.; Zhao, C.; McCann, J. ARAware: Assisting Visually Impaired People with Real-Time Critical Moving Object Identification. *Sensors* **2024**, *24*, 4282. [CrossRef] [PubMed]
4. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
5. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Springer International Publishing: New York, NY, USA, 2014.
6. Jiang, C.; Kuang, E.; Fan, M. How Can Haptic Feedback Assist People with Blind and Low Vision (BLV): A Systematic Literature Review. *ACM Trans. Access. Comput.* **2025**, *18*, 1–57. [CrossRef]
7. Kuriakose, B.; Shrestha, R.; Sandnes, F.E. Multimodal navigation systems for users with visual impairments—A review and analysis. *Multimodal Technol. Interact.* **2020**, *4*, 73. [CrossRef]
8. Kuriakose, B.; Shrestha, R.; Sandnes, F.E. DeepNAVI: A deep learning based smartphone navigation assistant for people with visual impairments. *Expert Syst. Appl.* **2023**, *212*, 118720. [CrossRef]
9. Rakkolainen, I.; Farooq, A.; Kangas, J.; Hakulinen, J.; Rantala, J.; Turunen, M.; Raisamo, R. Technologies for multimodal interaction in extended reality—A scoping review. *Multimodal Technol. Interact.* **2021**, *5*, 81. [CrossRef]
10. Jiménez, M.F.; Mello, R.C.; Bastos, T.; Frizera, A. Assistive locomotion device with haptic feedback for guiding visually impaired people. *Med. Eng. Phys.* **2020**, *80*, 18–25. [CrossRef]
11. See, A.R.; Costillas, L.V.M.; Advincula, W.D.C.; Bugtai, N.T. Haptic Feedback to Detect Obstacles in Multiple Regions for Visually Impaired and Blind People. *Sens. Mater.* **2021**, *33*, 1799–1808. [CrossRef]
12. Palani, H.P.; Fink, P.D.; Giudice, N.A. Comparing map learning between touchscreen-based visual and haptic displays: A behavioral evaluation with blind and sighted users. *Multimodal Technol. Interact.* **2021**, *6*, 1. [CrossRef]
13. Raynal, M.; Ducasse, J.; Macé, M.J.M.; Oriola, B.; Jouffrais, C. The FlexiBoard: Tangible and Tactile Graphics for People with Vision Impairments. *Multimodal Technol. Interact.* **2024**, *8*, 17. [CrossRef]
14. Tzimos, N.; Voutsakelis, G.; Kontogiannis, S.; Kokkonis, G. Evaluation of Haptic Textures for Tangible Interfaces for the Tactile Internet. *Electronics* **2024**, *13*, 3775. [CrossRef]
15. Billi, M.; Burzagli, L.; Catarci, T.; Santucci, G.; Bertini, E.; Gabbanini, F.; Palchetti, E. A unified methodology for the evaluation of accessibility and usability of mobile applications. *Univers. Access Inf. Soc.* **2010**, *9*, 337–356. [CrossRef]
16. Ballantyne, M.; Jha, A.; Jacobsen, A.; Hawker, J.S.; El-Glaly, Y.N. Study of accessibility guidelines of mobile applications. In Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia, Cairo, Egypt, 25–28 November 2018; Association for Computing Machinery: New York, NY, USA, 2018.
17. Park, E.; Han, S.; Bae, H.; Kim, R.; Lee, S.; Lim, D.; Lim, H. Development of automatic evaluation tool for mobile accessibility for android application. In Proceedings of the IEEE International Conference on Systems of Collaboration Big Data, Internet of Things Security (SysCoBIoTS), Casablanca, Morocco, 12–13 December 2019; IEEE: New York, NY, USA, 2019.
18. Acosta-Vargas, P.; Salvador-Ullauri, L.; Jadán-Guerrero, J.; Guevara, C.; Sanchez-Gordon, S.; Calle-Jimenez, T.; Lara-Alvarez, P.; Medina, A. Accessibility assessment in mobile applications for android. In *Advances in Human Factors and Systems Interaction*; Springer: Cham, Switzerland, 2020; pp. 279–288.
19. Mateus, D.A.; Silva, C.A.; Eler, M.M.; Freire, A.P. Accessibility of mobile applications: Evaluation by users with visual impairment and by automated tools. In Proceedings of the 19th Brazilian Symposium on Human Factors in Computing Systems, Diamantina, Brazil, 26–30 October 2020; Association for Computing Machinery: New York, NY, USA, 2020.
20. Acosta-Vargas, P.; Salvador-Acosta, B.; Salvador-Ullauri, L.; Villegas-Ch, W.; Gonzalez, M. Accessibility in native mobile applications for users with disabilities: A scoping review. *Appl. Sci.* **2021**, *11*, 5707. [CrossRef]
21. Zheng, Q.; Tian, X.; Yu, Z.; Elhanashi, Y.M.A.; Saponara, S. Robust automatic modulation classification using asymmetric trilinear attention net with noisy activation function. *Eng. Appl. Artif. Intell.* **2025**, *141*, 1098. [CrossRef]
22. Liu, F.; Zheng, Q.; Shu, T.X.F.; Jiang, W.; Wang, M.; Saponara, S. Rethinking the multi-scale feature hierarchy in object detection transformer (DETR). *Appl. Soft Comput.* **2025**, *175*, 113081. [CrossRef]
23. Alqahtani, D.K.; Cheema, M.A.; Toosi, A.N. Benchmarking deep learning models for object detection on edge computing devices. In Proceedings of the International Conference on Service-Oriented Computing, Tunis, Tunisia, 3–6 December 2024; Springer: Singapore, 2024.
24. Team, T.L. Easier Object Detection on Mobile with TF Lite. 2021. Available online: https://blog.tensorflow.org/2021/06/easier-object-detection-on-mobile-with-tf-lite.html (accessed on 1 May 2025).

25. Kee, E.; Chong, J.J.; Choong, Z.J.; Lau, M. Development of Smart and Lean Pick-and-Place System Using EfficientDet-Lite for Custom Dataset. *Appl. Sci.* **2023**, *13*, 11131. [CrossRef]
26. TensorFlow. Demo TensorFlow Lite Object Detection Android, GitHub. Available online: https://github.com/tensorflow/examples/blob/master/lite/examples/object_detection/android/README.md (accessed on 1 May 2025).
27. Kokkonis, G.; Minopoulos, G.; Psannis, K.; Ishibashi, Y. Evaluating Vibration Patterns in HTML5 for Smartphone Haptic Applications. In Proceedings of the 2nd World Symposium on Communication Engineering (WSCE), Nagoya, Japan, 20–23 December 2019; IEEE: New York, NY, USA, 2019.
28. Kentridge, R.W.; Heywood, C.A.; Weiskrantz, L. Attention without awareness in blindsight. *Proc. R. Soc. B Biol. Sci.* **1999**, *266*, 1805–1811. [CrossRef]
29. Gibson, J. The perception of the visual world. *Houghton Mifflin*. 1950. Available online: https://ia600805.us.archive.org/12/items/perceptionofvisu00jame/perceptionofvisu00jame.pdf (accessed on 1 May 2025).
30. Lund, A.M. Measuring usability with the use questionnaire. *Usability Interface* **2001**, *8*, 3–6.
31. Schrepp, M.; Hinderks, A.; Thomaschewski, J. Construction of a benchmark for the user experience questionnaire (UEQ). *Int. J. Interact. Multimed. Artif. Intell.* **2017**, *4*, 40–44. [CrossRef]
32. Microsoft. Seeing AI. 2023. Available online: https://www.seeingai.com/ (accessed on 1 May 2025).
33. Envision. Envision AI—Assistive Technology for the Blind and Visually Impaired. Available online: https://www.letsenvision.com/ (accessed on 1 May 2025).
34. Eyes, B.M. Be My Eyes—See the World Together. Available online: https://www.bemyeyes.com/ (accessed on 1 May 2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

The Innovative Design and Performance Testing of a Mobile Robot for the Automated Installation of Spacers on Six-Split Transmission Lines †

Jie Pan [1,2,3,*], Yongfeng Cheng [1], Chunhua Hu [1], Ming Jiang [1], Yong Ma [1], Fanhao Meng [1] and Qiang Shi [1]

[1] State Grid Electric Power Engineering Research Institute Co., Ltd., Beijing 100069, China; chengyongfeng@ei.sgcc.com.cn (Y.C.); huchunhua@ei.sgcc.com.cn (C.H.); jiangmings@163.com (M.J.); ustb_peter@126.com (Y.M.); meng87123@126.com (F.M.); 18733506855@163.com (Q.S.)
[2] China Electric Power Research Institute, Beijing 100055, China
[3] School of Mechanical Engineering and Automation, Beihang University, Beijing 100191, China
* Correspondence: panjie@buaa.edu.cn
† This is an extended version of our paper published in Pan, J., Hu, C., Jiang, M., Ma, Y., Meng, F., He, X., Liu, C. Development of a Spacer Bar Installing Robot for Six-Split Transmission Lines. In Proceedings of the 8th International Conference on Computing, Control and Industrial Engineering (CCIE2024), Singapore, 2024; pp. 11–17.

Abstract: The spacer is an important component of a transmission line and can effectively prevent wires from whipping each other and inhibit vibration. Given the complex installation conditions of multi-split lines, the installation of spacers is mainly achieved through manual work, which has the disadvantage of heavy labor intensity and a high risk factor. The robots that install two-split and four-split spacer bars cannot be applied to the complex operating conditions of six-split transmission lines. In order to improve the installation efficiency of spacers and reduce operating costs and risks, a new type of spacer-installing robot was researched based on the six-split transmission lines in this paper. Through the theoretical analysis of the wire's arc sag, the moving device of the robot was designed. In order to improve the operating efficiency of the robot, the storage and feeding device of the six spacers was designed. A planar arm with the ability to assemble the spacer was designed. The overall design of the robot was completed by integrating the design of each unit. Through the experimental test, the results indicated that the robot was capable of installing six spacers at once, the maximum moving slope was 15 degrees, and the error rate in the spacer installation was 2.33%, which matched the manual installation of the spacers. The robot provided new ideas for the design of new transmission line engineering equipment and expanded the scope of the application of robots in the power industry.

Keywords: spacer; installation; multi-split lines; robot; design

1. Introduction

Most high-voltage transmission lines are located in the field, which is frequently affected by the wind and snow, so it is a very important task to ensure the safety and stable operation of these transmission lines. The spacer is an important component of a transmission line and can effectively prevent wires from whipping each other, inhibit breeze vibration, and sub-file distance oscillation [1,2]. At present, the installation of spacers on high-voltage transmission lines mainly relies on workers carrying spacers to walk on overhead wires to achieve their installation, which poses high risks and inefficiency. Nevertheless, significant progress has been made in the study of high-voltage transmission line inspection robots [3,4], such as using the Unmanned Aerial Vehicle (UAV) for the

inspection of transmission lines [5,6], using wheeled robots in wire inspection [7], etc. This equipment can be adopted to replace the manual labor of workers to achieve automated routing inspection, but research into multi-split spacer-installing robots is still relatively limited [8–11]. At present, this field still relies on workers moving carefully on wires (as shown in Figure 1) to achieve spacer installation and even using helicopters to support workers in their installation of spacers, which requires more workers and equipment in the process. In the actual spacer installation, workers have to climb a transmission tower to the overhead wire position and then take a few spacers while moving slowly on the wires; this process is laborious and dangerous. When the workers reach the installation location, they can use their tools to install the spacers. It should be noted that the greater the number of transmission line splits, the more difficult the installation is. Currently, the spacer installation method has high risks, inefficiencies, and other shortcomings. In addition, workers are affected by high winds, rain, and snow during their work, which creates challenges to power construction. The current research and development of transmission line robots is focused on transmission line inspection [12–16]; it is still difficult for these robots to automatically install multi-split spacers.



Figure 1. Workers installing spacers.

With the application of robotics in the field of the power industry increasing, there have been two-split spacer- [17] and four-split spacer-installing robot [18] research reports. However, with the increase in the amount of wire splitting, spacer assembly difficulty has also increased [19–23]. For robots on two-split and four-split wires, the design for moving and installing devices is simpler, in addition to their reduced weight and size, leading to a more straightforward mechanical and control system design. This is due to the reduced number of wire splits, which facilitates easier adjustment in the wire spacing. But six-split lines have more spacer chucks and non-uniform spacing between wires, making the development of spacer-installing robots more challenging [24–28]. The high-voltage transmission lines are approximately 300 m~500 m in length, while the existing four-legged and climbing robots are difficult to adapt to the long-distance environment of the wires [29–31]. Meanwhile, UAVs are not yet able to automatically install multi-split spacers and can only achieve transport and inspection functions. As the demand for electrical energy increases, the scale of electrical construction will continue to expand; it is an important and difficult task to install spacers. Thus, there is an urgent need to develop a new type of construction equipment to improve the installation of spacers. This new equipment is needed to grant the ability to move accurately and to efficiently install spacers on the wires.

Based on the above analysis, there is a new spacer-installing robot for the six-split transmission lines, designed in this paper. This paper is focused on analyzing the effect of robot weight on the shape of the wire, the effect of the surface geometry of the wheels on the ability to move, the analysis of the motor control logic inside the robot, and the performance test of moving and installing spacer bars on real six-split transmission lines. The structure of this paper is divided into seven parts. The first part is the background and significance of the study, the second part is about the relevant research content of this study, the third part is the theoretical modeling and analysis process, the fourth part is the mechanism design method and process of the robot, the fifth part is the overall design of

the robot, the sixth part is the test and discussion, and the seventh part is the conclusion and outlook.

2. Related Research

Previous robot designs primarily focused on achieving movement along transmission lines, with limited research dedicated to operational functionality. This is because the robot must perform both movement and operational functionality on the six-split transmission lines, which requires stringent demands on its mechanism design, motion control, and workflow design. For instance, the robot's movement along the wire will cause deformation of the wire; thus, the theoretical analyses need to be made on the effect of gravity during movement. The encoder is a common position sensor for robots, but the spacer-installing robot requires precise positioning, so additional positioning sensors need to be added to the robot. Based on the need for mechanized construction, we have studied a six-split spacer installation robot, which was capable of slowly installing spacers under remote control [32]. However, critical aspects including design theory validation, mobility characterization, and automated installation performance evaluation remained uninvestigated. The actual capability of the robot was very limited, and it is not yet possible to apply it in real engineering.

It is worth noting that the robot needs to overcome the influence of the slope of the wire while moving, so it needs an in-depth theoretical study of the impact of the robot on the wire. Secondly, the mobility of the robot must be a prerequisite to ensure the operation, and further research on the mechanism design and motion control is needed. Finally, the performance test of the robot in practical engineering is an important aspect of presenting the real ability of the robot. Based on the above analysis, it is necessary to deepen further the theoretical modeling and mechanism design research work based on the previous research.

3. Modeling

3.1. The Impact of the Robot on the Wire

Due to the long distance of the transmission line and the complex moving environment, the robot made by the simple mechanical structure form is not qualified. The wheeled robot is capable of good mobility on the wire; therefore, the design of the mobile form of the robot will be wheeled. When the robot is moving on the wire, the wire is affected by the weight of the robot, which leads to a change in the radius of the wire, as shown in Figure 2. Therefore, it is necessary to make a theoretical analysis of the slope in the robot's moving direction on the wire.

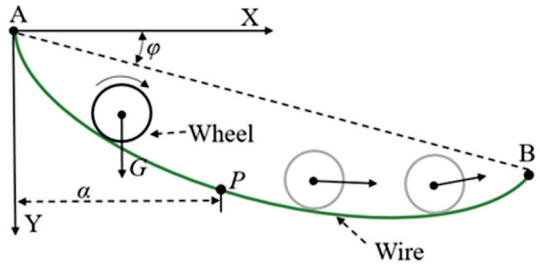


Figure 2. Effect of concentrated weight on the wire.

The shape of the suspension wire of steel-core aluminum stranded wire can be obtained from the equation of the suspension wire. Although the wire has a certain degree of stiffness, with the file distance spanning larger factors, the impact of gravity on the shape of the wire is much greater than the impact of stiffness, so the impact of stiffness on the shape of the wire can be ignored.

For the equal density wire, defining point A as the coordinate origin, according to the catenary equation [33], the value of the vertical coordinate y at point P is:

$$y = \frac{H}{\omega}\left[\mathrm{ch}\frac{l\omega}{2H} - \mathrm{ch}\frac{(l-2x)\omega}{2H}\right] \tag{1}$$

The x is the horizontal coordinate at point P, and the ch is the hyperbolic cosine function. The l is the length of the wire, H is the horizontal tension in the wire, and ω is the self-gravity force in unit length of the wire. $\sigma = H/s$, $g = \omega/s$, σ is the horizontal stress of the line; g is the ratio of gravity to the load of the line, and s is the cross-sectional area of the wire.

Therefore, Equation (1) can be expressed as:

$$y = \frac{\sigma}{g}\left[\mathrm{ch}\frac{lg}{2\sigma} - \mathrm{ch}\frac{(l-2x)g}{2\sigma}\right] \tag{2}$$

When there is a concentrated load on the wire, the value of the longitudinal coordinates of any point on the wire can be calculated as:

$$y = \left(\frac{l\omega}{2H\cos\varphi} + \frac{l-a}{l}\cdot\frac{Q}{H} \pm \frac{h}{l}\right)x - \frac{\omega}{2H\cos\varphi}x^2 \tag{3}$$

Q is the gravity of the concentrated load. φ is the height difference angle of the wire at the suspension points at both ends. a is the horizontal distance from point P to point A. When the concentrated load is located in the middle of the line, the arc sag value y_{max} reaches its maximum and can be expressed in the following form:

$$y_{\max} = \frac{l^2}{8H}\left(\frac{\omega}{\cos\varphi} + 2\frac{Q}{l}\right) \tag{4}$$

At this point, the value of moving angle θ reaches its maximum and satisfies the equation relationship:

$$\frac{l}{2}\tan\theta = y_{\max} \tag{5}$$

The maximum moving angle θ_{max} can be obtained by calculating Equations (4) and (5):

$$\theta_{\max} = \arctan\left[\frac{l}{4H}\left(\frac{\omega}{\cos\varphi} + 2\frac{Q}{l}\right)\right] \tag{6}$$

If the ends of the wire are equal, the value of φ is 0.

$$\theta_{\max} = \arctan\left[\frac{l}{4H}\left(\omega + 2\frac{Q}{l}\right)\right] \tag{7}$$

From Equation (7), it can be calculated that the maximum moving angle θ_{max} of the robot on the wire is mainly related to l, H, Q, and ω; the θ_{max} is inversely proportional to the value of H. Since the length of the wire l and ω are constant values after the wire has been built, the increasing weight of the robot will lead to a larger moving slope. Furthermore, the value of l is not consistent due to the terrain, but the larger the value of l, the larger the value of θ_{max}. Among the many factors, only Q can be adjusted by mechanical design. Based on this, a lightweight mechanism should be considered for the robot to enhance its moving performance.

3.2. Forward Analysis of Robot

The moving mechanism of the robot is composed of active and follower wheels, and the wheels are driven by DC motors. As shown in Figure 3, the driving torques of the traveling wheels are set to be M_1 and M_2, the positive pressures generated by the wire on the wheels are N_1 and N_2, the friction generated by the two wheels on the wire are F_1 and F_2, the weight of the robot is mg, and the rate of moving on the wire is v.

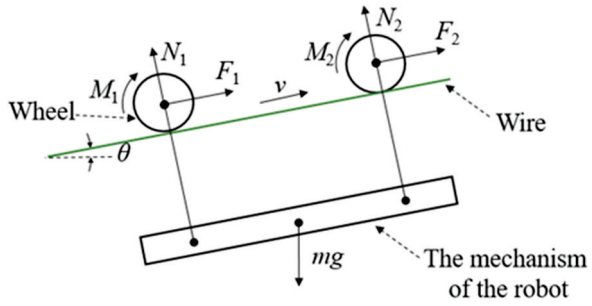


Figure 3. Force analysis of the moving mechanism.

Along the direction of gravity is the following equation:

$$N_1 + N_2 - mg\cos\theta = 0 \tag{8}$$

Along the direction parallel to the wire is the following equation:

$$F_1 + F_2 - mg\sin\theta = 0 \tag{9}$$

θ represents the angle of the wire relative to the horizontal ground.

The moment balance equation is established based on the center of the wheel as follows:

$$M_1 - M_f^1 - F_1 r = 0 \tag{10}$$

$$M_2 - M_f^2 - F_2 r = 0 \tag{11}$$

$$F_i = \mu N_i \tag{12}$$

$$M_f^i = N_i \delta \tag{13}$$

The values of i are 1 to 2, r is the radius of the driving wheel, μ is the coefficient of rolling friction between the wheel and the wire, and δ is the coefficient of rolling resistance between the wheel and the wire.

Since both wheels are made of the same material, the coefficient of rolling friction and the coefficient of rolling resistance between the wheels and the wire are the same. Associating Equations (8)–(13), the following can be calculated:

$$\begin{cases} M_1 + M_2 = rmg\sin\theta \\ \mu(N_1 + N_2) = mg\sin\theta \end{cases} \tag{14}$$

From Equation (14), the main factors affecting the movement of the robot on the wire are the friction coefficient μ, the rolling resistance coefficient δ, and the robot weight mg. To improve the moving ability of the robot, the friction between the robot wheels and the wire should be increased. Therefore, the moving ability of the robot is increased by adding a pressing wheel to the driving wheel of the robot, which increases the positive pressure of the wire on the driving wheel and thus increases the moving ability of the robot.

As shown in Figure 4, the positive pressures generated by the wire on the pressing wheels are N_3 and N_4. The F_3 and F_4 are the rolling friction forces generated by the wire on the surface of the pressing wheel, which drive the rotation of the pressing wheel. The values of F_3 and F_4 are controlled by the rolling resistance screw adjusting the distance between the active and pressing wheels.

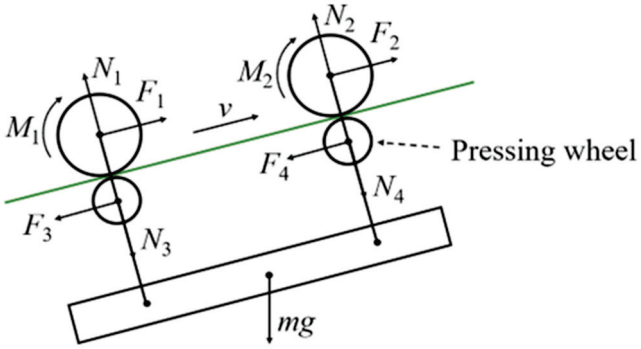


Figure 4. The force analysis of the moving mechanism under the action of the pressing wheel.

With added pressing wheels, the force analysis in the direction of gravity is:

$$N_1 + N_2 - N_3 - N_4 - mg\cos\theta = 0 \tag{15}$$

With the addition of compression sheave, the force analysis in the direction of the parallel wire is:

$$F_1 + F_2 - F_3 - F_4 - mg\sin\theta = 0 \tag{16}$$

N_i (i = 1~4) is the positive pressure from the wire applied to the four wheels; F_i (i = 1~4) is the friction force from the wire applied to the four wheels and $F_\mathrm{i} = \mu N_i$; θ is the angle between the wire and the horizontal ground; and m is the weight of the robot.

Thus, the condition for stable motion of the robot on the wire is:

$$F_1 + F_2 \geq mg\sin\theta + F_3 + F_4 \tag{17}$$

It can be calculated by bringing Equation (12) into Equation (17):

$$\mu(N_1 + N_2) \geq mg\sin\theta + F_3 + F_4 \tag{18}$$

From Equation (18), it can be seen that the positive pressure of the robot during its movement is not only related to the weight of the robot but also to the pressure exerted by the pressing wheels on the wire during the movement. Considering the possibility that the robot on the wire may slip in movement, it is necessary to design mechanism devices for the robot that can actively adjust the rolling friction generated by the active wheels on the wire; thus, the movement of the robot can be controlled by adjusting the force generated by the pressing wheels.

3.3. Wheel Cross-Section Analysis

Wheels are the important components for the robot to achieve mobility on the wires, and the design of the wheels has a significant impact on the mobility of the robot. Since the transmission line is located in the sky, the wheel is designed as a groove structure to increase the safety of the robot. The cross-sectional shape of the wheel is analyzed below. The cross-sectional shape of the wheel can be mainly classified as trapezoidal and curved, as shown in Figures 5 and 6. The friction generated by the two forms of wheels as they move on the wire is analyzed as below.

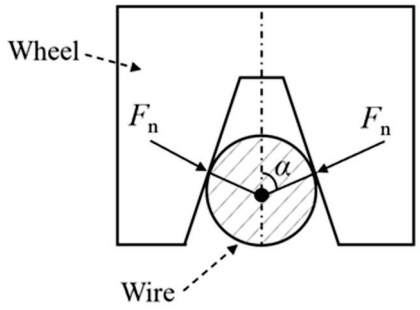


Figure 5. Force analysis of the wheel with a trapezoidal cross-section.

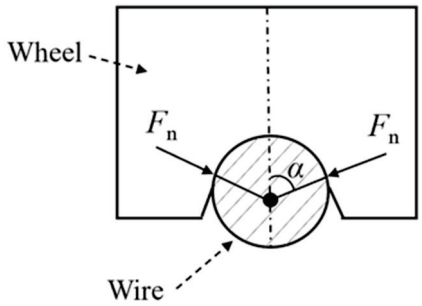


Figure 6. Force analysis of the wheel with a curved section.

Assuming that the wheel and wire are not deformed, the force relationship between the wheel and the wire is:

$$2F_n \cos \alpha_1 = W_1 \tag{19}$$

F_{n} is the pressure of the wheel on the surface of the wire and W_1 is the vertical downward force of F_{n}.

$$F_{f1} = 2\mu F_n \tag{20}$$

μ is the coefficient of static friction of the wheel surface on the wire. $F_{f\ 1}$ is the static friction.

Solving Equations (19) and (20), the following can be calculated:

$$F_{f1} = \frac{W_1 \mu}{\cos \alpha_1} \tag{21}$$

When the wheel is of a curved section, the contact type between the wheel and wire is the line contact. The positive pressure on the line should be calculated by using the mathematical integral.

$$2\int_0^{\alpha_2} F_{n2} \cos \alpha d\alpha = W_2 \tag{22}$$

F_{n2} is the pressure of the wheel on the surface of the wire, and W_2 is the vertical downward force of F_{n2}.

$$2F_{\mathrm{n2}} \sin \alpha_2 = W_2 \tag{23}$$

$$F_{\mathrm{n2}} = \frac{W_2}{2 \sin \alpha_2} \tag{24}$$

$$F_{f2} = 2\int_0^{\alpha_2} F_{\mathrm{n2}} \mu d\alpha \tag{25}$$

Calculating Equations (24) and (25), the following can be obtained:

$$F_{f2} = \frac{\alpha_2}{\sin \alpha_2} W_2 \mu \tag{26}$$

As the driving wheel load is the same, the following exists:

$$W_1 = W_2 \tag{27}$$

Calculating formulas (21), (26), and (27), the ratio k of the two friction forces is obtained as:

$$k = \frac{F_{f1}}{F_{f2}} = \frac{\sin \alpha_2}{\alpha_2 \cos \alpha_1} \tag{28}$$

If the boundary contact point of the wheel forms the same angle as the center of the wire circle, the following exists:

$$\alpha_1 = \alpha_2 \tag{29}$$

Therefore, Equation (28) can be expressed as:

$$k = \frac{\tan \alpha}{\alpha} \tag{30}$$

Therefore, under the condition that the weight of the robot and the material of the wheel are the same, the ratio of the driving force provided by the two types of driving wheels with trapezoidal and curved interfaces is determined by the contact angle α. Since $\tan \alpha > \alpha$, $k > 1$, and it can be concluded that the driving force provided by the driving wheel with a trapezoidal cross-section is greater than that of the driving wheel with an arc interface.

4. Mechanical Design

4.1. Subsection

Based on the analyses in Section 3.3, the forward wheel of the robot is designed as having a trapezoidal cross-section, and a pressing wheel is designed under the forward wheel to change the friction value between the forward wheels and the surface of the wires. By controlling the friction between the wheels separately, the forward and turning movements of the robot are controlled. Since the spacers are standard hexagonal shapes with fixed spacing values, the spacing of the six wires needs to be arranged into hexagonal shapes before installing the spacers, as shown in Figure 7.

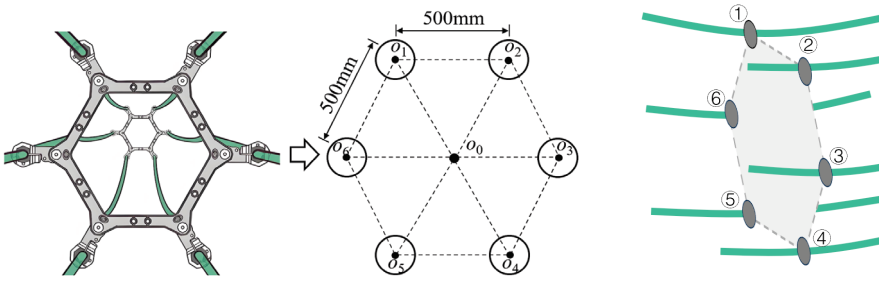


Figure 7. (**a**) The relative position between the clamps of the spacer; (**b**) the arrangement of wheels on the wires.

Based on the above analysis, the spacing of the six wheels in the robot's moving mechanism is designed as a hexagonal shape (The wheel numbers are ①~⑥), which in turn ensures that the spacing of the wires matches the spacing of the ends of the spacers, as shown in Figure 8a,b. The pressing wheels are used to adjust the rolling friction of the active wheel on the wire, thus achieving speed adjustment of the robot, as shown in Figure 8c. The center of mass of the robot will change during the moving process on the wire, and a large offset of the center of mass will cause the robot to be unstable. Therefore, it

is necessary to design follower wheels for the robot around the active wheel to improve the stability of the robot. When the center of gravity of the robot is designed between the active wheel and the follower wheel, the rotational torque generated on the wheels can effectively counterbalance each other, and thus the robot has a more stable movement ability on the wire than on the unfollowed wheels.

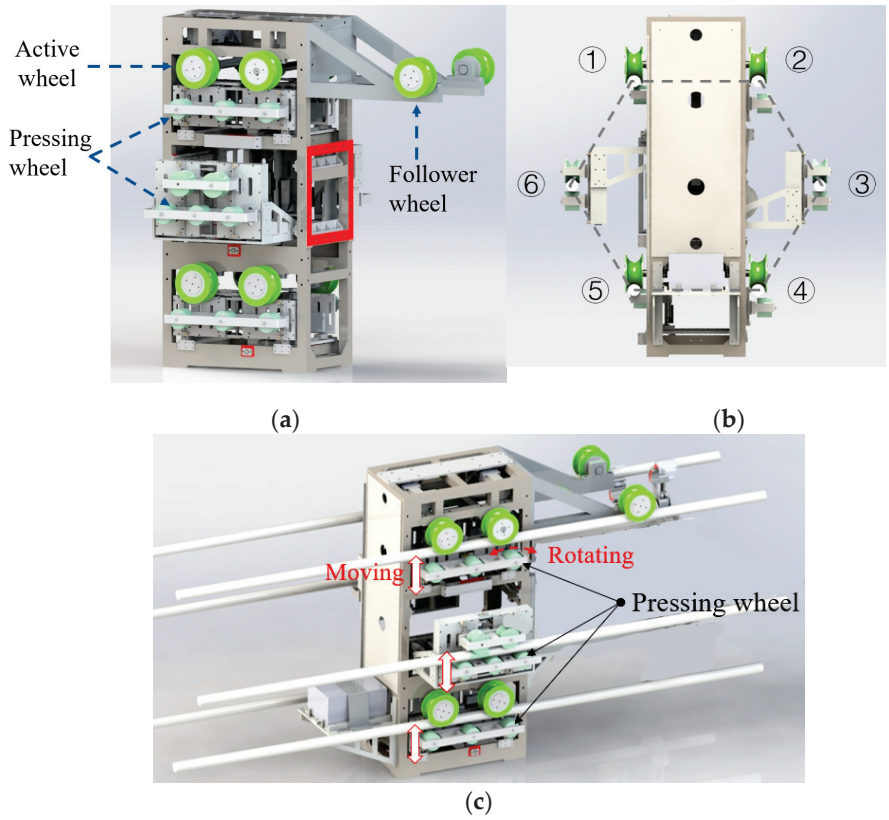


Figure 8. (**a**) The moving mechanism; (**b**) the effect of the six wheels on the wires; (**c**) the movement form of the pressing wheel.

4.2. Transfer Mechanism for Spacer

Besides the mobility function, the robot should also have the ability to store and transfer spacers to increase installation efficiency. Therefore, once the robot arrives at the target location on the wire, the internal transfer mechanism can transfer the spacers from the storage location to the aim location that can be lifted and installed. Considering that two spacers were installed at a spacing of about 60 m in length and the distance between two towers was 300 m~500 m in length, the storage mechanism was designed to be capable of storing six spacers. The storage and transfer mechanisms of the spacers were designed inside the robot, as shown in Figure 9. Firstly, the design of the spacer storage mechanism, which is composed of the upper and lower rails, can accommodate a maximum of six six-split spacers. In the design of the transfer mechanism, the spacer is transferred on the rail by controlling the ball screw, and there is a displacement sensor at the end of the rail, which can measure the transfer status of the spacer in real time. When the spacers are transferred to the end of the guide rail, the lifting device in the transfer mechanism will send the spacers to the pre-installed position on the upper part of the guide rail. The motion speed of the transfer mechanism is 30 mm/s. To prevent jamming or misalignment, the spacer is mounted on the upper and lower sliders of the transferring mechanism and is synchronously moved by the upper and lower guide rails to achieve precise transport.

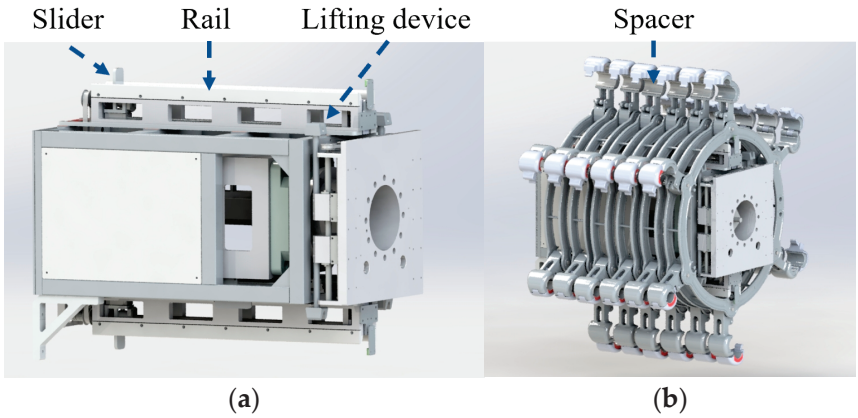


Figure 9. (**a**) Storage and transfer mechanism for spacers; (**b**) installation effect.

4.3. Installation Mechanism

The main task of the robot is to mount the spacer, so it is necessary to research the installation characteristics of the spacer. Currently, the installation of spacers on the wire can be divided into two ways: one is through the bolt constraint, the other is through the snap constraint. The first requires workers to repeatedly rotate the tool to achieve the installation, and the process is more complicated. The latter requires workers to press hard to achieve installation, and the process is relatively simple. As the robot is driven by motors, the second type of installation method is more compatible with the robot.

The installation of the spacer involves two parts: one is to contact the six spacer chucks with the six wires to complete the wires, and the other is to clamp the spacer chucks, so it is necessary to design a lifting mechanism and the installation mechanism for the spacer. The lifting mechanism of the robot mainly consists of a 3-degree-of-freedom planar mechanical arm, controlling the rotation of the arm to move the spacer from the top of the guide rail to the designated position, and then completing the spacer chuck on the wire. To install the spacer easily, the spacer chuck is of a snap type (Figure 10), clamping the chuck of the spacer to complete its assembly. The specific process is that the chuck drives the latch to rotate under the action of force F, and under the drive of the spring on one side of the latch, the latch can lock the chuck. The clamping mechanism of the spacer is mainly made of a 1-degree-of-freedom C-shaped gripper (Figure 11), the gripper on the head of the spacer is controlled to achieve the installation of the spacer.

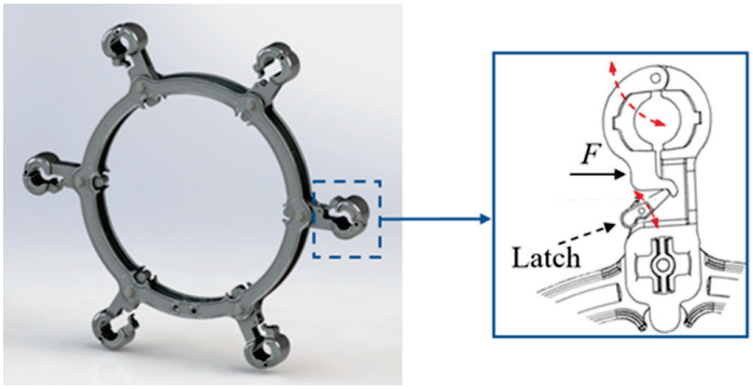


Figure 10. The chuck design of the spacer.

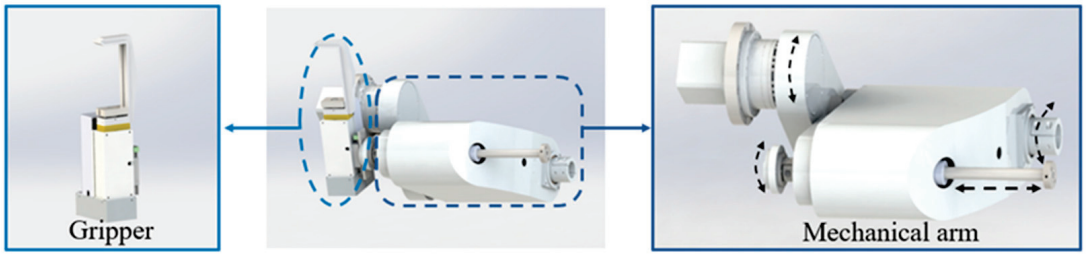


Figure 11. Design of spacer installation mechanism.

4.4. Control System

Besides the sensing signals, the most frequent communication signals between the controller and the base station are motion control signals. In the various mechanisms of the robot, all mechanism movements are driven by stepper motors. Therefore, precise control of the motion mechanism can be achieved by the precise driving of the motors by the robot. Since the robot is required to perform a variety of functions, the motors are controlled in a closed loop rather than an open loop, which can protect the lifetime of the motors and the transmission lines. The closed-loop control of existing motors is mainly achieved by encoder and current feedback signals. The encoder is used to provide feedback on speed and position information, while the current value is used to provide feedback on torque to prevent motor jamming. Except for the encoder, a Real Time Kinematic (RTK) was installed inside the robot to enhance its movement accuracy. The calculation of the robot's spatial position on wires is based on the fusion of the encoder and RTK information.

The primary motion control of the robot is the control of the motors, which contain the forward wheel drive motors and the gripper drive motor. Motor control involves coupled control of both displacement and output force, thus achieving force–position coupling. The displacement is controlled by the encoder feedback data to control the motion displacement of the robot and the output displacement of the gripper. The force control is by the current feedback data from the motor to control the output torque of the motor. The entire control system of the motor is shown in Figure 12.

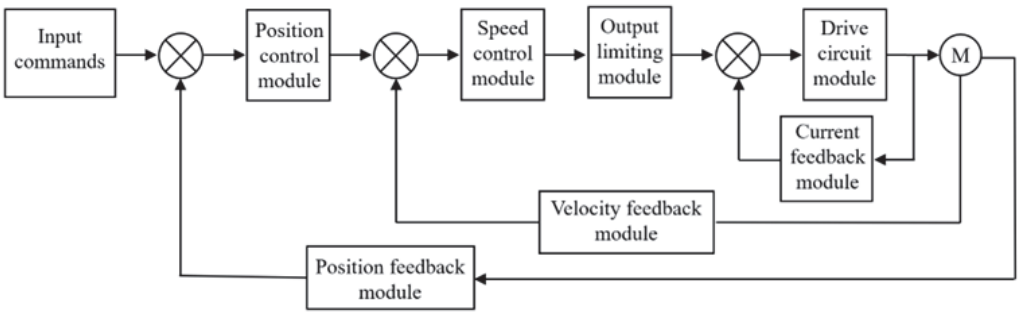


Figure 12. Control system of the motor.

5. Integration and Assembly

5.1. Integrated Design

Based on the above study of the robot's mechanism, each component was fabricated individually, and the prototype was assembled according to the 3D model. During the integrated design process, factors such as force distribution and operational workspace were carefully considered to prevent motion interference. In addition, since the robot is to be lifted onto the wire, the top of the robot should be designed to allow it to be lifted. Within the limits of structural rigidity, the parts of the robot should be small and lightweight. The parts should be bolted together to ensure the overall structural stability of the robot.

The robot contains the moving mechanism, spacer storage and transfer mechanism, and installation mechanism. In terms of the control system, the robot contains the motion control of the moving mechanism, the spacer transfer control, and the spacer assembly control. Combined with the component structure diagram of the robot and the existing machining process, the robot mechanism can be completed by machining and prototype assembly, as shown in Figure 13. The moving mechanism is mainly used to drive the robot to move and stop, and another role is to form the six split lines into regular hexagons to facilitate the installation of spacers. The storage and transfer mechanism was located inside the robot for the storage of six spacers and the transferring capability. The installation mechanism was designed at one end of the transfer mechanism and was driven by a flat mechanical arm to assemble the spacers.

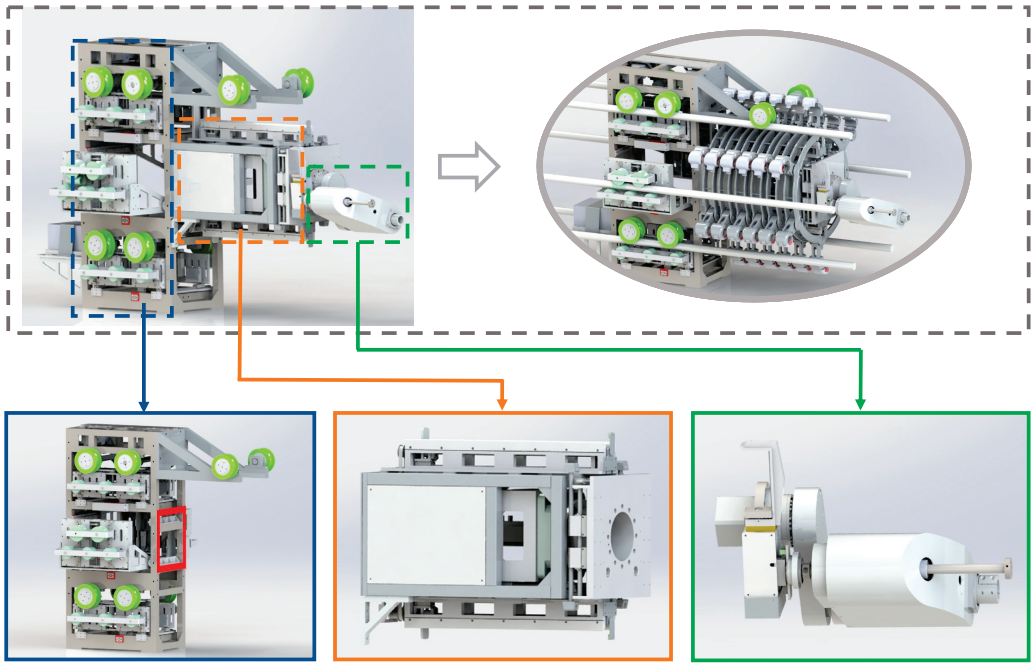


Figure 13. Integrated design of the robot.

5.2. Prototype Assembly

Based on the analyses made in the previous section, parts are machined separately for each part of the robot's mechanism, which used 3D printing, wire cutting, and machine tool machining. High-strength 304 stainless steel material was used for the main frame of the robot to ensure its structural rigidity. Based on the previous 3D drawings of each part of the robot for component assembly, the robot prototype and the composition of each part are shown in Figure 14. The robot contains a communication module, human–machine interaction, power supply module, advancement mechanism, storage and transfer mechanism, spacers, upper wire mechanism, spacer mounting mechanism, and so on. The robot prototype follows the relationship of each module specified in Section 4.1 during assembly.



Figure 14. Robot prototype and components.

5.3. Spacer Installation Process Design

In the process of installing spacers, workers can make micro-adjustments to the installation position according to actual needs. Unlike installation by workers, the installation of spacers by robots is performed strictly in accordance with the installation procedure, so it is necessary to design a correct installation procedure for the robots. During the installation of the spacer, workers would first wrap the chuck of the spacer around the surface of the wire and then use the installation tools to complete assembly. Therefore, the design of the installation mechanism for the robot can be based on the installation process of workers.

The process of assembling a single head of the spacer is shown in Figure 15. After the initial contact between the spacer and the wire, the gripper is actuated by the planar mechanical arm to achieve the installation of the spacer. The assembly process of the spacer can be roughly divided into three steps. First, the gripper pivots the chuck to complete the wrapping of the wire; second, the gripper is positioned in the assembly position of the chuck; third, the gripper assembles the chuck of the spacer. By repeating the above operations, the assembly of a single spacer can be completed.

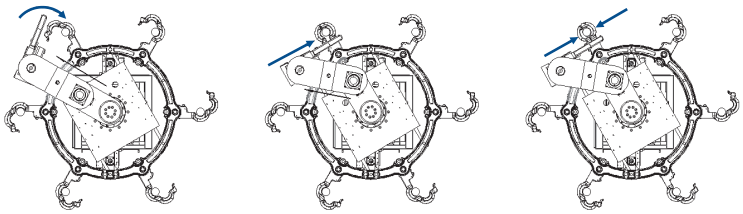


Figure 15. The process of assembling a spacer by the mechanical arm. (**a**) Contacting the chuck to the wire; (**b**) setting the chuck in place; (**c**) assembling.

5.4. Installation Process Design

The installation process of the spacers by the robot is shown in Figure 16. Firstly, the robot is put on the wire on the six split transmission wires and powered on to complete the initialization; secondly, the robot is controlled by the controller to move forward on the wires and arrive at the target position; next, the spacer inside the robot is transferred, and the rotating planar mechanical arm contacts the spacer with the six wires. Then, the gripper at the end of the mechanical arm clamps the six heads of the spacer in turn, which completes the installation; Through the above operations, the robot completes the

installation of the spacer at the given location. After the robot completes the installation of six spacers between two towers, workers move the robot to the next tower for a new installation.

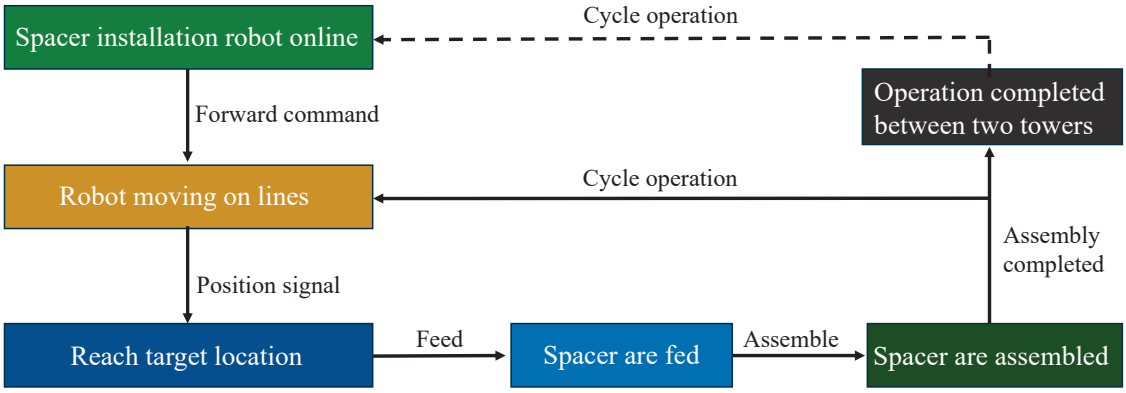


Figure 16. The process of installing spacers.

6. Tests and Discussion

6.1. Forward Capacity Test

After measurement, the weight of the robot without the six spacers is 510 kg, and the weight of one spacer is 20 kg. The robot was placed on simulated six-split transmission lines to test the robot's moving ability, as shown in Figure 17. The test tested the moving ability, the forward speed, and the positioning accuracy of the robot. The test results showed that the robot's moving ability was up to 15°, the maximum forward speed was 200 mm/s, and the positioning accuracy was ±10 mm. The mobility of the robot had matched the requirements of practical engineering. The main parameters of the robot are shown in Table 1.



Figure 17. The forward ability test on the six-split line.

In order to further test the performance of the robot, the robot's advancement system was tested on the ±800 kV DC transmission project line in Leshan City, Sichuan Province,

China, as shown in Figure 18. The test results indicated that the robot could achieve constant distance and stable movement under the control of the ground operator, which provided a reliable carrier for installing spacer. The movement process of the robot is shown in Video 1 in the Supplementary Materials.

Table 1. The main parameters of the robot.

Name	Parameters
Weight of robot G	510 kg
Weight of spacer G_1	20 kg
Maximum moving speed V	200 mm/s
Maximum slope moving capability	15°
Position accuracy	±10 mm
Dimension	2490 mm × 106 mm × 165 mm

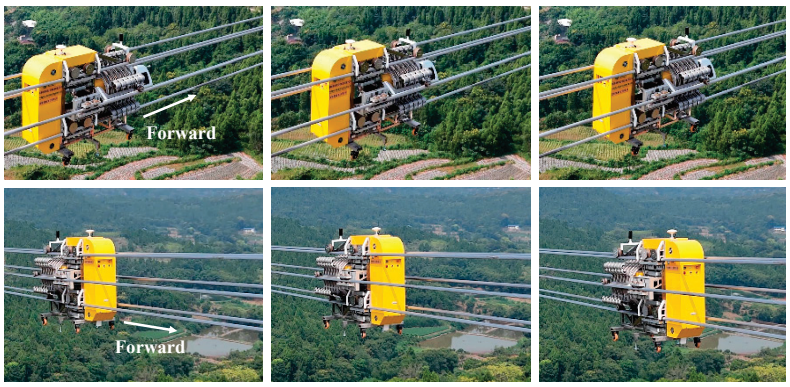


Figure 18. The movement test for the robot on the ±800kV transmission line.

6.2. Installation of Spacers

Four workers used an electric grinder to bring the robot online on the six-split line and then calibrated chucks ① to ⑥ on the spacers to test the performance of the robot's fully automated installation of the spacers. As shown in Figure 19, the robot is able to install the spacers, and the installation process is shown in video 2 and video 3 in the Supplementary Materials. It takes about 2 min to install a spacer chuck and about 12 min to install a Spacer. Through the test, the robot can complete the installation of six spacers at a time. Compared with manual installation, this robot does not require the adjustment of the wire spacing during installation, while workers need to adjust the wire spacing during installation, requiring longer working time and greater labor intensity. Currently, it takes workers more than 20 minutes to install a six-split spacer. When facing strong winds or a rainstorm, the manual time will be longer and more dangerous.

Further, the installation accuracy of the robot was tested. The robot was set to install spacers at intervals of 3000 mm, and after six spaces were installed, the installation accuracy of the spacers was measured. The test results are shown in Figure 20. The results indicated that the maximum deviation of the robot on the installation distance of the spacers was 70 mm, and the installation deviation ratio was 2.33%, which satisfied the installation distance accuracy requirement of the transmission line for the spacers (2.33% < 3%). It should be noted that this error is mainly caused by the robot's small movement distance. As the robot was equipped with an RTK, when moving over long distances, the robot can reset the accumulated error through periodic absolute positioning by using the RTK, so that the error does not increase linearly with distance. Although the 3000 mm is not the actual installation distance of the six spacers, which is limited by actual testing conditions, the measurement results of the robot can prove that it has relatively accurate mobility. The test

results indicated that the robot was able to install the spacers, and the installation deviation meets the engineering requirements, so the robot can replace workers to automate the installation of six-split spacers in the future.

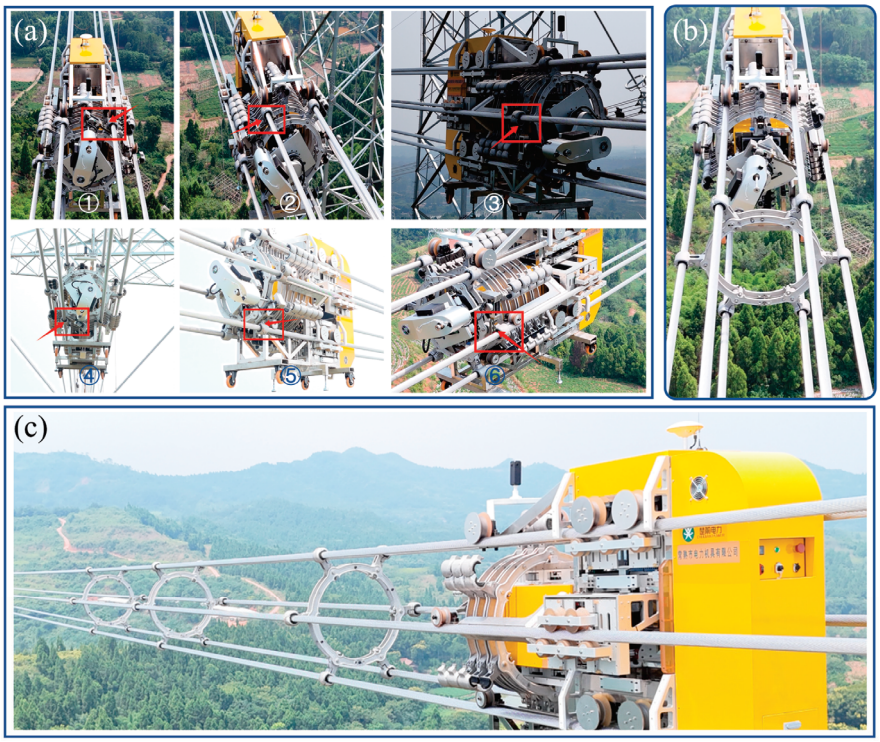


Figure 19. The spacer installation process. (**a**) chuck installation process; (**b**) spacer installation effect; (**c**) multiple spacer installation effect.

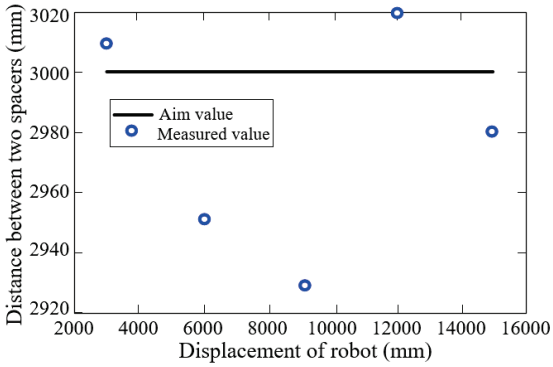


Figure 20. The installation distance deviation of spacers.

6.3. Discussion

Compared to previous robots [32], the robot developed in this paper performs better in terms of moving with high precision, moving with high stability, closed-loop control of motors, and automated installation of spacers. This robot can work with high precision to solve the challenges during real engineering operations, which are not available in previous robots. For example, it is due to the modeling and analysis of the surface geometry of the robot's active wheel that the robot can maintain stability for long periods while moving on the wire and during the installation of spacers.

However, there are still some challenges. When there are strong winds and vibrations in the installation environment, the relative distance between the wires will constantly change, which will cause some difficulties in the stable movement and precise operation of the robot. At present, robots can only perform installation operations when wind and line

vibrations are not severe. In addition, when the outdoor temperature is low (approximately 0 °C), the endurance time of the robot will be shortened, so battery protection measures have to be designed for robots in the future. At present, it mainly relies on workers using tools to transport the robots to the transmission lines, which is due to the high weight of the robot. In the future, it is necessary to design more advanced transportation tools to simplify the process and improve the working environment for workers.

7. Conclusions

The existing six-split spacer is mainly installed manually, which has problems such as high risk and inefficiency. The two-split and four-split spacer installation robot cannot adapt to the movement and installation needs of six-split transmission lines, and cannot perform six-split spacer installation. Based on this, this paper innovatively designs a spacer-installing robot for six-split transmission lines, which can replace the manual installation of spacers. Based on theoretical modeling, the force on the wire and the cross-section form of the moving wheel of the robot were studied to provide theoretical guidance for the design of the moving mechanism. Combined with a modular design, the moving mechanism, transfer mechanism, and assembly mechanism were designed. The control of the motor was designed based on force–position coupling. The robot's ability to move on the slope of the wire, move at a fixed distance, and install six spacers was tested, and the results indicated that the robot was able to move on a 15-degree slope on the wire, move under control, and install six spacers automatically on the wires, with an accuracy of the installation distance of the spacers of less than 3%, which is highly practical for engineering, and it can effectively replace manual work at high altitude, reduce construction risk and costs, and enhance construction efficiency.

As this robot system is heavy, it is possible to further research a lightweight design for the robot and spacer. In addition, the robot is not capable of crossing obstacles on the transmission line and can only install spacers in a single direction. Therefore, in the future, the robot needs to be designed to be able to cross obstacles. Besides the installation capability, the robot should be designed with a spacer removal capability to replace structurally damaged spacers and to enhance the robot's intelligence. In the future, the control system and perception system of the robot will be further researched to improve the intelligence level of the present robot.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/machines13050432/s1.

Author Contributions: Conceptualization, J.P. and Y.C.; methodology, J.P. and C.H.; software, M.J.; validation, J.P. and C.H.; formal analysis, Y.M.; investigation, J.P.; resources, M.J.; data curation, F.M. and Q.S.; writing—original draft preparation, J.P.; writing—review and editing, J.P.; visualization, J.P. and C.H.; supervision, F.M. and Q.S.; project administration, C.H.; funding acquisition, Y.C. and M.J. All authors have read and agreed to the published version of the manuscript.

Funding: This project is supported by the Science and Technology Project of the State Grid (5200-202220092A-1-1-ZN).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: Author Jie Pan was employed by the company of State Grid Electric Power Engineering Research Institute Co., Ltd. The remaining authors declare that the research was

conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Agarwal, H.; Mukherjee, S.; Chattaraj, S.; Khan, I.; Prasad, D. Development of quad spacer damper to suppress oscillation in 765 kV transmission line. In Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 1–3 July 2020; pp. 1–6.
2. Lee, H.K. Importance of the Spacer-Damper Installation in Overhead Lines. In Proceedings of the KIEE Conference, Kunming, China, 18–20 July 2005; The Korean Institute of Electrical Engineers: Seoul, Republic of Korea; pp. 46–48.
3. Disyadej, T.; Kwanmuang, S.; Muneesawang, P.; Promjan, J.; Poochinapan, K. Smart transmission line maintenance and inspection using mobile robots. *Adv. Sci. Technol. Eng. Syst. J.* **2020**, *5*, 493–500. [CrossRef]
4. Bahrami, M.R.; Abed, S.A. Mechanical challenges of electrical transmission lines inspection robot. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *709*, 022099. [CrossRef]
5. Li, X.; Li, Z.; Wang, H.; Li, W. Unmanned aerial vehicle for transmission line inspection: Status, standardization, and perspectives. *Front. Energy Res.* **2021**, *9*, 713634. [CrossRef]
6. Luo, Y.; Yu, X.; Yang, D.; Zhou, B. A survey of intelligent transmission line inspection based on unmanned aerial vehicle. *Artif. Intell. Rev.* **2023**, *56*, 173–201. [CrossRef]
7. Meng, D.; Lu, B.; Li, A.; Yin, J.; Li, Q. Pressure observer based adaptive dynamic surface control of pneumatic actuator with long transmission lines. *Appl. Sci.* **2019**, *9*, 3621. [CrossRef]
8. Siravuru, A.; Shah, S.V.; Krishna, K.M. An optimal wheel-torque control on a compliant modular robot for wheel-slip minimization. *Robotica* **2017**, *35*, 463–482. [CrossRef]
9. Moraitis, M.; Vaiopoulos, K.; Balafoutis, A.T. Design and implementation of an urban farming robot. *Micromachines* **2022**, *13*, 250. [CrossRef]
10. Gao, Y.; Song, G.; Li, S.; Zhen, F.; Chen, D.; Song, A. LineSpyX: A power line inspection robot based on digital radiography. *IEEE Robot. Autom. Lett.* **2020**, *5*, 4759–4765. [CrossRef]
11. Zorić, F.; Flegarić, S.; Vasiljević, G.; Bogdan, S.; Kovačić, Z. Autonomous Installation of Electrical Spacers on Power Lines Using Magnetic Localization and Special End Effector. *Machines* **2023**, *11*, 510. [CrossRef]
12. Richard, P.-L.; Morin, F.; Lepage, M.; Hamelin, P.; Lambert, G.; Sartor, A.; Hebert, C.; Pouliot, N. Inside lineranger: Mechanism design to optimize operation and performances of powerline inspection robot. In Proceedings of the 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022; pp. 8157–8163.
13. Qin, X.; Wu, G.; Lei, J.; Fan, F.; Ye, X.; Mei, Q. A novel method of autonomous inspection for transmission line based on cable inspection robot lidar data. *Sensors* **2018**, *18*, 596. [CrossRef]
14. Song, Y.; Wang, H.; Zhang, J. A vision-based broken strand detection method for a power-line maintenance robot. *IEEE Trans. Power Deliv.* **2014**, *29*, 2154–2161. [CrossRef]
15. Ma, L.; Zhang, H.; Li, Z.; He, W.; Wang, Z.; Guo, T. Research and application of automatic walking marking robot for overhead transmission line. In Proceedings of the International Conference on Computer Application and Information Security (ICCAIS 2023), Wuhan, China, 8 April 2024; SPIE: Bellingham, WA, USA, 2024; Volume 13090, pp. 1050–1057.
16. Guo, J.; Yang, Z.; Karkee, M.; Duan, J.; He, Y. Robotization of banana de-handing under multi-constraint scenarios: Challenges and future directions. *Artif. Intell. Agric.* **2025**, *15*, 1–11. [CrossRef]
17. Espinosa Peralta, P.; Palomino-Díaz, V.; Rodríguez Herrerías, P.; Rodríguez del Castillo, C.; Pedraza Cubero, C.M.; Yuste, J.P.; Aparicio Cillán, P.; Mesía López, R.; Águila Rodríguez, G.; Fernández Franco, V.P. ROSE: Robot for Automatic Spacer Installation in Overhead Power Lines. In *Iberian Robotics Conference*; Springer International Publishing: Cham, Switzerland, 2022; pp. 325–337.
18. Wu, Z.; Zhang, J.; Chen, H.; Huang, Z. Development on robot installation using quad spacer damper. In Proceedings of the International Conference on Artificial Intelligence, Virtual Reality, and Visualization (AIVRV 2021), Sanya, China, 16 December 2021; SPIE: Bellingham, WA, USA, 2021; Volume 12153, pp. 265–272.
19. Alhassan, A.B.; Zhang, X.; Shen, H.; Xu, H. Power transmission line inspection robots: A review, trends and challenges for future research. *Int. J. Electr. Power Energy Syst.* **2020**, *118*, 105862. [CrossRef]
20. Mostashfi, A.; Fakhari, A.; Ali Badri, M. A novel design of inspection robot for high-voltage power lines. *Ind. Robot Int. J.* **2014**, *41*, 166–175. [CrossRef]
21. Zhang, T.; Dai, J. Electric power intelligent inspection robot: A review. *J. Phys. Conf. Series.* **2021**, *1750*, 012023. [CrossRef]
22. Zhu, Q.; Guan, X.; Yu, B.; Zhang, J.; Ba, K.; Li, X.; Xu, M.; Kong, X. Overview of structure and drive for wheel-legged robots. *Robot. Auton. Syst.* **2024**, *181*, 104777. [CrossRef]
23. Pussente, G.A.N.; de Aguiar, E.P.; Marcato, A.L.M.; Pinto, M.F. UAV Power Line Tracking Control Based on a Type-2 Fuzzy-PID Approach. *Robotics* **2023**, *12*, 60. [CrossRef]

24. Debenest, P.; Guarnieri, M.; Takita, K.; Fukushima, E.F.; Hirose, S.; Tamura, K.; Kimura, A.; Kubokawa, H.; Iwama, N.; Shiga, F. Expliner-Robot for inspection of transmission lines. In Proceedings of the 2008 IEEE International Conference on Robotics and Automation, Pasadena, CA, USA, 19–23 May 2008; pp. 3978–3984.
25. Rui, G.; Feng, Z.; Lei, C.; Jun, Y. A mobile robot for inspection of overhead transmission lines. In Proceedings of the 2014 3rd International Conference on Applied Robotics for the Power Industry, Foz do Iguassu, Brazil, 14–16 October 2014; pp. 1–3.
26. Zhu, A.; Tu, Y.; Zheng, W.; Shen, H.; Zhang, X. Design and implementation of high-voltage transmission line inspection and foreign bodies removing robot. In Proceedings of the 2018 15th International Conference on Ubiquitous Robots (UR), Honolulu, HI, USA, 26–30 June 2018; pp. 852–856.
27. Gulzar, M.A.; Kumar, K.; Javed, M.A.; Sharif, M. High-voltage transmission line inspection robot. In Proceedings of the 2018 International Conference on Engineering and Emerging Technologies (ICEET), Lahore, Pakistan, 22–23 February 2018; pp. 1–7.
28. Chen, M.; Cao, Y.; Tian, Y.; Li, E.; Liang, Z.; Tan, M. A passive compliance obstacle-crossing robot for power line inspection and maintenance. *IEEE Robot. Autom. Lett.* **2023**, *8*, 2772–2779. [CrossRef]
29. Seok, K.H.; Kim, Y.S. A state of the art of power transmission line maintenance robots. *J. Electr. Eng. Technol.* **2016**, *11*, 1412–1422. [CrossRef]
30. Toussaint, K.; Pouliot, N.; Montambault, S. Transmission line maintenance robots capable of crossing obstacles: State-of-the-art review and challenges ahead. *J. Field Robot.* **2009**, *26*, 477–499. [CrossRef]
31. Pouliot, N.; Montambault, S. LineScout Technology: From inspection to robotic maintenance on live transmission power lines. In Proceedings of the 2009 IEEE International Conference on Robotics and Automation, Kobe, Japan, 12–17 May 2009; pp. 1034–1040.
32. Pan, J.; Hu, C.; Jiang, M.; Ma, Y.; Meng, F.; He, X.; Liu, C. Development of a Spacer Bar Installing Robot for Six-Split Transmission Lines. In Proceedings of the 8th International Conference on Computing, Control and Industrial Engineering (CCIE2024), Singapore, 21–22 June 2024; pp. 11–17.
33. Hatibovic, A. Derivation of equations for conductor and sag curves of an overhead line based on a given catenary constant. *Period. Polytech. Electr. Eng. Comput. Sci.* **2014**, *58*, 23–27. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Digital Twin for Flexible Manufacturing Systems and Optimization Through Simulation: A Case Study

Adriana Florescu

Department of Engineering and Industrial Management, Faculty of Technological Engineering and Industrial Management, Transilvania University of Brasov, Bd. Eroilor 29, 500036 Brasov, Romania; fota.a@unitbv.ro

Abstract: The research presented in this paper aligns with the advancement of Industry 4.0 by integrating intelligent machine tools and industrial robots within Flexible Manufacturing Systems (FMS). Primarily, a development approach for Digital Twin (DT) is presented, beginning from the design, sizing, and configuration stages of the system and extending through its implementation, commissioning, operation, and simulation-based optimization. The digitization of current industrial processes entails the development of applications based on modern technologies, utilizing state-of-the-art tools and software. The general objective was to create a digital replica of a process to propose optimization solutions through simulation and subsequently achieve virtual commissioning. The practical nature of the research is reflected in the design and implementation of a Digital Twin for a real physical system processing a family of cylindrical parts within an existing experimental FMS. A digital model of the system was created by defining each individual device and piece of equipment from the physical system, so the virtual model operates just like the real one. By implementing the Digital Twin, both time-based and event-based simulations were performed. Through the execution of multiple scenarios, it was possible to identify system errors and collisions, and propose optimization solutions by implementing complex, collaborative-robot equipment where multiple interactions occur simultaneously.

Keywords: flexible manufacturing systems; digital-twins; simulation; Industry 4.0

1. Introduction

Human creativity, correlated with the desire for a permanent increase in the standard of living, constitutes the basis of societal development. Recognized as a new paradigm in the field of manufacturing, the concept of Industry 4.0 presents an opportunity to re-engineer and revitalize production through a new perspective focused on the digitization of current production systems. In order to effectively respond to new market demands, companies must rethink their entire production workflow by implementing digital technologies. This digital transformation will improve organizational competitiveness [1], while strengthening companies' ability to make optimal decisions.

With the rapid technological evolution driven by Industry 4.0, industrial companies need up-to-date strategies to respond to various challenges, including greater uncertainty and variability in global markets, heightened competition, pressures for innovation, and accelerated product launch times, leading to reduced delivery times. Additionally, customer behavior is shifting toward a demand for mass customized products.

The transformation to Industry 4.0 is grounded in three core directions: industrial–technological, economics in business models, and social [2].

The current stage of development is characterized by the transition from Industry 4.0 with new technologies already implemented in some companies, toward a sustainable and resilient society based on the new paradigm of Society 5.0 [3–5]. A new concept, "The Resilient Operator 5.0", integrated into smart manufacturing systems represents an alternative vision [6,7] for the development of Industry 5.0. Presently, we are seeing avant-garde studies and research focused on the development and implementation of

artificial intelligence, human–intelligent machine collaboration, and the integration of biotechnologies, as well as quantum computers and quantum technologies—a stage some authors refer to as Industry 6.0 [8].

The implementation of flexibility and automation concepts, together with the rapid progress of information technology, has created the essential prerequisites for achieving the transition to the digitalization of production systems within the Industry 4.0 framework. The development of manufacturing systems capable of managing flexibility requirements in achieving a wide typological diversity of products, based on customer demand, is now a key competitive factor for companies across industrial sectors. Enhanced production system flexibility boosts manufacturing performance, leading to increased competitiveness. In the research study conducted by Brettel [9], relationships between flexibility and performance are analyzed, and a systematic, logical diagram of empirical reviews is developed to connect the model to advancements in Industry 4.0 technologies [10].

Production systems have undergone dynamic evolution, requiring increasing flexibility, the trend being towards complete automation and the development of flexible robotic systems [11,12]. Flexible production can be viewed from a strategic perspective as essential for the development of advanced production systems capable of offering customers high-quality, varied, and customized products. This justifies the need for greater process flexibility and versatile resources, to respond to the variability and uncertainties inherent in the operation of production systems [13]. The flexible production system has proven to be the foundation of Industry 4.0, capable of managing changes both in technology and in the adoption of new management methods. It has the agility to reconfigure operating instructions in order to introduce new products into manufacturing or modify existing ones within a specified time period at the request of customers. Practice has shown that the flexible automated manufacturing system [14] is a pivotal element in the ongoing industrial evolution.

Increasing the competitive advantage and performance of companies, especially in the industrial sector, requires continuous innovation and the adaptation of production systems and business processes to accommodate new customized products and services. By developing intelligent platforms that integrate all operations in the value-creating chain and interconnected systems, people, machines, and products, the industry enters a new stage of development characterized by "intelligent and autonomous production" [15].

The experience and development of large industrial companies [16–18] demonstrate the feasibility of gradually implementing advanced production systems characterized by flexibility and automation, to which new technologies in the field of digitization, connectivity, and artificial intelligence can be added today. Iterative and agile product development processes enable complex processes to be managed within flexible, intelligent systems under conditions of uncertainty. These conditions are generated by the variability in product demand in the market and the focus on diverse, mass-customized products, as well as variations in the internal production environment: rapid changes in manufacturing, the simultaneous production of several types of random items from the same family, and functional blockages that occur in the operation of dynamic systems like Flexible Manufacturing Systems (FMS). Along the way, companies have identified that optimization begins by simplifying existing processes, then continuing with the integration of modern technologies and the adoption of flexible automation [12] solutions to effectively respond to changing market demands.

An element of novelty in the development of smart manufacturing in recent years has been the introduction of the notion of the "Digital Twin" [19,20] as a software solution for the virtual creation and simulation of a real physical system, in order to analyze the performance and optimize the functioning of complex dynamic systems. The Digital Twin (DT), along with the technologies of Industry 4.0 (the Internet of Things, Big Data Analytics, and Cloud Computing), enables advanced analysis in the field of system optimization through the integration of monitoring and control functions. DT is an exceptionally useful tool in industrial practice, particularly in real-time monitoring and total predictive mainte-

nance, with new data being constantly updated and facilitating accurate and immediate interventions in the system.

Current concepts, such as "Digital Twins" [21,22] and "virtual commissioning" [23], studied in specialized literature, are still in the early stages, as there is no general method for their design and implementation. Currently, there are studies and research focused on designing models and architectures of Digital Twins, as highlighted by authors [24–26]. Efforts are being made to standardize these approaches; however, practical application poses several challenges. Additionally, well-documented studies, such as [27–29], suggest that the use of a Digital Twin results in significant differences based on the fields of applicability, as noted by [30–33]. As a result, it is timely to initiate future research activities to establish relevant industrial applications and to investigate and demonstrate the wide range of applications and benefits of these new technologies.

The present scientific research contributes to the development of intelligent manufacturing systems by implementing and applying new knowledge, methodologies regarding the design and use of relatively new concepts such as "Digital Twins (DT)" and virtual commissioning and the optimization of DT simulation, based on the current state of the reviewed literature.

Even though classic simulation tools are still used in the design of production systems, working with static data, they do not capture the system in its dynamic evolution. Events asynchronous in time that cause interruptions in the production flow or delayed reactions from operators can only be captured by simulation in a real-time environment. To help overcome these limitations that the production field still faces and to reduce this gap, the paper aims to study and configure a digital model of the Digital Twin type in the manufacturing industry, with the goal of creating a dynamic and autonomous manufacturing system.

The use of advanced technologies like Digital Twins throughout the design, manufacturing, and maintenance phases, by developing technical solutions for collaboration between humans and smart machines, creating human–machine interface systems, and enabling remote maintenance, brings significant transformations in managing dynamic processes in real time. Expected results can only be achieved by using established, state-of-the-art simulation software.

These new approaches could become elements of resilience and flexibility for the physical production environment; "agility could be the key to the resilience of the manufacturing industry" [34], digital technologies being the main pillars. A manufacturer can virtually redesign a product, process or production system and even simulate its performance in the real world. DT creates a virtual representation of both the physical elements and the dynamics of how an IoT, Internet of Things, connected product functions and interacts in its environment throughout its entire life cycle.

Structured in four sections, the scientific article begins by presenting the current context of the development of advanced production systems and an analysis of the specialized literature on the Industry 4.0 and Digital Twins approach, highlighting the opportunity for application and their benefits (Section 1). The proposed research methodology and framework are presented in Section 2. The applied research (Section 3) aims to design and implement the "Digital Twin" of a Flexible Manufacturing System, conducted through physical experimentation, and to find optimization solutions within a company, through its virtual commissioning and dynamic simulation. In Section 4, conclusions and future research directions are discussed.

2. Materials and Methods

The main objective of the research consists in approaching a flexible manufacturing system from the perspective of implementing Industry 4.0 (I4.0) technologies and moving to Industry 5.0 (I5.0) by designing a Digital Twin and offering optimization solutions through simulation. Subsequently, the behavior of the real physical system was tested, by creating a virtual commissioning model (predictive maintenance).

In terms of simulation, the main contribution is represented by the correspondence between the real and the digital universe. The fact that physical objects can be transposed into computer objects through digitization is currently the most powerful concept for simulation, which allows new processes to be tested and studied before they are implemented.

The applied case study focused on the development of a company in the manufacturing industry, and a project to digitize a production system aligned with the management's requirements. Many companies that have taken the first step toward digitalizing their production systems see sustainable development as an opportunity by adding new dimensions of digitalization and connectivity. The goal is to streamline processes and increase competitive advantage, especially given the shortening of product life cycles and the demand for mass customization according to customer requirements.

2.1. Industry 4.0 Approaches

Initially perceived as an "industrial paradigm", the concept of Industry 4.0 (I4.0) first introduced in Germany in 2011 has revolutionized industries worldwide. It has since expanded globally under various names tailored to each country's specific context: France—"Industrie du Futur", UK—"Catapult", USA—"Smart Manufacturing", China—"Made in Chaina 2025", and Japan—"Industrial Value-Chain Initiative".

Currently, through advanced technologies, Industry 4.0 actively contributes to societal transformation by digitizing and connecting systems and incorporating elements of artificial intelligence. The technological advancements in Industry 4.0 are built upon key pillars: Cyber-Physical Systems (CPS), Big Data Analytics, the Internet of Things, Cloud Computing, Human–Machine Interaction (HMI), Robotics, and Artificial Intelligence.

A synthetic analysis of the transition from Industry 4.0 to Industry 5.0 within the current specialized literature—covering concepts, technologies, models, and areas of applicability—has led to the identification of several reference papers (Table 1).

Table 1. Summary of reviewed papers: transition from I4.0 to I 5.0 and Society 5.0.

Research Areas	Literature Review
Industry 4.0 initiative Industry 4.0 concept	Quin, J. et al. (2016) [1]; Ulrich, S. (2013) [35]; Bauernhansel, T. (2014) [36]; Berger, R. (2014) [37]; Burmeister, C. et al. (2016) [38].
Basic conceptual models in Industry 4.0	Dombrovski (2017) [39]; Sony (2018) [40]; Schumacher, A. (2019) [41]; Santos, R.C. et al. (2020) [42]; Amaral, A. (2021) [43]; Zoubek, M. et al. (2021) [44].
Industry 4.0 technologies	Ryalat, M. et al. (2023) [45]; Gupta, B.B. et al. (2021) [46]; Culot, G. et al. (2020) [47]; Pei, E. et al. (2020) [48]; Frank, A. G. et al. (2019) [49].
Industry 4.0 implementation in companies and SMEs	Vrchota, J. et al. (2019) [50]; Gajdzik, B. et al. (2021) [51]; Moeuf, A. et al. (2020) [52]; Lodgaard, E. et al. (2022) [53]; Grufman, N. et al. (2020) [54]; Vinodh, S. et al. (2020) [55]; Schönfuß, B. et al. (2021) [56].
Industry 4.0 sustainability	Ghobakhloo, M. (2020) [3]; De Sousa Jabbour, A.B.L et al. (2018) [57]; Fritzsche, K. et al. (2018) [58]; Leong, W.D. et al. (2020) [59]; Sharma, M. et al. (2023) [60].
Industry 5.0 Society 5.0	Mourtzis, D. et al. (2024) [4]; Pereira, A.G. et al. (2020) [5]; Romero, D., & Stahre, J. (2021) [6]; Lu, Y., et al. (2022) [7]; Adel, A. (2022) [61]; Aslam, F. et al. (2020) [62]; Polonara, M. et al. (2024) [63]; Borboni, A. et al. (2023) [64]; Li, C. et al. (2023) [65]; Prassida, G.F. et al. (2022) [66]; Kopp, T. et al. (2021) [67]; Vanderborght, B. (2020) [68].

If until now Industry 4.0 has provided tools for technological development, we are currently witnessing a transformation toward Industry 5.0, which emphasizes human and social development within Society 5.0 [61,62].

In this context, there is a need to develop new models of intelligent production, motivating and empowering people to work alongside intelligent machines in Industry 4.0. The innovative technologies of Industry 4.0 are driving an increasing number of industrial companies to transform into smart factories by reconfiguring and optimizing flexible manufacturing systems.

2.2. Digital Twin Concept

The concept of the Digital Twin (DT), closely associated with the development of Industry 4.0 and various fields of application, appears in the reviewed literature with multiple definitions and approaches [69,70]. A recent scientific study [71] highlights the development trends of this key technology for the industry in the coming years, emphasizing the digital transformation from "Industry 4.0—Smart Factory" to "Industry 5.0—Human Robot Interface" and eventually to "Industry 6.0—Cognitive Manufacturing". In accordance with the latest ideas on industrial development [28], Digital Twins (DTs) endowed with elements of Artificial Intelligence (AI) and utilizing machine learning algorithms are the foundation for the development of autonomous factories that can continuously and dynamically self-adapt.

The authors of the research presented in [72] believe that an essential characteristic of a "Digital Twin" is the "connection between digital and physical systems", with the ability to provide real-time information about the physical system. The key technologies underlying the configuration and modeling of a DT were also highlighted in [73], including modeling based on physical systems, data-based modeling, big data cybernetics, platform infrastructure, and human–machine interface (HMI). Table 2 presents some representative research tracing the concept of Digital Twins from its emergence to the present, covering its evolution, characteristics, implementation of digital models, and applicability.

Table 2. Summary of reviewed papers: "Digital Twins".

Research Areas	Literature Review
"Digital Twin" concept	Bilberg, A., & Malik, A. A. (2019) [74]; Kritzinger, W. et al. (2018) [75]; Wagg, D. et al. (2020) [76].
Methodologies, conceptual models, DT configurations	Psarommatis, F. et al. (2022) [22]; Kusiak, A. (2022) [23]; Heindl, W. et al. (2022) [24]; Perno, M. et al. (2022) [25]; Riedelsheimer, T. et al. (2021) [26]; Gupta, B.B. et al. (2021) [46]; Culot, G. et al. (2020) [47]; Segovia, M. et al. (2022) [62].
Research on the applicability of the DT	Semeraro, C. et al. (2023) [18]; Wang, K. J. et al. (2021) [19]; Yangguang L. et al. (2021) [20]; Kusiak, A. (2022) [21]; Singh, M. et al. (2020) [29]; Psarommatis, F. et al. (2022) [77]; Huang, H. et al. (2021) [78]; Piromalis, D. et al. (2022) [79]; Leng, J. et al. (2021) [80].
Research on the different effects of the DT in various fields of activity	Fett, M. et al. (2023) [28]; Singh, M. et al. (2022) [29]; Kousi, N. et al. (2021) [30]; Zohdi, T.I. (2021) [31]; Zandi, K. et al. (2019) [32]; Li, W. et al. (2020) [33]; Attaran, M. et al. (2023) [81]; Wang, K. et al. (2022) [82]; Agnusdei, G.P. et al. (2021) [27]; Fera, M. et al. (2020) [83]; Malik, A. A et al. (2020) [84].
DT integration in Flexible Manufacturing Systems and collaborative robots	Pereira, J.A.P. and Campilho, R.D.S.G. et al. (2022) [11]; Sousa, V.F.C. et al. (2022) [12]; Adel, A. et al. (2022) [61]; Polonara, M. et al. (2024) [63]; Makris, S. (2020) [85]; Koesters, A. et al. (2024) [86]; Resman, M. et al. (2021) [87]; Kaiser, J. et al. (2023) [88].

In practice, there are several software solutions for implementing Digital Twin models, with Siemens offering the most services for building a "Digital Twin" (Process Simulate module—Tecnomatix, Siemens PLM [89]) and including a Human–Machine Interface (Process Simulate Human). Considering the various existing approaches to the "Digital Twin" concept and the lack of a standard methodology for constructing Digital Twins, it is

timely to initiate research in this direction, particularly focusing on applicability in Flexible Manufacturing Systems.

Digital Twin models applied to advanced production systems and cyber-physical systems can mirror their state and allow simulation in a specific environment, enabling real-time monitoring and control of processes. These models form the very basis for the development of total predictive maintenance by creating models for virtual commissioning [90].

By collecting relevant data from the analyzed system through simulation and predictions based on virtual models, a Digital Twin can act in all phases of the product life cycle, from design to manufacturing, testing, operation, exploitation, and maintenance (Figure 1).

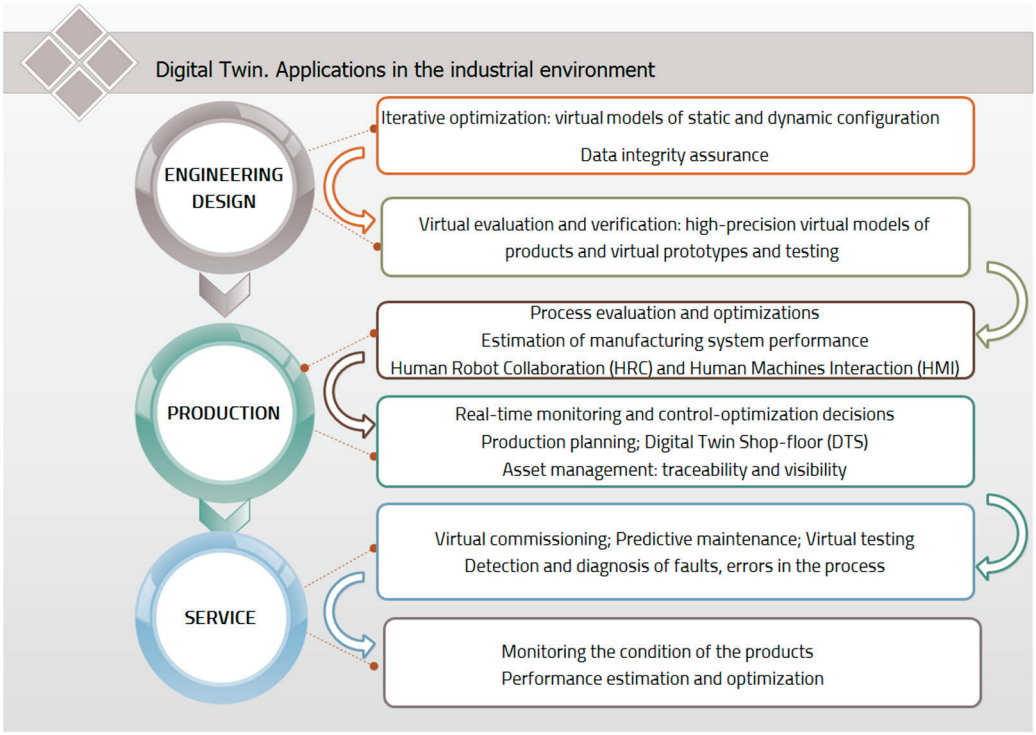


Figure 1. Digital Twin applications in industrial environment.

A Digital Twin can represent the real state of a manufacturing system as a 3D simulation model with real-time updates, requiring a suitable simulation model, such as the generic simulation model of the developed Flexible Manufacturing System.

3. Results of Digital Twin Design and Dynamic System Simulation

The DT simulation research method for the flexible manufacturing system involves replacing the real, dynamic, and complex system with its model or a simulator. This approach involves using simulation languages and/or software to perform experiments and obtain information on the evolution of the system's behavior over real time.

3.1. Process Description: Case Study

The sequential process carried out by the physical manufacturing system implements an automated flow of operations for cylindrical parts (discs, flanges, rings, and circular shafts of machine tool components). Figure 2 shows parts supply, transport, transfer, drilling, turning, inspection, grinding, the storage of good parts, and the transfer of scrap parts to rework stations [14]. The requirements for improving the cycle time and increasing the productivity in the processing of parts from this benchmark family were taken into account.

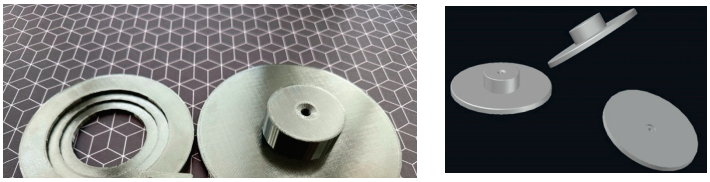


Figure 2. Examples of machined landmark types.

The phases of the processing process within the experimental FMS, physically existing in the laboratory, previously designed by the author [14], consist of the following:

- The automatic supply of cylindrical semi-finished products from a warehouse;
- Transporting the parts on the conveyor to the drilling station;
- The transfer of the piece to the machining station through drilling;
- The drilling processing of the semi-finished product, with the possibility of setting the drilling depth and advance speed on the computer;
- Transporting the drilled piece to the next station, for processing by turning;
- The turning of cylindrical semi-finished products;
- Processing by rectification and final control;
- The transportation of processed parts, with a conveyor, to the product warehouse;
- The final control and selection of good parts for delivery and their storage in the warehouse of finished products;
- The identification and evacuation of scrap parts, which do not fall within the allowed tolerances, using the SCADA (Supervisory Control and Data Acquisition) software application, CX–Supervisor (Release 3.1 by OMRON 2010);
- Controlling the process through a programmable logic controller (PLC) and process monitoring and control software;
- The actuation of equipment and devices in the system by using nine pneumatic actuation cylinders, the existence of a regulator filter and electrovalves, a stepper motor and related driver, three presence sensors (optical or inductive), and laser sensors.

The 3D model of the real system is shown in Figure 3. The original (Supervisory Control and Data Acquisition (SCADA) software application, previously developed by the author for the same system, was designed for production tracking by real-time monitoring and controlling of asynchronous events that may occur in the flexible manufacturing process.

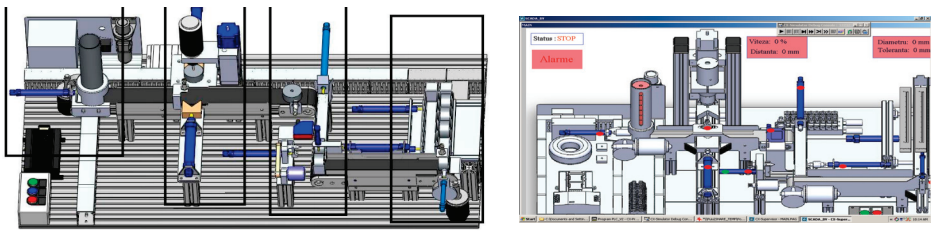


Figure 3. 3D model of the automated system—FMS and process monitoring and control application.

Creating a digital representation of the factory involves digitizing flexible lines, workstations, machine tools, devices, and individual equipment, resulting in a virtual model with similar functionality to the real system. Virtual commissioning, implemented with Process Simulate—version V15.1.2. by Siemens [89,90], is useful for reducing costs, product launch time, and other resources involved in the actual commissioning of certain processes or even the entire plant.

3.2. Digital Twin Design for an FMS: Case Study

The research conducted toward the integration of Flexible Manufacturing Systems into Industry 4.0 had the general objective of creating a digital copy of a process, aimed at proposing optimization solutions and subsequently implementing it in virtual operation.

A first step was to design a "Digital Twin" for an existing manufacturing system, taking into account process improvement requirements. The potential for applying the concept of virtual commissioning in the system studied, within a manufacturing company, was also analyzed.

Digital Twin Configuration Methodology for a Flexible Manufacturing System

The Digital Twin design methodology for the studied system follows the stages outlined in Figure 4.

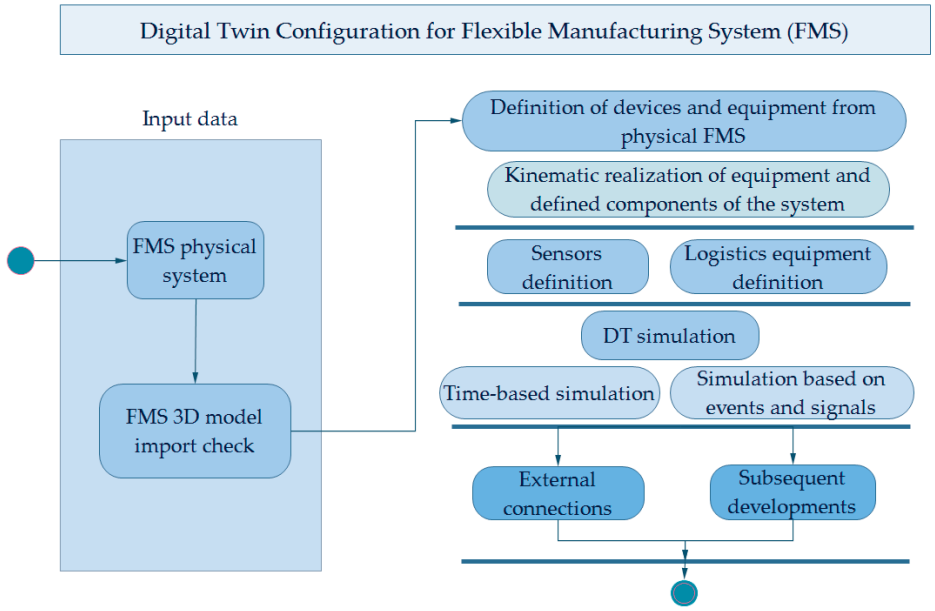


Figure 4. Digital Twin design stages for Flexible Manufacturing Systems (FMS).

Figure 5 summarizes some intermediate sequences from the Process Simulate application used in the Digital Twin design for the physical model of the existing FMS.

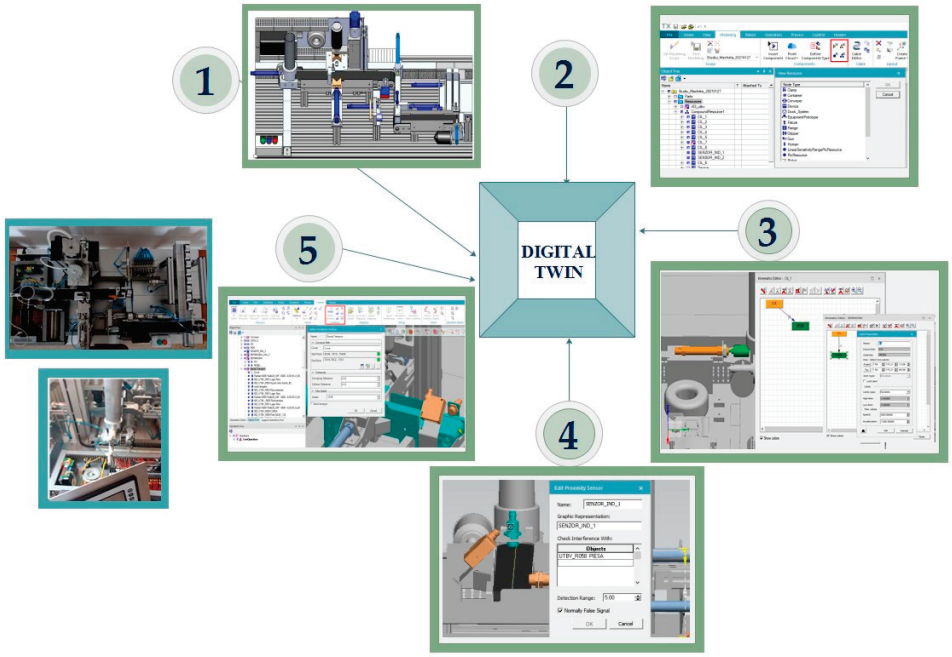


Figure 5. Digital Twin design sequences: case study application.

- Step 1—Importing the 3D model of the system. The CAD (3D) model of the automated system was taken in STEP format and imported into Process Simulate using the "Convert and Insert CAD Files" function.
  In this way, the original model was converted to JT format, the native format recognized by Process Simulate.
- Step 2—Defining devices and equipment based on the existing physical system model. After the conversion of the 3D model was completed, the stage followed where each piece of equipment was isolated into an individual device. In Process Simulate, new resources can be created to represent each piece of equipment as realistically as possible.
- Step 3—Kinematic realization for the defined equipment. Each piece of equipment defined in Process Simulate can have one or more kinematic torques defined. They make the connections between the moving elements, such as pneumatic cylinders for actuating the equipment and the devices integrated into the experimental physical system. Thus, the device can move in such a way as to reproduce the movements of real devices. The fixed and moving parts of the device were defined, as well as the types of motion: rotation, translation, or combined motion (Kinematics Editor Command). The parameters considered in the definition of these kinematic couples are the type of movement, the axis or vector of movement, the minimum and maximum limits of the movement, the velocity, and the acceleration (Joint Type Command). Except for the movement type, all mentioned parameters can be changed at any time. Predefined positions of the device (Pose Editor command) where the kinematic torque has a certain value have been set. Here, the name and value of the kinematic couple can be assigned for the respective position. Defining these elements is very important, as they will be used in all aspects of the project. For example, in Figure 6, the values set for a pneumatic actuation cylinder (pneumatic cylinder 7) from the experimental FMS system are presented in detail. Similarly, the positions for each kinematic couple of the system were defined.

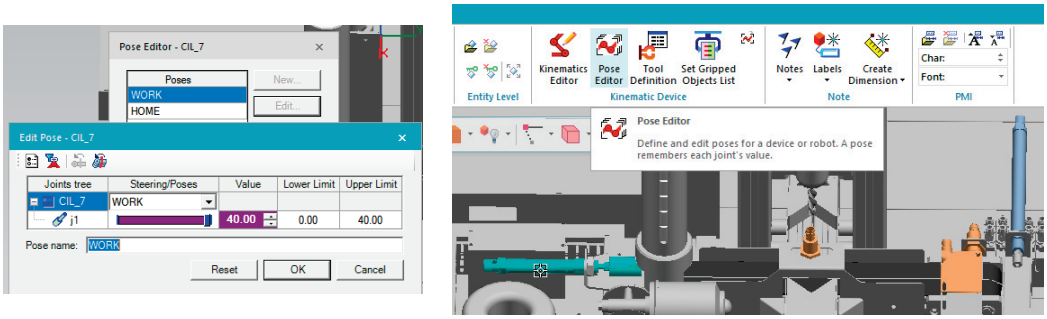


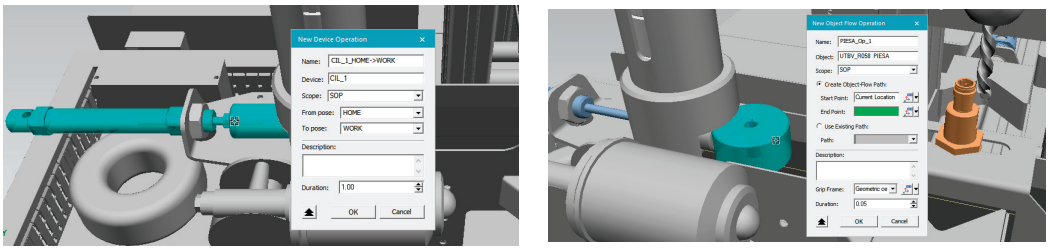
Figure 6. Kinematic implementation for the defined equipment: program sequence.

- Step 4—Defining the sensors. In Process Simulate, sensors are resources capable of detecting the presence of an object or a specific property of that object. The most commonly used types are proximity sensors and photoelectric sensors. Property sensors are used to detect certain characteristics assigned to the object, such as weight, color, type, barcode, etc. A real-life analogy would be RFID (Radio Frequency Identification) systems, where a special tag can be written or read by a device.
- Step 5—Conveyor belt definition. To create a conveyor belt in Process Simulate, the 3D resource was selected, and the "Define Conveyer" command was used. The conveyor belts of the system were defined, specifically a curve that dictates the direction and orientation of the transport, with the latter characteristic being reversible. The conveyor belt's characteristics were set as follows: tolerance refers to the precision required for placing objects on its surface, and feed speed is represented in units of [mm/s] and can be adjusted at any time depending on the process.

3.3. DT Simulation Results

Time-based simulation aims to perform a predefined sequence of operations to present the ideal process or solution. The operations used are simple and predefined, and their goal is to move the device from one position to another or to move parts along a predetermined trajectory. To carry out this simulation, commands such as "New Device Operation" or "New Object-Flow Operation" were used.

Sequences regarding the setting of the movement parameters and the trajectory are presented in Figure 7. "Device-type" operations allow the movement of a piece of equipment from one position (pose) to another within a predefined time, set in Figure 7a. The name is arbitrary, and the parameters of the operation can be changed at any time. "Object-Flow" operations are created in a similar way, the difference being that in this case, the operation is actually a path that the selected object will follow. The start and end points of the trajectories can be defined, or a predefined trajectory can be selected, as shown in Figure 7b.



(**a**) (**b**)

Figure 7. Settings: (**a**) movement parameter settings and (**b**) trajectory setting.

In the end, a Digital Twin for the analyzed system was created, as a complex sequence of operations that will identically reproduce the real process (Figure 8).

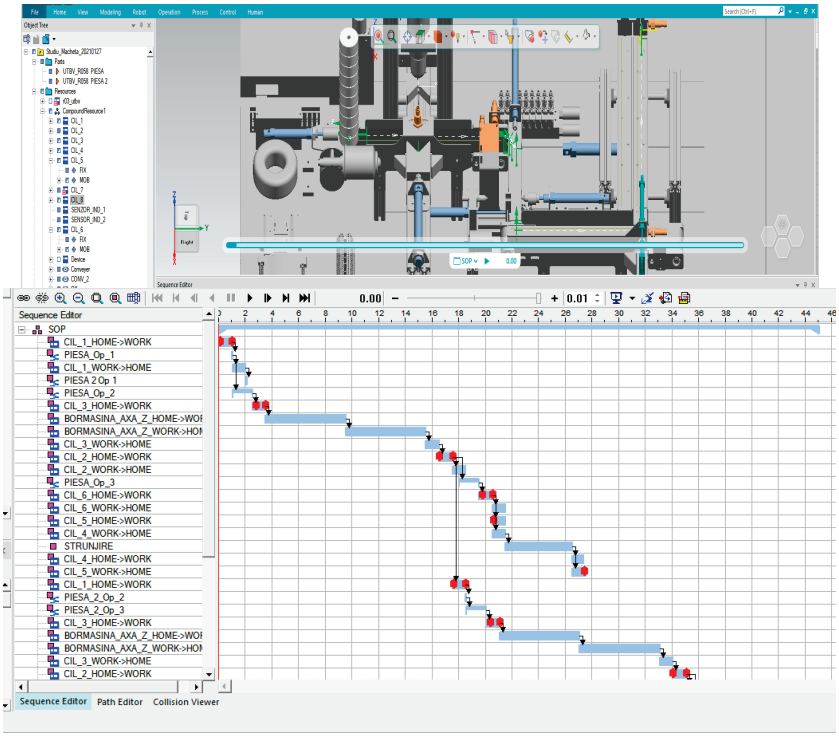


Figure 8. Digital Twin. Time-based simulation—Process Simulate.

Knowing and managing the parameters of each device is necessary to carry out the virtual commissioning phase of the project—experimental FMS.

The time-based simulation led to a number of significant results:

- Everyone involved in the project can have a visual representation of the production line. This results in a better understanding of the process.
- The simulation allowed for the identification of collisions in the system between the devices or equipment that compose it and the possibility of correcting them.
- Several process variants were simulated to choose the optimal solution.

These simulation variants took into account the identification of blockages, functional errors, the recording of production times in different process sequences, the load levels of the workstations, and even the interactions between man and machine or device. The simulation analysis allowed the visualization of statistical data for each workstation and equipment. The appearance of blockages (narrow places, waiting lines) was observed at the unloading workstation, at the exit from the system (Figure 9), the blockage time being very high (81.53%).

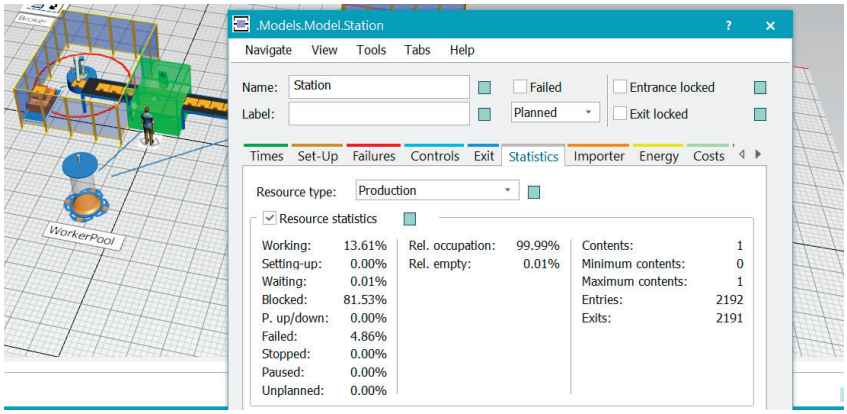


Figure 9. Simulation scenario. Analysis of bottlenecks.

In order to eliminate the asynchronies in the system, respectively the blockages, measures were taken to optimize the operation, for example, arranging a workstation in parallel and integrating a robot on the exit area. Thus, the working times of the system improved (to 95.03%) and it was possible to eliminate the malfunctions that appeared in the system (Figure 10).

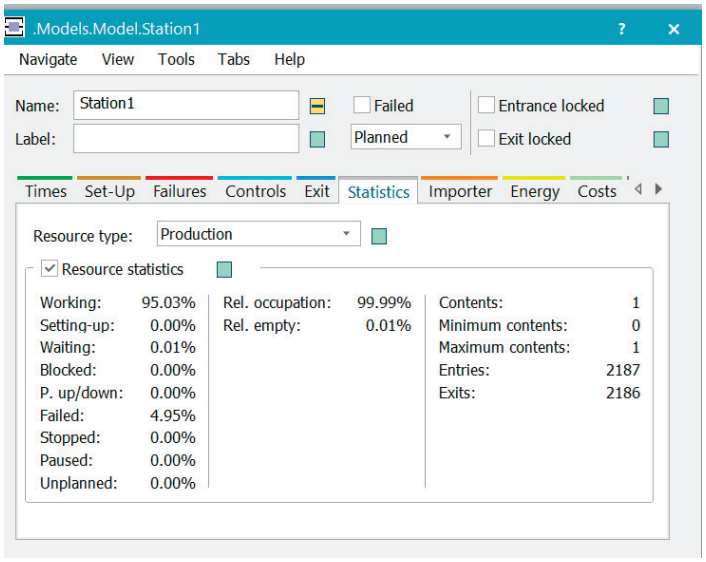


Figure 10. Simulation statistical resources.

Through the created model, the dynamic flows of the process were evaluated, including the manufacturing technological operations, but also the industrial logistics activities—the handling and transport/transfer of semi-finished products.

The grapho-analytical model (Figure 11) developed for the system analyzed by simulation and a digital twin contains a number of optimization parameters and criteria, objective functions, and performance indicators and offers the optimal synthesis in the environment and in real time.

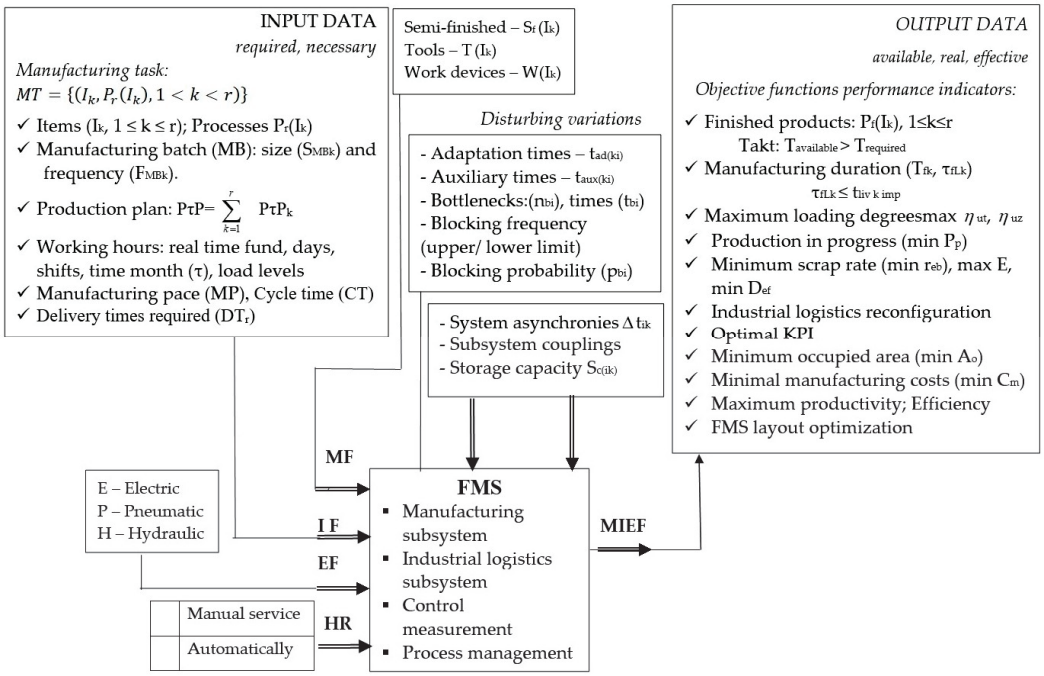


Figure 11. Graphical model of manufacturing system parameters.

In order to carry out the modeling and simulation used in the design and performance analysis of a manufacturing system, the developed manufacturing model [91] was used, which contains the input data, variables and output data, conditions, and restrictions imposed. The system is characterized by inputs–outputs and information links. Functional or organizational blockages, blockage times—t_{bi}, auxiliary times—$t_{aux(ki)}$, adaptation times—$t_{ad(ki)}$, the number of causes of blockages—n_{bi}, the frequency of occurrence of blockages—f_{bi}, the blocking probability—pi, and the size of the asynchronies—Δt_{ik}, have a major influence on the evolution of the system. According to the diagram in Figure 11, the three flows interact in the manufacturing system, material flow—MF, informational flow—IF, and energetic flow—EF, in order to create the finished product, under imposed economic conditions, the human resource—HR for servicing, control, and manufacturing.

3.4. Optimization Solutions

3.4.1. Correcting the Collisions Identified in the System

Following the previously performed time-based simulation analysis, the mechanical interactions between the equipment could be better observed. With the Digital Twin of the system, the ways to avoid and eliminate the identified collisions were analyzed. The DT simulation allowed the identification of the collision between the pneumatic actuation cylinder 4 and a device in the automated system (Figure 12). Procedurally, the correct solution was to reconfigure the 3D model and run the simulation again. Through the data provided by the Digital Twin in real time, constant monitoring was done during the simulation to check if the equipment was in proximity or in collision. When the equipment are at a very small distance from each other, they will be colored yellow (Figure 13), and when the distance reaches (0.00 mm), the equipment will be colored red (Figure 12).

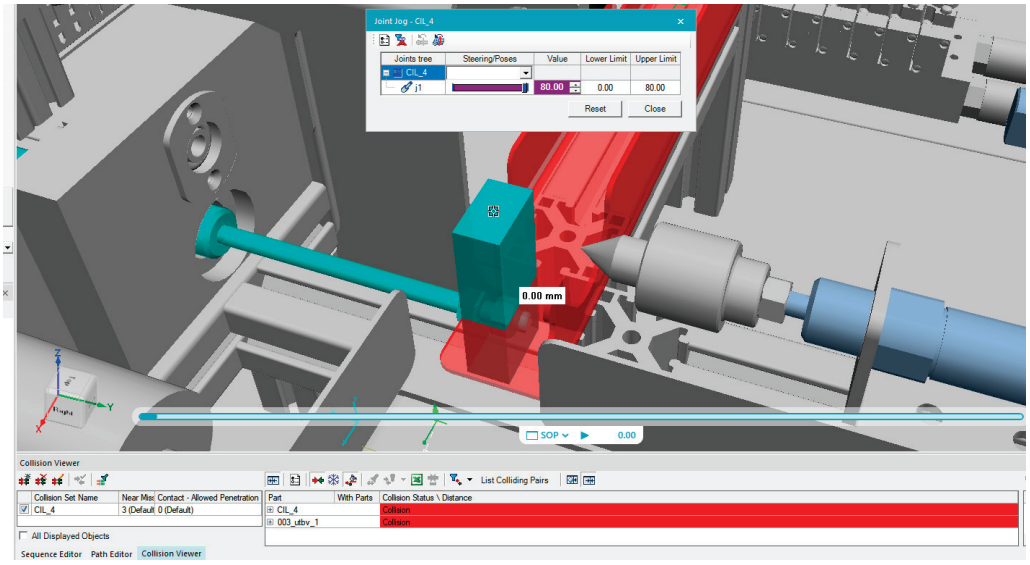


Figure 12. Identification of mechanical interactions: collisions in the system.

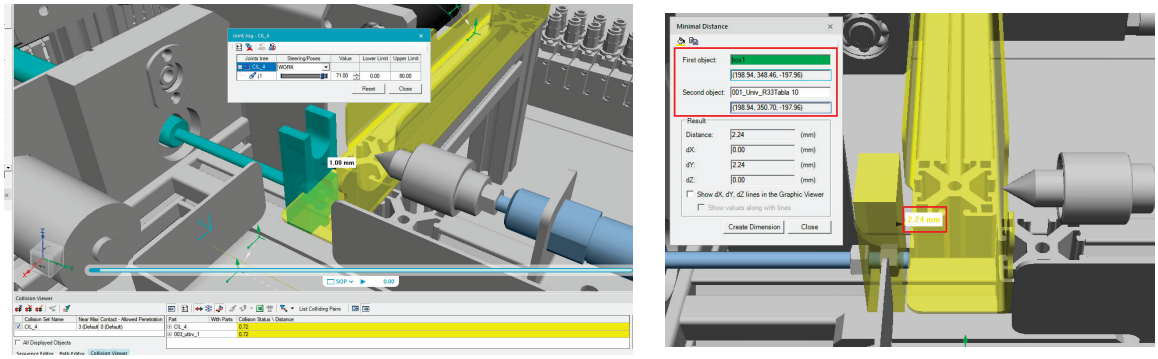


Figure 13. Optimization solution: correcting system collisions.

As an optimization solution for correcting operational errors, (Figure 12), it was possible to limit the stroke of the pneumatic cylinder and redesign the shape of the part at the end of the cylinder rod. At the same time, a validation of the equipment in terms of dimensions was carried out by performing the exact calculation of the minimum distance between the objects of the system.

3.4.2. Configuring the Virtual Commissioning Environment

Process Simulate offers several possibilities for connecting the system to an external control environment for the purpose of virtual commissioning (Virtual Commissioning) [73]: OPC DA/UA (data access/unified access); Siemens PLCSIM Classic; Siemens PLCSIM Advanced; Siemens SIMIT; Siemens Simulation Unit; and WinMod. The path to the OPC Server, the location of the signals on the server, and the type of mapping of the signals between Process Simulate and the PLC will be chosen (the mapping/association can be done according to its name and address). The images in Figure 14 show the sequences from the configuration application for the analyzed case study, the FMS—experimental system designed and built in the research laboratory.

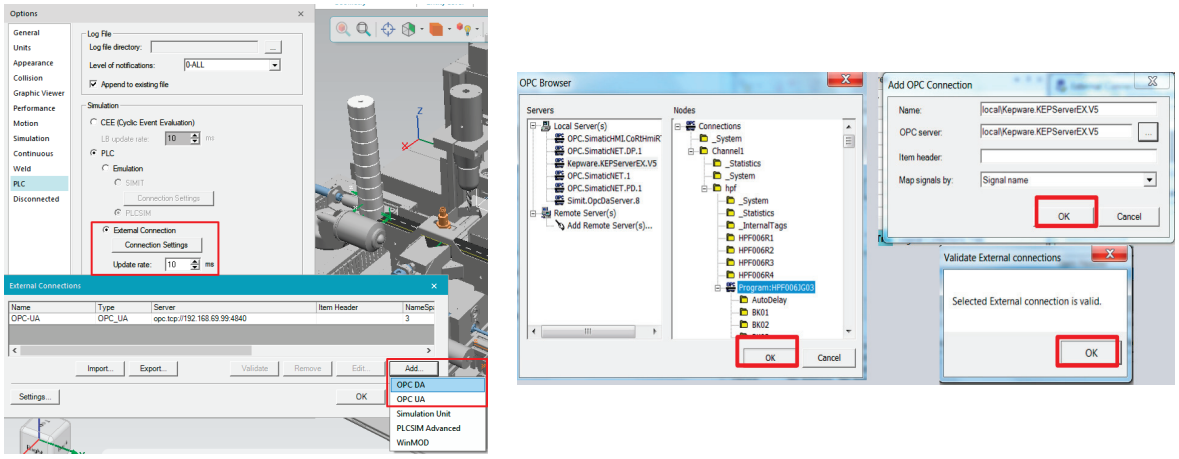


Figure 14. The configuration sequence for the virtual commissioning of the system.

A newly developed research direction has taken shape through the possibilities of adding new elements to an existing process, as well as other existing functionalities in Process Simulate that could become of interest in the future for the digitization of processes within certain companies in the industrial environment.

3.4.3. Integration of Collaborative Robotics

The proposed simulation optimization solutions describe the possibility of adding new elements to the existing process, as well as other existing functionalities in Process Simulate that may become of interest in the future. Several work scenarios were analyzed to determine if the process can be modified, extended, or optimized [74].

The solution to optimize the existing process by introducing the collaborative robot is presented in Figure 15. The interactions between the equipment and the possibility of inserting the robot at the entrance or exit of the FMS were tested. Several scenarios were tested by simulation.

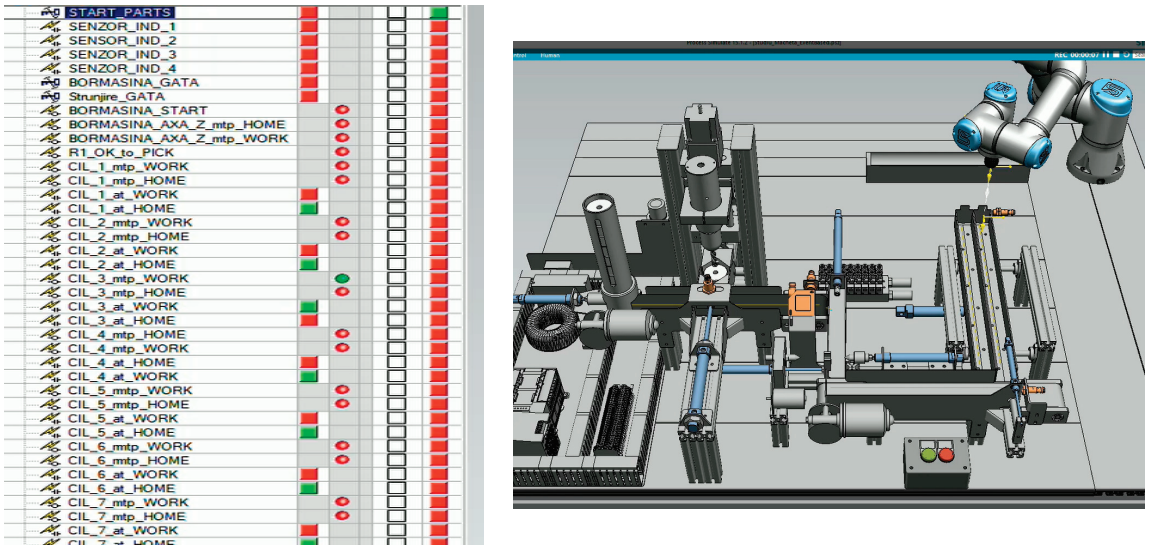


Figure 15. Proposed optimization solutions—collaborative robot.

Through discrete-event simulation, the state of the simulation resources was checked at predetermined intervals (between 10 and 100 ms) (Figure 13). The simulation then runs based on a predetermined sequence of operations, and the triggering events, signal, and state changes necessary to move from one operation to another were identified. Essentially,

every 10 ms, the status of each device, each sensor, the position of objects on conveyor belts, the position of robots, etc., is checked, and all of this information is updated both visually and logically.

An important result of the research is the realization of the DT model, which enabled the development of a digital copy that mirrors reality, where we send commands to actuate the devices and receive constant, real-time information about their state. This is the major difference between regular simulation and DT simulation.

The designed digital model allowed for the testing of the interactions between the equipment and the possibility of introducing the robot at the entrance or exit of the FMS. Integrating the robot at the parts evacuation station at the exit was chosen based on previous simulation variants, which identified bottlenecks and long waiting times (Figure 10).

3.4.4. Human–Machine Interface (HMI)

The DT model, configured through the HMI solution, serves as a foundation for human–machine resource allocation and human operator assistance. Through the "Human" module in the Process Simulate application (Figure 16), several scenarios could be tested, resulting in optimization solutions for the process analyzed in FMS [91]:

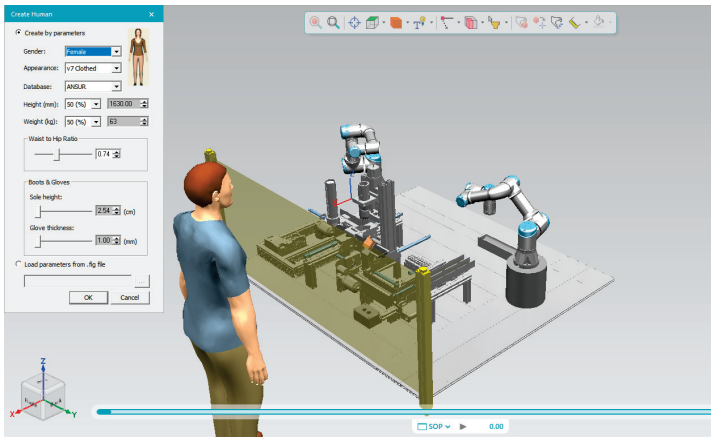


Figure 16. Human–machine interaction: simulation sequence [91].

- Cycle time validation for human operations;
- The testing of safety equipment: light barriers, laser scanners, emergency buttons, etc.;
- Carrying out studies on ergonomics, accessibility to equipment, and tools.

The main measurable indicators in the process relate to downtime during unplanned equipment shutdowns within the system and the frequency of their occurrence. Additionally, the created digital model, together with HMI development, enables real-time monitoring of process parameters, such as failure rates, operational errors, and human errors. Minimizing these factors positively impacts product quality and the safety of human operators.

The operator views the data in real time by connecting to the workstation in the industrial process through a user interface screen. This allows real-time monitoring of process parameters, such as equipment speed and the pressure variation in the pneumatic actuation cylinders, to ensure the correct system operation. The output data can be displayed in the form of reports diagrams, which are useful for operators in the supervision of, and intervention in, the process.

At the same time, the application of HMI technology identifies the training and development needs of operators. The HMI integration solution, as an advanced technology in Industry 4.0 that creates physical and cognitive connections between humans, machines, and intelligent devices, constitutes a future research direction.

3.4.5. Implementation of Automated Guided Vehicle Systems (AGVS)

Transport systems using AGVs serve manufacturing lines where a system based on conveyor belts or conveyors is not an optimal solution. Thus, it was possible to expand the functionality of the process by introducing several workstations and allowing the transport of parts between the stations to be carried out through a system of AGVs.

Figure 17 shows the sequence of the virtual model of the flexible manufacturing system—experimental FMS.

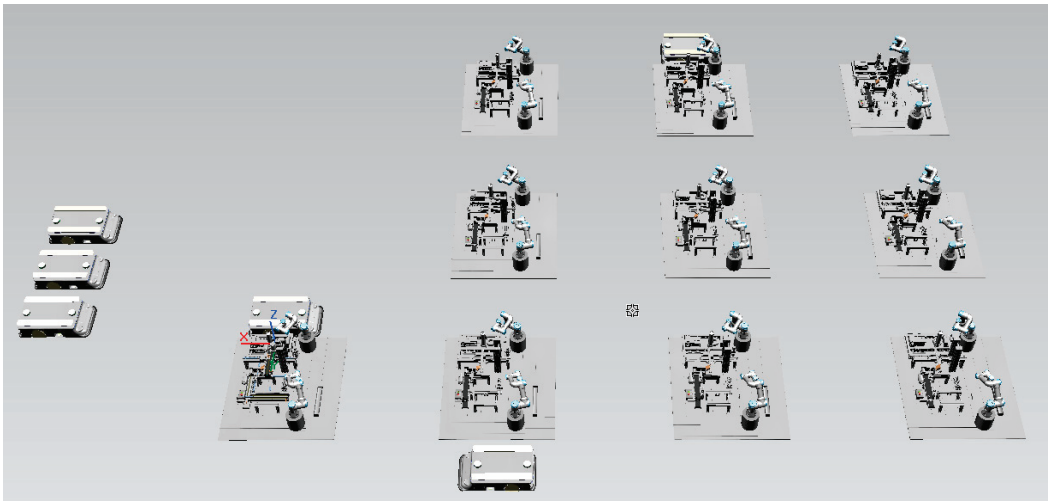


Figure 17. The AGV optimization solution. Simulation scenario sequence.

In this scenario, a series of AGVs supplying the facilities with parts and another series where the completed parts are stored were considered. Robots installed in the vicinity of the facility will be able to transfer parts between the work table and the AGV. Through the digital representation of the real process, it was possible to detect problems in operation and create scenarios through simulation to find the implementation solution, all in real time, thereby reducing the physical time of commissioning.

In this research, all of the resources and technologies needed to implement a new contemporary concept in Industry 4.0, virtual commissioning, have been detailed, resulting in an extended model that can be applied in a real factory and allows the detection of any anomalies in operation before connecting to any real equipment.

4. Conclusions

By integrating into the current flexible production systems, the components and technologies of Industry 4.0, collaborative robots, intelligent machines, additive manufacturing, Digital Twins, and simulation, a new qualitative leap can be made towards the mass customization of products at the request of customers. The direction of applied research on the integration of flexible production systems in Industry 4.0 is part of the current concerns at the international level and is associated with upward development in multinational companies in the country and has allowed the achievement of important results of an applicative nature. Through the analysis of the Digital Twin (DT) concept, a leading simulation concept for Industry 4.0, its applications in the industrial environment could be identified. This resulted in the possibility of applying DT starting from the design, dimensioning, and configuration stages of the system, all the way through to its implementation, commissioning, and operation. Production processes and operational efficiency have been optimized, and new opportunities have emerged for integration into the evolving paradigm of Industry 5.0. A novel element of this research is the realization of the Digital Twin of the analyzed Flexible Manufacturing System, using the Process Simulate software package from Tecnomatix/Siemens, version 15.1.2 [89], in a manufacturing company. An important step has been taken in human resources by training operators to acquire new skills in working with

collaborative robots and human–machine interaction (HMI Concept), thus preparing the company for the transition to Industry 5.0.

Research in the field of implementing a "Digital Twin" in the manufacturing process began by digitizing the experimental system, obtaining results regarding the following:

- Proposing solutions to optimize the process through digitization using Industry 4.0-specific technologies with implications for production and management;
- The implementation of collaborative robots, simulation studies on interactions with the human operator (HMI—human–machine interface), and ergonomic studies;
- The implementation of the AGVS (automated guided vehicle system) concept.

The applied scientific research was carried out through concrete activities, with the final result being the software design of the digital copy: defining devices and equipment based on the experimental physical model and the kinematic realization; defining the sensors; performing time-based simulation; identifying mechanical interactions between equipment and highlighting potential collisions; the optimization solution for avoiding collisions by analyzing several scenarios; realizing the simulation based on discrete events, in which the event sequences validate both the equipment from a constructive point of view and the operation logic of the process; and utilizing the optimization solution by implementing complex, robot-type equipment, where many interactions occur simultaneously.

Expanding the research by disseminating the results and applying the studied concepts and designed models within a manufacturing company, through a scientific research contract, was an immediate priority. A future research direction resulting from the current study is the virtual commissioning of the existing physical experimental system by implementing the new virtual commissioning concept. The result will be an extended model that can be applied in a real factory, allowing for the detection of any anomalies in operation before connecting to any real equipment, as well as the reduction of physical commissioning time, with favorable implications in production management. Further developments were proposed for the short term, through other studies and implementations in several companies based on specific case studies. This research contributes to a technical advance and evolution in the way production systems interact with their economic and social environment, promoting interconnectivity and adaptability in a more applied way.

Funding: This research received no external funding.

Data Availability Statement: The data used to support the findings of the current study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Qin, J.; Liua, Y.; Grosvenor, R. A Categorical Framework of Manufacturing for Industry 4.0 and Beyond. *Procedia CIRP* **2016**, *52*, 173–178. [CrossRef]
2. Chiarello, F.; Trivelli, L.; Bonaccorsi, A.; Fantoni, G. Extracting and mapping industry 4.0 technologies using Wikipedia. *Comput. Ind.* **2018**, *100*, 244–257. [CrossRef]
3. Ghobakhloo, M. Industry 4.0, digitization, and opportunities for sustainability. *J. Clean. Prod.* **2020**, *252*, 119869. [CrossRef]
4. Mourtzis, D.; Angelopoulos, J.; Panopoulos, N. A Literature Review of the Challenges and Opportunities of the Transition from Industry 4.0 to Society 5.0. *Energies* **2022**, *15*, 6276. [CrossRef]
5. Pereira, A.G.; Lima, T.M.; Charrua-Santos, F. Society 5.0 as a result of the technological evolution: Historical approach. *Adv. Intell. Syst. Comput.* **2020**, *1018*, 700–705.
6. Romero, D.; Stahre, J. Towards the resilient operator 5.0: The future of work in smart resilient manufacturing systems. *Procedia CIRP* **2021**, *104*, 1089–1094. [CrossRef]
7. Lu, Y.; Zheng, H.; Chand, S.; Xia, W.; Liu, Z.; Xu, X.; Wang, L.; Qin, Z.; Bao, J. Outlook on human-centric manufacturing towards Industry 5.0. *J. Manuf. Syst.* **2022**, *62*, 612–627. [CrossRef]
8. Almusaed, A.; Yitmen, I.; Almssad, A. Reviewing and Integrating AEC Practices into Industry 6.0: Strategies for Smart and Sustainable Future-Built Environments. *Sustainability* **2023**, *15*, 13464. [CrossRef]

9. Brettel, M.; Klein, K.; Friederichsen, N. The relevance of manufacturing flexibility in the context of Industrie 4.0. In Proceedings of the 48th CIRP Conference on Manufacturing Systems, Ljubljana, Slovenia, 12–14 June 2016; Elsevier: Amsterdam, The Netherlands, 2015; Volume 41, pp. 105–110. [CrossRef]
10. Schuh, G.; Potente, T.; Wesch-Potente, C.; Weber, A.R.; Prote, J. Collaboration Mechanisms to increase Productivity in the Context of Industrie 4.0. In Proceedings of the 2nd CIRP Robust Manufacturing Conference, Bremen, Germany, 7–9 July 2014; Volume 19, pp. 51–56.
11. Pereira, J.A.P.; Campilho, R.D.S.G.; Silva, F.J.G.; Sánchez-Arce, I.J. Robotized cell design for part assembly in the automotive industry. *Proc. Inst. Mech. Eng. Part C* **2022**, *236*, 8807–8822. [CrossRef]
12. Sousa, V.F.C.; Silva, F.J.G.d.; Campilho, R.D.S.G.; Pinto, A.G.; Ferreira, L.P.; Martins, N. Developing a novel fully automated concept to produce bowden cables for the automotive industry. *Machines* **2022**, *10*, 290. [CrossRef]
13. Mishra, R.; Pundir, A.K.; Ganapathy, L. Manufacturing Flexibility Research: A Review of Literature and Agenda for Future Research. *Glob. J. Flex. Syst. Manag.* **2014**, *15*, 101–112. [CrossRef]
14. Florescu, A.; Barabas, S.A. Modeling and Simulation of a Flexible Manufacturing System—A Basic Component of Industry 4.0. *Appl. Sci.* **2020**, *10*, 8300. [CrossRef]
15. Arnold, C.; Kiel, D.; Voigt, K.I. How the Industrial Internet of Things Changes Business Models in Different Manufacturing Industries. *Int. J. Innov. Manag.* **2016**, *20*, 1640015. [CrossRef]
16. Tao, F.; Zhang, H.; Liu, A.; Nee, A.Y.C. Digital Twin in Industry: State-of-the-Art. *IEEE Trans. Ind. Inform.* **2019**, *15*, 2405–2415. [CrossRef]
17. Zheng, T.; Ardolino, M.; Bacchetti, A.; Perona, M.; Zanardini, M. The impacts of Industry 4.0: A descriptive survey in the Italian manufacturing sector. *J. Manuf. Technol. Manag.* **2020**, *31*, 1085–1115. [CrossRef]
18. Veile, J.W.; Kiel, D.; Muller, J.M.; Voigt, K.-I. Lessons learned from Industry 4.0 implementation in the German manufacturing industry. *J. Manuf. Technol. Manag.* **2020**, *31*, 977–997. [CrossRef]
19. Semeraro, C.; Lezoche, M.; Panetto, H.; Dassisti, M. Digital twin paradigm: A systematic literature review. *Comput. Ind.* **2021**, *130*, 103469. [CrossRef]
20. Wang, K.J.; Lee, Y.H.; Angelica, S. Digital twin design for real-time monitoring—A case study of die cutting machine. *Int. J. Prod. Res.* **2021**, *59*, 6471–6485. [CrossRef]
21. Lu, Y.; Liu, Z.; Min, Q. A digital twin-enabled value stream mapping approach for production process reengineering in SMEs. *Int. J. Comput. Integr. Manuf.* **2021**, *34*, 764–782. [CrossRef]
22. Psarommatis, F.; May, G. A literature review and design methodology for digital twins in the era of zero defect manufacturing. *Int. J. Prod. Res.* **2022**, *61*, 5723–5743. [CrossRef]
23. Kusiak, A. Predictive models in digital manufacturing: Research, applications, and future outlook. *Int. J. Prod. Res.* **2022**, *61*, 6052–6062. [CrossRef]
24. Heindl, W.; Stary, C. Structured Development of Digital Twins—A Cross-Domain Analysis towards a Unified Approach. *Processes* **2022**, *10*, 1490. [CrossRef]
25. Perno, M.; Hvam, L. Developing a Framework for Scoping Digital Twins in the Process Manufacturing Industry. In *SPS2020*; Säfsten, K., Elgh, F., Eds.; IOS Press: Amsterdam, The Netherlands, 2020; ISBN 9781643681467.
26. Riedelsheimer, T.; Gogineni, S.; Stark, R. Methodology to develop Digital Twins for energy efficient customizable IoT-Products. *Procedia CIRP* **2021**, *98*, 258–263. [CrossRef]
27. Agnusdei, G.P.; Elia, V.; Gnoni, M.G. Is Digital Twin Technology Supporting Safety Management? A Bibliometric and Systematic Review. *Appl. Sci.* **2021**, *11*, 2767. [CrossRef]
28. Fett, M.; Wilking, F.; Goetz, S.; Kirchner, E.; Wartzack, S. A Literature Review on the Development and Creation of Digital Twins, Cyber-Physical Systems, and Product-Service Systems. *Sensors* **2023**, *23*, 9786. [CrossRef]
29. Singh, M.; Srivastava, R.; Fuenmayor, E.; Kuts, V.; Qiao, Y.; Murray, N.; Devine, D. Applications of Digital Twin across Industries: A Review. *Appl. Sci.* **2022**, *12*, 5727. [CrossRef]
30. Kousi, N.; Gkournelos, C.; Aivaliotis, S.; Lotsaris, K.; Bavelos, A.C.; Baris, P.; Michalos, G.; Makris, S. Digital Twin for Designing and Reconfiguring Human–Robot Collaborative Assembly Lines. *Appl. Sci.* **2021**, *11*, 4620. [CrossRef]
31. Zohdi, T.I. A digital twin framework for machine learning optimization of aerial firefighting and pilot safety. *Comput. Methods Appl. Mech. Eng.* **2021**, *373*, 113446. [CrossRef]
32. Zandi, K.; Ransom, E.H.; Topac, T.; Chen, R.; Beniwal, S.; Blomfors, M.; Shu, J.; Chang, F.-K. A Framework for Digital Twin of Civil Infrastructure—Challenges & Opportunities. *Struct. Health Monit.* **2019**. [CrossRef]
33. Li, W.; Rentemeister, M.; Badeda, J.; Jöst, D.; Schulte, D.; Sauer, D.U. Digital twin for battery systems: Cloud battery management system with online state-of-charge and state-of-health estimation. *J. Energy Storage* **2020**, *30*, 101557. [CrossRef]
34. Delollite Development LLC. 2021 Manufacturing Industry Outlook. Available online: https://www2.deloitte.com/us/en/pages/energy-and-resources/articles/manufacturing-industry-outlook.html (accessed on 18 January 2022).
35. Ulrich, S. *Industrie 4.0—Beherrschung der Industriellen Komplexität mit SysLM*; Springer-Vieweg: Berlin, Germany, 2013.
36. Bauernhansel, T.; ten Hompel, M.; Vogel-Heuser, B. *Industrie 4.0 in Produktion, Automatisierung, und Logistik*; Springer-Vieweg: Berlin, Germany, 2014.
37. Berger, R. *Industry 4.0 The New Industrial Revolution How Europe Will Succeed*; Roland Berger Strategy Consultants GmbH: München, Germany, 2014.

38. Burmeister, C.; Lüttgens, D.; Piller, F.T. Business Model Innovation for Industrie 4.0: Why the Industrial Internet Mandates a New Perspective on Innovation. *Die Unternehm.* **2016**, *70*, 124–152. [CrossRef]
39. Dombrowski, U.; Richter, T. The Lean Production System 4.00 Framework—Enhancing Lean Methods by Industrie 4.0. In *Advances in Production Management Systems. Smart Manufacturing for Industry 4.0, Proceedings of the IFIP International Federation for Information Processing 2018, Seoul, Republic of Korea, 26–30 August 2018*; Springer International Publishing: Basel, Switzerland, 2018; Volume 536, pp. 473–481; ISBN 978-3-319-99706-3.
40. Sony, M. Industry 4.0 and lean management: A proposed integration model and research propositions. *Prod. Manuf. Res.* **2018**, *6*, 416–432. [CrossRef]
41. Schumacher, A.; Nemeth, T.; Sihn, W. Roadmapping towards industrial digitalization based on an Industry 4.0 maturity model for manufacturing enterprises. *Procedia CIRP* **2019**, *79*, 409–414. [CrossRef]
42. Santos, R.C.; Martinho, J.L. An Industry 4.0 maturity model proposal. *J. Manuf. Technol. Manag.* **2020**, *31*, 1023–1043. [CrossRef]
43. Amaral, A. A Framework for Assessing Manufacturing SMEs Industry 4.0 Maturity. *Appl. Sci.* **2021**, *11*, 6127. [CrossRef]
44. Zoubek, M.; Poor, P.; Broum, T.; Basl, J.; Simon, M. Industry 4.0 Maturity Model Assessing Environmental Attributes of Manufacturing Company. *Appl. Sci.* **2021**, *11*, 5151. [CrossRef]
45. Ryalat, M.; ElMoaqet, H.; AlFaouri, M. Design of a Smart Factory Based on Cyber-Physical Systems and Internet of Things towards Industry 4.0. *Appl. Sci.* **2023**, *13*, 2156. [CrossRef]
46. Gupta, B.B.; Tewari, A.; Cvitić, I.; Peraković, D.; Chang, X. Artificial intelligence empowered emails classifier for Internet of Things based systems in Industry 4. *Wirel. Netw.* **2021**, *28*, 493–503. [CrossRef]
47. Culot, G.; Nassimbeni, G.; Orzes, G.; Sartor, M. Behind the definition of Industry 4.0: Analysis and open questions. *Int. J. Prod. Econ.* **2020**, *226*, 107617. [CrossRef]
48. Pei, E.; Loh, G.H.; Nam, S. Concepts and terminologies in 4D printing. *Appl. Sci.* **2020**, *10*, 4443. [CrossRef]
49. Frank, A.G.; Dalenogare, L.S.; Ayala, N.F. Industry 4.0 Technologies: Implementation Patterns in Manufacturing Companies. *Int. J. Prod. Econ.* **2019**, *210*, 15–26. [CrossRef]
50. Vrchota, J.; Volek, T.; Novotná, M. Factors Introducing Industry 4.0 to SMES. *Soc. Sci.* **2019**, *8*, 130. [CrossRef]
51. Gajdzik, B.; Grabowska, S.; Saniuk, S. A Theoretical Framework for Industry 4.0 and Its Implementation with Selected Practical Schedules. *Energies* **2021**, *14*, 940. [CrossRef]
52. Moeuf, A.; Lamouri, S.; Pellerin, R.; Tamayo-Giraldo, S.; Tobon-Valencia, E.; Eburdy, R. Identification of Critical Success Factors, Risks and Opportunities of Industry 4.0 in SMEs. *Int. J. Prod. Res.* **2020**, *58*, 1384–1400. [CrossRef]
53. Lodgaard, E.; Torvatn, H.; Sørumsbrenden, J. Future competence at shopfloor in the era of Industry 4.0—A case study in Norwegian industry. *Procedia CIRP* **2022**, *107*, 961–965. [CrossRef]
54. Grufman, N.; Lyons, S.; Sneiders, E. Exploring Readiness of SMEs for Industry 4.0. *Complex Syst. Inform. Model. Q. CSIMQ* **2020**, *25*, 54–86. [CrossRef]
55. Vinodh, S.; Antony, J.; Agrawal, R.; Douglas, J.A. Integration of continuous improvement strategies with Industry 4.0: A systematic review and agenda for further research. *TQM J.* **2020**, *33*, 441–472. [CrossRef]
56. Schönfuß, B.; McFarlane, D.; Hawkridge, G.; Salter, L.; Athanassopoulou, N.; de Silva, L. A catalogue of digital solution areas for prioritising the needs of manufacturing SMEs. *Comput. Ind.* **2021**, *133*, 103532. [CrossRef]
57. De Sousa Jabbour, A.B.L.; Jabbour, C.J.C.; Foropon, C.; Godinho Filho, M. When titans meet—Can industry 4.0 revolutionise the environmentally-sustainable manufacturing wave? The role of critical success factors. *Technol. Forecast. Soc. Chang.* **2018**, *132*, 18–25. [CrossRef]
58. Fritzsche, K.; Niehoff, S.; Beier, G. Industry 4.0 and Climate Change—Exploring the Science-Policy Gap. *Sustainability* **2018**, *10*, 4511. [CrossRef]
59. Leong, W.D.; Teng, S.Y.; How, B.S.; Ngan, S.L.; Rahman, A.A.; Tan, C.P.; Ponnambalam, S.; Lam, H.L. Enhancing the adaptability: Lean and green strategy towards the Industry Revolution 4.0. *J. Clean. Prod.* **2020**, *273*, 122870. [CrossRef]
60. Sharma, M.; Luthra, S.; Joshi, S.; Kumar, A.; Jain, A. Green logistics driven circular practices adoption in industry 4.0 Era: A moderating effect of institution pressure and supply chain flexibility. *J. Clean. Prod.* **2023**, *383*, 135284. [CrossRef]
61. Adel, A. Future of Industry 5.0 in society: Human-centric solutions, challenges and prospective research areas. *J. Cloud Comput.* **2022**, *11*, 40. [CrossRef]
62. Aslam, F.; Aimin, W.; Li, M.; Ur Rehman, K. Innovation in the era of IoT and industry 5.0: Absolute innovation management (AIM) framework. *Information* **2020**, *11*, 124. [CrossRef]
63. Polonara, M.; Romagnoli, A.; Biancini, G.; Carbonari, L. Introduction of Collaborative Robotics in the Production of Automotive Parts: A Case Study. *Machines* **2024**, *12*, 196. [CrossRef]
64. Borboni, A.; Reddy, K.V.V.; Elamvazuthi, I.; AL-Quraishi, M.S.; Natarajan, E.; Azhar Ali, S.S. The Expanding Role of Artificial Intelligence in Collaborative Robots for Industrial Applications: A Systematic Review of Recent Works. *Machines* **2023**, *11*, 111. [CrossRef]
65. Li, C.; Zheng, P.; Yin, Y.; Pang, Y.M.; Huo, S. An AR-assisted deep reinforcement learning-based approach towards mutual-cognitive safe human-robot interaction. *Robot. Comp. Integr. Manuf.* **2023**, *80*, 102471. [CrossRef]
66. Prassida, G.F.; Asfari, U. A conceptual model for the acceptance of collaborative robots in industry 5.0. *Procedia Comp. Sci.* **2022**, *197*, 61–67. [CrossRef]

67. Kopp, T.; Baumgartner, M.; Kinkel, S. Success factors for introducing industrial human-robot interaction in practice: An empirically driven framework. *Int. J. Adv. Manuf. Technol.* **2021**, *112*, 685–704. [CrossRef]
68. Vanderborght, B. *Unlocking the Potential of Industrial Human–Robot Collaboration: A Vision on Industrial Collaborative Robots for Economy and Society*; Publications Office of the European Union: Luxembourg, 2020.
69. Negri, E.; Fumagalli, L.; Macchi, M. A Review of the Roles of Digital Twin in CPS-based Production Systems. *Procedia Manuf.* **2017**, *11*, 939–948. [CrossRef]
70. Gejo-García, J.; Reschke, J.; Gallego-García, S.; García-García, M. Development of a System Dynamics Simulation for Assessing Manufacturing Systems Based on the Digital Twin Concept. *Appl. Sci.* **2022**, *12*, 2095. [CrossRef]
71. Rahmani, R.; Jesus, C.; Lopes, S.I. Implementations of Digital Transformation and Digital Twins: Exploring the Factory of the Future. *Processes* **2024**, *12*, 787. [CrossRef]
72. Segovia, M.; Garcia-Alfaro, J. Design, Modeling and Implementation of Digital Twins. *Sensors* **2022**, *22*, 5396. [CrossRef] [PubMed]
73. Rasheed, A.; San, O.; Kvamsdal, T. Digital Twin: Values, Challenges and Enablers from a Modeling Perspective. *IEEE Access* **2020**, *8*, 21980–22012. [CrossRef]
74. Bilberg, A.; Malik, A.A. Digital twin driven human–robot collaborative assembly. *CIRP Ann.* **2019**, *68*, 499–502. [CrossRef]
75. Kritzinger, W.; Karner, M.; Traar, G.; Henjes, J.; Sihn, W. Digital Twin in manufacturing: A categorical literature review and classification. *IFAC-PapersOnLine* **2018**, *51*, 1016–1022. [CrossRef]
76. Wagg, D.; Worden, K.; Barthorpe, R.; Gardner, P. Digital Twins: State-of-The-Art Future Directions for Modelling and Simulation in Engineering Dynamics Applications. *ASCE-ASME J. Risk Uncertain. Eng. Syst. Part B Mech. Eng.* **2020**, *6*, 030901. [CrossRef]
77. Psarommatis, F.; May, G. A standardized approach for measuring the performance and flexibility of twins. *Int. J. Prod. Res.* **2022**, *61*, 6923–6938. [CrossRef]
78. Huang, H.; Yang, L.; Wang, Y.; Xu, X.; Lu, Y. Digital Twin-driven online anomaly detection for an automation system based on edge intelligence. *J. Manuf. Syst.* **2021**, *59*, 138–150. [CrossRef]
79. Piromalis, D.; Kantaros, A. Digital Twins in the Automotive Industry: The Road toward Physical-Digital Convergence. *Appl. Syst. Innov.* **2022**, *5*, 65. [CrossRef]
80. Leng, J.; Wang, D.; Shen, W.; Li, X.; Liu, Q.; Chen, X. Digital twins-based smart manufacturing system design in Industry 4.0: A review. *J. Manuf. Syst.* **2021**, *60*, 119–137. [CrossRef]
81. Attaran, M.; Celik, B.G. Digital Twin: Benefits, use cases, challenges, and opportunities. *Decis. Anal. J.* **2023**, *6*, 100165. [CrossRef]
82. Wang, K.; Wang, Y.; Li, Y.; Fan, X.; Xiao, S.; Hu, L. A review of the technology standards for enabling digital twin. *Digit. Twin* **2022**, *2*, 4. [CrossRef]
83. Fera, M.; Greco, A.; Caterino, M.; Gerbino, S.; Caputo, F.; Macchiaroli, R.; D'Amato, E. Towards digital twin implementation for assessing production line performance and balancing. *Sensors* **2020**, *20*, 97. [CrossRef]
84. Malik, A.A.; Masood, T.; Bilberg, A. Virtual reality in manufacturing: Immersive and collaborative artificial-reality in design of human-robot workspace. *Int. J. Comput. Integr. Manuf.* **2020**, *33*, 22–37. [CrossRef]
85. Makris, S. *Cooperating Robots for Flexible Manufacturing*; Springer: Berlin/Heidelberg, Germany, 2020.
86. Koesters, A.; Koetz, F.; Bock, M.; Fett, M.; Breimann, R.; Kirchner, E. Methodical Development of a Digital Twin for an Industry Valve. *Machines* **2024**, *12*, 674. [CrossRef]
87. Resman, M.; Protner, J.; Simic, M.; Herakovic, N. A Five-Step Approach to Planning Data-Driven Digital Twins for Discret Manufacturing Systems. *Appl. Sci.* **2021**, *11*, 3639. [CrossRef]
88. Kaiser, J.; McFarlane, D.; Hawkridge, G.; André, P.; Leitão, P. A review of reference architectures for digital manufacturing: Classification, applicability and open issues. *Comput. Ind.* **2023**, *149*, 103923. [CrossRef]
89. Siemens Tecnomatix—Process Simulate V15.1.2. Available online: https://docs.plm.automation.siemens.com/tdoc/tecnomatix/15.1.2/tecnomatix_eMS/#uid:index (accessed on 7 February 2024).
90. Siemens PLM Virtual Commissioning. Web p. Available online: https://www.plm.automation.siemens.com/en/products/tecnomatix/launch-production/virtual-commissioning (accessed on 22 April 2024).
91. Florescu, A.; Barabas, S. Development Trends of Production Systems through the Integration of Lean Management and Industry 4.0. *Appl. Sci.* **2022**, *12*, 4885. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Near-Zero Parasitic Shift Rectilinear Flexure Stages Based on Coupled n-RRR Planar Parallel Mechanisms

Loïc Tissot-Daguette [1,*], Célestin Vallat [1], Marijn Nijenhuis [2], Florent Cosandier [1] and Simon Henein [1]

[1] Micromechanical and Horological Design Laboratory (Instant-Lab), École Polytechnique Fédérale de Lausanne (EPFL), CH-2000 Neuchâtel, Switzerland; celestin.vallat@epfl.ch (C.V.); florent.cosandier@epfl.ch (F.C.); simon.henein@epfl.ch (S.H.)
[2] Precision Engineering, University of Twente, 7522 NB Enschede, The Netherlands; m.nijenhuis@utwente.nl
* Correspondence: loic.tissot-daguette@epfl.ch

Abstract: Flexure-based linear stages have become prevalent in precision engineering; however, most designs suffer from parasitic shifts that degrade positioning accuracy. Conventional solutions to mitigate these parasitic motions often compromise support stiffness, reduce motion range, and increase structural complexity. This study presents a novel family of flexure-based rectilinear-motion stages using coupled n-RRR planar parallel mechanisms, achieving extremely low parasitic shifts while addressing the forementioned limitations. Four design variants are selected and analyzed via Finite Element Method (FEM) simulations, evaluating parasitic shifts, stroke, and support stiffness. The most precise configuration, a 4-RRR rectilinear stage having kinematic chains coupled via two Watt linkages, exhibits a lateral shift smaller than 0.258 µm and an in-plane parasitic rotation smaller than 12.6 µrad over a 12 mm stroke. Experimental validation using a POM prototype confirms the high positioning precision and support stiffness properties. In addition, a silicon prototype incorporating thermally preloaded buckling beams is investigated to reduce its translational stiffness. Experimental results show a translational stiffness reduction of 98% in the monostable configuration and 112% in the bistable configuration (i.e., negative stiffness), without support stiffness reduction. These results highlight the potential of the proposed mechanisms for a wide range of precision applications, offering a scalable and high-accuracy solution for micro- and nano-positioning systems.

Keywords: compliant mechanisms; flexures; rectilinear translation stage; parasitic shift; support stiffness

1. Introduction

Flexure-based linear stages are frequently employed in precision mechanics to provide translational guiding functions with high precision. The lack of solid friction, clearance, backlash, wear, and their ability for monolithic manufacturing makes them extremely advantageous in micro- and nanopositioning systems [1–4], micro-electromechanical system (MEMS) sensors and actuators [5], micromanipulation [6,7], metrology [8,9], and horological oscillators [10]. It is often required that flexure linear guides describe a motion that is as close as possible to a pure translation. However, most designs exhibit parasitic shifts in addition to the intended rectilinear motion. For planar configurations, this defect is generally quantified by measuring the lateral in-plane deviation of the stage and its in-plane rotation. Most compliant stages, such as the well-known parallel leaf spring stage, have parasitic shifts in second order with respect to the primary displacement, due to the shortening of its flexures [11]. These parasitic motions are frequently insignificant for

small motion amplitudes, but they become a major source of error with larger deflections, degrading positioning accuracy.

Several solutions exist to minimize the parasitic shifts in flexure-based linear stages. One method is to substitute the ideal joints of straight-line linkages by flexure hinges [8,12,13]. Such flexure-based mechanisms can typically exhibit a straightness error smaller than 200 nm over a 35 mm translational stroke [8]. Another way to create flexure rectilinear stages is to compensate for the deflection shortening of the flexures with the motion of the rigid links [14,15]. It is also possible to symmetrize an existing flexure linear stage such that its platform is forced to translate rectilinearly [16,17]. However, such a structure is overconstrained, leading to substantial tensile stress in the flexures, which lowers its range of motion and results in highly nonlinear force–displacement characteristics. Conversely, linear stages can also be underconstrained for rectilinear motion of the platform [1,18]. While offering large strokes, it leads to low lateral stiffness and potentially uncontrolled vibration modes.

The support stiffness over the translational stiffness ratio of flexure-based rectilinear stages usually needs to be high [19]. As a result, the rectilinear motion of the platform is less sensitive to external loads exerted on the mechanism. Since planar structures are considered in this paper, the term support stiffness mainly refers to the in-plane lateral translation stiffness and the in-plane rotation stiffness. The out-of-plane stiffness, being strongly dependent on the structure depth and other factors (e.g., the number and the distribution of parallel kinematic chains in the structure), is the topic of future work.

In this article, a novel family of rectilinear kinematic structures is presented. These mechanisms are constituted of n parallel kinematic chains, each based on three serial revolute joints (RRR), which are rotationally coupled at the base revolute joints to reduce the mobility of the platform to one degree of freedom (DOF) in translation (preprint [20] and patent pending [21]). Since this translation involves very low (or even zero, in some cases) parasitic shifts, we name the members of the family n-RRR rectilinear stages. Their planar kinematics can be implemented using flexures to create fully compliant rectilinear stages. The high support stiffness, resulting from the parallel kinematics, implies high positioning accuracy in the presence of loads in their support directions and facilitates their dynamic control (high undesired eigenfrequencies). Both large strokes and highly linear force–displacement characteristics can be accomplished, because these structures can be made kinematically isostatic (exact constrained). Fortunately, the mechanisms can be compact and symmetrical to reduce possible sensitivity to temperature gradients [22].

The main contributions of this work are summarized as follows:

1. Presentation of a new family of flexure-based mechanisms that achieve rectilinear motion with minimum parasitic shifts.
2. Performance study based on Finite Element Method (FEM) models to evaluate and compare different configurations within the family, in terms of motion range, parasitic motion, and support stiffness.
3. Validation of the models by experiments performed on a mesoscale plastic mockup of one of the design variants.
4. Investigation of integrating buckled beams in a 4-RRR rectilinear stage silicon prototype to effectively reduce its stiffness along the primary degree of freedom without compromising support stiffness.

The proposed designs offer an innovative solution to the limitations of existing flexure-based mechanisms, paving the way for improved performance in high-precision applications that require large strokes and minimized parasitic shift properties. It is also possible to use such stages in series to create a multi-degrees-of-freedom stage with minimum cross-axis coupling errors [12], or redundant dual-stages with minimal interference behavior [23].

Sections 2 and 3 present the kinematics of this new family of mechanisms and their flexure implementations, respectively. An analytical model is derived in Section 4 to help in designing such flexure rectilinear mechanisms. Section 5 introduces the FEM modeling used to compare the performance of the selected design variants within the family. Section 6 presents the experimental method carried out on a plastic mockup to verify the parasitic motions and the stiffness characteristics. The analytical, FEM, and experimental results are compared in Section 7. An application example is described in Section 8, where buckling beams are incorporated into a silicon prototype for a reduction in the translational stiffness. Concluding remarks and future directions for this research are provided in Section 9.

2. Kinematics of the n-RRR Rectilinear Stage Family

The rectilinear stages presented in this article are based on n-RRR planar parallel manipulators with kinematic chain coupling to reduce the mobility of the platform to 1 DOF in translation. Depending on the number of kinematic chains n, the individual architectures are then given names such as TRIOSTAGE ($n = 3$) and QUADRISTAGE ($n = 4$). Based on these particular cases, it is straightforward to conceive other rectilinear stages of the same family having other values of n. Depending on the linkages used to couple kinematic chains, the order of the parasitic shift with respect to the primary displacement is modified. We consider three types of couplers as follows: parallelogram linkage (Type I), Watt's linkage (Type II), and Kempe's Reversor linkage (Type III). Note that it is possible to conceive rectilinear stages having different types of couplers at once; we call them hybrid rectilinear stages. The name of each mechanism topology is based on the type of coupler (PARA-, WATT-, and KEMPE-) and the number of kinematic chains (TRIO- and QUADRI-).

In Section 2.1, the kinematics of the rectilinear stage family are discussed on the basis of the generic Type 1 TRIOSTAGE. In Section 2.2, several design variants are presented. In Section 2.3, a selection of particular designs for further analysis is presented.

2.1. Generic Type I TRIOSTAGE

2.1.1. Topology

A Type I TRIOSTAGE is presented in Figure 1. It is composed of a mobile platform (PF) connected via three main revolute joints (A_1, A_2, and A_3) to three main rigid links (ML1, ML2, and ML3). These three main links are connected via three intermediate revolute joints (B_1, B_2, and B_3) to three secondary rigid links (SL1, SL2, and SL3), respectively. The three secondary links are connected to the fixed base via three secondary revolute joints (C_1, C_2, and C_3). The secondary links SL1 and SL2 are connected to each other via a rigid coupling link CL1 and two coupling revolute joints D_{11} and D_{12}. The secondary links SL2 and SL3 are connected to each other via a rigid coupling link CL2 and two coupling revolute joints D_{21} and D_{22}.

Remark 1. *The mobile platform (PF) and the fixed base are functionally interchangeable parts in all rectilinear stage mechanisms, which means that switching them does not affect the mechanism's kinematic characteristics.*

Remark 2. *A Type I TRIOSTAGE has exactly the same topology as a Type I TRIOVOT (described in [24]). However, since the geometric conditions are different (Section 2.1.2), a Type I TRIOSTAGE is a rectilinear stage with near-zero parasitic shift and a Type I TRIOVOT is a pivot with zero parasitic shift.*

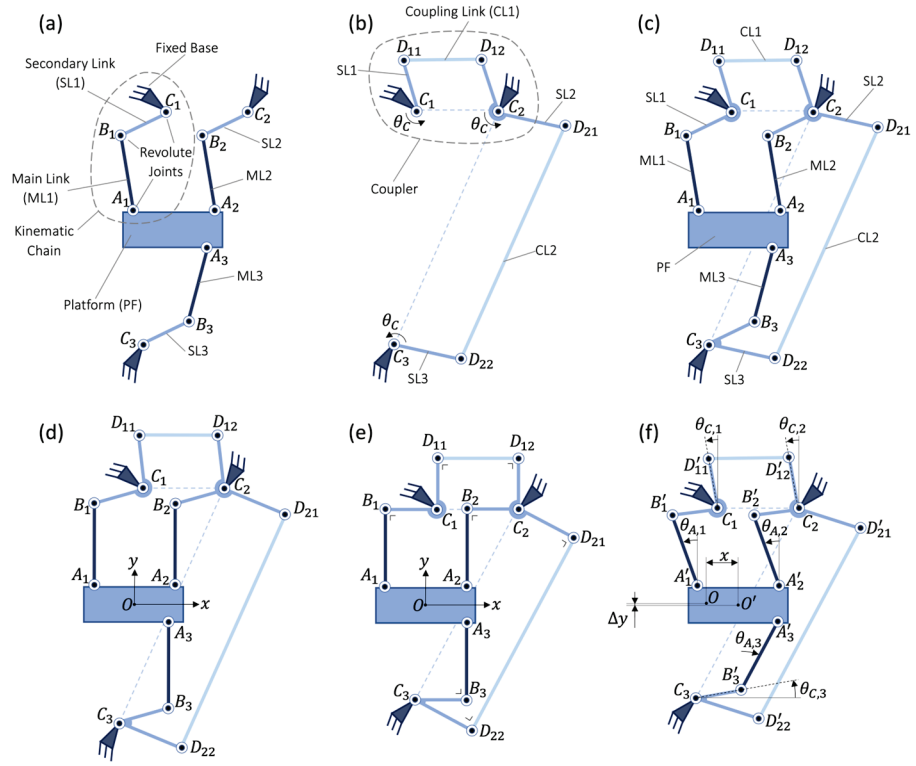


Figure 1. Type I TRIOSTAGE. (**a**) 3-RRR sub-mechanism (3-DOF). (**b**) Coupling sub-mechanism (1-DOF). (**c**) Type I TRIOSTAGE satisfying exclusively Conditions 1 to 3 in an arbitrary position. (**d**) Mechanism (**c**) in neutral position. (**e**) Type I TRIOSTAGE satisfying all Conditions 1 to 6 in neutral position (named PARA-TRIOSTAGE). (**f**) Mechanism (**e**) translated along the x-axis.

2.1.2. Geometric Conditions

Type I TRIOSTAGE is defined by the topology given in Section 2.1.1 complying with the following conditions:

Condition 1. *In neutral position, the triangles $A_1B_1C_1$ and $A_2B_2C_2$ are congruent and have the same orientation.*

Condition 2. *In neutral position, the triangles $A_2B_2C_2$ and $A_3B_3C_3$ are congruent and rotated 180 degrees with respect to each other.*

Condition 3. *The quadrilaterals $C_1D_{11}D_{12}C_2$ and $C_2D_{21}D_{22}C_3$ are parallelograms.*

2.1.3. Kinematic Properties

The x-axis is defined as the axis perpendicular to the line segment A_1B_1 that passes by the center of the line segment A_2A_3 when the linear stage is in neutral position. The y-axis is perpendicular to the x-axis and passes by the center of the line segment A_1A_2 when the linear stage is in neutral position. The center O is the intersection of the axes x and y. The center O' is the point lying on the platform where the rectilinear stage is deformed coinciding with O in neutral position. The projection of the line segment OO' on the y-axis is called the parasitic translation (Δy).

If Conditions 1 to 3 are respected, Type I TRIOSTAGE has a single DOF which is a primary translation of the mobile platform along x in addition to a small parasitic translation along y. The platform has no in-plane parasitic rotation because $A_2B_2C_2$ remains a translate of $A_1B_1C_1$, even if the mechanism is deformed, thanks to the coupling of kinematic chains 1 and 2.

It is remarkable to note that, as long as singularities are avoided, the mechanism is isostatic, i.e., not under- or overconstrained. Indeed, when the couplers (Figure 1b) are attached to the 3-RRR planar parallel mechanism (Figure 1a), the platform mobility decreases from 3 DOF (all in-plane DOF) to 1 DOF in translation along x (Figure 1c). Each coupler independently blocks 1 DOF of the platform as follows: CL1 prevents the in-plane rotation and CL2 limits the translation along y. Advantageously, ML1, ML2, and ML3 extend perpendicular to the x-axis when the platform is in a neutral position, which maximizes the in-plane support stiffness of the stage.

2.1.4. Particular Cases

Some auxiliary geometric conditions can be added to improve the performance of Type I TRIOSTAGE as follows:

Condition 4. *In neutral position, the lines drawn by C_1B_1 and B_1A_1 (respectively, C_2B_2 and B_2A_2, and C_3B_3 and B_3A_3) are orthogonal. Advantage: When the platform is translated in the vicinity of the neutral position, the rotation amplitude of the secondary links tends to zero. This minimizes the lateral shift Δy and the stiffness constant along x of the mechanism when based on flexures.*

Condition 5. *In neutral position, the line segment $D_{11}D_{12}$ is orthogonal to C_1D_{11} as well as to C_2D_{12} (respectively, $D_{21}D_{22}$ is orthogonal to C_2D_{21} as well as to C_3D_{22}). Advantage: The forces transmitted through the coupling links are minimized. This maximizes the support stiffness along y and around the out-of-plane direction z of the platform when the mechanism is based on flexures.*

Condition 6. *In neutral position, the line segment A_1A_2 is parallel to x, and A_2A_3 is parallel to y. Advantage: This enables the transmission of forces between ML1, ML2 and ML3 through the platform, with minimized distance.*

A Type I TRIOSTAGE satisfying all the listed conditions (1 to 6) is shown in neutral and deformed positions, respectively, in Figure 1e,f. When the platform is displaced along the x-direction, it causes a rotation of secondary links in an anticlockwise direction about their respective frame hinges with an equal angle $\theta_{C,1} = \theta_{C,2} = \theta_{C,3}$, thanks to the angular coupling of the parallelogram linkages (see Figure 1b). In that case, there is no parasitic rotation in the xy-plane since main links ML1 and ML2 rotate equally ($\theta_{A,1} = \theta_{A,2}$). However, the platform has a slight parasitic translation of the platform parallel to the y-axis, because ML2 and ML3 do not rotate equally ($\theta_{A,2} \neq \theta_{A,3}$). Geometrically, the amplitude of this parasitic translation is of fifth order with respect to the primary displacement x:

$$\Delta y = -\frac{1}{8}\frac{1}{r_A^3 r_C}x^5 + \mathcal{O}(x^6) \tag{1}$$

where r_A and r_C are the lengths of the line segments A_iB_i and B_iC_i, respectively, for all kinematic chains i. For comparison, a classic parallel leaf spring stage has a second-order parasitic translation, i.e., $\Delta y \propto x^2$ [11].

2.2. Design Variants

Design alternatives to Type I TRIOSTAGE, where the topology or Conditions 1 to 3 are modified (e.g., supernumerary kinematic chains or coupling links), are presented for their additional benefits. The considered design variants are listed below, with their respective advantages and potential drawbacks:

Mirrored kinematic chain: Condition 2 is modified such that triangles $A_2B_2C_2$ and $A_3B_3C_3$ are congruent and are mirror symmetric with respect to the x-axis.

Watt's linkage coupler (Type II): In order to couple two mirrored kinematic chains, Condition 3 is modified as follows: the coupler formed by $C_2D_{21}D_{22}C_3$ is replaced by a Watt linkage, as shown in Figure 2a,b. This fulfills the coupling of mirrored secondary links SL2 and SL3, forcing them to rotate with approximately equal angle magnitude but in opposite directions. If Condition 5 is satisfied, a Watt linkage coupler transmits the angular displacement between C_2 and C_3 with a fourth-order error about the neutral position of the mechanism:

$$\theta_{C,3} = \theta_{C,2} + \frac{1}{2}\frac{r_D}{w_D}\theta_{C,2}^4 + \mathcal{O}\left(\theta_{C,2}^6\right) \tag{2}$$

where r_D is the length of the line segments C_2D_{21} and C_3D_{22}, and w_D is the length of the line segments $D_{21}D_{22}$. This transmission defect leads to a lateral shift Δy of the platform. If, in addition, Condition 4 is respected, this parasitic shift is minimized and of eighth order with respect to x:

$$\Delta y = \frac{1}{64}\frac{r_D}{w_D}\frac{1}{r_A^4 r_C^3}x^8 + \mathcal{O}(x^{10}) \tag{3}$$

It can be seen that the order of Δy is greatly reduced in comparison to a Type I TRIOSTAGE (Equation (1)). Note also that the advantage of transmitting minimized forces through the coupling links is preserved with Condition 5.

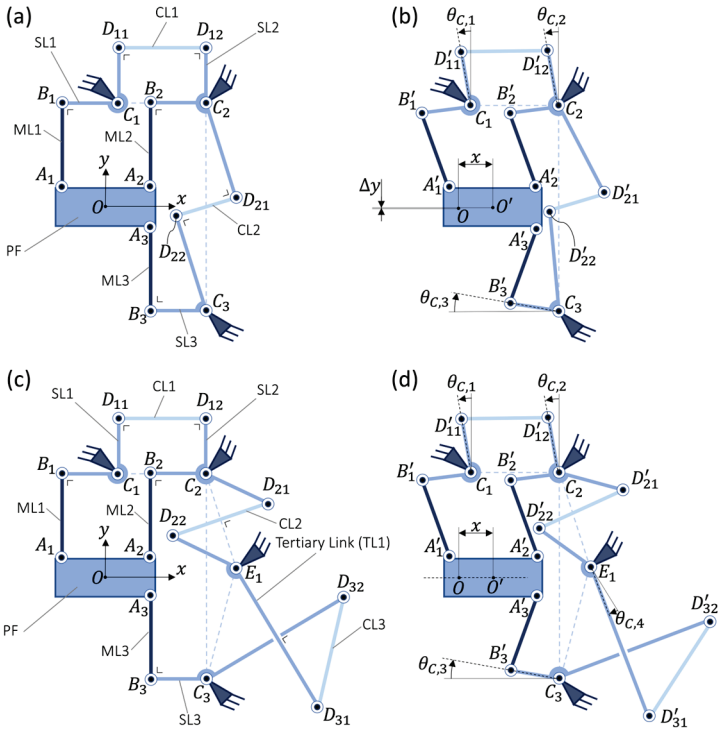


Figure 2. Hybrid TRIOSTAGES. Type I + II TRIOSTAGE: TRIOSTAGE based on parallelogram and Watt's linkage couplers in (**a**) neutral and (**b**) deformed position. Type I + III TRIOSTAGE: TRIOSTAGE based on parallelogram and Kempe's linkage couplers in (**c**) neutral and (**d**) deformed position.

Kempe's linkage coupler (Type III): Instead of using Watt's linkage (see kinematic variant above), it is possible to employ a linkage called Kempe's Reversor [25] to couple two mirrored kinematic chains. This linkage, which is composed of two antiparallelograms $C_2D_{21}D_{22}E_1$ and $E_1D_{31}D_{32}C_3$ (as shown in Figure 2c,d), can perfectly reverse the angular displacement of two pivot joints. Since $\theta_{C,1} = \theta_{C,2} = \theta_{C,3}$, the platform has no parasitic shift in rotation nor in translation (i.e., $\Delta y = 0$). This mechanism is thus a novel perfect straight-line rectilinear mechanism. Condition 5 is modified for Kempe's linkage couplers

as follows: in neutral position, $C_2D_{21}E_1D_{22}$ and $E_1D_{32}D_{31}C_3$ form squares in order to minimize the forces transmitted through the coupling links.

Supernumerary kinematic chains: The number n of kinematic chains ($n = 3$ in the case of TRIOSTAGE structures) can be increased (e.g., $n = 4$ for QUADRISTAGE structures). This can lead to increased load capacity in out-of-plane directions, but also an increased mechanism footprint.

Supernumerary coupling links: The number m of coupling links ($m = 2$ in the case of Type I TRIOSTAGE) can be increased. Note that additional coupling links (i.e., $m > 2$) induce overconstraints but lead to increased load capacity in in-plane directions. In general, at least two couplers must be used, and at least one of these must connect two 180° rotated or two mirrored kinematic chains (in order to block the degree of freedom in the y direction).

2.3. Selected n-RRR Rectilinear Stage Architectures

Based on the generic Type I TRIOSTAGE (Section 2.1) and its kinematic variants (Section 2.2), we select four n-RRR rectilinear stage structures of interest. The performance will be compared when embodied with flexures (Section 3).

We first consider a Type I TRIOSTAGE where Conditions 1 to 6 are fulfilled (Figure 1e,f), which we name PARA-TRIOSTAGE. Then, we design QUADRISTAGE mechanisms (i.e., 4-RRR rectilinear stages) with two symmetrical couplers arranged in pairs. The three different types of couplers are tested, and we call the respective mechanisms as follows: PARA-QUADRISTAGE (based on Type I couplers), WATT-QUADRISTAGE (based on Type II couplers), and KEMPE-QUADRISTAGE (based on Type III couplers). These arrangements enable symmetrical behavior in both translational directions, which is not the case for any TRIOSTAGE. To further reduce the lateral shift Δy at the platform center O', reflection and central symmetries are selected if coupled kinematic chains involve odd and even orders of lateral shift, respectively. The advantages of such symmetry are discussed for each mechanism in Sections 2.3.1–2.3.3.

In this selection, we do not consider n-RRR rectilinear stages with more than $n = 4$ kinematic chains for reasons of compactness.

2.3.1. PARA-QUADRISTAGE

The PARA-QUADRISTAGE (Figure 3) satisfies Conditions 2 to 6 for all its kinematic chains and couplers. Furthermore, its two couplers are arranged with a reflection symmetry around the y-axis in neutral position, see Figure 3a. This kind of symmetry allows the lateral shift Δy, which is of fifth order at the center points of $A'_1A'_4$ and $A'_2A'_3$ (Equation (1)), to be averaged at the center O' such that the order of the lateral shift is increased. However, the counterpart is that the platform has a fifth-order parasitic rotation compared to PARA-TRIOSTAGE which has no parasitic rotation.

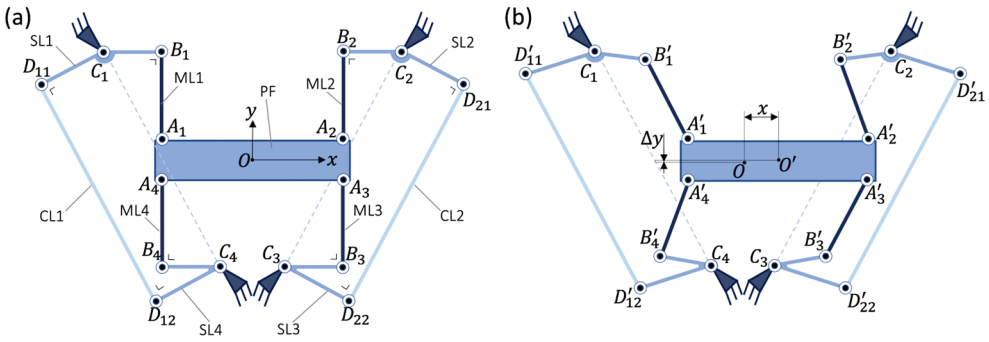


Figure 3. PARA-QUADRISTAGE: 4-RRR rectilinear stage with two parallelogram linkage couplers (**a**) in neutral and (**b**) in deformed position.

2.3.2. WATT-QUADRISTAGE

Conditions 2 to 6 are respected for all the kinematic chains and couplers of the WATT-QUADRISTAGE (where Condition 3 is changed to comply with Type II couplers). Its two couplers are arranged with a central symmetry around O in neutral position; see Figure 4. This kind of symmetry allows the lateral shift, which is of eighth order at the center points of $A'_1A'_4$ and $A'_2A'_3$ (Equation (3)), to be averaged at the center O' to increase the order of the lateral shift. However, the platform exhibits an eight-order parasitic rotation in comparison to the hybrid Type I+II TRIOSTAGE (Figure 2a) which has no parasitic rotation.

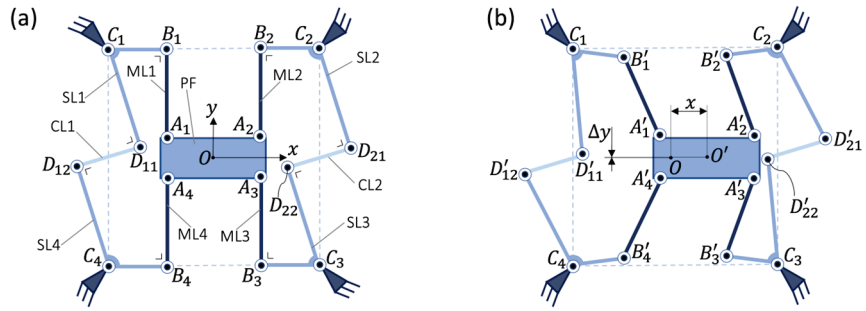


Figure 4. WATT-QUADRISTAGE: 4-RRR rectilinear stage with two Watt's linkage couplers (**a**) in neutral and (**b**) in deformed position.

2.3.3. KEMPE-QUADRISTAGE

Conditions 2 to 6 are respected for all the kinematic chains and couplers of the KEMPE-QUADRISTAGE (where Condition 3 and 5 are modified to comply with Type III couplers). Since Type III couplers do not result in parasitic shift in the platform, the two couplers can be arranged with either reflection or central symmetries in neutral position, without altering the rectilinear motion. It is chosen that the two couplers are mirror symmetric about the y-axis to reduce the size of the mechanism; see Figure 5. As for the Type I + III TRIOSTAGE (Figure 2c), the KEMPE-QUADRISTAGE is a perfect straight-line rectilinear mechanism.

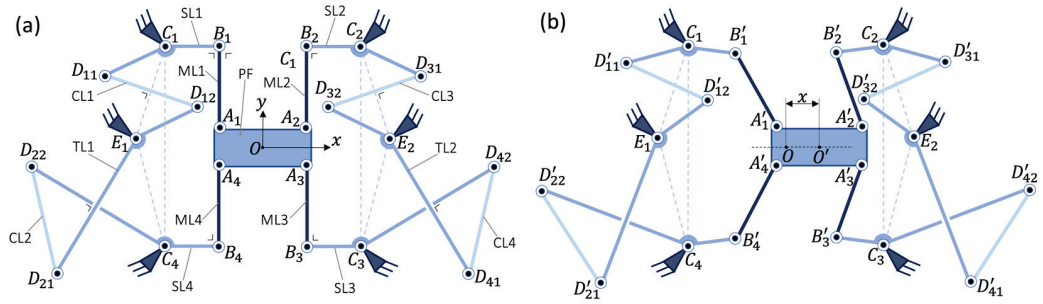


Figure 5. KEMPE-QUADRISTAGE: 4-RRR rectilinear stage with two Kempe's linkage couplers (**a**) in neutral and (**b**) in deformed position.

3. Flexure Implementation

In this section, we provide planar monolithic flexure implementations of the PARA-TRIOSTAGE, PARA-QUADRISTAGE, WATT-QUADRISTAGE, and KEMPE-QUADRISTAGE (Figure 6). For all these rectilinear stages, the revolute joints at the extremities of the main links are each implemented by leaf springs, called main flexures. The base revolute joints are implemented by Remote Center of Compliance (RCC) pivots, called secondary flexures. The revolute joints at the extremities of the coupling links are replaced by leaf springs, called coupling flexures. The center of each leaf spring and the rotation axis of each RCC pivot coincide with the point location of their respective kinematic revolute joints; see Figure 6. The distances r_A, r_B, r_C, w_D, and r_D all geometrically retain the same definition.

The other rigid parts of the flexure pivots keep the same name as their corresponding kinematic architectures illustrated in Figures 1 and 3–5. Filets are placed at the extremities of each flexure to mitigate stress concentration.

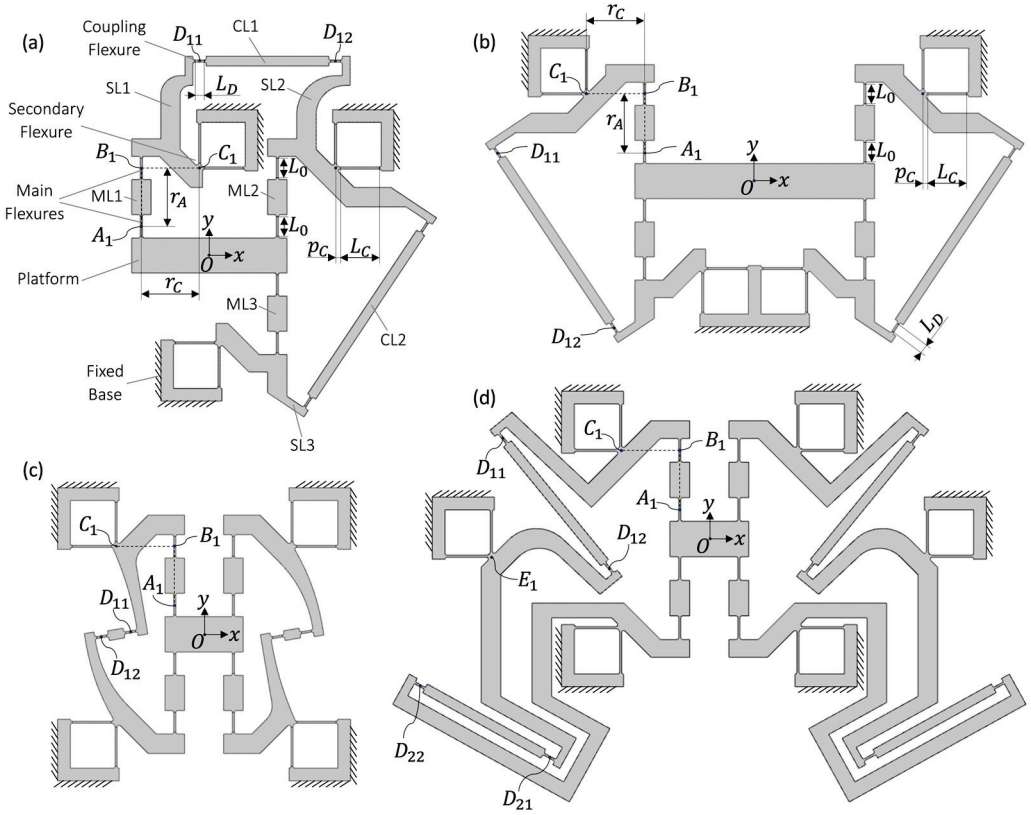


Figure 6. Flexure implementation of (**a**) the PARA-TRIOSTAGE, (**b**) the PARA-QUADRISTAGE, (**c**) the WATT-QUADRISTAGE, and (**d**) the KEMPE-QUADRISTAGE. The flexure rectilinear stages are all presented at equilibrium position (i.e., undeformed).

4. Analytical Model

This section presents models for the flexure-based versions of the selected n-RRR rectilinear stages for the purpose of designing such structures. The models will be validated in Section 7 by comparing the results with FEM and experimental data. For the calculations, we consider that all the kinematic chains of a given rectilinear stage have the same dimensions. To further simplify the design, the main flexures have the same length L_0 and thickness h_0. The platform is supposed to move rectilinearly along the x-axis and sufficiently close to its neutral position. In this case, the rotation angle θ_C of the secondary flexures (and conversely the coupling flexures) can be assumed equal for all the kinematic chains. We compute the rotation magnitude θ_C with respect to the displacement x in Section 4.1. Then the translational stiffness and the admissible stroke of the platform are expressed in Sections 4.2 and 4.3, respectively.

4.1. Deflection of the Secondary and Coupling Flexures

The rotation angle θ_C is related to the axial shortening λ (directed along the y-axis) of the main flexures and the main links; see Figure 7. For small deformations, we have the following approximation:

$$\theta_C \cong \frac{\lambda}{r_C} \tag{4}$$

The value of λ is calculated based on a pseudo-rigid-body model (PRBM) [26] of the main flexures (Figure 7). As it is assumed that the instant rotation axis of each flexure

lies at their center, we select a PRBM radius factor of $\gamma = 3/4$ [27], Equation (35.2). The length of the rigid body is $R = r_A + (2\gamma - 1)L_0 = L_0/2 + r_A$. When the platform deflects by a displacement x (Figure 7b), the rigid body rotates by an angle β, leading to the following shortening:

$$\lambda \cong \frac{1}{2}\frac{x^2}{R} = \frac{x^2}{L_0 + 2r_A} \tag{5}$$

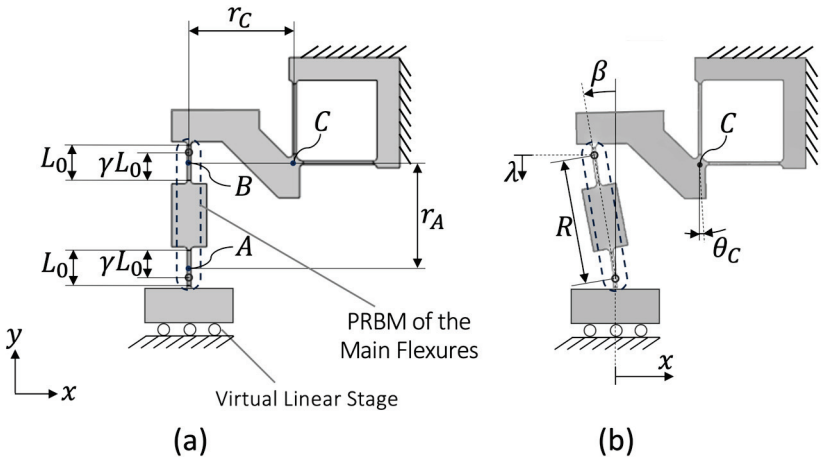


Figure 7. Pseudo-rigid-body model of one of the kinematic chains of the flexure-based n-RRR rectilinear mechanisms in (**a**) neutral position and (**b**) translated by a displacement x.

4.2. Translational Stiffness

When Condition 4 is respected, the stiffness constant K_x is independent of the stiffness of the secondary and coupling flexures. Indeed, the latter rotate at second order with respect to the translation displacement x; see Equations (4) and (5). It can then be noticed that the deformation of each main flexure largely matches that of a so-called four prismatic notch hinge stage [11]. In this case, the translational stiffness of the latter mechanism (derived in [11], Equation (4.70)) is used to evaluate the x-axis stiffness constant K_x of an n-RRR rectilinear stage:

$$K_x = \frac{nEb_0h_0^3}{\xi(3 - 3\xi + \xi^2)L_{\text{tot}}^3} \tag{6}$$

where $L_{\text{tot}} = L_0 + r_A$, $\xi = 2L_0/L_{\text{tot}}$, E is Young's modulus, and b_0 is the out-of-plane depth of the stage.

4.3. Translational Stroke

Since the deformation of the secondary and coupling flexures is significantly smaller than that of the main flexures, it is assumed that the maximum internal stress $\sigma_{\max}$ is located in the main flexures. This assumption can be verified analytically (using Equations (4) and (5)), or through FEM modeling (which is carried out in Section 5). In this case, the admissible stroke also corresponds to that of a four prismatic notch hinge stage, whose formula is provided by [11], Equation (4.71) as:

$$x_{\text{adm}} = \frac{\xi(3 - 3\xi + \xi^2)L_{\text{tot}}^2\sigma_{\text{adm}}}{3Eh_0} \tag{7}$$

where σ_{adm} is the maximum engineering tolerated value of stress that can be applied to the material (e.g., the yield or fatigue strength) while accounting for a safety factor. Likewise, the maximum stress in the mechanism for a given displacement x can be determined by simply inverting Equation (7):

$$\sigma_{\max} = \frac{3Eh_0x}{\xi(3 - 3\xi + \xi^2)L_{\text{tot}}^2} \tag{8}$$

5. Finite Element Models for Performance Comparison

In order to evaluate and compare the performance of the selected flexure rectilinear stages (i.e., the PARA-TRIOSTAGE, the PARA-QUADRISTAGE, the WATT-QUADRISTAGE, and the KEMPE-QUADRISTAGE), we perform FEM simulations. While the optimal design is strongly application-dependent, to provide a basis for comparison of the selected structures, the following parameters are chosen or optimized:

- The rectilinear stages are monolithic and made of the same material: polyoxymethylene (POM) having Young's modulus E of 3 GPa.
- The out-of-plane width b_0 of the mechanisms is 5 mm.
- The thickness of all flexures is 1 mm.
- All the kinematic chains are identical (the dimensions are given in Table 1).
- The coupling flexures have length L_D = 5 mm.
- The length of the coupling links depends on the ideal kinematics of the mechanism.
- The platform is a rectangle with a height (along y) equal to 20 mm.
- The mechanism size is minimized while avoiding collision between the different parts of the mechanism.

Table 1. Design parameters of the selected flexure rectilinear stages (see Figure 6).

	Parameter	Value (mm)
Main Flexures	L_0	10
	r_A	30
Secondary Flexures	p_C	2.5
	L_C	20
	r_C	30
Coupling Flexures	L_D	5

To evaluate the characteristics of the different rectilinear stage mechanisms, three 2D nonlinear static FEM studies were carried out using the Solid Mechanics module of COMSOL Multiphysics 6.2:

Study 1: A displacement is applied along x to the platform while the mechanism base is fixed (Figure 8). The displacement x is varied from -6 mm to 6 mm with equal steps of 0.5 mm. The stiffness constant K_x is evaluated as the linear term of a fifth degree polynomial fitted to the force–displacement FEM data along the x-direction. The polynomial degree is selected to avoid over- and underfitting the data. During this study, we also evaluate the maximum von Mises stress $\sigma_{\max}$ and the parasitic shifts Δy and $\Delta\theta_z$ along the lateral axis y and around the out-of-plane axis z, respectively. Note that the reference frame (x,y) corresponds to the center of the platform O at neutral position. This reference frame remains attached to the stationary base of the mechanism, even when the platform moves; see Figure 8.

Study 2: A lateral displacement of 0.1 mm is applied along y to the platform, while its motion is free in the other directions and the mechanism base is fixed. The reaction force is extracted in order to compute the lateral support stiffness K_y.

Study 3: A rotation of 1 deg is applied around z to the platform, while its motion is free in the other directions and the mechanism base is fixed. The reaction moment is extracted in order to compute the in-plane rotation support stiffness K_{Rz}.

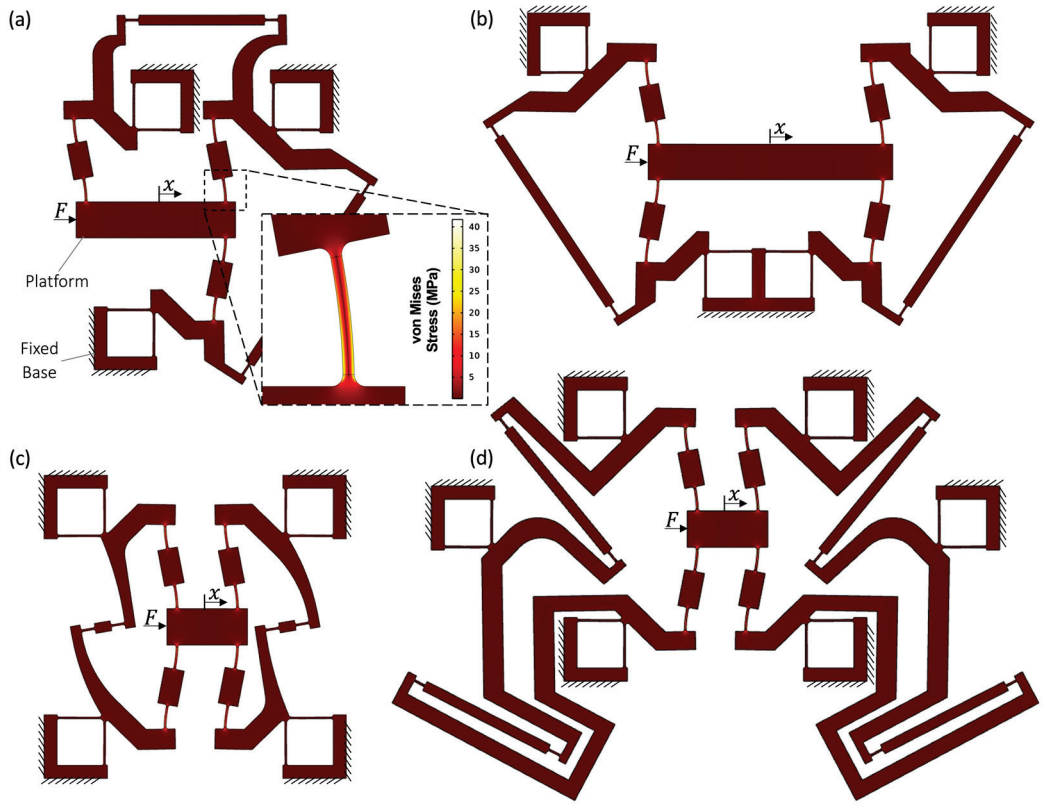


Figure 8. FEM simulations of (**a**) the PARA-TRIOSTAGE, (**b**) the PARA-QUADRISTAGE, (**c**) the WATT-QUADRISTAGE, and (**d**) the KEMPE-QUADRISTAGE, all translated by a displacement of $x = 6$ mm.

For all the studies, we verify that the structures do not buckle to ensure that the evaluated load−deformation characteristics correspond to the linear deformation of the flexures. The flexures are meshed with quadrilateral solid elements such that 4 elements are distributed over their thickness, and 40, 80, and 20 elements are distributed over the length of the main, secondary, and coupling flexures, respectively. The rigid links are meshed with free triangular elements. Plane stress approximation is selected for simulating the 2D FEM model. The mechanism base is fixed at the hatched boundaries illustrated in Figure 8. A rigid connector is attached to the left and right sides of the mechanism platform to apply displacements along x, y, or around z. The FEM results are presented in Section 7.

6. Experimental Validation

In order to validate the performance of the new architectures, an experimental investigation was performed on a mockup of the WATT-QUADRISTAGE. This structure was chosen for its superior motion straightness in comparison to the other selected n-RRR rectilinear structures, as demonstrated by the FEM models (see Section 7.4). The mockup of the WATT-QUADRISTAGE was fabricated in POM using a laser cutting machine (manufacturing tolerances of $\pm$ 0.1 mm). The dimensions of the demonstrator correspond to the values used in the FEM simulations. A test bench was designed to experimentally evaluate the planar parasitic shifts in the platform and, at the same time, the translational and lateral in-plane stiffnesses.

The setup is composed of a manual translation stage used to push the platform by a specific displacement via a rigid connector. A force sensor (Type 9207, Kistler, Winterthur, Switzerland) measures the mechanism's elastic restoring force (Figure 9). The orientation of the mechanism with respect to the setup can be modified in order to measure the restoring force along x or y to evaluate the translational and lateral stiffnesses, respectively. The

mockup plane is perpendicular to the direction of gravity to avoid measuring gravity-related forces.

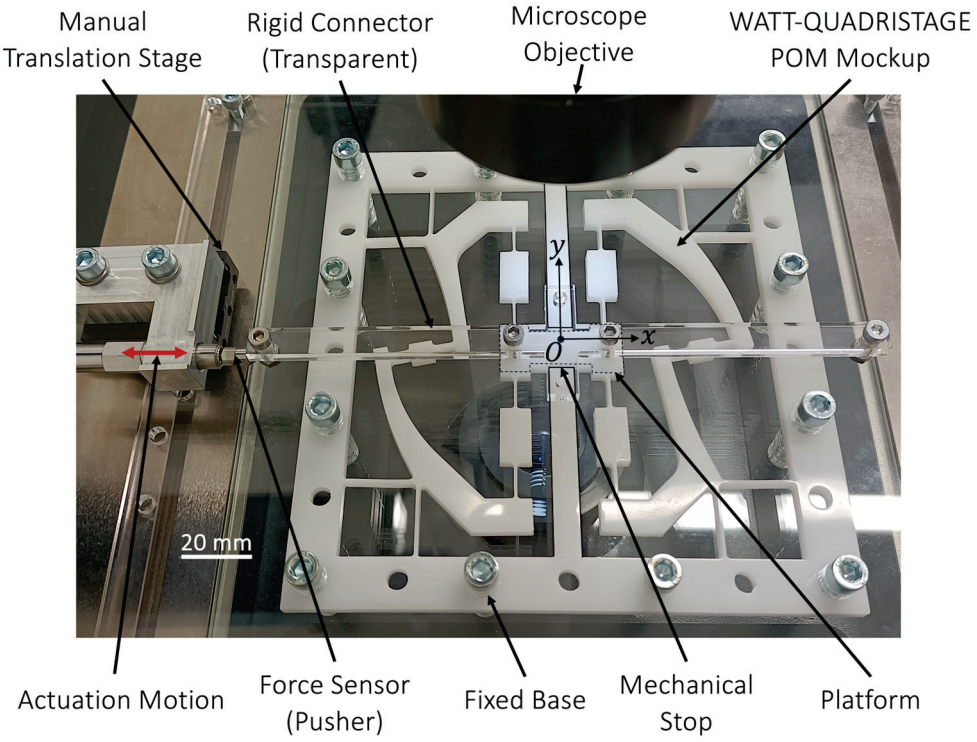


Figure 9. Experimental setup to characterize the parasitic shifts and the stiffness of the WATT-QUADRISTAGE POM mockup.

When the mechanism is displaced along the x-axis, the parasitic shifts in the platform are assessed via an optical microscope (Opensys 300 MZ, Marcel Aubert, Nidau, Switzerland) located above the setup. Two measuring marks are placed on the mockup platform and on the fixed base (serving as reference), in order to measure its translation displacement x, its lateral shift Δy, and its rotation shift $\Delta\theta_z$.

The translational stiffness and the parasitic shifts are evaluated by incremental displacement of 0.5 mm in both directions along x until one of the mechanical stops (located at $x = \pm 6$ mm) is reached by the platform. Since the parasitic shift may be disturbed by contact with the mechanical stops, the measurements are omitted at $x = \pm 6$ mm. The support stiffness is measured when the mechanism is translated along y by applying a fixed displacement of 0.1 mm. The measurement uncertainty is assumed to be ± 0.02 N for the measured forces, ± 1 µm for the displacements x and Δy, and ± 20 µrad for the rotation shift $\Delta\theta_z$. The measurements are always performed when the system is static. Five repetitions were performed to average the measurements.

7. Results and Discussion

In this section, the FEM performance results of the selected n-RRR rectilinear stage family are compared in terms of translational and support stiffnesses, parasitic translation and rotation shifts, and internal stress. The analytical model is verified by comparing it to the FEM results (Tables 2 and 3). The experimental results of the POM mockup are also presented in this section to physically demonstrate the performance of the new family of rectilinear flexure stages and to validate the models.

Table 2. Translational and support stiffnesses of the selected flexure rectilinear stages, obtained from the analytical model (AM), the simulations (FEM), and the experiment (EXP).

	Translational Stiffness			**Support Stiffness**		
	K_x **(N/mm)**			K_y **(N/mm)**		K_{Rz} **(Nm/rad)**
	AM	**FEM**	**EXP**	**FEM**	**EXP**	**FEM**
PARA-TRIOSTAGE	0.804	0.692	-	62.2	-	164
PARA-QUADRISTAGE	1.07	0.917	-	195	-	588
WATT-QUADRISTAGE	1.07	0.922	0.928	127	196	29.5
KEMPE-QUADRISTAGE	1.07	0.921	-	49.7	-	12.0

Table 3. Maximum stress and maximum parasitic shift magnitude of the selected flexure rectilinear stages, obtained from the analytical model (AM), the simulations (FEM), and the experiment (EXP).

	Maximum Stress		**Maximum Parasitic Shift**			
	$\sigma_{\max}$ **(MPa)**		$\Delta y_{\max}$ **(μm)**		$\Delta\theta_{z,\max}$ **(μrad)**	
	AM	**FEM**	**FEM**	**EXP**	**FEM**	**EXP**
PARA-TRIOSTAGE	38.6	41.7	36.5	-	742	-
PARA-QUADRISTAGE	38.6	41.7	2.04	-	221	-
WATT-QUADRISTAGE	38.6	43.3	0.258	2.82	12.6	50.4
KEMPE-QUADRISTAGE	38.6	42.6	8.78	-	74.8	-

7.1. Translational Stiffness

Table 2 shows that the translational stiffness K_x is nearly equal for the mechanisms having four kinematic chains (i.e., all QUADRISTAGE mechanisms), because their kinematic chains all have identical dimensions. In contrast, the PARA-TRIOSTAGE has a translational stiffness which is 25% lower because it has three kinematic chains instead of four. This rectilinear stage can thus be actuated with less force in comparison to the QUADRISTAGE mechanisms. It can also be noticed that the stiffness constant K_x is always lower in the FEM models by ~14% with respect to the analytical model. This discrepancy could be attributed to the deformation of the rigid links, including the filets located at the flexure extremities, which are not considered in the analytical model. This assumption was confirmed by an additional FEM simulation of the mechanisms, in which every component that is not a flexure was modeled as a rigid component with infinite stiffness. In that case, the simulated values of K_x are equal to the ones of the analytical model with less than 0.4% relative variation for all the selected mechanisms.

The stiffness constant K_x predicted by the analytical and FEM models is verified experimentally with the WATT-QUADRISTAGE POM mockup. A stiffness of $K_x = 0.928$ N/mm is obtained by linear regression of its measured force–displacement characteristics. This result is in good agreement with the FEM data, since the relative difference is lower than 1%.

7.2. Support Stiffness

The FEM results show that the mechanism with the maximum values of support stiffnesses along y and around z is the PARA-QUADRISTAGE (Table 2). This shows that the parallelogram couplers can efficiently transmit forces from one secondary link to another in order to provide higher support stiffness.

The values K_y and K_{Rz} of the PARA-TRIOSTAGE are 3.1 and 3.6 times lower in comparison to the PARA-QUADRISTAGE, respectively. This illustrates that, in contrast to

a same-orientation coupling (Condition 1), the kinematic chains with a 180-degree relative orientation (Condition 2) can yield larger support stiffnesses.

The WATT-QUADRISTAGE is 1.5 and 20 times less stiff along y and around z in comparison to the PARA-QUADRISTAGE. It is highly probable that the support stiffness of the WATT-QUADRISTAGE could be increased if the distance C_1D_{11} and the distance A_1A_2 are increased. Indeed, these modifications would imply a reduction in the force transmitted through the coupling links when loads along y or around z are applied.

The mechanism with the lowest values for K_y and K_{Rz} is the KEMPE-QUADRISTAGE architecture. This low support stiffness could be explained by the geometry of the relatively large secondary and coupling connections (required to ensure a planar structure) compared with those in the other structures (Figure 8). The large bodies introduce additional compliance, reducing coupling efficiency and hence support stiffness.

A support stiffness of $K_y = 196$ N/mm is measured on the WATT-QUADRISTAGE prototype. This result is in a large discrepancy with the FEM data, since the measured value is 35% higher than FEM value. This can be explained by the poor manufacturing tolerances of the laser-cut POM mockup which resulted in flexure thickness or length being noticeably greater or lower than anticipated.

Remark 3. *Since this paper aims to provide a basic performance comparison of the presented mechanisms, their support stiffness is only evaluated when they are in neutral position (i.e., undeflected). However, it should be understood that the support stiffness of flexure-based mechanisms usually tends to decrease as the mechanism deformation increases, due to geometric stiffness nonlinearities [19,28]. Characterizing the support stiffness of n-RRR rectilinear stages in deflected state is one of the future research directions.*

7.3. Maximum Stress

The analytical and FEM values of the maximum stress for a range of motion of $x = \pm 6$ mm are reported in Table 3. In the FEM models, the maximum von Mises stress is situated in the main flexures of the mechanisms (Figure 8) and is obtained at the maximum amplitude (i.e., at $x = \pm 6$ mm), as expected. Since all mechanisms have identical main flexure dimensions, the analytical model and the FEM model give, respectively, equal and nearly equal maximum stress values (Table 3). The analytical value is consistently slightly lower than the simulated value. This is primarily due to the stress concentration located near the filets. In addition, the analytical model does not take into account the tension in the main flexures that is produced by the deformation of the secondary and coupling flexures.

Since the simulated maximum stress value never exceeds the yield strength of the material ($\sigma_y \cong 65$ MPa), we can predict that the admissible range of motion x_{adm} of the mechanisms is higher than 6 mm. When testing the POM mockup of the WATT-QUADRISTAGE, we observed that the mechanism can indeed be deflected to this stroke and return to its neutral position without noticeable plastic deformation.

7.4. Parasitic Shifts

The parasitic shift in the different mechanisms is shown in Figure 10 as a function of the displacement x. For the total range of x, the maximum absolute values of the lateral and rotation shifts, i.e., Δy_{max} and $\Delta\theta_{z,\text{max}}$, are reported in Table 3. High rectilinear-motion performance is observed across the different mechanisms, although some perform better than others.

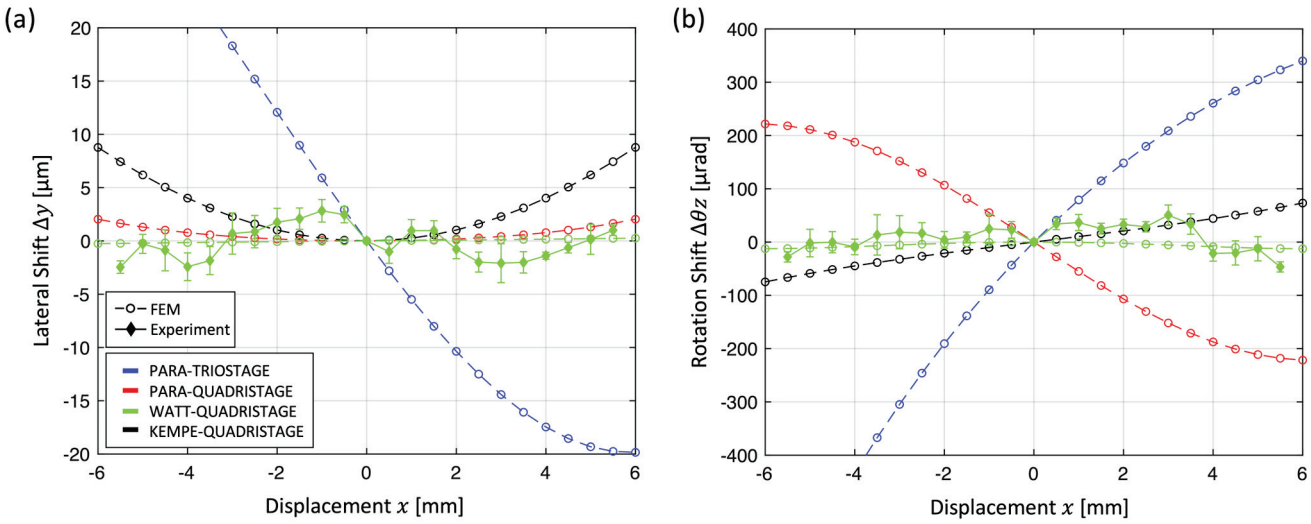


Figure 10. FEM and experimental results of the selected rectilinear stages: (**a**) lateral and (**b**) rotational parasitic shifts as a function of the translational displacement x.

In comparison to the other mechanisms having four kinematic chains, the PARA-TRIOSTAGE (with three kinematic chains) shows the lowest rectilinear-motion performance. Indeed, the values of $\Delta y_{\max}$ and $\Delta\theta_{z,\max}$ of the latter are approximately one order of magnitude higher than the other mechanisms. The asymmetry of the PARA-TRIOSTAGE (with two kinematic chains on the right side and one on the left side as shown in Figure 8a) seems to drastically affect the straightness of the stage. This illustrates the advantage of using symmetric configurations based on two pairs of coupled kinematic chains, such as the PARA-QUADRISTAGE, WATT-QUADRISTAGE, and KEMPE-QUADRISTAGE.

The simulations demonstrate lower parasitic shifts for the WATT-QUADRISTAGE in comparison to the PARA-QUADRISTAGE, which is consistent with what is expected from their ideal kinematics (Section 2). However, the KEMPE-QUADRISTAGE exhibits a higher value for $\Delta y_{\max}$ than the two other 4-RRR rectilinear stages. This FEM result is in contradiction with the ideal kinematics of the KEMPE-QUADRISTAGE having zero parasitic shift. The possible reason for this disparity is that the secondary link rotation is not effectively transmitted throughout the flexure-based Kempe linkage couplers, because the rigid links are not sufficiently stiff. Furthermore, the force transmitted through the coupling links is not aligned with the longitudinal axis of the coupling flexures (due to the modification of Condition 3 for this kind of coupler). These hypotheses are further discussed in Section 7.5.

The mechanism exhibiting the lowest maximum parasitic shifts $\Delta y_{\max}$ and $\Delta\theta_{z,\max}$, is the WATT-QUADRISTAGE (Table 3). Advantageously, it is also the mechanism requiring the least amount of space, see Figure 8. This structure was therefore selected for experimental testing. The measurements of the parasitic shifts carried out on the manufactured WATT-QUADRISTAGE prototype are also plotted in Figure 10 to be compared with the FEM data. The mean values (points) and the standard deviations (error bars) of the measured lateral platform deviation and rotation are displayed as a function of x. There are noticeable differences in the results between the experiment and the FEM model. Indeed, the measured maximum parasitic translation $\Delta y_{\max}$ is one order of magnitude larger than what the simulations predict (Table 3). This disparity can be attributed to the manufacturing tolerances of the POM mockup, which may cause flexure deformation that differs significantly from the FEM model. Furthermore, we can observe that the standard deviation is large compared to the mean values, which demonstrates that it is difficult to precisely characterize the parasitic shift in the POM prototype with the proposed test

bench. In addition, the FEM model does not consider gravity effects which may affect the actual trajectory of the stage. Nevertheless, this experiment demonstrates some preliminary results of the performance of these rectilinear stages, as it reveals maximum parasitic shifts of 2.82 µm laterally and 50.4 µrad in rotation over a 11 mm translation stroke. Further experimental validation is foreseen by measuring the straightness error of titanium alloy prototypes fabricated by Electrical Discharge Machining (EDM) using high-resolution laser displacement sensors.

7.5. Performance Ranking and Further Optimizations

In summary, the FEM results show that the WATT-QUADRISTAGE performs significantly better than the other selected mechanisms in terms of rectilinearity and space occupation. It exhibits a maximum parasitic translation of $\Delta y_{\max} = 0.258$ µm and a maximum parasitic rotation of $\Delta\theta_{z,\max} = 50.4$ µrad for a range of motion of $x_{\text{tot}} = 12$ mm. This leads to a normalized straightness error of $\Delta y_{\max}/x_{\text{tot}} = 2.15 \times 10^{-5}$. The mechanism with the highest support stiffness is the PARA-QUADRISTAGE, with a lateral stiffness of $K_y = 195$ N/mm and rotation stiffness $K_{Rz} = 588$ Nm/rad. It leads to an elevated lateral to translational stiffness ratio of $K_y/K_x = 213$. The presented structures have approximately equivalent maximum stresses and thus equivalent admissible strokes.

Using other manufacturing technologies, the *n*-RRR rectilinear mechanisms could be further optimized by decreasing the thickness ratio between the flexures and the frame parts. This would make the frame parts stiffer relative to the flexures, in such a way that the frame parts deform considerably less and approximate rigid links better. The deflection of the flexure-based stages would then correspond more to their ideal joint kinematics. This could improve rectilinearity as well as increase the support stiffness over translational stiffness ratio. To verify this hypothesis, we simulated the mechanisms in Figure 8, considering that the frame parts are infinitely stiff and thus do not deform. We observed that the support stiffness and the parasitic shift was, respectively, increased and decreased by approximately one order of magnitude for every mechanism. As such, the performance ranking of the mechanisms remains the same as discussed above. It is surprising that the PARA-QUADRISTAGE and the WATT-QUADRISTAGE, which have theoretically higher parasitic shifts than KEMPE-QUADRISTAGE, still perform better than the latter. For the KEMPE-QUADRISTAGE to improve performance, it is therefore necessary to optimize the location and orientation of flexures. In subsequent studies, it would be worth implementing parametric or topology optimization in order to minimize the parasitic shift and maximize the support stiffness of this mechanism.

8. Application Example: Stiffness-Reduced Silicon Rectilinear Stage

Since the presented *n*-RRR rectilinear mechanisms can have higher translation stiffness (due to the larger number of kinematic chains) compared to existing flexure linear stages, we investigate a method to reduce the translational stiffness of such mechanisms while maintaining a high support stiffness [29]. In this section, we test the ability of tuning the translational stiffness of a WATT-QUADRISTAGE rectilinear stage by incorporating buckling beams. To provide applicability at MEMS scale, the mechanism is fabricated in silicon using DRIE (Deep Reactive Ion Etching) and the preloading of the buckling beams is performed by thermal oxidation. Furthermore, Preloading Chevron Mechanisms (PCMs) [30,31] are employed to increase the operating stroke of the stage by enhancing the residual stress effects.

8.1. Stiffness Reduction Concept

Figure 11 presents the silicon WATT-QUADRISTAGE mechanism preloaded by $n_b = 4$ PCM-preloaded buckling beams. These buckling beams apply a nearly constant moment $M_C = Pr_L + M_L$ to the secondary links. The potential energy $E_{pot,buckling}$ denotes the energy associated with the preloading introduced by the buckling beams in the flexure mechanism. Since the secondary links, to which the buckled beams are attached, always rotate in the same direction (because $\theta_C \propto x^2$), the energy stored in the buckled beams consistently decreases as the moving stage undergoes rectilinear displacement. This energy reduction with respect to the stage motion can be modeled as an effective negative stiffness, denoted as K_{neg}, which reduces the positive stiffness K_x of the stage. The potential energy $E_{pot,buckling}$ can be computed as the work of the moment M_C applied by the n_b buckled beams as follows:

$$E_{pot,buckling} = \frac{1}{2} K_{neg} x^2 = n_b M_C \theta_C \tag{9}$$

where from Equations (4) and (5), the expression for the effective negative stiffness can be derived as:

$$K_{neg} = \frac{n_b (Pr_L + M_L)}{r_C (r_A + L_0/2)} \tag{10}$$

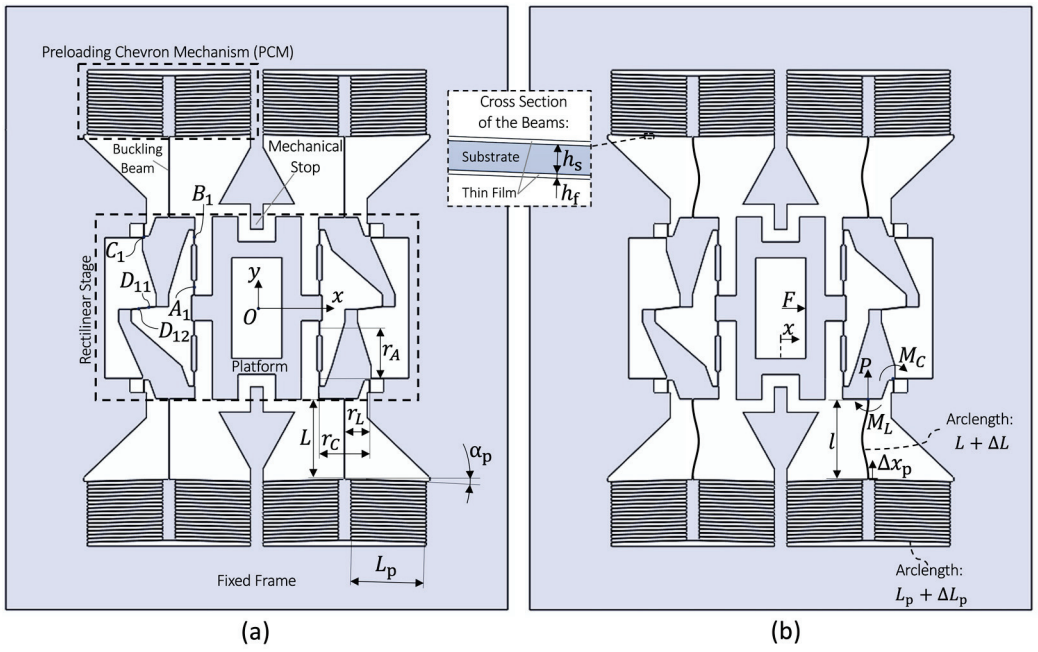


Figure 11. WATT-QUADRISTAGE rectilinear stage shown in neutral position (**a**) before and (**b**) after thermal oxidation.

In the vicinity of the neutral position, the buckled beams have fixed–fixed boundary conditions, leading to the following axial load and reaction moment [32]:

$$P = \frac{4\pi^2 (EI)_{tot}}{L^2}, \; M_L = \frac{4\pi (EI)_{tot}}{L} \sqrt{\frac{\Delta x_{tot}}{L}} \tag{11}$$

where $(EI)_{tot}$ is the total flexural rigidity of the beams, accounting for the flexural rigidities of both the substrate and the thin film (see inset in Figure 11), and $\Delta x_{tot} = \Delta L + \Delta x_p$. It can interestingly be noted that if the buckled beams deflect with an outward buckling direction (as opposed to the inward buckling direction illustrated in Figure 11b), then the sign of M_L is reversed. Considering both the positive stiffness of the stage flexures K_x

(Equation (6)) and the negative stiffness provided by the buckling beams (Equation (10)), the total stiffness $K_{\text{tot}} = F/x$ is given by:

$$K_{\text{tot}} = K_x - K_{\text{neg}} \tag{12}$$

Note that the total stiffness can either be positive (stiffness adjustment), near-zero (zero-force properties), or negative (bistable behaviors).

8.2. FEM and Experimental Validations

The mechanism is designed, manufactured, and tested experimentally. The FEM and experimental methods are as described in [30]. The flexures of the rectilinear stage as well as the buckling beams all have the same cross-section. Their silicon core has a thickness h_{s} = 20 μm and an out-of-plane depth b_{s} = 380 μm. The thickness of the thermal oxide layer deposited all around the silicon core is equal to h_{s} = 3 μm. The PCMs each have 50 preloading beams. The beam lengths and other parameter values are provided in Table 4. The same method as in [30] is used to extract the residual stress value in the oxide. A residual stress of −290 MPa (compressive) is measured in the fabricated wafer.

Table 4. Design parameters of the silicon WATT-QUADRISTAGE mechanism (see Figure 11).

	Parameter	Value
Main Flexures	L_0	0.6 mm
	r_A	2 mm
Secondary Flexures	p_C	50 μm
	L_C	0.5 mm
	r_C	2 mm
Coupling Flexures	L_D	0.5 mm
Buckling Beams	L	3.2 mm
	r_L	1 mm
Preloading Chevron Mechanism	L_{p}	3 mm
	α_{p}	1.5 deg

Photographs of the manufactured prototypes are shown in Figure 12. We can see that the buckling beams of the mechanisms can indeed either buckle inwards or outwards. An equivalent prototype without buckling beams is also tested to assess the stiffness reduction. The measured force–displacement characteristics of the different mechanisms are presented in Figure 13, along with analytical and FEM data.

The stiffness reduction method is validated since the stiffness is greatly reduced (even to a negative value) when incorporating buckling beams in the WATT-QUADRISTAGE structure. Indeed, the measurements, which are in good agreement with the models, show that the translational stiffness is reduced by 98% (monostable prototype) and 112% (bistable prototype) when buckling beams are incorporated with outward and inward buckling directions, respectively. Stiffness reduction can be very interesting for low-voltage actuators or highly sensitive MEMS devices [33–35], whereas bistable properties could be advantageous for MEMS optical switches or energy harvesters [36,37].

Technically, the buckling direction is switched by manually pushing the buckling beams with a small needle. In practical systems, MEMS actuators (e.g., comb-drive or thermal actuators) might be integrated to modify the buckling direction of the different buckled beams in situ, if triggering and resetting the mechanism bistability is required. Note that the monostable state of the mechanism is limited to a displacement of ±0.35 mm around the neutral position. Indeed, when this stroke is exceeded, all the buckled beams

snap to the inward buckling configuration, making the mechanism bistable. In that case, the mechanism cannot be returned to its monostable configuration without actively changing the buckling direction. Mechanical stops could be employed to limit the translation stroke to mitigate snap-through of the mechanism from monostable to bistable.

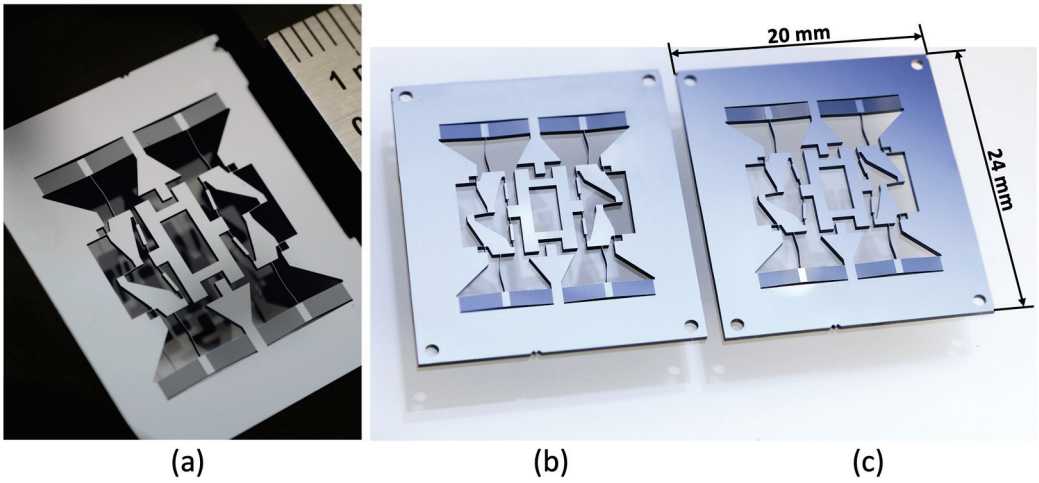


Figure 12. Photographs of preloaded silicon WATT-QUADRISTAGE. (**a**) Close-up view of one of the prototypes. The scale on the upper-right corner displays millimetric graduations. (**b**) If the buckling beams are buckled outwards, the mechanism is monostable. (**c**) If they are buckled inwards, the mechanism is bistable.

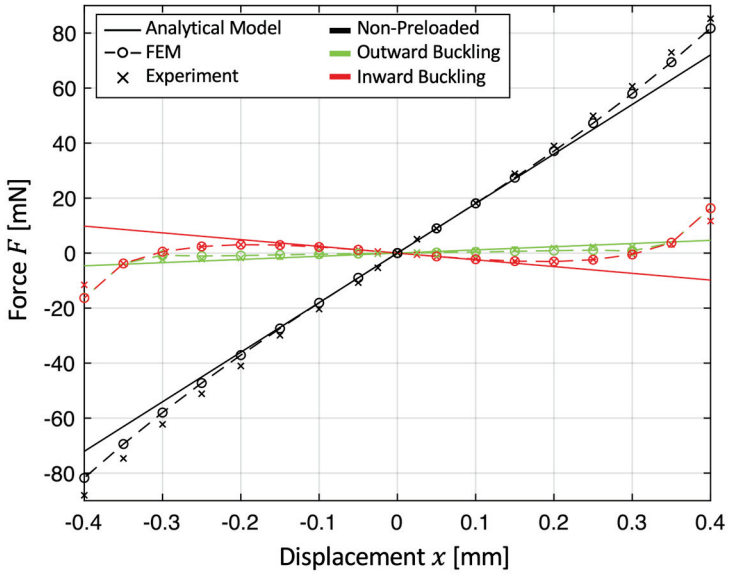


Figure 13. Force–displacement characteristics of the WATT-QUADRISTAGE silicon mechanisms.

Remark 4. *It can be seen that the FEM and experimental force–displacement curves exhibit a small nonlinear load-stiffening effect with displacement, whereas the analytical model has a linear curve with constant stiffness (Figure 13). This nonlinearity is considered acceptable, because it is small compared to overconstrained flexure stages with similar dimensions, where the stiffness increases significantly with displacement, strongly reducing the motion range [11].*

For both monostable and bistable configurations of the mechanism, FEM simulations show parasitic shifts less than 0.66 nm along y and 6.7 µrad around z for a total stroke of ± 0.5 mm. Due to the small scale of these mechanisms, we did not perform any experiment to validate these parasitic shift values. The lateral stiffness of the platform is

K_y = 1086 N/mm, regardless of whether the mechanism is preloaded by buckling beams (inward or outward orientations) or not. This leads to ratios between support and translational stiffness K_y/K_x of 6.0×10^3 and 2.6×10^5 when the mechanism is, respectively, non-preloaded and preloaded by outward buckling beams. With these values, we can notice the importance of employing stiffness reduction along the primary axis in order to increase the ratio between support and translational stiffness [19].

9. Conclusions

In this paper, we present a novel family of *n*-RRR planar mechanisms where the kinematic chains are coupled to reduce the mobility of the platform to a 1-DOF rectilinear translation. The following three types of couplers are investigated: parallelogram, Watt, and Kempe linkages, which lead to different orders of parasitic shifts. Based on the ideal-joint kinematics, we select four design variants (named PARA-TRIOSTAGE, PARA-QUADRISTAGE, WATT-QUADRISTAGE, and KEMPE-QUADRISTAGE) and propose flexure-based implementations. We discuss the performance of these novel compliant rectilinear stages having the same geometric constraints and material (POM). The comparison, based on FEM models, covers the parasitic shifts, the maximum internal stress, and the support stiffness. In summary, the WATT-QUADRISTAGE shows the minimum parasitic shift and the lowest space consumption when compared to the other structures. Indeed, its platform exhibits a theoretical straightness inaccuracy of 0.258 µm and a maximum parasitic rotation of 12.6 µrad for a total stroke of 12 mm. An experimental study performed on a POM mockup validated the low parasitic shifts and high support stiffness properties of such rectilinear-motion mechanisms. With such performance, these scalable and easy-to-manufacture rectilinear mechanisms can be used for a large spectrum of applications.

We also demonstrate that the translational stiffness of such rectilinear mechanisms can be greatly reduced by integrating post-buckled beams into the structure. Indeed, fully compliant silicon prototypes of the WATT-QUADRISTAGE were built, thermally oxidized to provide buckling, and tested experimentally. The measurements show that the translational stiffness was reduced by 98% when incorporating buckled beams, while the support stiffness remains the same. The same mechanism becomes bistable if the buckling beams are buckled in the opposite direction (stiffness reduction of 112%).

Further research will include topology optimization to further minimize the parasitic shift and maximize the support stiffness of this new family of kinematics-based flexure rectilinear mechanisms. We will validate the geometric optimization with experiments made on titanium alloy prototypes fabricated by wire-cut EDM.

Author Contributions: Conceptualization, all authors; methodology, L.T.-D., C.V., F.C. and S.H.; software, L.T.-D.; validation, L.T.-D. and C.V.; formal analysis, L.T.-D. and C.V.; investigation, L.T.-D.; resources, S.H.; data curation, S.H.; writing—original draft preparation, L.T.-D.; writing—review and editing, all authors; visualization, L.T.-D., C.V. and S.H.; supervision, S.H.; project administration, S.H.; funding acquisition, S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors thank Michel Despont and Yves Pétremand from CSEM for the manufacturing of the silicon parts and the discussions related to the design. We also thank Quentin Gubler for his help in carrying out the design, the FEM modeling, and the experiment of the POM mockups.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xu, Q. Design and Development of a Compact Flexure-Based XY Precision Positioning System With Centimeter Range. *IEEE Trans. Ind. Electron.* **2014**, *61*, 893–903. [CrossRef]
2. Zhang, H.; Wu, Z.; Xu, Q. Design of a New XY Flexure Micropositioning Stage With a Large Hollow Platform. *Actuators* **2020**, *9*, 65. [CrossRef]
3. Choi, Y.; Sreenivasan, S.V.; Choi, B.J. Kinematic design of large displacement precision XY positioning stage by using cross strip flexure joints and over-constrained mechanism. *Mech. Mach. Theory* **2008**, *43*, 724–737. [CrossRef]
4. Zhang, Z.; Yan, P.; Hao, G. A Large Range Flexure-Based Servo System Supporting Precision Additive Manufacturing. *Engineering* **2017**, *3*, 708–715. [CrossRef]
5. Algamili, A.S.; Khir, M.H.M.; Dennis, J.O.; Ahmed, A.Y.; Alabsi, S.S.; Hashwan, S.S.B.; Junaid, M.M. A Review of Actuation and Sensing Mechanisms in MEMS-Based Sensor Devices. *Nanoscale Res. Lett.* **2021**, *16*, 16. [CrossRef]
6. Hao, G.; Zhu, J. Design of a Monolithic Double-Slider Based Compliant Gripper with Large Displacement and Anti-Buckling Ability. *Micromachines* **2019**, *10*, 665. [CrossRef]
7. Beroz, J.; Awtar, S.; Bedewy, M.; Tawfick, S.; Hart, A.J. Compliant microgripper with parallel straight-line jaw trajectory for nanostructure manipulation. In Proceedings of the ASPE 2011 Annual Meeting, Denver, CO, USA, 13–18 November 2011.
8. Cosandier, F.; Eichenberger, A.; Baumann, H.; Jeckelmann, B.; Bonny, M.; Chatagny, V.; Clavel, R. Development and integration of high straightness flexure guiding mechanisms dedicated to the METAS watt balance Mark II. *Metrologia* **2014**, *51*, S88–S95. [CrossRef]
9. Darnieder, M.; Pabst, M.; Wenig, R.; Zentner, L.; Theska, R.; Fröhlich, T. Static behavior of weighing cells. *J. Sens. Sens. Syst.* **2018**, *7*, 587–600. [CrossRef]
10. Rubbert, L.; Bitterli, R.; Ferrier, N.; Fifanski, S.; Vardi, I.; Henein, S. Isotropic springs based on parallel flexure stages. *Precis. Eng.* **2016**, *43*, 132–145. [CrossRef]
11. Cosandier, F.; Henein, S.; Richard, M.; Rubbert, L. *The Art of Flexure Mechanism Design*; EPFL Press: Lausanne, Switzerland, 2017.
12. Gräser, P.; Linß, S.; Harfensteller, F.; Torres, M.; Zentner, L.; Theska, R. High-precision and large-stroke XY micropositioning stage based on serially arranged compliant mechanisms with flexure hinges. *Precis. Eng.* **2021**, *72*, 469–479. [CrossRef]
13. Hricko, J. Design of Compliant Micro-Stage Based on Peaucellier–Lipkin Straight-Line Mechanism. In Proceedings of the RAAD 2014 23rd International Conference on Robotics in Alpe-Adria-Danube Region, Smolenice Castle, Slovakia, 3–5 September 2014.
14. Tissot-Daguette, L.; Schneegans, H.; Gubler, Q.; Baur, C.; Henein, H. Rectilinear translation four-bar flexure mechanism based on four Remote Center Compliance pivots. In Proceedings of the Conference of the European Society for Precision Engineering and Nanotechnology, Geneva, Switzerland, 30 May–3 June 2022.
15. Zhao, H.; Bi, S.; Yu, J. A novel compliant linear-motion mechanism based on parasitic motion compensation. *Mech. Mach. Theory* **2012**, *50*, 15–28. [CrossRef]
16. Hubbard, N.B.; Wittwer, J.W.; Kennedy, J.A.; Wilcox, D.L.; Howell, L.L. A novel fully compliant planar linear-motion mechanism. In Proceedings of the ASME 2004 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Salt Lake City, UT, USA, 28 September–2 October 2004.
17. Liu, Z.; Zhang, Z.; Yan, P. A self-adjusting stiffness center design for large stroke compliant XY nanomanipulators. *Mech. Sci.* **2018**, *9*, 41–50. [CrossRef]
18. Li, Y.; Xu, Q. Design and Robust Repetitive Control of a New Parallel-Kinematic XY Piezostage for Micro/Nanomanipulation. *IEEE/ASME Trans. Mechatron.* **2012**, *17*, 1120–1132. [CrossRef]
19. Brouwer, D.M.; Meijaard, J.P.; Jonker, J.B. Large deflection stiffness analysis of parallel prismatic leaf-spring flexures. *Precis. Eng.* **2013**, *37*, 505–521. [CrossRef]
20. Vallat, C.; Tissot-Daguette, L.; Cosandier, F.; Henein, S. Rectilinear Flexure-based Translation Stage Family with Minimized Parasitic Shift, High Support Stiffness and Large Range of Motion. 2025. Available online: https://engrxiv.org/preprint/view/4458/version/6077 (accessed on 10 June 2025).
21. Tissot-Daguette, L.; Vallat, C.; Cosandier, F.; Henein, S. One Degree of Freedom Linear Stages with Minimised Parasitic Shift. Patent EP24218366.3, 9 December 2024.
22. Hao, G.; Li, H.; He, X.; Kong, X. Conceptual design of compliant translational joints for high-precision applications. *Front. Mech. Eng.* **2014**, *9*, 331–343. [CrossRef]
23. Xu, Q. Design and Development of a Flexure-Based Dual-Stage Nanopositioning System with Minimum Interference Behavior. *IEEE Trans. Autom. Sci. Eng.* **2012**, *9*, 554–563. [CrossRef]
24. Tissot-Daguette, L.; Cosandier, F.; Thalmann, E.; Henein, S. Near-Zero Parasitic Shift Flexure Pivots Based on Coupled n-RRR Planar Parallel Mechanisms. *ASME J. Mech. Robot.* **2024**, *16*, 111006. [CrossRef]
25. Kempe, A.B. On a General Method of describing Plane Curves of the nth degree by Linkwork. *Proc. Lond. Math. Soc.* **1875**, *1*, 213–216. [CrossRef]
26. Howell, L.L. *Compliant Mechanisms*; Wiley: New York, NY, USA, 2001.

27. Pei, X.; Yu, J.; Zong, G.; Bi, S. An effective pseudo-rigid-body method for beam-based compliant mechanisms. *Precis. Eng.* **2010**, *34*, 634–639. [CrossRef]
28. Nijenhuis, M.; Meijaard, J.P.; Mariappan, D.; Herder, J.L.; Brouwer, D.M.; Awtar, S. An Analytical Formulation for the Lateral Support Stiffness of a Spatial Flexure Strip. *ASME J. Mech. Des.* **2017**, *139*, 051401. [CrossRef]
29. de Jong, J.; Theans, S.; Epping, L.; Brouwer, D.M. Improving Support Stiffness of Flexure Mechanisms by Statically Balancing. In Proceedings of the 36th Annual Meeting of the American Society for Precision Engineering, Minneapolis, MN, USA, 1–5 November 2021.
30. Tissot-Daguette, L.; Cosandier, F.; Gubler, Q.; Petremand, Y.; Despont, M.; Henein, S. Residual Stress Chevron Preloading Amplifier for Large-Stroke Stiffness Reduction of Silicon Flexure Mechanisms. *J. Micromech. Microeng.* **2025**, *35*, 025003. [CrossRef]
31. Tissot-Daguette, L.; Henein, S. Flexure Based Mechanism with Reduced Stiffness. Patent EP24199124.9, 9 September 2024.
32. Tissot-Daguette, L.; Schneegans, H.; Thalmann, E.; Henein, S. Analytical modeling and experimental validation of rotationally actuated pinned–pinned and fixed–pinned buckled beam bistable mechanisms. *Mech. Mach. Theory* **2022**, *174*, 104874. [CrossRef]
33. de Laat, M.L.C.; Pérez Garza, H.H.; Herder, J.L.; Ghatkesar, M.K. A review on in situ stiffness adjustment methods in MEMS. *Micromech. Microeng.* **2016**, *26*, 063001. [CrossRef]
34. Hussein, H.; Wang, C.; Esteves, R.A.; Kraft, M.; Fariborzi, H. Near-zero stiffness accelerometer with buckling of tunable electrothermal microbeams. *Microsyst. Nanoeng.* **2024**, *10*, 43. [CrossRef] [PubMed]
35. Xiong, R.; Xu, X.; Liu, Y.; Du, S.; Jin, L.; Chen, F.; Wu, T. A miniaturized MEMS accelerometer with anti-spring mechanism for enhancing sensitivity. *Microsyst. Nanoeng.* **2025**, *11*, 42. [CrossRef]
36. Xu, R.; Akay, H.; Kim, S.G. Buckled MEMS Beams for Energy Harvesting from Low Frequency Vibrations. *Research* **2019**, *2019*, 2581241. [CrossRef]
37. Yang, Y.J.; Liao, B.T.; Kuo, W.C. A novel 2 $\times$ 2 MEMS optical switch using the split cross-bar design. *J. Micromech. Microeng.* **2007**, *17*, 875. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Multibody Simulation of 1U CubeSat Passive Attitude Stabilisation Using a Robotic Arm

Filippo Foiani [1], Giulia Morettini [1,*], Massimiliano Palmieri [1], Stefano Carletta [2], Filippo Cianetti [1] and Marco Dionigi [1]

[1] Department of Engineering, University of Perugia, Via G. Duranti 93, 06125 Perugia, Italy; filippo.foiani@dottorandi.unipg.it (F.F.); massimiliano.palmieri@unipg.it (M.P.); filippo.cianetti@unipg.it (F.C.); marco.dionigi@unipg.it (M.D.)
[2] School of Aerospace Engineering, Sapienza University of Rome, Via Salaria 851, 00138 Rome, Italy; stefano.carletta@uniroma1.it
* Correspondence: giulia.morettini@unipg.it

Abstract: Robotics plays a pivotal role in contemporary space missions, particularly in the development of robotic manipulators for operations in environments that are inaccessible to humans. In accordance with the trend of integrating multiple functionalities into a single system, this study evaluates the feasibility of using a robotic manipulator, termed a C-arm, for passive attitude control of a 1U CubeSat. A simplified multibody model of the CubeSat system was employed to assess the robotic arm's functionality as a gravity gradient boom and subsequently as a passive magnetic control mechanism by utilising a permanent magnet at its extremity. The effectiveness of the C-arm as a gravitational boom is constrained by size and weight, as evidenced by the simulations; the pitch angle oscillated around $\pm 40°$, while roll and yaw angles varied up to 30° and 35°, respectively. Subsequent evaluations sought to enhance pointing accuracy through the utilisation of permanent magnets. However, the absence of dissipative forces resulted in attitude instabilities. In conclusion, the integration of a robotic arm into a 1U CubeSat for passive attitude control shows potential, especially for missions where pointing accuracy can tolerate a certain range, as is typical of CubeSat nanosatellite missions.

Keywords: C-arm for CubeSats; passive attitude control; multibody attitude stabilization simulation; 1U CubeSat attitude control

1. Introduction

Robotics plays a pivotal role in contemporary space missions. The development of specialized robots has enabled humans to explore and operate in previously inaccessible environments. Of these applications, one of the most significant is the field of manipulators. These are robotic arms that are employed primarily for space activity, docking [1], close-proximity operations [2], assembly, and maintenance in both orbital and planetary surface environments, allowing for the performance of tasks that are infeasible or excessively risky for humans. [3]

Among the various types of manipulators employed in the context of different space missions, it is necessary to highlight the first robotic arm deployed in space: the Remote Manipulator System (RMS) designed specifically for the Space Shuttle, also known as Canadarm-1. This system made its inaugural flight aboard the Space Shuttle Columbia (STS-2) in November 1981. The RMS was primarily designed for the capture and manoeuvring of cargo in free flight. However, it has been employed in a wide variety of other

applications as well, including the assembly of the International Space Station, support for extra-vehicular activities, and orbital inspections. Its versatility has made it an experimental platform for materials studies and enabled the performance of unique tasks, including ice removal during shuttle re-entry missions [4]. It is also worth mentioning the manipulators developed for the assembly, maintenance, and scientific research activities of the International Space Station (ISS). These include the Mobile Servicing System (MSS), which comprises the Space Station Remote Manipulator System (SSRMS), Special Purpose Dextrous Manipulator (SPDM), and Mobile Base System (MBS); the Japanese Experiment Module Remote Manipulator System (JEMRMS); and the European Robotic Arm (ERA) [5].

The use of robotic arms in satellites has a more recent history. In particular, the first example of a satellite equipped with a manipulator was the Engineering Test Satellite (ETS-VII), which was launched on 27 November 1997 to conduct automated rendezvous docking and space robot technology experiments [6]. Other pertinent missions involving large satellites equipped with robotic manipulators include the Orbital Express mission developed by the Defense Advanced Research Projects Agency (DARPA) [7]. In this mission, the first autonomous transfer of an additional battery and backup computer to a target spacecraft was successfully executed. Another notable mission is the Spacecraft for the Universal Modification of Orbits (SUMO), which aims to combine stereo-photogrammetry with robotic manipulators in order to facilitate the retrieval of objects from a spacecraft [8].

Although manipulators are a common feature of large satellites, their extension to microsatellites has been slower. For example, we can take a case application on CubeSat-type satellites. These are small satellites based on standard units (commonly referred to as 'U') with dimensions of $10 \times 10 \times 10$ cm [9,10]. The applications of CubeSats extend to a range of disciplines, including earth monitoring, astrophysics, remote sensing, technology demonstration, and space science [11]. The development of manipulators for these types of satellites currently remains in the planning stage due to the limited space in which this complex technology must be implemented.

Among the most relevant applications are those proposed by Silva et al. [12], who suggested a robotic arm solution integrated into the payload of a CubeSat 2U. This 3-DoF (Degrees of Freedom) manipulator was designed with the objective of assisting or replacing astronauts during Extra-Vehicular Activities (EVA) on the International Space Station. Liu et al. [13] proposed an innovative concept for a deployable robotic arm, designated Cubot. The manipulator is designed for on-orbit servicing, with a particular focus on removal of space debris and maintenance of the space station. The robotic arm, which folds into a space of 1U, comprises three active and three passive joints that lock when the arm is deployed, reducing the effective degrees of freedom from six to three. Örger et al. [14] presented a concept for a CubeSat equipped with two robotic arms, each with 5 degrees of freedom. Their design is intended for on-orbit servicing, with a particular focus on inspection and control of spacecraft through cameras mounted on the arms. In [15], McCormick et al. presented the REMORA project, which aims to implement a robotic arm that can be adapted to CubeSats with 6/12 U configurations. Their article presents an interesting case study of a satellite equipped with a 5-DoF robotic arm that can be folded into a 2U configuration. The objective of the satellite is to mitigate the risk of collisions with large debris in Low-Earth Orbit (LEO). The article describes the mission design phases, starting with the launch of the satellite and ending with the docking of the debris object. It also presents the robotic payload and the design of the arm in order to illustrate the choice of end-effector type. In [16], McCormick et al. illustrate the feasibility study of a small satellite platform with Cluster-Forming On-Board Robotic Manipulators (CFORM), i.e., a robotic arm designed for docking. This study is of particular interest due to the high packing capacity of the system, which can fit into a 0.5 U CubeSat. Zeis et al. [17] presented

a very interesting modular robotic arm for small satellites called the Little Inspection and Servicing Arm (LISA). This prototype is equipped with 24 degrees of freedom and takes up only 1U when folded. The most significant scientific outcome in the field of CubeSats with manipulators is represented by the RSat project [18]. This work is of particular interest; it differs from those previously discussed in that it was successfully launched into orbit in 2018 as part of NASA's Educational Launch of Nanosatellites (ELaNa) Mission 19. The RSat is a 3U CubeSat equipped with two robotic manipulators for on-orbit servicing. The RSat project is part of the larger Intelligent Space Assembly Robot (ISAR) project, which aims to advance in-orbit assembly technology to the point where a constellation of satellites can perform semi-autonomous or autonomous assembly operations in space [19].

The current technological landscape demonstrates growing interest in the development of miniaturised robotic systems for CubeSats, as evidenced by the numerous design solutions presented in the scientific literature. Though mostly still in the on-orbit testing phase, these prototypes demonstrate the ability to perform a range of operations, including component maintenance and repair, sample collection, and establishing links between multiple satellites. Their applications are primarily focused on tasks such as in-orbit servicing, assembly, or inspection.

To date, however, no dedicated applications have been developed leveraging a robotic arm to control the attitude of CubeSat satellites, a function that is typically managed by standard attitude control systems [20–22]. Specifically, the innovative concept of utilizing a robotic manipulator as an integrated passive component for spacecraft attitude stabilization remains largely unexplored in the scientific literature.

Our study addresses this research gap by investigating the feasibility of employing a robotic manipulator for passive attitude stabilization, specifically, a C-arm designed for a 1U CubeSat. We propose a novel approach in which the C-arm is initially utilized as a gravity gradient boom and subsequently enhanced with a permanent magnet at its extremity to function as a passive magnetic control device. The primary objective is to ascertain the viability and performance of the C-arm in these unconventional roles, particularly for nanosatellite missions where mission constraints often favour simplicity and power efficiency. To achieve this, a simplified multibody model of the entire CubeSat system, including the C-arm, is developed and analysed using numerical simulations to quantify key attitude parameters such as oscillation amplitudes (e.g., maximum angular deviations) and frequencies/periods under various initial conditions. This preliminary investigation aims to lay the foundation for more detailed and complex analyses assessing the coupled dynamics of the C-arm and the CubeSat in an orbital environment.

To effectively achieve the objectives outlined in this study, this paper is meticulously structured as follows. Section 2 lays the groundwork by detailing the comprehensive design and seamless integration of the C-arm. Moving forward, Section 3 thoroughly describes the critical satellite and environmental models, including a clear explanation of the simplified multibody representation used in our analysis. The core of our initial findings is presented in Section 4, which delves into the analysis and compelling results of passive attitude stabilization via gravity gradient. Building on this, Section 5 introduces and rigorously evaluates the significant effect of Earth's magnetic field on attitude control. Finally, Section 6 provides a concise summary of the key conclusions, candidly discusses the inherent limitations of this preliminary study, and outlines promising directions for future research.

2. C-Arm System Design and Integration

The deployable C-arm serves as the primary payload of a 1U CubeSat. It can operate independently as a stand-alone satellite; alternatively, it can be integrated as a module into

satellites with larger form factors such as 3U or 6U, with the potential for further expansion. Depending on the type of end-effector, the C-arm can be adapted to perform various on-orbit activities such as spacecraft maintenance, on-orbit assembly, deployment of CubeSat clusters in formation, and berthing. Regardless of its primary function, the robotic arm is also used unconventionally as a tool for passive attitude control. It is worth clarifying that even though the robotic arm is an active device (i.e., it requires electric power for motion), it is not moved when used as an attitude control device (i.e., it is preset in its extended position); hence, the resulting attitude control strategy is passive. First, we evaluate its effectiveness as a boom for gravity gradient control. Later, we additionally evaluate the use of passive magnetic attitude control.

2.1. Design Requirements and Goals

Due to the fundamental physical constraints associated with the arm, which has a maximum storage volume of $(10 \times 10 \times 5)$ cm (corresponding to 0.5 U (half of a CubeSat unit), one of the principal objectives for the manipulator is to be compatible with the 1U form factor of a CubeSat [23]. This is depicted in Figure 1a. Another crucial limitation is the mass, which must be less than 0.7 kg. Furthermore, the manipulator design is oriented towards maximizing extension within the given geometric constraints. The precision of the end-effector is contingent upon the motor characteristics and the selected control law, which is determined by the mission requirements of the arm. The arm can extend to a length of approximately 23.67 cm and is estimated to weigh around 200 g.

The mechanical design of the C-arm is presented in Figure 1b. The shoulder joint comprises two motors designated A and B. Motor A serves as the rotating base of the arm, enabling angular rotation. The two motors are connected by a specially designed small aluminium plate, designated as F. Motor B is connected to link 1 by plate E, which has been designed to minimise the volume of the arm when it is stowed. Motor D is integrated into link 2 and drives the movement of link 3, as presented in Figure 1c. Link 3 is not equipped with a dedicated end-effector, although it is capable of accommodating a variety of elements, including a gripper for a more conventional utilisation of the manipulator or other features such as sensors or small permanent magnets.

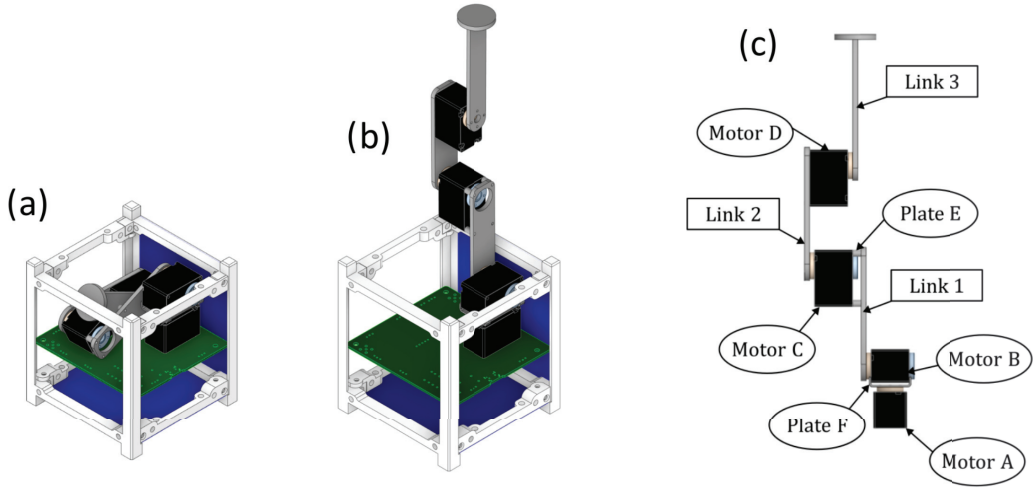


Figure 1. The C-arm robotic manipulator: (**a**) stowed configuration on a 1U CubeSat; (**b**) extended configuration, illustrating the full deployment length; and (**c**) mechanical design schematic with key components.

2.2. Component and Material Selection

The motors selected to drive the joints were chosen based on a preference for using flight heritage components. The primary challenge was to find a compact motor that could enable dexterous and precise movements for the robotic arm at a low cost. To meet these requirements, Commercial Off-The-Shelf (COTS) elements were used. The chosen motor type was the XC330-M288-T from the ROBOTIS Dynamixel-X series that is part of the Dynamixel-X series produced by ROBOTIS, Seoul, South Korea.

These servomotors are frequently used in robotic manipulators due to their integrated features, which include a DC motor, encoder, microcontroller, sensors, gear reduction, and network connection in every model.

This motor type is a coreless DC motor known for its improved efficiency, reduced weight and size, low inertia, precise motion, and reduced power consumption. The motor type is equipped with a full metal spur gear that provides a gear ratio of 288.35:1. It also incorporates an integrated position sensor, specifically the ams AS5601 from ams AG, Premstätten, Austria, contactless absolute encoder, which provides $0.0881^\circ \times 4069$ steps of positioning accuracy.

The microcontroller used is the ARM CORTEX-M0+ from STMicroelectronics, Geneva, Switzerland, which is controlled by a PID controller. The use of a servomotor is advantageous, as it eliminates the need for a separate brake mechanism to hold the joints in place.

3. Satellite and Environmental Model

The conventional approach to modelling satellite orbital and attitude dynamics is to solve the equations of motion while accounting for a range of environmental effects. The dominant force acting on a satellite orbiting in LEO is the gravitational attraction of the Earth. Minor effects are those associated with atmospheric drag, solar radiation pressure, and the luni-solar perturbation. A comprehensive overview of these factors can be found in [24]. Attitude dynamics for a microsatellite can be conveniently modeled assuming that the spacecraft is a rigid body that follows the Euler dynamic equations of motion [25]. Environmental factors exerting influence on the satellite's attitude include gravity gradient, aerodynamic, magnetic, and solar radiation pressure torques. Wertz et al. (2012) provided a comprehensive analysis of the modelling, mitigation, and management of these factors to ensure precise attitude determination and control [26].

By solving these equations with the inclusion of environmental effects, it is possible to accurately predict and manage the behavior of satellites in space, ensuring their successful operation and fulfillment of the assigned mission. Mathematical models are typically implemented in software such as Matlab version R2025a or Python version 3.13.0 for the purpose of studying the motion of satellites and developing control techniques. In this context, a CubeSat is represented as a 6-DoF rigid body with the capacity to rotate about its principal axes of inertia.

In order to describe the satellite orbit and attitude, three commonly used reference frames are employed. The first reference frame is the Earth-Centered Inertial (ECI, $\mathcal{T}_{ECI}$) frame, which is an inertial frame with the origin set at the center of the Earth with the x-axis pointing towards the vernal equinox, the z-axis parallel to the Earth's spin axis, and the y-axis completing the orthogonal frame.

In order to define the attitude propagation of the satellite, it is necessary to introduce two reference systems. The first of these is integral with the satellite but fixed with respect to the orbit, while the second describes the satellite's current behaviour. The former is designated as the Orbit ($\mathcal{T}_{\mathcal{O}}$) frame, centered at the spacecraft's center of mass with the z-axis pointing towards the Earth's center, the y-axis parallel and opposite to the orbit's angular momentum vector, and the x-axis completing the left-handed triplet. It is worth noting that

for a circular orbit the x-axis is always parallel to the orbital speed. The remaining reference frame is designated the Body ($\mathcal{T}_\mathcal{B}$) frame. This frame is situated at the spacecraft's center of mass, and its axes are aligned with the primary inertial axes of the satellite. The various reference frames are illustrated in Figure 2.

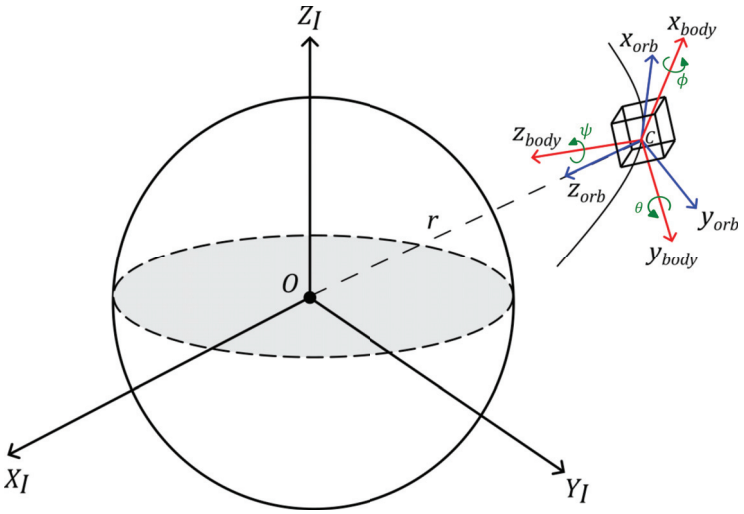


Figure 2. Reference frames and Euler angles for satellite attitude dynamics: the Earth-Centered Inertial (ECI) frame ($\mathcal{T}_{ECI}$) is illustrated in black, the Orbit (T_O) frame in blue, and the Body (T_B) frame in red. The attitude of T_B relative to T_O is described by the three Euler angles: roll (ϕ), pitch (θ), and yaw (ψ).

The orbit is determined by orbital mechanics, while the rotation can be described in an inertial frame centered on the center of mass, as shown in (2). Here, $\mathbf{H}_C$ represents the total angular momentum, while its derivative with respect to time is the sum of the external perturbation torques and control torques $\mathbf{M}_C$. This work exclusively models perturbative effects arising from the action of gravity, specifically the gravity gradient torque ($\mathbf{M}_g$) and the effects of the Earth's magnetic field that generate the magnetic disturbance torque ($\mathbf{M}_m$):

$$\mathbf{M}_C = \mathbf{M}_g + \mathbf{M}_m. \tag{1}$$

Only these disturbance effects were modelled, as our simulations were carried out at an altitude where drag effects can be ignored.

$$\mathbf{M}_C = \frac{d\mathbf{H}_C}{dt} \tag{2}$$

Recalling the classical expression of the time derivative of angular momentum

$$\mathbf{M}_C = \mathbb{I}_B \dot{\boldsymbol{\omega}}_B + \boldsymbol{\omega}_B \times \mathbb{I}_B \boldsymbol{\omega}_B, \tag{3}$$

Euler's equation, designated as Equation (3), is formulated with respect to the $\mathcal{T}_\mathcal{B}$ reference frame. The angular velocity of the $\mathcal{T}_\mathcal{B}$ frame with respect to $\mathcal{T}_{ECI}$ is as follows:

$$\boldsymbol{\omega}_B = \boldsymbol{\omega}_{OI} + \boldsymbol{\omega}_{BO} \tag{4}$$

where $\boldsymbol{\omega}_{BO}$ is the angular velocity of the satellite body axis $\mathcal{T}_\mathcal{B}$ with respect to the $\mathcal{T}_\mathcal{O}$ reference frame and $\boldsymbol{\omega}_{OI}$ is the angular velocity of $\mathcal{T}_\mathcal{O}$ with respect to $\mathcal{T}_{ECI}$.

The angular displacement of the satellite ($\mathcal{T}_\mathcal{B}$) from the desired attitude, represented here by ($\mathcal{T}_\mathcal{O}$), can be expressed in terms of the three Euler angles: roll (ϕ), pitch (θ) and yaw (ψ), as in Figure 2.

3.1. CubeSat and C-Arm Multibody Model

In addition to the utilisation of precise mathematical models to characterize the dynamics of the satellite and evaluate the performance of the orbit control system, multibody simulation can be a valuable tool for investigating the coupled evolution of the satellite's orbit and attitude without the necessity for a two-phase approach, including the execution of complex tasks such as robotic manipulator movements. The equations of motion are solved directly, and any kind of perturbation can be readily implemented. In this context, the complex system comprising the CubeSat satellite and C-arm is modelled in ADAMS/View and depicted in Figure 1. As illustrated in Figure 1, the satellite and C-arm were decomposed into three rigidly constrained bodies. These bodies serve to identify the three main components depicted in Figure 3. The cube represents the structure of the 1U CubeSat, the cylinder simplifies the C-arm in full extension, and the final sphere represents the potential end-effector of the C-arm.

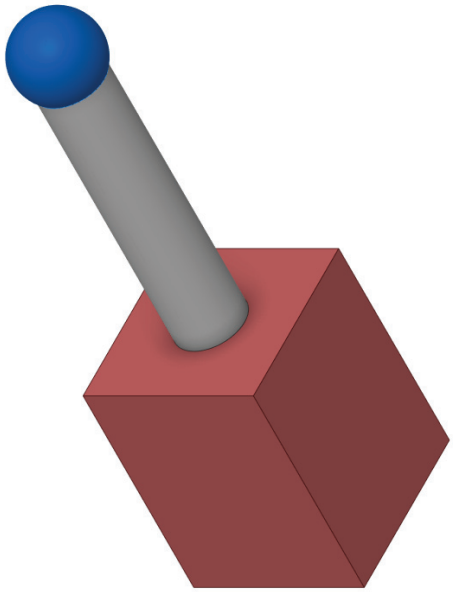


Figure 3. Simplified model of 1U CubeSat and C-arm for multibody implementation in ADAMS software version 2024.2.

The three rigid bodies depicted in Figure 3 are constrained from moving independently with a fixed joint. This configuration results in a system that behaves as a rigid body with a total of six degrees of freedom. To ensure that the mass properties of the real system are accurately represented, the mass and inertia moments synthesized in Table 1 were respectively assigned to each body at the corresponding center of mass.

Table 1. Geometric characteristics of a simplified model for multibody implementation in ADAMS.

Body	Dimension [m]	Mass [kg]	Inertia [kgm^2]
Cube	$L_z = 0.13$, $L_x = L_y = 0.1$	0.4	$I_x = I_y = 8.97 \times 10^{-4}$, $I_z = 6.7 \times 10^{-4}$
Cylinder	$L_z = 0.18$, $d = 3.69 \times 10^{-2}$	0.2	$I_x = I_y = 5.57 \times 10^{-4}$, $I_z = 3.4 \times 10^{-5}$
Sphere	$d = 5 \times 10^{-2}$	0.7	$I_x = I_y = I_z = 1.75 \times 10^{-4}$
Total	$L_z = 0.335$, $L_x = L_y = 0.1$	1.3	$I_x = I_y = 2.70 \times 10^{-2}$, $I_z = 9 \times 10^{-4}$

It is important to note that this multibody model represents a preliminary simplification to assess the fundamental feasibility of the C-arm's influence on attitude dynamics. While it accurately captures the overall mass and inertia properties, it intentionally simplifies the internal complexities of the robotic arm, including individual joint flexibilities and detailed kinematic constraints beyond fixed connections. This approach allows us to obtain a foundational understanding of the coupled system's behavior under the influence of environmental torques before proceeding to more intricate and computationally demanding models.

3.2. Orbit Description and Initial Condition of Motion

This paper presents a case study that evaluates the evolution in time of a satellite's attitude in the presence of a gravity gradient. The presented simulations are based on the orbital parameters described in Table 2. To guarantee the satellite's movement along the prescribed orbit in the multibody simulation, the gravitational attraction force between the CubeSat and the Earth was incorporated into the ADAMS/View model. Equation (5) illustrates the translational equation of motion, with the solution referred to as the classical solution to the two-body problem:

$$\mathbf{F}_g = -G\frac{Mm_s}{|\mathbf{r}_s|^3}\mathbf{r}_s = m_s\ddot{\mathbf{r}}_s \tag{5}$$

where $G = 6.67 \times 10^{-11}$ $\mathrm{m^3\ kg^{-1}\ s^{-2}}$ is the gravitational constant, M and m_s are the respective masses of the Earth and the satellite, $\mathbf{r_s}$ is the satellite's position in $\mathcal{T}_{ECI}$ and the double dot indicates the double derivative in time. As previously stated in Section 3, the satellite's orbit is subject to environmental disturbances; however, for the sake of simplicity these are not considered in this particular case study.

The CubeSat attitude is employed as the study parameter, which is fully delineated by the Euler angles of roll (ϕ), pitch (θ), and yaw (ψ) describing the deviation of the $\mathcal{T}_\mathcal{B}$ with respect to the $\mathcal{T}_\mathcal{O}$. Furthermore, we work under the hypothesis that the initial angular rates reported in Table 3 are equal to zero. The initial velocity conditions are presented in Table 3.

Table 2. Orbital parameters.

Orbital Parameters	Symbols	Value
Inclination	i	30°
Right Ascension of the Ascending Node	Ω	40°
Argument of periapsis	ω	265°
Eccentricity	e	0
Semimajor Axis	a	6918 km

Table 2 illustrates that a circular posigrade orbit has been selected, with randomly generated values for the Right Ascension of the Ascending Node (RAAN) and the argument of the perigee.

Table 3. Attitude initial conditions.

$\omega_{b_{x_0}}$	$\omega_{b_{y_0}}$	$\omega_{b_{z_0}}$
0°/s	0°/s	0°/s

4. Study of the Gravitational Action on Attitude

The CubeSat experiences a gradient of gravitational force along its body, with the greatest attraction occurring on the masses that are closer to the Earth. This phenomenon leads to a gravity gradient torque, which is dependent on the altitude, attitude, and inertia characteristics of the satellite. In particular, the torque, denoted as $\mathbf{M}_g$, can be expressed with respect to the body-fixed frame $\mathcal{T}_\mathcal{B}$, as shown in Equation (6) in the cited work by Sidi [27]:

$$[\mathbf{M}_g]_{\mathcal{T}_\mathcal{B}} = \frac{3\mu}{R^3}\begin{bmatrix} (I_z - I_y)\ sin\phi\ cos\phi\ cos^2\theta \\ (I_z - I_x)\ cos\phi\ sin\theta\ cos\theta \\ (I_x - I_y)\ sin\phi\ sin\theta\ cos\theta \end{bmatrix} \tag{6}$$

where μ is the Earth gravitational parameter, R is the orbit radius, I_x, I_y, I_z are the principal moments of inertia of the satellite, and ϕ, θ are the roll and pitch angle of the spacecraft, respectively.

4.1. Multibody Implementation

The model presented in Section 3.1 must include gravitational forces. These forces are applied at the barycentre of each of the three rigid bodies, implicitly defining the torque induced by the gravity gradient on the whole satellite. This torque counteracts any displacement of the CubeSat from its equilibrium state. This induces a counter-rotation around the global Center of Mass (CoM). However, in order to accurately represent the effect of the gravitational force on the entire satellite, it is necessary to include the effect of the gravitational gradient on each rigid body by adding additional torque contributions. Therefore, based on Equation (6), gravitational gradient torques are applied to the center of mass of each rigid body, as illustrated in Figure 4.

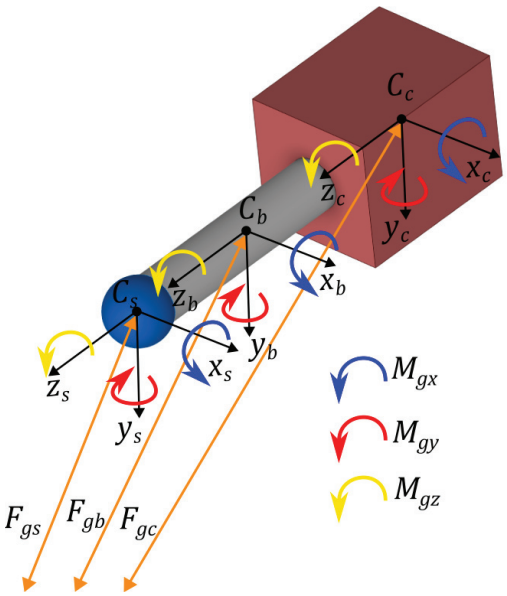


Figure 4. Representation of gravitational forces F_{gs}, F_{gb}, F_{gc} and gravity torque components M_{gx}, M_{gy}, M_{gz} acting on the simplified model.

4.2. Results of Gravity Gradient Action

The results of our simulations conducted under the initial conditions of Table 4 are presented in Figure 5. The propagation of the pitch angle does not affect the evolution of roll or yaw; thus, these remain at zero values in Figure 5. This behaviour is consistent with theoretical predictions for gravity gradient stabilization in a circular orbit, where the equilibrium position is typically along the local vertical and any initial pitch displacement

leads to oscillations around this equilibrium. If not perturbed, the roll and yaw motions tend to remain stable or oscillate around zero in the absence of coupling from pitch or initial angular rates; however, as further illustrated below, the inherent coupling in the Euler equations becomes evident when initial roll or yaw perturbations are introduced, leading to more complex dynamic interactions.

Table 4. Study on the influence of the pitch propagation θ analysed with the initial conditions.

Roll ϕ	Pitch θ	Yaw ψ
0°	0°	0°
0°	5°	0°
0°	10°	0°
0°	15°	0°
0°	20°	0°

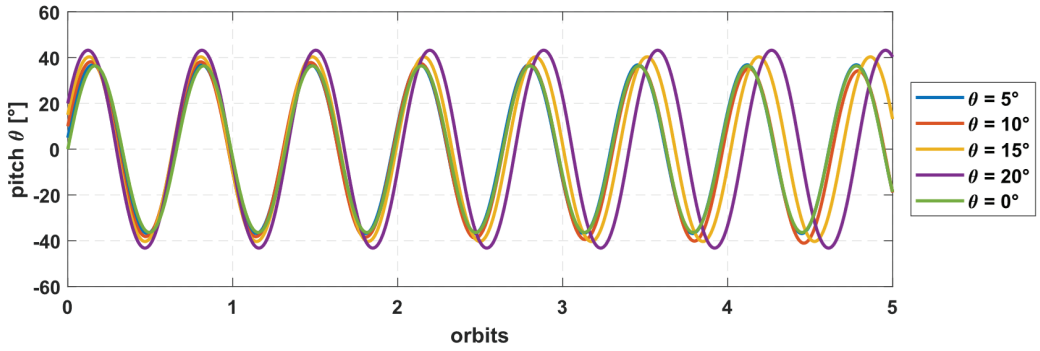


Figure 5. Pitch propagation with different initial conditions for the pitch angle θ.

When only an initial pitch angle perturbation is applied (Table 4), the simulation results in Figure 5 reveal that the pitch angle consistently oscillates around the equilibrium position ($\theta = 0°$). Quantitatively, the maximum amplitude of these oscillations is observed to be approximately $\pm 40°$ across the tested initial conditions (ranging from 0° to 20°). The oscillation frequency remained consistently low, characteristic of passive gravity gradient stabilization. Notably, under these specific initial conditions, the roll (ϕ) and yaw (ψ) angles remain at or very close to zero, demonstrating minimal coupling in the absence of initial perturbations in these axes. This behaviour is consistent with theoretical predictions for gravity gradient stabilization in a circular orbit, where the equilibrium position is typically along the local vertical.

Introducing initial roll perturbations (Table 5) significantly alters the system's dynamics. As shown in Figure 6a, the roll angle itself oscillates, with maximum amplitudes reaching approximately $\pm 20°$ for initial values up to 20°. A critical observation is the strong coupling between the roll and yaw angles. Figure 6c quantitatively demonstrates that an initial roll perturbation directly induces significant oscillations in the yaw angle, reaching maximum amplitudes of up to $\pm 25°$. Conversely, the pitch angle (Figure 6b) shows only minimal influence from initial roll perturbations, with its amplitude and phase largely unaffected

Table 5. Study on the influence of roll propagation ϕ analysed with the initial conditions.

Roll ϕ	Pitch θ	Yaw ψ
5°	0°	0°
10°	0°	0°
15°	0°	0°
20°	0°	0°

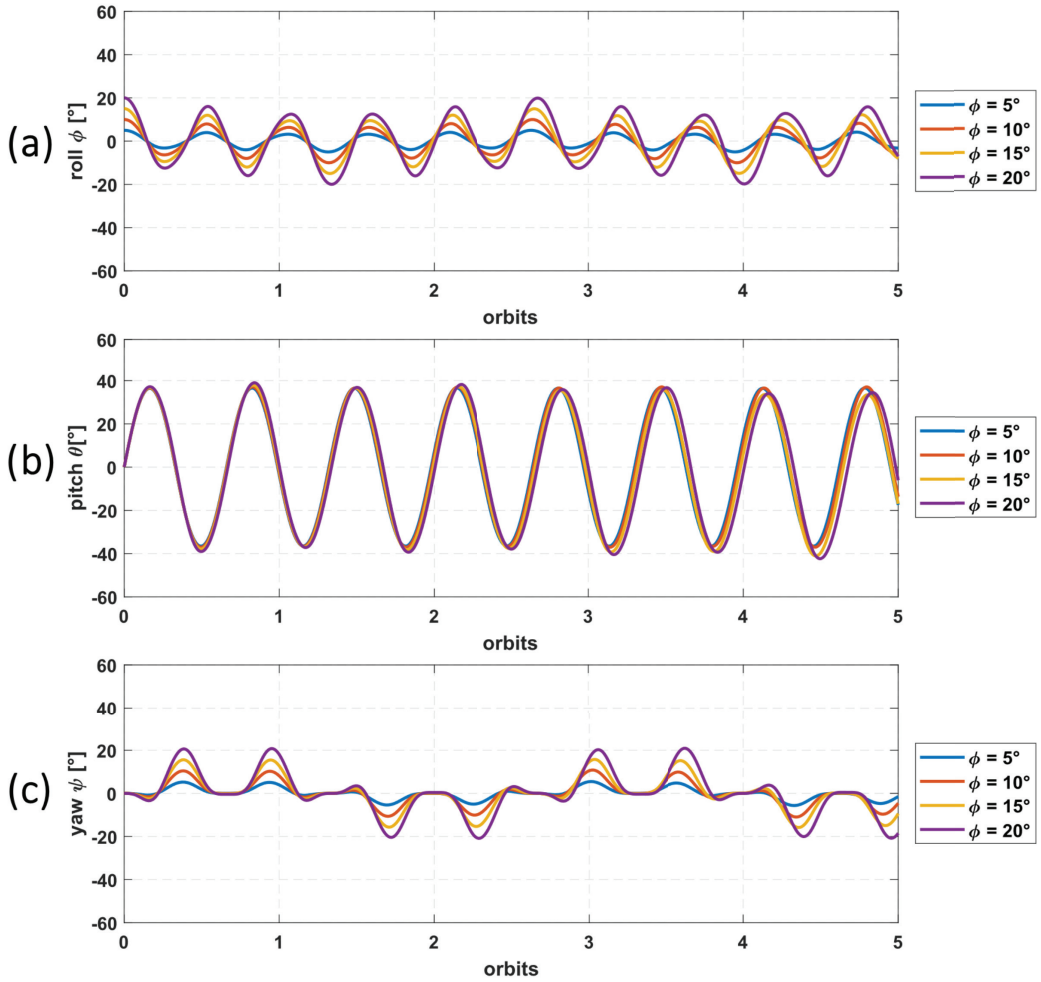


Figure 6. Propagation of attitude angles for roll (**a**), pitch (**b**), and yaw (**c**) with different initial conditions of roll angle ϕ.

The yaw angle similarly oscillates around its initial value when the system is subjected to initial yaw perturbations (Table 6), as seen in Figure 7c, with maximum amplitudes up to 25°. Figure 7a illustrates that these yaw perturbations induce oscillations in the roll angle. However, the amplitude of the induced roll oscillations (up to approximately $\pm 15°$) is quantitatively smaller than the amplitude of yaw oscillations induced by initial roll perturbations. The pitch angle (Figure 7b) again exhibits minimal change, confirming its relative isolation from direct roll and yaw influences under these specific initial conditions.

Although the results to this point demonstrate the manner in which the effect of the gravity gradient varies with the initial attitude angles, the case studies analysed thus far are of limited representativeness for a real-life situation in which attitude angles can have different values at the same time. Consequently, numerical simulations were conducted with the particular initial attitude conditions shown in Table 7.

Table 6. Study on the influence of yaw propagation ψ analysed with the initial conditions.

Roll ϕ	Pitch θ	Yaw ψ
0°	0°	5°
0°	0°	10°
0°	0°	15°
0°	0°	20°

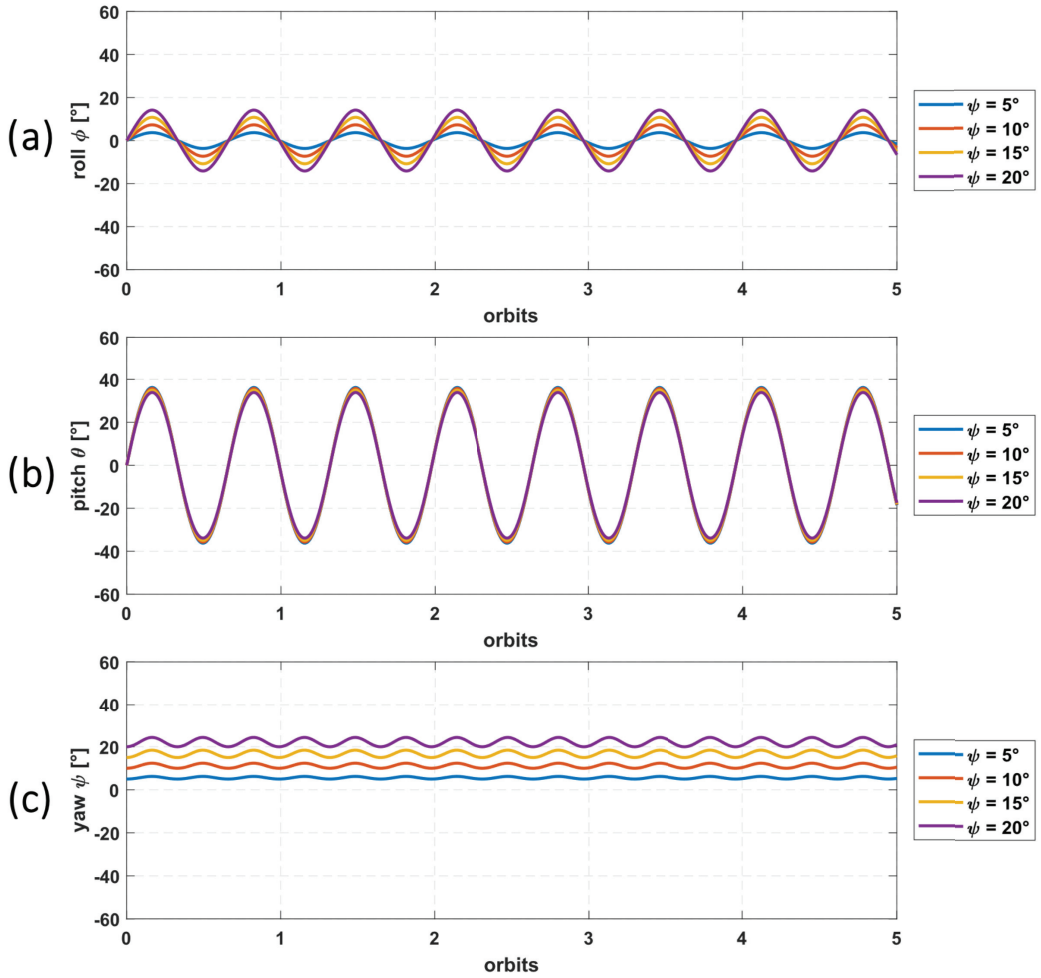


Figure 7. Propagation of attitude angles for roll (**a**), pitch (**b**), and yaw (**c**) with different initial conditions of ψ.

Table 7. Study on the influence of Euler angle propagation analysed with the initial conditions.

Case	Roll ϕ	Pitch θ	Yaw ψ
a	5°	5°	5°
b	10°	10°	10°
c	15°	15°	15°
d	20°	20°	20°
e	15°	20°	5°

To represent more realistic scenarios where multiple attitude angles can be perturbed simultaneously, simulations were conducted with combined initial conditions (Table 7). Figure 8 presents these results. While the behaviour becomes more complex due to the simultaneous coupling, the observed maximum amplitudes for pitch, roll, and yaw remain within the ranges previously identified (pitch $\approx \pm 40°$, roll $\approx \pm 20°$, yaw $\approx \pm 25°$). These simulations further underscore the inherent nonlinearity and strong coupling between the roll and yaw axes under gravity gradient stabilization.

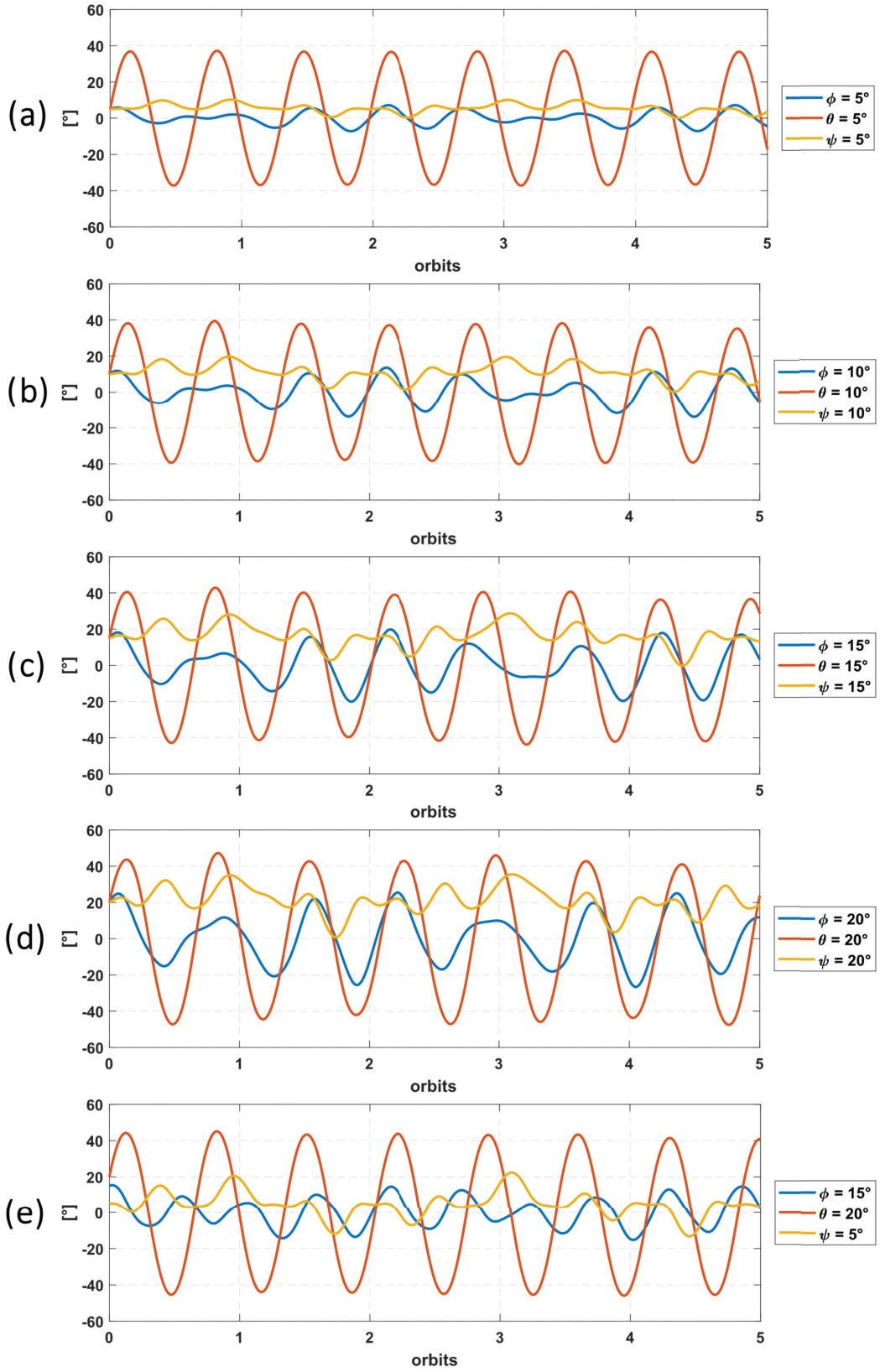


Figure 8. This figure demonstrates how the attitude angles change in response to different initial conditions of roll, pitch, and yaw. Each figure corresponds to cases (**a**–**e**) detailed in Table 7.

In summary, it is observed that as the initial attitude angles change, the behaviour of the CubeSat with C-arm changes when subjected to the effect of the gravity gradient alone. In particular, the simulation results demonstrate how the pitch angle always oscillates around $\theta = \pm 40^{\circ}$ when the initial conditions vary. The roll and yaw angles exhibit a more

variable trend, with a maximum of $\phi \approx 20^\circ$ and $\psi \approx 25^\circ$ in case (e). The next step is to incorporate the effect of the Earth's magnetic field on the satellite's attitude, with the aim of exploiting a permanent magnet to reduce the pitch angle for improved pointing.

5. Implementing the Action of the Earth's Magnetic Field on Attitude

The outcomes of the investigation conducted in Section 4.2 employing passive gravity gradient control demonstrates that the robotic manipulator is unable to orient the CubeSat to the predefined target using only this tool.

This subsequent step ascertains whether the incorporation of a magnetic passive control system can enhance the satellite's attitude. The objective is to utilize the manipulator to orient permanent magnets in accordance with the CubeSat's desired orientation. The magnets are positioned in a slot specifically designed for this purpose and located in link 3 of the robotic manipulator, as illustrated in Figure 9.

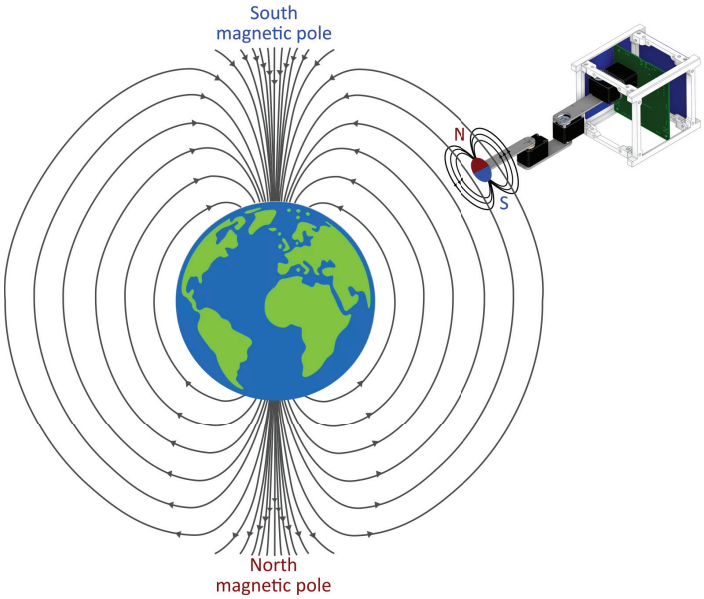


Figure 9. Schematic illustration of the alignment magnet concept using a permanent magnet located in link 3 of the robotic manipulator.

The selection of permanent magnets was based on their reliability in maintaining uniaxial attitude control. This is due to their high magnetic moment, which aligns the satellite with the tangent of the Earth's magnetic field. For communication purposes, uniaxial orientation is sufficient.

Based on previous research [28], four cylindrical neodymium magnets with external and internal diameters of 7 mm and 3.4 mm, respectively, and a height of 11.5 mm are placed on the corresponding plate to create a magnetic field in link 3 of the C-arm. Each magnet has a dipole moment of $m_M = 0.3095$ Am^2 along the axis of symmetry, while the radial component is negligible.

The magnetic torque is a consequence of the interaction between the Earth's magnetic field and the satellite's magnetic dipole moment.

$$\mathbf{M}_M = \mathbf{m}_M \times \mathbf{B} \tag{7}$$

In addition to the dipole moment of the permanent magnet, the effect of residual magnetic disturbances in the CubeSat must be considered. The effect of all these magnetic moments is denoted as $\mathbf{m}_m$ in Equation (7).

The Earth's magnetic field, designated as **B** in Equation (7), is represented by the International Geomagnetic Reference Field-13 (IGRF-13) model. It was necessary to recreate the model in the ADAMS multibody software, although the geomagnetic field model is fully defined in [29] and there are numerous programs and software that implement it.

The IGRF describes the main geomagnetic field $\mathbf{B}(r, \theta, \phi, t)$ as the gradient of a scalar potential $\mathbf{B} = -\nabla V$ and the potential (8) as a finite series expansion in terms of spherical harmonic coefficients g_n^m, h_n^m, also known as Gauss coefficients:

$$V(r, \theta, \phi, t) = a \sum_{n=1}^{N} \left(\frac{a}{r}\right)^{n+1} \sum_{m=0}^{n} (g_n^m(t) \cos m\phi + h_n^m(t) \sin m\phi) P_n^m(\cos\theta) \tag{8}$$

where r, θ, ϕ refer to coordinates in a geocentric spherical coordinate system ($\mathcal{T}_{ECEF}$), with r as the radial distance from the Earth's center and θ, ϕ respectively representing the geocentric latitude and longitude. A reference radius $a = 6371.2$ km is chosen to approximate the mean radius of the Earth. Here, $P_n^m(\cos\theta)$ are Schmidt semi-normalised associated Legendre functions of degree n and order m. The parameter N indicates the maximum degree of spherical harmonic expansion. For the objectives predicted for this work, the series expansion was stopped at the first order, obtaining the potential of the dipole model of the geomagnetic field (Equation (9)):

$$V(r, \theta, \phi, t) = \frac{a^3}{r^2}\Big[\, g_1^0 P_1^0(\theta) + \Big(\, g_1^1 \cos\phi + h_1^1 \sin\phi\Big) P_1^1(\theta)\Big]. \tag{9}$$

The components of **B** are provided in Equation (10) with respect to a geocentric non-inertial coordinate system. The Gauss coefficients for the year 2023 are shown in Table 8.

$$\begin{aligned} B_r &= -\frac{\partial V}{\partial r} = 2\Big(\frac{a}{r}\Big)^3\Big[\, g_1^0 \cos\theta + \Big(\, g_1^1 \cos\phi + h_1^1 \sin\phi\Big)\sin\theta\Big] \\ B_\theta &= -\frac{1}{r}\frac{\partial V}{\partial \theta} = \Big(\frac{a}{r}\Big)^3\Big[\, g_1^0 \sin\theta - \Big(\, g_1^1 \cos\phi + h_1^1 \sin\phi\Big)\cos\theta\Big] \\ B_\phi &= -\frac{1}{r\sin\theta}\frac{\partial V}{\partial \phi} = \Big(\frac{a}{r}\Big)^3\Big[\, g_1^1 \sin\phi - h_1^1 \cos\phi\Big] \end{aligned} \tag{10}$$

Table 8. Gauss coefficients for the year 2023.

Coefficient	Value [nT]
g_1^0	−29,399.1
g_1^1	−1443.5
h_1^1	4626.6

Equation (10) expresses the Earth's magnetic field in a non-inertial geocentric frame of reference. It should be noted for the sake of completeness that to calculate the magnetic external torque it is first necessary to transform the **B** component to the body reference frame $\mathcal{T}_\mathcal{B}$. The magnetic moment of the magnets is designated as $\mathbf{M}_M$, as illustrated in Figure 10, and calculated in accordance with the methodology outlined in Equation (7).

5.1. Multibody Implementation

To incorporate the resulting moment caused by the interaction between the Earth's magnetic field and the satellite's dipole moment, an update to the model presented in Section 4.1 is required. In an initial step, we assessed the magnitude of the magnetic perturbations acting on the satellite. The satellite is equipped with electronic components, including circuit boards, robotic arm motors, and batteries, which are sources of magnetic

perturbations and are modelled as a dipole moment. The study parameter we employed was the dipole moment ($\mathbf{m}_m$), with the reference values provided in Table 9. The parameters were selected with a degree of conservatism, using values generally in excess of the actual order of magnitude of the CubeSat's dipole moments. The moment acting on the body due to this perturbation, denoted as $\mathbf{M}_m$, is applied to the center of mass (CoM) of the CubeSat and calculated in a manner analogous to that described in Equation (7).

Table 9. Values of **m** in Am^2.

	$\mathbf{m}_m$ **Am²**		
	m_x	m_y	m_z
Test 1	1×10^{-3}	1×10^{-3}	1×10^{-3}
Test 2	0.1×10^{-3}	0.1×10^{-3}	0.1×10^{-3}
Test 3	0.05×10^{-3}	0.05×10^{-3}	0.05×10^{-3}

It was anticipated that the simulations presented in Section 5.2 would be illustrated only with the perturbation defined in Test 2, since the effect of the magnet is stronger than that of the residual dipole effects. Therefore, the results of every test of Table 9 are very similar. After defining the action of the magnetic perturbation torque and the action of the permanent magnet torque, we integrated these torques into the multibody model, as illustrated in Figure 10.

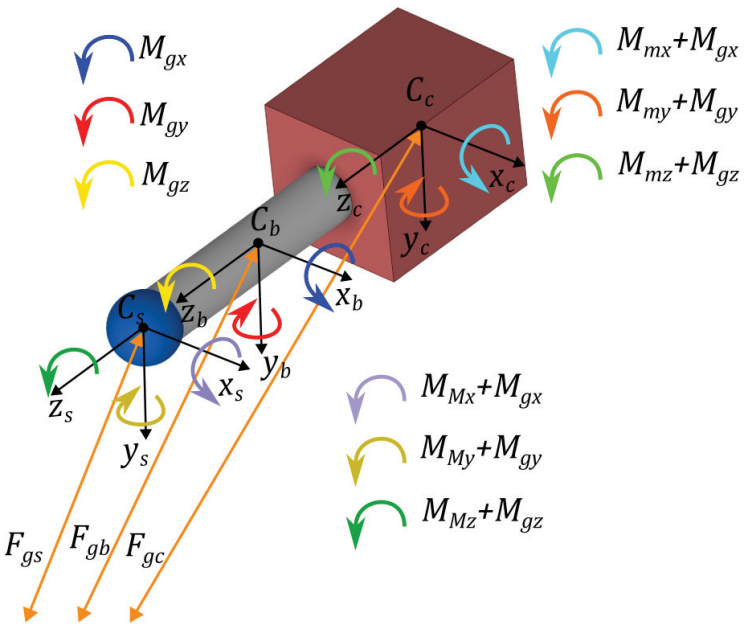


Figure 10. Representation of gravitational forces F_{gs}, F_{gb}, F_{gc}, gravity torque components M_{gx}, M_{gy}, M_{gz}, magnetic disturbance torque components M_{mx}, M_{my}, M_{mz}, and magnetic torque components of permanent magnets M_{Mx}, M_{My}, M_{Mz} acting on the simplified model.

Simulations were performed using the orbit described in Section 4.1, as previously; however, the initial conditions of the attitude (provided in Table 10) differed from those used in the case of gravitational control. In this case, we assumed ideal initial conditions of motion at equilibrium.

Table 10. Initial attitude conditions.

ϕ_0	θ_0	ψ_0	$\omega_{b_{x_0}}$	$\omega_{b_{y_0}}$	$\omega_{b_{z_0}}$
0°	0°	0°	$0^\circ/\mathrm{s}$	$0^\circ/\mathrm{s}$	$0^\circ/\mathrm{s}$

In this way, the magnetic moment effect was incorporated into the simplified multi-body model. At this stage, the model incorporates the effects of both the gravity gradient torque and the magnetic moment.

5.2. Final Results

The simulations utilizing a permanent magnet revealed a complex and predominantly unstable attitude behaviour. While the magnet is intended to provide a pointing mechanism, the results (particularly those shown in Figure 11c) indicate the induction of unstable rotation around the z-axis (yaw angle ψ). Quantitatively, yaw oscillations frequently exceeded 35°. Similarly, pitch oscillations continued to exhibit amplitudes in the range of $\pm 40^\circ$ and roll oscillations frequently exceeded 30°, as is visible in Figure 11a,b. This widespread instability across multiple axes stems from the tight coupling of rotations inherent in the Euler equations (Equation (3)). Consequently, instability originating or amplified in any one axis is propagated, causing the pitch and roll angles to exceed the desired limits.

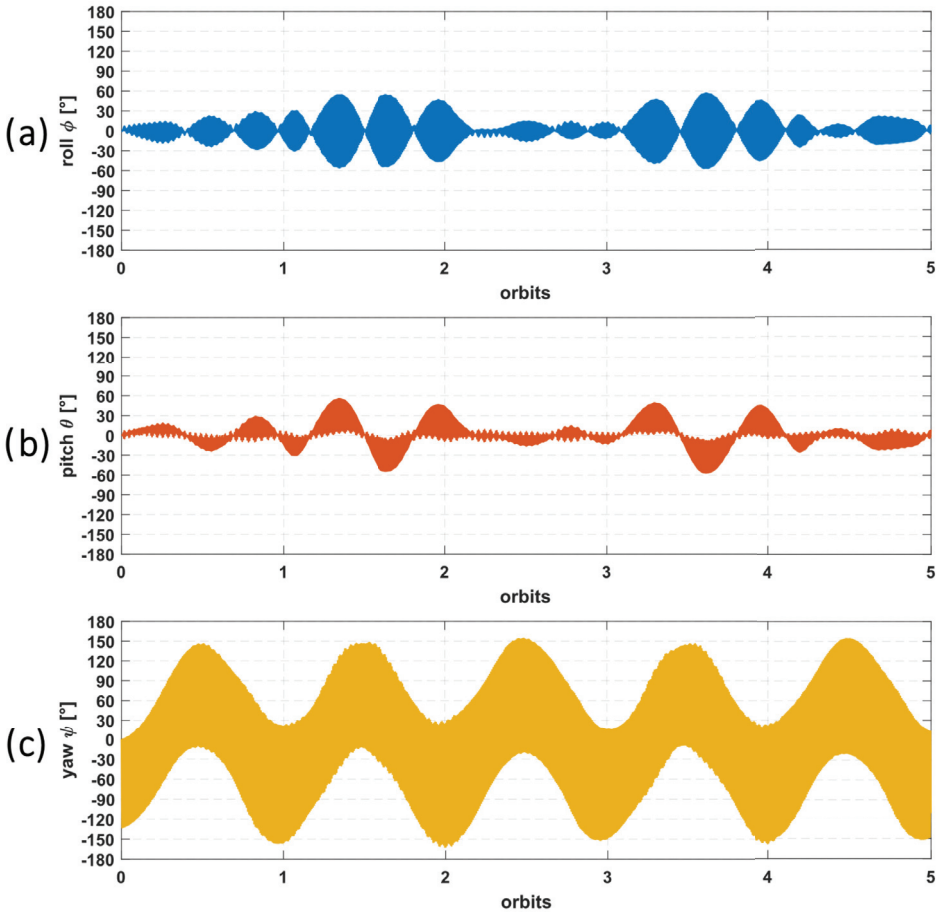


Figure 11. Propagation of the angles of attitude ϕ, θ, ψ when motion is performed with permanent magnets placed on the sphere and magnetic perturbation torques acting on the cube.

A critical and distinct quantitative observation from these simulations is the emergence of significantly higher-frequency oscillations superimposed on the underlying low-

frequency gravity gradient motion, which is particularly noticeable in Figure 11a,c. These rapid oscillations, with periods observed to be on the order of a few tens of seconds (markedly shorter than the orbital period of approximately 90 min), are a direct consequence of the permanent magnet's continuous and rapid attempt to align with the dynamic local geomagnetic field. This active interaction introduces additional kinetic energy into the system, emphasizing the critical need for robust damping mechanisms that can dissipate this induced energy and achieve stable pointing.

6. Conclusions and Future Work

The objective of this study was to evaluate the effectiveness of utilising a robotic arm as a passive device to control the attitude of a microsatellite. In particular, a robotic arm that is compatible with a 1U CubeSat, designated the C-arm, was developed in this study. The compact dimensions of the robotic arm imposed notable constraints in terms of size and weight (the latter reaching 200 g). This was particularly evident in the maximum extension length of the arm, which when retracted occupied 0.5 U. Subsequently, simulations of attitude propagation conducted using a simplified multibody satellite model demonstrated the efficacy of this innovative approach as a tool for achieving the objectives of this research activity. The effectiveness of two passive actions on the satellite's attitude control was evaluated: the gravitational effect produced by the extended arm, and the action of the Earth's magnetic field when a permanent magnet was introduced at the end of the C-arm.

The results of simulations that exclusively considered the gravitational effect demonstrated that despite the small size of the robotic arm, its influence on the temporal evolution of the satellite's attitude motion as a consequence of the torque exerted by the gravitational gradient is not negligible. Our results showed that even though the effects of this external torque are negligible for a 1U CubeSat (and for compact microsatellites in general), this is not the case when the C-arm linked with the satellite is completely deployed. In particular, we found that when the initial conditions vary, the pitch angle always oscillates at a low frequency around $\theta = \pm 40^\circ$. The roll and yaw angles exhibit a more variable trend but always at low frequency, with maximum values lower than the pitch angle at $\phi \approx 30^\circ$ and $\psi \approx 35^\circ$. These results, while significant for a 1U CubeSat, suggest that a purely passive gravity gradient approach with this C-arm configuration may be suitable for missions with less demanding pointing accuracy requirements, as is typical of many nanosatellite applications.

In a subsequent step, we incorporated the effect of the Earth's magnetic field on the satellite's attitude. This was achieved by exploiting a permanent magnet to reduce the pitch angle and achieve improved pointing. The interaction provides a magnetic dipole moment parallel to the longitudinal axis of the robotic arm (in an extended position), ensuring a three-dimensional attitude motion even when the initial value of the Euler angles and angular rates are null. The pitch, roll, and yaw angles do not exhibit significant maximum value variations; however, the magnetic interaction results in high-frequency oscillations of all the Euler angles, which overlap with the low-frequency oscillations associated with the gravity gradient torque. Consequently, the aforementioned desired attitude is established in a shorter time frame. Notably, we observed a temporary reduction in pitch oscillations with the use of magnets.

In conclusion, the results of this study prove that integration of a robotic arm into a CubeSat has a strong influence on the attitude motion of the satellite, and as such can be considered as a valid and effective tool to passively control the attitude of the CubeSat with medium/high accuracy of the pitch angle.

Limitations and Future Developments of the Study

This preliminary study offers valuable insights into using a robotic arm for passive CubeSat attitude control. However, it is essential to acknowledge several inherent simplifications which define its current scope. Our multibody model of the C-arm simplifies its structure (representing it as a rigid cylinder and sphere) and neglects detailed joint flexibilities. Crucially, the model also excludes significant dissipative forces such as atmospheric drag, solar radiation pressure, and internal damping, which are vital for achieving stable and settled attitudes in real-world scenarios. In addition, this work focused exclusively on passive control and explored a limited range of initial conditions and scenarios.

These limitations highlight the foundational nature of this research. Future work will directly address these aspects. We plan to refine our multibody models to incorporate more detailed CubeSat and arm characteristics, including joint flexibilities. In addition, we will introduce mechanisms such as soft magnetic devices for dissipating rotational kinetic energy in order to achieve stable oscillations. Further research will expand the analysis in this paper to include other environmental disturbances such as atmospheric drag and solar radiation pressure as well as to explore active control strategies using the robotic arm, potentially integrating magnetic coils or controlled movements for precise pointing and disturbance rejection. Lastly, we aim to investigate a wider range of orbital conditions and C-arm configurations in order to assess its robustness and adaptability.

Author Contributions: Conceptualization, F.F., G.M. and M.D.; data curation, F.F.; formal analysis, F.F. and G.M.; funding acquisition, M.D.; investigation, G.M.; methodology, F.F., G.M. and S.C.; project administration, M.D.; software, F.F. and M.P.; supervision, F.C. and M.D.; validation, M.P.; visualization, S.C. and F.C.; writing—original draft, F.F. and G.M.; writing—review and editing, G.M. and S.C. All authors have read and agreed to the published version of the manuscript.

Funding: The research presented in this paper has been funded by Università degli Studi di Perugia within the UNIPG-ASI Agreement "Accordo ASI Unipg N. 2019-2-HH.0: Realizzazione di attività di Ricerca e Sviluppo competenze innovative".

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Nomenclature

a	Semi-major axis; Earth's reference radius
B	Earth's magnetic field vector
B_r, B_θ, B_ϕ	Components of Earth's magnetic field in spherical coordinates
CoM	Center of Mass
C-Arm	CubeSat Robotic Arm
d	Diameter of cylindrical/spherical body
DoF	Degrees of Freedom
e	Eccentricity
ECI	Earth-Centered Inertial frame
ERA	European Robotic Arm
ETS-VII	Engineering Test Satellite VII
F_g	Gravitational force
F_{gb}	Gravitational force on cylinder (arm)
F_{gc}	Gravitational force on CubeSat body
F_{gs}	Gravitational force on sphere (end-effector)
G	Gravitational constant
g_n^m, h_n^m	Gauss coefficients for geomagnetic model
H_c	Total angular momentum of the satellite

i	Inclination of the orbit
$\mathbb{I}_B$	Satellite's inertia tensor
I_x, I_y, I_z	Principal moments of inertia of the satellite
IGRF-13	International Geomagnetic Reference Field-13
ISS	International Space Station
JEMRMS	Japanese Experiment Module Remote Manipulator System
L_x, L_y, L_z	Dimensions of CubeSat body
LEO	Low-Earth Orbit
LISA	Little Inspection and Servicing Arm
MBS	Mobile Base System
M_c	Total external perturbation and control torques
M_E	Mass of the Earth
M_g	Gravity gradient torque
M_{gx}, M_{gy}, M_{gz}	Components of gravity gradient torque in body frame
M_m	Magnetic disturbance torque
m_m	Satellite's total magnetic dipole moment
m_x, m_y, m_z	Components of residual magnetic dipole moment (from disturbances)
M_{mx}, M_{my}, M_{mz}	Components of magnetic disturbance torque in body frame
M_M	Magnetic torque from permanent magnet
M_{Mx}, M_{My}, M_{Mz}	Components of magnetic torque from permanent magnet
M_s	Satellite mass
MSS	Mobile Servicing System
N	Maximum degree of spherical harmonic expansion
P_n^m	Schmidt semi-normalised associated Legendre functions
R	Orbit radius
RMS	Remote Manipulator System
r_s	Satellite position vector
$\ddot{r}_s$	Satellite acceleration vector
RSat	Robotic Satellite project
SPDM	Special-Purpose Dextrous Manipulator
SSRMS	Space Station Remote Manipulator System
SUMO	Spacecraft for the Universal Modification of Orbits
t	Time
T_B	Body frame
$\mathcal{T}_{ECI}$	Earth-centered inertial frame
T_O	Orbit frame
V	Geomagnetic scalar potential
ϕ	Roll angle
ϕ_0	Initial roll angle
θ	Pitch angle
θ_0	Initial pitch angle
ψ	Yaw angle
ψ_0	Initial yaw angle
μ	Earth gravitational parameter
Ω	Right Ascension of the Ascending Node (RAAN)
ω	Argument of periapsis
ω_B	Angular velocity of body frame with respect to ECI
$\dot{\omega}_B$	Angular acceleration of body frame with respect to ECI
ω_{BO}	Angular velocity of the satellite body axis with respect to the orbit frame
$\omega_{bx0}, \omega_{by0}, \omega_{bzo}$	Initial angular velocity components in body frame
ω_{OI}	Angular velocity of the orbit frame with respect to the Earth-centered inertial frame

References

1. Lion, L.; Caon, A.; Olivieri, L.; Branz, F.; Francesconi, A. Kinematic tests on a docking mechanism for microsatellites. *CEAS Space J.* **2024**, *16*, 445–455. [CrossRef]
2. Basana, F.; Pavanello, Z.; Branz, F.; Francesconi, A.; Borelli, G.; Invernizzi, D.; Massari, M.; Lovera, M. Satellite and robotic arm combined control for spacecraft close-proximity operations. *CEAS Space J.* **2024**, *17*, 309–335. [CrossRef]
3. Sąsiadek, J. Space Robotics and its Challenges. In *Aerospace Robotics: Selected Papers from I Conference on Robotics in Aeronautics and Astronautics*; Sąsiadek, J., Ed.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 1–8. [CrossRef]
4. Hiltz, M.; Rice, C.; Boyle, K.; Allison, R. CANADARM: 20 Years of Mission Success Through Adaptation. In Proceedings of the 6th International Symposium on Artificial Intelligence and Robotics & Automation in Space: I-SAIRAS 2001, St-Hubert, QC, Canada, 18–22 June 2001.
5. Laryssa, P.; Lindsay, E.; Layi, O.; Marius, O.; Nara, K.; Aris, L.T. International Space Station Robotics: A Comparative Study of ERA, JEMRMS and MSS. In Proceedings of the 7th ESA Workshop on Advanced Space Technologies for Robotics and Automation (ASTRA), ESTEC, Noordwijk, The Netherlands, 19–21 November 2002.
6. Oda, M. ETS-VII: Achievements, Troubles and Future. In Proceedings of the 6th International Symposium on Artificial Intelligence and Robotics & Automation in Space: I-SAIRAS 2001, St-Hubert, QC, Canada, 18–22 June 2001.
7. Friend, R.B. Orbital Express program summary and mission overview. In *Proceedings of the Sensors and Systems for Space Applications II*; Howard, R.T., Motaghedi, P., Eds.; International Society for Optics and Photonics: Bellingham, WA, USA, 2008; Volume 6958, p. 695803. [CrossRef]
8. Bosse, A.B.; Barnds, W.J.; Brown, M.A.; Creamer, N.G.; Feerst, A.; Henshaw, C.G.; Hope, A.S.; Kelm, B.E.; Klein, P.A.; Pipitone, F.; et al. SUMO: Spacecraft for the universal modification of orbits. In *Proceedings of the Spacecraft Platforms and Infrastructure*; SPIE: Bellingham, WA, USA, 2004; Volume 5419, pp. 36–46. [CrossRef]
9. Agenzia Spaziale Italiana (ASI). Tecnologia e Ingegneria di Micro e Nanosatelliti. Available online: https://www.asi.it/tecnologia-ingegneria-micro-e-nanosatelliti/ (accessed on 27 March 2024).
10. Millan, R.M.; von Steiger, R.; Ariel, M.; Bartalev, S.; Borgeaud, M.; Campagnola, S.; Castillo-Rogez, J.C.; Fléron, R.; Gass, V.; Gregorio, A.; et al. Small satellites for space science: A COSPAR scientific roadmap. *Adv. Space Res.* **2019**, *64*, 1466–1517. [CrossRef]
11. Poghosyan, A.; Golkar, A. CubeSat evolution: Analyzing CubeSat capabilities for conducting science missions. *Prog. Aerosp. Sci.* **2017**, *88*, 59–83. [CrossRef]
12. Au, A.S.; Peña, R.; Martinez, J.M.; Greig, A.D.; Quintana, J.; Flores-Abad, A.; Choudhuri, A. Elastically Actuated Small Robotic Arm for CubeSat Platforms. In Proceedings of the ASCEND 2021, Las Vegas, NV, USA & Virtual, 15–17 November 2021. [CrossRef]
13. Liu, J.; Zhao, P.; Chen, K.; Zhang, X.; Zhang, X. 1U-Sized Deployable Space Manipulator for Future On-Orbit Servicing, Assembly, and Manufacturing. *Space Sci. Technol.* **2022**, *2022*, 9894604. [CrossRef]
14. Örger, N.C.; Karyot, T.B. A symmetrical robotic arm design approach with stereo-vision ability for CubeSats. In Proceedings of the 2013 6th International Conference on Recent Advances in Space Technologies (RAST), Istanbul, Turkey, 12–14 June 2013; pp. 961–965. [CrossRef]
15. McCormick, R.; Austin, A.; Wehage, K.; Backus, S.; Miller, R.; Leith, J.; Bradley, B.; Durham, P.; Mukherjee, R. REMORA CubeSat for large debris rendezvous, attachment, tracking, and collision avoidance. In Proceedings of the 2018 IEEE Aerospace Conference, Big Sky, MT, USA, 3–10 March 2018; pp. 1–13. [CrossRef]
16. McCormick, R.; Austin, A.; Cubrich, L.; Marth, B.; Mukherjee, R.; Wenger, T.; Patel, M.; Ho, K. Development of miniature robotic manipulators to enable SmallSat clusters. In Proceedings of the 2017 IEEE Aerospace Conference, Big Sky, MT, USA, 4–11 March 2017; pp. 1–15. [CrossRef]
17. Zeis, C.; de Alba-Padilla, C.A.; Schroeder, K.U.; Grzesik, B.; Stoll, E. Fully Modular Robotic Arm Architecture Utilizing Novel Multifunctional Space Interface. *IOP Conf. Ser. Mater. Sci. Eng.* **2022**, *1226*, 012096. [CrossRef]
18. Hanlon, E.A.; Keegan, B.P.; Lange, M.E.; Pittman, J.K.; Wenberg, D.L.; Roser, J.G.; Kang, J.S. Leveraging the Autonomous Mobile On-Orbit Diagnostic System to Initiate a Doctrinal Shift in Spacecraft Operations. In *Space Operations: Contributions from the Global Community*; Springer International Publishing: Cham, Switzerland, 2017; pp. 53–86. [CrossRef]
19. Wenberg, D.L.; Hanlon, E.A.; Rubiocastaneda, B.; Lai, T.; Kang, J. Intelligent Space Assembly Robot: Design and Ground Testing. In Proceedings of the AIAA SPACE and Astronautics Forum and Exposition, Orlando, FL, USA, 12–14 September 2017. [CrossRef]
20. Ruf, O.; von Arnim, M.; Kempf, F.; Haber, R.; Elsner, L.; Dauner, J.; Dombrovski, S.; Kramer, A.; Schilling, K. Advanced test environment for automated attitude control testing of fully integrated CubeSats on system level. *CEAS Space J.* **2024**, *16*, 491–510. [CrossRef]
21. Eggert, M.; Bötsch-Zavřel, L.; Fugmann, M.; Holeczek, C.; Koller, M.T.; Kranz, M.; Lengowski, M.; Löffler, T.; Loidold, L.-M.; Maier, P.; et al. Qualification of the attitude control system of the E-band CubeSat EIVE. *CEAS Space J.* **2024**, *17*, 293–308. [CrossRef]
22. Henna, H.; Toubakh, H.; Kafi, M.R.; Sayed-Mouchaweh, M.; Djemai, M. Hybrid supervision scheme for satellite attitude control with sensor faults. *CEAS Space J.* **2023**, *16*, 753–767. [CrossRef]

23. Girardello, C.; Tajmar, M.; Scharlemann, C. GREATCUBE+: Conceptual design tool for CubeSat's design. *CEAS Space J.* **2024**, *16*, 375–392. [CrossRef]
24. Curtis, H.D. *Orbital Mechanics for Engineering Students: Revised Reprint*; Butterworth-Heinemann: Oxford, UK, 2020.
25. Wafi, I.E.; Haloua, M.; Guennoun, Z.; Moudden, Z. Hybrid adaptive sliding mode attitude control for earth pointing nanosatellites. *CEAS Space J.* **2024**, *16*, 307–318. [CrossRef]
26. Wertz, J.R. *Spacecraft Attitude Determination and Control*; Springer Science & Business Media: New York, NY, USA, 2012; Volume 73.
27. Sidi, M.J. *Spacecraft Dynamics and Control: A Practical Engineering Approach*; Cambridge Aerospace Series; Cambridge University Press: Cambridge, UK, 1997.
28. Carletta, S.; Nascetti, A.; Gosikere Matadha, S.S.; Iannascoli, L.; Baratto de Albuquerque, T.; Maipan Davis, N.; Schirone, L.; Impresario, G.; Pirrotta, S.; Brucato, J.R. Characterization and Testing of the Passive Magnetic Attitude Control System for the 3U AstroBio CubeSat. *Aerospace* **2022**, *9*, 723. [CrossRef]
29. Alken, P.; Thébault, E.; Beggan, C.D.; Amit, H.; Aubert, J.; Baerenzung, J.; Bondar, T.N.; Brown, W.J.; Califf, S.; Chambodut, A.; et al. International Geomagnetic Reference Field: The thirteenth generation. *Earth Planets Space* **2021**, *73*, 49. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Review

Recent Advances in Multi-Agent Reinforcement Learning for Intelligent Automation and Control of Water Environment Systems

Lei Jia [1,†] and Yan Pei [2,*,†]

[1] Graduate School of Computer Science and Engineering, University of Aizu, Aizuwakamatsu 965-8580, Fukushima, Japan
[2] Computer Science Division, University of Aizu, Aizuwakamatsu 965-8580, Fukushima, Japan
* Correspondence: peiyan@u-aizu.ac.jp
† These authors contributed equally to this work.

Abstract: Multi-agent reinforcement learning (MARL) has demonstrated significant application potential in addressing cooperative control, policy optimization, and task allocation problems in complex systems. This paper focuses on its applications and development in water environmental systems, providing a systematic review of the theoretical foundations of multi-agent systems and reinforcement learning and summarizing three representative categories of mainstream MARL algorithms. Typical control scenarios in water systems are also examined. From the perspective of cooperative control, this paper investigates the modeling mechanisms and policy coordination strategies of MARL in key tasks such as water supply scheduling, hydro-energy co-regulation, and autonomous monitoring. It further analyzes the challenges and solutions for improving global cooperative efficiency under practical constraints such as limited resources, system heterogeneity, and unstable communication. Additionally, recent progress in cross-domain generalization, integrated communication–perception frameworks, and system-level robustness enhancement is summarized. This work aims to provide a theoretical foundation and key insights for advancing research and practical applications of MARL-based intelligent control in water infrastructure systems.

Keywords: reinforcement learning; multi-agent systems; cooperative control; intelligent environmental systems; Air-space-ground-water integrated framework

1. Introduction

A multi-agent system (MAS) is a complex system composed of multiple agents equipped with capabilities for computation, perception, communication, learning, and execution. Through localized sensing and interactive collaboration, these agents achieve coordinated control and optimal decision-making for global tasks. Compared with single-agent systems, a MAS offers greater flexibility and robustness in addressing large-scale, complex problems characterized by dynamics, uncertainty, and high-dimensional state spaces. MASs have thus found widespread application in various domains, including cooperative robotics, intelligent transportation, unmanned aerial vehicle (UAV) formation control, and power system scheduling. In recent years, with the continuous advancement of intelligent technologies, multi-agent cooperative control has become a core topic across several interdisciplinary fields, such as artificial intelligence, automation, game theory, and control theory, and is recognized as one of the key theoretical challenges in next-generation AI [1,2].

Traditional research on multi-agent cooperative control has primarily focused on consensus-based approaches, emphasizing the interplay between agent dynamics and information exchange mechanisms. These studies aim to address distributed consensus control under complex constraints, such as limited communication, dynamic topologies, and actuator saturation. However, as agent objectives become increasingly diverse and autonomous, the conventional single-objective coordination paradigm is no longer sufficient to capture complex interactions in real-world scenarios. To address this limitation, an increasing number of studies have adopted a game-theoretic framework, modeling agents as rational decision-makers engaged in strategic interactions driven by individual objectives. Through strategy optimization under multi-objective trade-offs, system-level coordination is achieved. This integration of game theory and control theory has shown promising results in robust control, optimal control, and cooperative estimation, providing a unified modeling foundation and solution framework to describe autonomy, interaction, and cooperation in MASs [3,4].

On the other hand, reinforcement learning (RL) has emerged as a powerful tool for cooperative MAS research due to its capability to optimize adaptive policies in model-free environments. Multi-agent reinforcement learning (MARL) further combines the representational power of deep learning with the policy optimization ability of RL, enabling agents to learn effective strategies under incomplete information and unstable environments. Building upon this foundation, deep MARL methods have been widely applied in domains such as autonomous formation control, intelligent manufacturing, and traffic management, demonstrating strong generalization and scalability [5,6].

With the rapid advancement of complex systems research driven by large-scale models, high-performance computing, and big data, water environment systems have emerged as a critical application domain for multi-agent deep reinforcement learning (MADRL) methods. These systems typically exhibit the following key characteristics: (1) an inherently multi-agent structure consisting of components such as pump stations, monitoring sites, control gates, and unmanned underwater or surface platforms; (2) strong spatiotemporal heterogeneity and nonlinear coupling in system dynamics; (3) the need to coordinate multiple sub-objectives, including ensuring water supply security, reducing energy consumption, and responding to extreme events; and (4) high levels of uncertainty, data sparsity, and stringent real-time requirements [7,8].

Against this backdrop, MADRL offers novel theoretical tools and practical approaches for modeling and optimizing water environment systems. Representative application scenarios include the following:

- Intelligent pump station scheduling: Modeling multiple pump stations as cooperative agents, MADRL integrates predictive management with adaptive control to achieve optimal trade-offs between energy consumption and water supply efficiency.
- Urban flood control: Under extreme rainfall conditions, drainage nodes must collaboratively optimize retention and discharge strategies. MADRL enables the learning of disaster response policies in simulated environments.
- Underwater monitoring and detection systems (USV/UGV/AUV): Teams of autonomous surface or underwater vehicles execute long-term monitoring and pollutant source tracking through cooperative path planning and area coverage optimization [9,10].
- Joint scheduling of multiple water sources and watershed management: In large-scale basins with coupled water sources, MADRL supports multi-objective water resource allocation and coordination.

- **Emergency pollution control:** In response to sudden water quality events, multiple sensing and control nodes collaborate to optimize the timing and spatial extent of emergency interventions.

To date, a growing body of literature has systematically reviewed MADRL approaches from various perspectives. For example, Yang et al. explored multi-agent strategy evolution mechanisms within a game-theoretic (especially meta-game) framework [11]; Shi et al. investigated experience sharing and policy generalization among agents through transfer learning [12]; Han et al. provided a comprehensive review of hierarchical multi-agent reinforcement learning algorithms [13]; and Luo et al. examined the differences between online and offline learning paradigms in the context of game-theoretic modeling [14]. Furthermore, challenges such as scalability, heterogeneity, and stability in large-scale multi-agent systems have spurred research into emerging directions such as "massively multi-agent systems" and "scalable MARL," driving the application of graph neural networks, attention mechanisms, and mean-field theory in MARL [15–17].

The integration of multi-agent deep reinforcement learning (MADRL) with water environment systems is experiencing rapid advancement, yet several persistent challenges remain. These include the complexity of cooperative game modeling, limited policy stability, and the lack of interpretability and deployability in real-world scenarios. To address these issues, this paper introduces MADRL as a pivotal research paradigm for intelligent control in water environment systems, highlighting its unique strengths in enabling collaborative mechanisms and dynamic control optimization among multiple control units.

This work presents a comprehensive overview of the theoretical foundations, learning architectures, representative applications, and critical challenges of MADRL in water-related domains. Special emphasis is placed on addressing complex conditions—such as partial observability, non-stationarity, and large-scale spatial distribution—through policy coupling, task-level coordination, and inter-agent information sharing. MADRL demonstrates strong potential to enhance system-wide coordination and efficiency, particularly in clustered control tasks such as multi-pump station scheduling, upstream–downstream regulation, and multi-point compliance of water quality metrics. By building unified frameworks for learning, MADRL opens new pathways for intelligent scheduling in complex systems. Its growing maturity not only propels innovation in multi-agent control theory and algorithms but also provides a technological foundation for more resilient, efficient, and sustainable management of water resources.

From the perspective of multi-agent coordination and reinforcement learning integration, this paper systematically explores the typical applications, core techniques, and development trends of MADRL in the context of water environment systems. The remaining sections of this article are organized as follows. In Section 2, this paper introduces the coordination mechanisms and game-theoretic foundations for multi-agent systems, providing the theoretical underpinnings for distributed decision-making; Section 3 reviews the evolution of MADRL, focusing on training paradigms and modeling frameworks in dynamic and partially observable environments; Section 4 proposes a cross-domain collaborative control architecture tailored to water systems, addressing structural heterogeneity and operational complexity; Section 5 presents MADRL-based applications in four key water system scenarios, namely, urban drainage control, distributed water resource management, hydropower regulation, and autonomous monitoring with path planning; Section 6 investigates the generalization capabilities of MADRL in complex system settings and its integration with digital twin technologies; Section 7 outlines future research directions for MADRL-driven intelligent control in water systems; and Section 8 concludes this paper by summarizing the key findings and offering critical research insights.

2. Theoretical Fundamental and Modeling Mechanisms of Multi-Agent DRL

Benefiting from the powerful nonlinear approximation and high-dimensional feature extraction capabilities of deep neural networks, DRL has emerged as a mainstream algorithm in intelligent decision-making research [18]. In DRL, neural networks are used to approximate value functions and policies, thereby enabling the handling of continuous and high-dimensional state–action spaces that are otherwise intractable. This allows for end-to-end learning in complex environments [19].

As illustrated in Figure 1, the general framework of DRL involves two tightly integrated components: feature representation and policy optimization. First, deep learning techniques are employed to extract meaningful representations from high-dimensional input data, providing an accurate depiction of the current environmental state. Then, reinforcement learning algorithms use the extracted features and cumulative rewards to optimize decision-making policies, mapping the current state to appropriate actions [20].

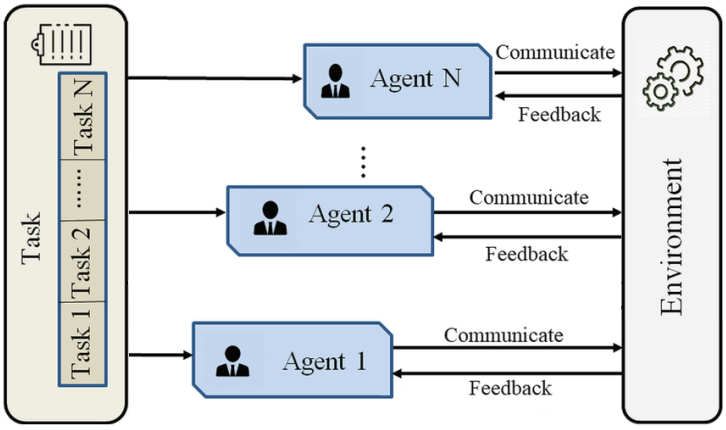


Figure 1. Structure of multi-agent reinforcement learning.

2.1. Problem Formulation and Modeling Framework

In water environment systems engineering, many control processes exhibit the Markov property, i.e., the system's state transitions depend solely on the current and historical states, independent of future anticipated states. This memoryless nature makes water environment control problems naturally amenable to modeling as Markov Decision Processes (MDPs). As the foundational framework of reinforcement learning, MDPs offer a theoretical basis for structuring the interaction between decision-making agents and complex, dynamic environments [21,22].

With the increasing deployment of sensor networks and the Internet of Things (IoT), environmental systems now generate high-dimensional, spatiotemporally heterogeneous data related to water quality, quantity, and pollutant dispersion. Traditional control approaches often face significant performance limitations when dealing with such complex and dynamic systems. In contrast, deep reinforcement learning (DRL) has demonstrated remarkable capabilities in handling high-dimensional state spaces and continuous control problems, positioning it as a promising tool for intelligent water environment management.

Under this modeling paradigm, an individual control unit in the water environment system—such as a treatment plant, pump station, sluice gate, or sub-watershed—can be abstracted as an intelligent agent. Its interaction with the surrounding environment can be formalized as a Markov Decision Process (MDP), defined by the tuple $\mathrm{MDP} = (\mathcal{S}, \mathcal{A}, \mathcal{P}_{sa}, \pi, \gamma)$, where

- $\mathcal{S}$ denotes the system state space (e.g., water quality parameters, flow conditions, meteorological inputs);
- $\mathcal{A}$ represents the set of available control actions (e.g., switching pumps on/off, adjusting chemical dosages, regulating flow rates);
- $\mathcal{P}_{sa}$ is the state transition probability function;
- π is the agent's decision-making policy;
- γ is the discount factor used to balance immediate and future rewards.

When multiple control units operate concurrently, the water environment control problem evolves into a finite-horizon Multi-Agent Markov Decision Process (MAMDP), represented as MAMDP $= (\mathcal{S}, \mathcal{A}, N, \mathcal{P}_{sa}, \pi, \gamma)$, where N denotes the number of interacting agents. In such settings, agents may engage in both information cooperation and resource competition, with the shared objective of optimizing joint control policies under system constraints to achieve goals such as pollution mitigation, water quality assurance, and coordinated resource allocation.

2.2. Identification of Key State Variables in Water Environment System Regulation

In the context of intelligent control for water environment systems, the system typically comprises multiple observation nodes or control units, each of which can access partial state information from its local environment, forming the overall state observation space $s \in \mathcal{S}$. Based on the observed state, an intelligent agent executes an optimal control action to achieve specific objectives such as water quality maintenance, hydrological balance, or emergency pollution response. This section reviews the commonly used categories of state observations in current research.

(1) Pollutant concentration.

Pollutant indicators represent the most fundamental type of state information in water environment control. Typical variables include dissolved oxygen (DO), ammonium nitrogen (NH_4^+), total phosphorus (TP), pH, and temperature. These measurements are often collected from multiple spatial nodes and organized into spatial-variable matrices or time-series vectors, which reflect either localized or system-wide water quality conditions. Such representations are critical for assessing pollution levels and determining regulatory requirements.

(2) Flow velocity, discharge, and water level.

Information regarding flow velocity, discharge, and water level—obtained via hydrodynamic simulation or real-time sensor measurements—serves as a key basis for control decisions, especially in scheduling operations such as pump activation or gate regulation. These features are typically represented using grid-based hydrodynamic models in the form of multi-dimensional tensors or numerical fields, allowing the spatiotemporal dynamics of water movement within the system to be effectively captured.

(3) Operational status of control facilities.

The real-time operational status of control facilities (e.g., pumps, sluice gates, aerators, chemical dosing systems) also constitutes an important component of the system's state space. This category of information is often encoded as vectors that represent the on/off states, operational intensity, or duration of various devices. Such representations facilitate the reduction of regulatory conflicts and enhance operational safety and efficiency.

Beyond the typical observation types listed above, the selection of state features can be tailored to specific control objectives. Depending on the task, state representations may consist of a single modality or leverage multi-modal data fusion to construct joint state

spaces, thereby enhancing the system's perception and responsiveness in complex aquatic environments [23].

2.3. Action Space Design and Classification in Intelligent Control

Agents select and execute optimal control actions A_m based on the current observed state $s \in \mathcal{S}$. These actions aim to optimize system objectives such as water quantity regulation, hydraulic stability, and pollutant control. The design of the action space and its corresponding policy structure directly affects the decision-making flexibility of the agent and the safety and stability of the control system. At present, action spaces in water environment systems are generally categorized into two types:

(1) Fixed action sequences.

In this control strategy, the action set is denoted in Equation (1):

$$A_m = \{A_{m1}, A_{m2}\} \tag{1}$$

with clearly defined switching rules and a predetermined sequence:

- When the agent executes action A_{m1}, the system maintains the current control state without adjustment (e.g., keeping the current pump speed, gate open, or aeration intensity);
- When executing action A_{m2}, the system switches to the next predefined control mode or scheduling scheme (e.g., sequentially activating different water treatment units or rotating pump operations according to a fixed schedule).

To ensure operational safety and prevent system lag or environmental load accumulation, each control mode is typically assigned a maximum duration threshold $\Delta t_{\max}$. Once the threshold is reached, the system automatically transitions to the next state. This approach is widely applied in standardized water treatment processes and automated control systems, particularly in scenarios with well-defined task flows and stable operational logic.

(2) Variable action sets.

In this control strategy, the action set is defined in Equation (2):

$$A_m = \{A_{m1}, A_{m2}, \ldots, A_{mn}\} \tag{2}$$

where the set includes all feasible combinations of control actions within the system, such as different pump speeds, gate opening levels, and on/off states of multiple aeration or chemical dosing units, as follows:

- At each decision step t, the agent selects an action $a_t \in A_m$ based on a learned policy $\pi(a_t \mid s_t)$, which defines a probability distribution over all candidate actions.
- This approach provides greater flexibility in control and is more adaptable to non-stationary water quality disturbances or multi-objective regulatory tasks.

However, due to the high dimensionality and combinatorial complexity of the action space, this method may introduce risks such as discontinuous control behavior, abrupt system load changes, water body disturbances, or resource conflicts. Therefore, in practical implementations, it is often necessary to introduce the following:

- **Action constraints** $\mathcal{C}(A_m)$: to limit the feasible action set;
- **Penalty functions** $R_{\text{penalty}}(a_t)$: to discourage unsafe or costly actions;
- **Safety filtering mechanisms**: to override or restrict high-risk action execution.

This classification of action space design serves as a theoretical and practical foundation for applying RL-based intelligent controllers in real-world water environment systems.

By carefully structuring the action space, it is possible to balance decision-making precision and system robustness, enabling safe and adaptive control in dynamic and uncertain aquatic environments [24].

2.4. Reward Mechanism Design in Intelligent Regulation of Water Environment Systems

In reinforcement learning (RL)-based intelligent control of water environment systems, the agent receives a scalar feedback signal—known as the reward—from the environment after executing a specific control action. The reward function not only quantifies the positive or negative impact of an action on the system's control objectives but also serves as the core driver for policy optimization in the learning process. A well-designed reward function can effectively guide the agent to learn control strategies that contribute to water quality improvement, energy efficiency, and operational effectiveness, as detailed in Table 1.

Table 1. Multi-agent reinforcement learning-based modeling for water environment regulation.

State Variable S	Action Variable A	Reward Function R	Control Objective
Water quality status	Chemical dosing rate	Penalty for threshold exceedance of pollutants	Control nutrient levels to prevent eutrophication
Time correlation parameters	Associated control operations	Penalty based on recovery time	Evaluate the system's recovery efficiency after pollution events
Water level/flow indicators	Gate or valve scheduling control	Penalty for deviation from target water level	Maintain water levels within the target range
Operational status of devices	Regulatory actions of the agent	Penalty for energy consumption and operational cost	Optimize resource utilization to enhance economic efficiency

To accommodate the multi-objective nature of water environment regulation, reward function design typically incorporates the following categories of indicators:

(1) Duration or cumulative degree of pollutant concentration violations.

This indicator measures the time duration or cumulative extent to which key water quality parameters (e.g., dissolved oxygen [DO], ammonium nitrogen [NH_4^+], total phosphorus [TP]) exceed predefined safety thresholds. Persistent exceedance indicates serious pollution and ecological risk and should be penalized via a negative reward. A typical formulation is denoted in Equation (3):

$$R_1 = -\sum_{i=1}^{n} \max(C_i - C_{\text{th}}, 0) \cdot \Delta t \tag{3}$$

where C_i is the pollutant concentration at observation point i, C_{th} is the specified threshold value, and Δt is the time step duration.

(2) Energy consumption or operational cost of control actions.

Operational facilities (e.g., pump stations, aerators, dosing systems) in water environments often incur significant energy consumption and maintenance costs. To enhance economic efficiency, control cost terms are frequently introduced as penalties in the reward function denoted in Equation (4):

$$R_2 = -\sum_{j=1}^{m} P_j \cdot t_j \tag{4}$$

where P_j is the power consumption of device j and t_j is the operational duration of device j.

(3) Water quality recovery time or target achievement time.

In pollution events or under fluctuating water quality, the speed of recovery to a safe state is a key indicator of control performance. A time-based reward function can be formulated as in Equation (5):

$$R_3 = -T_{\text{restore}} \tag{5}$$

where T_{restore} is the time required for the system to return from a threshold-violating to a compliant state; a shorter T_{restore} yields a higher reward.

(4) Water level and volume balance indicators.

For systems involving water resource scheduling or flood control, hydraulic variables such as water level and flow must also be considered. One approach is to penalize deviations from a target water level range, denoted in Equation (6):

$$R_4 = -\sum_t \left| h_t - h_{\text{target}} \right| \tag{6}$$

where h_t is the observed water level at time t and h_{target} is the desired/target water level, reflecting constraints for flood mitigation or ecological flow maintenance.

(5) Multi-objective integration via weighted aggregation.

In practical applications, water environment control systems often involve multiple concurrent objectives. To balance factors such as water quality, energy consumption, and response efficiency, a composite reward function is often constructed as a weighted sum denoted in Equation (7):

$$R = \lambda_1 R_1 + \lambda_2 R_2 + \lambda_3 R_3 + \lambda_4 R_4 \tag{7}$$

where λ_i denotes the weighting coefficients reflecting the relative importance of each control objective.

The choice of weights directly influences the agent's behavioral strategy. These weights can be manually tuned based on domain knowledge or adaptively learned through meta-optimization techniques.

3. Training Paradigms in Multi-Agent DRL

Multi-agent systems (MASs) consist of multiple autonomous agents and are well-suited for controlling complex, distributed, and dynamically coupled environments. In the context of water environment systems, functional components such as pump stations, sluice gates, aeration devices, and sensor nodes can be abstracted as agents. These agents perceive local environmental states, perform control actions, and—when necessary—interact or coordinate with other agents to collectively achieve system-level objectives such as water quality stabilization, energy minimization, or pollution mitigation. The training paradigms of multi-agent reinforcement learning are summarized in Table 2.

Table 2. Comparative analysis of multi-agent learning paradigms.

Learning Paradigm	Core Concept	Evaluation	Representative Algorithms
Independent learning	This approach ignores interactions among agents and learns individual policies independently.	Offers high scalability but is prone to non-stationarity due to the lack of inter-agent coordination.	DRUQN, DLCQN, DDRQN WDDQN
Centralized learning	Treats all agents as a unified entity, enabling joint policy optimization.	Immune to environmental non-stationarity but prone to curse of dimensionality.	—
Value function decomposition	Decomposes global value function to optimize the the overall objective.	Solves credit assignment, struggles in non-stationary or complex environments.	QMIX, WQMIX, QPLEX, Qatten
Centralized value function	Trains a value network with global information to guide independent policy learning.	Reduces problems of high dimensionality and non-stationarity problem.	MADDPG, MATRPO, MAPPO, MATD3

3.1. Independent Learning Paradigm

The independent learning paradigm (ILP) is one of the most fundamental strategies in MADRL. Under this paradigm, each agent independently optimizes its own policy, treating the behaviors of other agents as part of the environment dynamics—without explicitly modeling their strategies or states. This approach is particularly suitable for scenarios where inter-agent communication is limited or costly. In water environment systems, geographically distributed and structurally independent units such as pump stations, gates, and aerators naturally align with the assumptions of this paradigm.

As shown in Figure 2a, ILP allows each agent to update its policy independently, without relying on coordination or information sharing. This design offers good parallelism and scalability, making it especially applicable to small-scale, discrete state–action space scenarios such as sectional chemical dosing or localized water quality compliance strategies. However, due to the non-stationarity introduced by independently evolving policies, ILP may suffer from convergence instability and degraded learning performance.

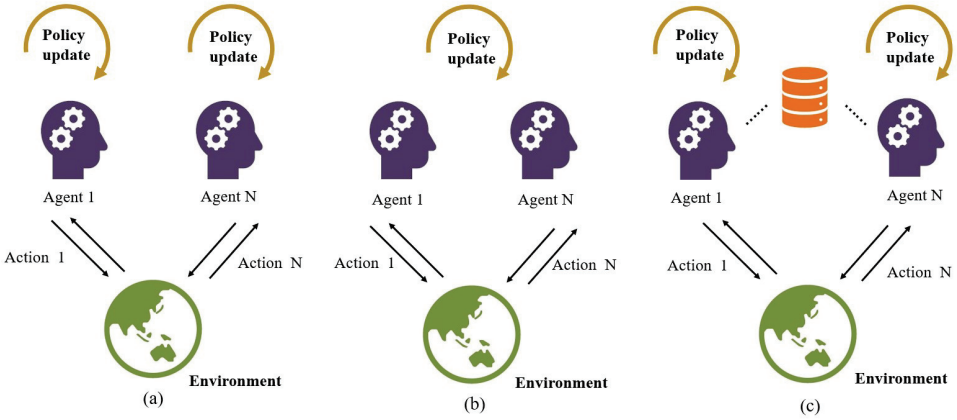


Figure 2. Schematic illustration of three representative learning paradigms in multi-agent deep reinforcement learning. (**a**) distributed training and distributed execution; (**b**) centralized training and centralized execution; (**c**) centralized training and distributed execution.

DLCQN (Decentralized Learning with Coupled Q-Networks) introduces a policy adaptivity mechanism that allows agents to autonomously switch between independent and coordinated behaviors. This switching is guided by negative reward signals and observation dependencies. In river network systems with upstream–downstream hydrody-

namic coupling, DLCQN enables local agents to dynamically adjust their strategy coupling strength based on hydraulic linkages, improving system adaptability. Palmer et al. proposed the Lenient Deep Q-Network (LDQN) to address early-stage policy oscillations that occur in synchronous multi-agent learning. By assigning tolerance to early negative feedback, this method avoids premature penalization and helps mitigate early policy collapse [25]. It is well-suited for uncertain scenarios such as pollution source tracking and emergency response dispatching. Building on this, Zheng et al. proposed the Weighted Double Q-Network (WDQN), which combines lenient reward strategies with double Q-learning updates. WDQN significantly alleviates Q-value overestimation and improves the robustness of policy updates under stochastic disturbances, making it valuable in high-risk and dynamic water regulation tasks [26].

In addition to DQN-based methods, some researchers have extended the independent learning paradigm to partially observable environments using recurrent neural network (RNN) architectures. This is particularly relevant in water environment systems where sensors may fail intermittently, monitoring stations may be sparsely distributed, or measurements are collected with low temporal resolution. In the context of water environment systems, such approaches can be extended to complex tasks such as multi-source pollution tracing and multi-section collaborative regulation, where observations may be incomplete or delayed. These RNN-based independent learners improve the robustness and adaptability of control strategies under conditions of limited or unreliable environmental information.

3.2. Centralized Learning Paradigm

The centralized learning paradigm is a fundamental architecture in multi-agent reinforcement learning (MARL), where the policy optimization processes of all agents are managed and trained by a central learner. As shown in Figure 2b, the central controller has access to the local states, actions, and rewards of all agents, as well as the global state of the system, thereby enabling comprehensive environmental awareness and a system-level perspective. In complex water environment systems—such as cross-regional water quality regulation, upstream–downstream coordination in river networks, and joint operation of reservoir clusters—control units are highly coupled and interdependent. These tasks require precise coordination and joint decision-making among agents. The centralized learning paradigm provides theoretical support for learning such collaborative strategies by aggregating global information, making it especially suitable for multi-objective optimization scenarios requiring a high degree of cooperation. However, as the number of agents increases, the joint state–action space grows exponentially, leading to the well-known curse of dimensionality. This significantly degrades learning efficiency, hinders convergence to global optima, and limits the paradigm's scalability and applicability to large-scale intelligent water control systems.

3.3. Centralized Training with Decentralized Execution in Water Environment Systems

The centralized training and decentralized execution (CTDE) paradigm effectively combines the strengths of global coordination and local autonomy. As shown in Figure 2c, this approach allows agents to exchange state, action, and reward information during training, facilitating the centralized optimization of system-wide objectives. Once deployed, however, each agent operates independently using only its own local observations and experience, without relying on others' internal states or policies.

This learning paradigm is particularly well-suited to intelligent water environment systems, which are characterized by delayed information flow, limited observability, and spatially distributed control nodes. In such contexts, fully centralized control is often

impractical due to communication constraints, while isolated learning may fail to achieve coordinated and globally optimal behaviors.

By leveraging global insights during training and enabling autonomous action at runtime, this hybrid framework offers a scalable, practical solution for real-world deployment. It has thus emerged as a key enabler for developing intelligent, adaptive, and resilient environmental control systems.

3.3.1. Value Decomposition Approaches for Distributed Control

Methods based on value function decomposition seek to reduce the complexity of multi-agent learning by factorizing the global value function into a set of local components. Within this framework, each agent updates its policy independently, relying solely on local observations and rewards without requiring full awareness of other agents' strategies. This design alleviates the training instability associated with non-stationary environments typical in independent learning settings while enabling effective credit assignment for each agent's contribution to overall system performance—an essential capability in collaborative water resource management. Representative algorithms and their applicability to water environment systems are summarized below.

Value decomposition networks (VDNs) decompose the global Q-value linearly into the sum of individual agents' local Q-values. The contribution of each agent to the global return is quantified via gradient backpropagation, as shown in Figure 3. Due to its simplicity and training efficiency, the VDN is particularly applicable to water systems in which each control unit has clearly defined local objectives and weak coupling with others.

1. **Multi-point water quality control** involves maintaining indicators such as dissolved oxygen (DO), ammonium nitrogen (NH_4^+), or total phosphorus (TP) within target thresholds across multiple monitoring sections;
2. **Zonal aeration strategy optimization** uses independently operating aeration units to locally regulate oxygenation while contributing to system-wide water quality balance [27].

QMIX introduces a nonlinear mixing network that improves the model's ability to approximate complex interactions while maintaining a monotonicity constraint between the global Q-value and individual agents' local Q-values [28]. This structure is well-suited for coordinated control tasks in complex water systems, such as total pollutant load regulation or basin-wide discharge compliance. However, the performance of QMIX may degrade in tasks that involve non-monotonic coordination dynamics, such as pollutant dispersion control coupled with water level regulation.

QPLEX introduces a dueling network structure for modeling agent-specific advantage functions, which constrains the mixing network to be a monotonic function of individual agent Q-values to guarantee consistency with decentralized execution. QPLEX relaxes this constraint by introducing an advantage-based factorization framework. Specifically, it decomposes the joint action-value function as

$$Q_{\text{tot}}(\mathbf{s}, \mathbf{a}) = V(\mathbf{s}) + \sum_{i=1}^{n} A^i(\mathbf{s}, a^i) \tag{8}$$

where $V(\mathbf{s})$ is a shared, state-dependent baseline and $A^i(\mathbf{s}, a^i)$ is the advantage function for agent i. The mixing network then aggregates these agent-specific advantages using a monotonic function, allowing the model to capture more expressive joint action dependencies while still preserving the guarantees of decentralized execution.

This formulation effectively addresses the representational limitations of QMIX in non-monotonic environments, such as competitive or partially cooperative tasks.

Qatten integrates a multi-head attention mechanism into the value decomposition framework to more effectively capture the varying significance of individual agents under different global state conditions. In contrast to conventional additive or monotonic mixing architectures, Qatten employs an attention-based mixing network, where the attention weights are computed as

$$\alpha_i = \text{softmax}(f(\mathbf{h}_i, \mathbf{g})) \tag{9}$$

Here, $\mathbf{h}_i$ represents the latent feature of agent i, typically derived from its local observation, and $\mathbf{g}$ denotes a global context vector extracted from the shared environment state. These attention scores α_i are then used to dynamically modulate each agent's contribution to the joint action-value function.

By leveraging this adaptive weighting mechanism, Qatten enables the model to focus more effectively on contextually relevant agents, thereby enhancing the expressiveness of the value function and improving credit assignment in cooperative tasks, especially in scenarios involving role specialization or situational importance.

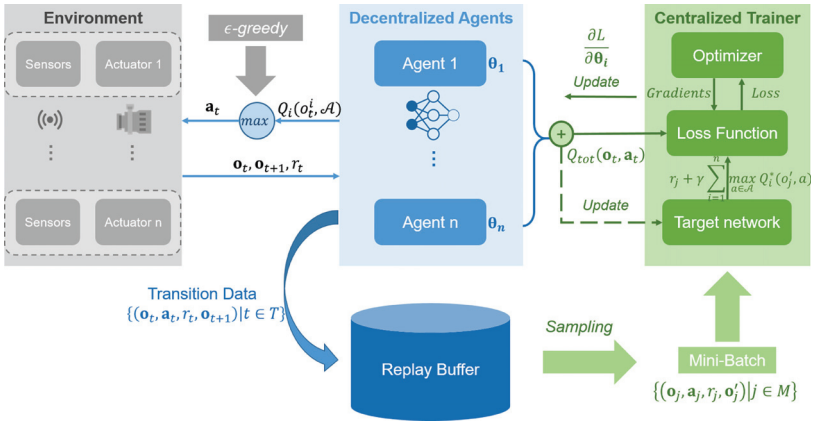


Figure 3. Schematic illustration of the training process in the value decomposition network (VDN) algorithm.

Value decomposition-based MARL methods, due to their localized information dependency, scalable architecture, and training stability, provide a solid theoretical foundation and technical pathway for distributed intelligent control in complex water environments. These methods are particularly effective under challenging conditions such as high-dimensional state spaces, non-stationary dynamics, and weakly coupled or delayed inter-agent communication. However, most existing value decomposition algorithms are designed for discrete action spaces, which limits their applicability in continuous control tasks—such as pump speed modulation, water level adjustment, or continuous chemical dosing.

3.3.2. Applications of Centralized Value Function Methods in Multi-Agent Regulation

Compared with value decomposition-based methods, centralized value function (CVF) methods leverage global information during training to construct a unified critic network, which guides the independent optimization of each agent's actor policy network. These methods typically adopt the actor–critic architecture, enabling decoupled learning between policy and value functions. This paradigm offers a balance between policy flexibility and training stability and is particularly suitable for water environment systems where control units exhibit strong state coupling and complex inter-agent interactions [29,30].

Representative algorithms and their application scenarios in water environment regulation are as follows:

MADDPG (Multi-Agent Deep Deterministic Policy Gradient) was the first to extend deterministic policy gradient methods to multi-agent settings by combining a centralized critic with decentralized actors [31]. It enables agents to learn deterministic policies while utilizing shared global information. MADDPG is applicable to scenarios such as multi-pump flow distribution, reservoir group water level coordination, and upstream–downstream joint discharge optimization. However, its scalability is limited in large-scale systems due to the curse of dimensionality and difficulties in credit assignment.

MATD3 (Multi-Agent Twin Delayed DDPG) improves upon MADDPG by introducing twin centralized critics and delayed policy updates, effectively addressing the overestimation bias in value functions. It is particularly suited for fine-grained control tasks involving continuous variables, such as pump frequency modulation, water level gradients, or chemical dosing rates, and offers strong performance in terms of both control precision and training stability.

HATRPO and HAPPO were proposed to overcome the monotonic policy improvement limitations of traditional TRPO/PPO in cooperative multi-agent settings. By decomposing the advantage function and employing a sequential update mechanism, these algorithms enhance policy expressiveness and coordination capabilities without requiring parameter sharing or joint value function assumptions. They are particularly effective for heterogeneous control scenarios, such as functionally differentiated pump stations or asynchronous operation of distributed treatment plants [32].

Centralized value function-based methods, by leveraging global information integration, demonstrate strong adaptability in controlling water environment systems characterized by strongly coupled regulation nodes, complex collaborative tasks, and widespread continuous action spaces. In particular, algorithms based on the actor–critic architecture exhibit distinct advantages in terms of policy stability, sample efficiency, and multi-objective optimization.

4. Cross-Domain Collaborative Control and Decision Architectures in Water Environment Systems

With the rapid advancement of sensor networks, intelligent control, and environmental modeling technologies, intelligent control of water environment systems is evolving from traditional point-based regulation toward large-scale, multi-agent, and cross-regional collaborative management paradigms. By integrating the global scheduling capability of shore-based control centers, the environmental sensing capability of in situ monitoring nodes, and the local execution capability of distributed agents (e.g., autonomous monitoring buoys, intelligent pumping stations), intelligent water systems can effectively support critical tasks such as pollution source tracking, dissolved oxygen regulation, water quality early warning, and coordinated emergency response. These capabilities position such systems as a core technological framework for achieving precision water quality management, basin-wide coordinated scheduling, and ecological restoration. Current research on intelligent water environment systems primarily focuses on the following key technical domains:

1. Task deployment and optimization;
2. Information perception and transmission mechanisms;
3. Path planning and agent scheduling;
4. Multi-agent cooperative control strategies.

Meanwhile, ongoing breakthroughs in reinforcement learning, graph neural networks (GNNs), adaptive control, and multi-agent systems provide viable pathways for advancing water environment systems toward large-scale, multi-source collaboration, heterogeneous device integration, and adaptation to complex physical environments. This transformation is gradually pushing water environment systems from automation in monitoring toward

ultimately self-organizing ecosystems, establishing itself as a prominent research frontier in smart water management and watershed governance.

In this context, this section presents a comprehensive review of the current state of the art and emerging trends of these key technologies in water environment systems. In addition, the interconnections among various technical modules are mapped and analyzed. The technical architecture is illustrated in Figure 4, and the representative algorithms are described in Table 3.

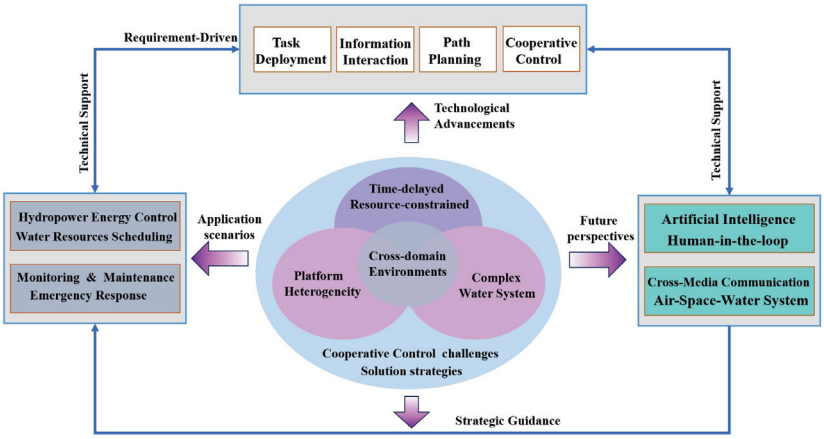


Figure 4. Logical structure and technical pathways of cross-domain collaborative control architecture.

Table 3. Representative algorithms for cross-domain intelligent cooperative control in water environment systems.

Technical Category	Algorithm Type	Technical Characteristics	Model Limitations
Task deployment	Heuristic algorithm and game-theoretic strategy Self-organizing mapping algorithm	Construct goal allocation Adaptive capability of edge intelligence	Ignore intrinsic correlation between complex tasks Lack of coordination in task execution
Information exchange	Local topology control by autonomous learning Graph-theoretic or Q-learning-based	Flexible topology configuration Capability-aware routing	Difficult to ensure network connectivity Difficult to balance transmission rate
Path planning	Robust optimization or state estimation Federated learning or centralized learning	Multi-sensor fusion Edge computing algorithms	Data heterogeneity across cross-domain platforms High computational cost and parameter space
Cooperative control	Centralized or distributed architecture Federated learning algorithm	Architecture-based coordination Centralized/unified cross-domain training	Communication overhead and latency Data heterogeneity and high control cost

4.1. Strategies of Task Deployment in Collaborative Control

Building on the preceding discussion of training paradigms and policy optimization mechanisms in multi-agent systems, this section focuses on the practical layer of task deployment in intelligent water environment control. It provides a systematic overview of centralized, distributed, and hybrid task scheduling strategies, as well as their architectural designs and representative applications.

In intelligent water systems, task deployment plays a critical role in the efficient organization and dynamic coordination of heterogeneous control resources—such as pumping stations, sluice gates, unmanned monitoring vessels, and sensor nodes—with respect to diverse operational objectives including emergency pollution response, real-time water

quality regulation, and multi-source supply coordination. By constructing rational task allocation schemes and control frameworks, the overall system performance in terms of responsiveness and resource efficiency can be significantly improved. This is particularly crucial under complex scenarios involving sudden pollution events, water quality fluctuations, or basin-scale coordinated regulation.

According to the underlying collaboration architecture, current task deployment strategies in intelligent water systems can be broadly categorized into centralized and distributed approaches.

(1) Centralized task deployment strategies.

In centralized deployment architectures, a shore-based control center assumes responsibility for generating global task plans, allocating system resources, and issuing execution commands to individual control nodes. This paradigm is suitable for scenarios with tight time constraints, high task complexity, and a need for high-precision coordination, such as

- Multi-section pollutant source tracking;
- Eutrophication emergency remediation;
- Unified water quality standard enforcement across river basins.

For instance, Liu et al. proposed a task-oriented intelligent control framework for centralized coordination across multiple water resource nodes [33]. Chen et al. developed a multi-source cooperative model incorporating heuristic optimization and coalition game theory to optimize joint monitoring task assignment between surface buoys and unmanned vessels [34]. Zhang et al. introduced a variable task value model that considers both static goals and dynamic feedback to achieve centralized allocation for maximum pollution reduction benefit [35].

While centralized strategies offer strong consistency and command controllability, they often suffer from limitations in adaptability and scalability, particularly under conditions of high task heterogeneity, environmental uncertainty, incomplete information, or unstable communication. These constraints can result in inflexible resource allocation, delayed node responses, and low execution robustness.

(2) Distributed task deployment strategies.

To address the limitations of centralized methods, distributed task deployment has garnered increasing attention. In this paradigm, individual control units or edge agents autonomously perceive, decide, and collaborate, executing tasks and managing resources at the regional level without relying on global communication. This makes it particularly suitable for large-scale, dynamically changing, and weakly coupled water environments with limited data transmission capability [36,37].

Han et al. proposed a self-organizing task mapping algorithm based on belief functions, enabling dynamic task chain configuration from sensing nodes to target control zones [38]. Davis et al. used a marginal cost-based method to optimize conflict-free routing of mobile monitoring robots in water quality surveillance [39].

Although distributed strategies enhance system resilience, local responsiveness, and edge intelligence utilization, they may lead to task conflicts, resource fragmentation, and insufficient coordination if global planning mechanisms are not properly integrated.

4.2. Communication Mechanisms for Multi-Agent Collaboration in Intelligent Water Environment Systems

In intelligent water environment systems, the collaborative operation and optimization of control strategies across multi-source heterogeneous nodes fundamentally rely on efficient and reliable communication mechanisms. This is particularly critical in representa-

tive applications such as watershed scheduling, pollution tracking, lake management, and smart drainage systems, where multi-level coordination among underwater sensors, surface unmanned vehicles, and land-based control platforms is required. These components collectively constitute a complex air–land–water integrated communication network operating across diverse physical media and dynamic environments. Consequently, designing communication architectures and cooperative mechanisms adapted to the hydrophysical characteristics of water environments has become a foundational task for enabling system-wide intelligent operation [40].

Currently, the dominant communication modalities in such systems include radio frequency (RF) communication, optical wireless communication, and underwater acoustic (UWA) communication. Each medium exhibits distinct trade-offs in terms of data rate, coverage, energy consumption, and environmental robustness and must be tailored to specific spatial and operational requirements.

4.2.1. Intelligent Communication Mechanisms for Collaborative Regulation in Water Environment Systems

In complex environments, communication mechanisms are regarded as a key enabler for achieving collaboration and efficient decision-making in multi-agent systems. In the context of intelligent regulation of water environment systems, agents—such as pumps, valves, sensor nodes, and monitoring drones—must share experiences, intentions, and local observations through communication technologies. This information exchange allows each agent to develop a more comprehensive understanding of both the system state and the behavior of other agents, ultimately facilitating the coordinated completion of complex tasks such as pollution control, water resource allocation, and flood warning, as illustrated in Figure 5.

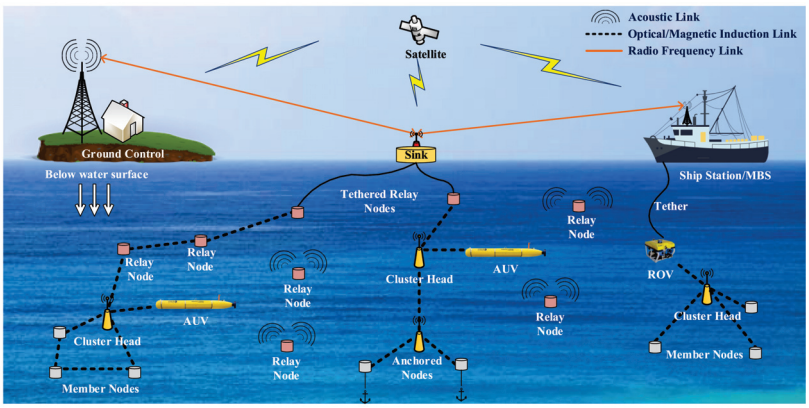


Figure 5. Schematic diagram of a multi-agent communication system.

The core of multi-agent communication mechanisms involves three essential questions: when to communicate, with whom, and what information to exchange. In practical tasks within intelligent water regulation, this is reflected in situations such as follows: When should upstream sensing nodes transmit pollutant concentration data to downstream control units? Should control modules communicate based on dynamic factors in water quality models? Research on communication mechanisms has provided transferable technical frameworks and insights for water environment systems.

DIAL (Differentiable Inter-Agent Learning) offers an end-to-end communication mechanism that, while simple, is limited to discrete communication settings—thus restricting its direct applicability in complex water resource management scenarios [41]. Abdallah proposed a centralized communication architecture in which the hidden states of all agents are averaged to enhance local observation capabilities [42]. In the context of water systems,

this can be analogized to aggregating data from multiple sensor nodes to assist decision-making. However, the method suffers from semantic information loss, which limits its performance when dealing with multi-source heterogeneous environmental data.

In highly complex water system coordination tasks, communication optimization has also been modeled as a game-theoretic problem. For instance, Xue evaluated the marginal contribution of communication content to effectively prune irrelevant information [43]. This idea can be applied to assess the value of communication among multiple nodes in river network scheduling, providing more valuable transmission paths under resource-constrained conditions.

4.2.2. Graph Topologies for Air–Water–Land Multi-Agent Collaboration in Water Environment Systems

In the study of intelligent control for complex water environment systems, the problem of coordination and interaction among multiple control units has emerged as one of the core challenges in multi-agent modeling and optimal control. Coordination graphs, which explicitly represent inter-agent relationships, have gained significant attention in distributed control and intelligent optimization in recent years. By decomposing the global value function over a graph structure, this framework models system-wide utility through a higher-order abstraction.

In this setting, nodes correspond to control units, while edges or hyperedges represent local reward functions defined over the joint action–observation spaces of interacting agents. Once the coordination graph is constructed, distributed constrained optimization algorithms are typically employed to solve for optimal control actions. Control units perform multiple rounds of local information exchange and policy updates along graph edges [44].

A critical challenge in water environment systems lies in learning dynamic, sparse, and task-dependent coordination graphs that can efficiently support distributed decision-making among control agents. Existing research still faces several limitations. For instance, sparse coordination Q-function methods attempt to automatically learn sparse graph structures for value function factorization, but the learned structures are often static and require substantial prior knowledge. The MACO environment and the CASEC algorithm provide testing platforms and context-aware coordination graph modeling methods, aiming to reduce value estimation error through graph representation learning, and have demonstrated promising results across multiple tasks [45]. Recent studies have also focused on structural advances in coordination graphs, such as nonlinear topology modeling and self-organizing polynomial structures, to further enhance adaptability in complex water control scenarios.

On the other hand, several approaches adopt implicit graph construction strategies by leveraging attention mechanisms to prune redundant connections and improve both the efficiency and interpretability of the learned graph structures. For example, DICG constructs an implicit coordination graph via a graph neural network that dynamically infers inter-agent coordination relationships, thereby improving cooperative performance in multi-agent systems [46]. Although such methods increase the flexibility of graph modeling, they still encounter scalability and stability issues when applied to large-scale, multi-scale, and highly coupled control scenarios in water environment systems.

Therefore, developing scalable, robust, and task-adaptive coordination graph learning and optimization mechanisms in high-dimensional and dynamic water environments remains an open and critical research direction [47].

4.2.3. Communication-Aware Collaboration and Adaptive Swarm Control Strategies

Most existing task scheduling designs for perception platforms remain centered on single-agent control, often neglecting the impact of dynamic communication channel

conditions on overall system performance in water environments. To address the control challenges of multi-node systems under communication constraints, researchers have gradually begun to integrate channel state modeling with task coordination mechanisms, aiming to explore stable operational strategies for distributed intelligent systems operating in communication-limited environments.

Cao et al. proposed a communication quality estimator based on integral reinforcement learning (IRL). By considering both channel fading and uncertainties in node dynamics, a communication-aware cooperative control framework was constructed, which significantly improves the convergence efficiency of monitoring platform formation tasks, illustrated in Figure 6 [48]. The state update model is formulated in Equation (10):

$$X_i = AX_i + B\pi_i + B(\hat{\pi}_i - \pi_i) + G_i + N_i \tag{10}$$

where π_i denotes the current learned policy, $\hat{\pi}_i$ is a reference policy, and A, B, G_i, and N_i represent system transition dynamics and disturbance terms. This model enables the neural network to estimate system evolution under conditions of node mobility and channel fluctuations.

Yan et al. developed a terminal sliding mode controller with a dual-power reaching law, effectively reducing control chattering and enhancing system stability under wind-induced disturbances. Furthermore, considering channel constraints such as path loss and multipath fading, a value iteration-based RL controller was proposed to predict the communication quality of the target node. This controller achieves an integrated design that combines communication estimation with task coordination control, offering a more robust solution for multi-platform water environment monitoring tasks [49,50].

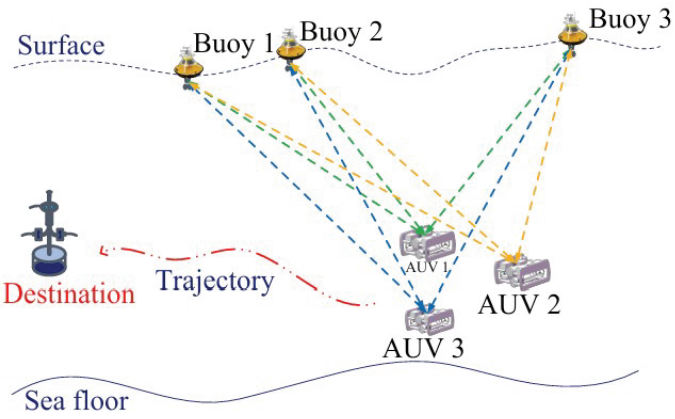


Figure 6. Formation control under joint communication–perception feedback in water environment systems.

4.2.4. Cross-Regional Integrated Sensing and Control System Architectures

In practical water environment governance, challenges such as large spatial coverage of water bodies, strong task heterogeneity, and diverse information sources are frequently encountered. To address these issues, recent studies have proposed the construction of cross-regional, multi-platform unified perception and regulation frameworks. By integrating ground stations, unmanned aerial vehicles (UAVs), autonomous monitoring buoys, and underwater sensor nodes, a multi-layer heterogeneous sensing network can be established to enable task data sharing and distributed control execution.

Zhu et al. proposed a cooperative protocol design for cross-regional unmanned platforms, aiming to address the communication heterogeneity between aerial and surface platforms and thereby improve overall task scheduling efficiency [51]. Chen et al. introduced blockchain and federated learning mechanisms to ensure data synchronization and

security within the system [52]. In addition, a threshold signature scheme was employed to prevent tampering of monitoring data, thereby enhancing the verifiability and resilience of the system against adversarial attacks.

Meanwhile, related studies utilized network simulation platforms such as NS-3 to evaluate the performance of cross-platform collaborative systems, assessing key metrics including data throughput, command transmission delay, and system redundancy tolerance. These simulations provide theoretical support for real-world deployment. Through strategies such as proactive deployment, dynamic aquatic area partitioning, and task buffering at edge nodes, the system can adapt to diverse water environment task scenarios and has demonstrated strong collaborative performance in applications such as pollution response, emergency dispatching, and water–ecological joint regulation.

4.3. Path Planning and Adaptive Control for Water Infrastructure Systems in Complex Environments

In the simulation and intelligent control of water infrastructure systems, improving task reliability and operational safety requires integrated consideration of target locations, hydro-geographic terrain features, and obstacle distributions. Based on the sensed environmental state, cooperative scheduling between path planning and dynamic obstacle avoidance enables control agents to adapt to environmental uncertainties—such as sudden hydrological events, equipment malfunctions, or structural changes—by dynamically adjusting control strategies, thereby enhancing system robustness and response efficiency.

4.3.1. Multi-Platform Collaborative Mapping for Water Environment Infrastructure Systems

To ensure accurate navigation and stable task scheduling in large-scale water environment systems, robust and high-resolution map construction has become a critical prerequisite, especially in integrated control frameworks spanning aerial, surface, and underwater domains. Recent advances have explored diverse perception strategies incorporating visual, LiDAR, acoustic, and multi-sensor fusion techniques to address this fundamental challenge.

Du et al. developed a sparse, trackable mapping method based on environmental feature matching and key-frame extraction, which has been widely adopted for localization and navigation in autonomous platforms [53,54].

However, such vision-based methods often encounter limitations in complex 3D environments—such as fluctuating water surfaces, low-light scenes, and turbid underwater conditions—where sparse features and unstable registration severely degrade mapping quality. To enhance robustness under these conditions, Matsuki et al. constructed a dense mapping framework based on camera poses and selected key-frames [55], while Alireza et al. applied full-pixel sampling of image intensity gradients by minimizing photometric error, removing the reliance on conventional feature detectors [56].

In an effort to incorporate semantic understanding into the mapping process, Zhong proposed a sparse semantic mapping method that integrates semantic priors into deep odometry networks. While effective in structured environments, most of these models are still designed under static or ideal conditions and fail to address the challenges posed by dynamic natural disturbances—such as wind fields, water flows, and tidal shifts—which lead to image mismatches and position drift in multi-agent collaborative mapping [57].

To address these issues, recent work has turned attention toward robust mapping under environmental perturbations. For instance, Billings et al. incorporated visual feature extraction with robust optimization techniques to mitigate the impact of external noise and enhance the reliability of visual map matching [58].

4.3.2. Target Situation Awareness in Intelligent Water Environment Systems

Target Situation Awareness (TSA) technologies aim to support decision-making units in dynamic air–space–surface–underwater environments by fusing, analyzing, and modeling multi-source information to enable accurate judgment and adaptive strategy formulation. This technology is particularly well-suited for mission execution in unstructured water environments—such as oceans, reservoirs, and waterways—including applications such as pollutant tracking, illegal vessel detection, and emergency response. Existing research has proposed a variety of TSA approaches, aiming to construct an interconnected and intelligent perception framework to overcome challenges such as underwater noise, weak communication, and discontinuous sensing.

(1) Multi-platform collaborative perception based on edge computing.

To address issues such as limited underwater bandwidth, low acoustic communication rates, and frequent data loss, researchers have proposed heterogeneous edge computing architectures for distributed data processing. Brecko et al. introduced a federated learning framework on edge platforms that enhances convergence speed and perception accuracy while preserving data locality [59]. However, this approach suffers from high local training costs relative to aggregation, leading to suboptimal resource utilization.

To improve system-level resource scheduling, Chen et al. proposed a hybrid learning scheme that integrates centralized and federated paradigms and solves the joint optimization iteratively [60]. Lindsay et al. developed a cooperative system combining drones, surface vehicles, and underwater agents [61]. Through multi-domain sensing fusion, this system enhances real-time situational awareness and performs well in applications such as reservoir gate control, water quality monitoring, and target tracking.

(2) Multi-source perception enhanced by intelligent learning.

Traditional sensor-reliant systems face limitations in detecting non-cooperative targets and in image degradation scenarios. Braca et al. fused active and passive observations with Bayesian inference for underwater target recognition [62]. Chen et al. addressed underwater image degradation by combining blur estimation with physical image formation models to restore clarity [63].

To address the computational demands of deep models, Wang et al. proposed a frequency-interleaved progressive network to learn frequency-domain features for super-resolution enhancement [64]. Guo et al. built a 3D multi-static sonar network to reduce localization error [65]. For system-level validation, Wang et al. introduced a lightweight YOLOX-s model with transfer learning to detect vessels at multiple scales, achieving a balance between detection accuracy and runtime performance [66].

Research on Temporal–Spatial Awareness (TSA) in water environment systems is evolving from traditional physical sensing toward intelligent and integrated perception. Recent advancements include cross-platform information fusion, edge inference under weak communication conditions, semantic-aware sensing, and model compression. Future efforts should prioritize the system-level integration and validation of TSA methods, as well as their tight coupling with control strategies and mission planning, in order to develop closed-loop intelligent water environment systems.

4.3.3. Dynamic Monitoring and Target Tracking Mechanisms

In intelligent regulation tasks of water environment systems, complex spatial structures, hydrodynamic processes, and pollutant transport pathways impose increasing demands on monitoring systems. Achieving efficient and dynamic perception and tracking of target regions—such as pollution sources, water quality anomaly zones, and ecologically sensitive areas—has become a critical challenge in water environment management. In recent years, researchers have introduced multi-source sensing, machine learning, and intel-

ligent control into aquatic monitoring networks, driving advancements in target situation recognition, path optimization, and task coordination.

(1) Dynamic perception path planning and local tracking control.

During practical monitoring operations, limitations such as communication range, water disturbances, and system latency make traditional fixed-node deployment insufficient for fine-grained sensing in spatially heterogeneous areas. In response, dynamic and adaptive path planning frameworks have been proposed to enhance the match between mobile sensing units and target regions, enabling flexible and responsive coverage. Typically, the spatiotemporal delay and localization error between sensor nodes and control points are formulated as an optimization problem, which can be solved using supervised, unsupervised, or semi-supervised learning frameworks, illustrated in Figure 7.

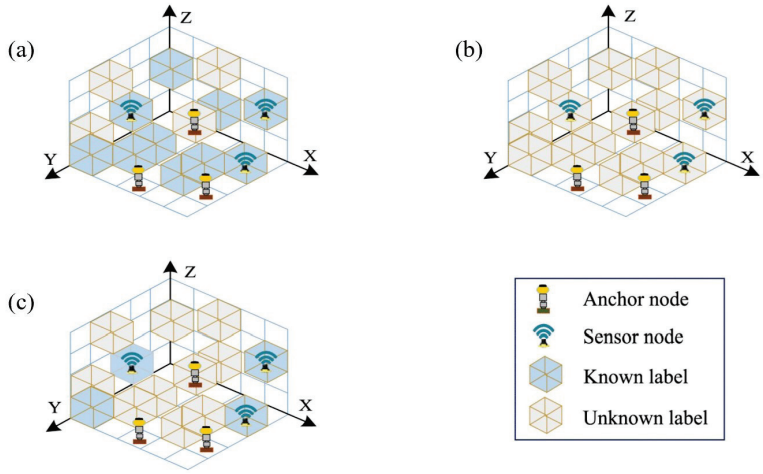


Figure 7. Scenario illustration of target localization in complex water systems. (**a**) supervised learning; (**b**) unsupervised learning; (**c**) semi-supervised learning.

For example, Yan et al. proposed a target localization method based on broad learning, integrating incremental learning for rapid model parameter adjustment [67]. While this improved computational efficiency, it did not fully compensate for errors introduced by node mobility and clock drift. In water environments, asynchronous clocks and depth-dependent acoustic propagation are modeled in Equation (11):

$$T = \hat{t} + \theta, \quad C(z) = \alpha z + b \tag{11}$$

where T denotes the local node clock, $\hat{t}$ is the actual time, and θ is the clock drift. $C(z)$ represents the sound speed at depth z, with α and b denoting the vertical gradient and surface-layer sound speed, respectively. Based on this, Han et al. developed an adaptive non-singular fast terminal sliding mode controller to track pollution plumes or critical water regions within a finite time horizon [68].

Furthermore, to handle uncertainties caused by environmental factors such as water density gradients, wind fields, and temperature layers, Naoki et al. proposed a robust control framework that integrates sliding mode control with disturbance observers [69]. Zhao et al. employed minimum rigidity graph theory and velocity observers to predict the target evolution path, balancing monitoring accuracy and operational cost [70].

(2) Multi-source platform collaborative sensing and task allocation.

Although individual sensor path planning can provide localized monitoring capabilities, large-scale and multi-regional water environments demand efficient multi-platform sensing coordination. Soriano et al. introduced a cross-platform relative localization scheme

that fuses aerial visual feedback with surface/underwater sensing data. This approach maintains stability even under conditions of data loss or inconsistency [71].

To meet the needs of distributed monitoring in time-varying target regions, Zeng et al. integrated predictor structures with neural dynamic surface controllers, enabling collaborative multi-point sensing under unknown disturbances and model uncertainties. This method demonstrates strong adaptability in scenarios such as reservoir pollution tracing and algal bloom detection [72].

Despite these advances in improving system responsiveness and target recognition, real-world deployments still face challenges such as communication instability, sensor energy limitations, and weak model transferability. Future research should focus on heterogeneous data fusion mechanisms, dynamic modeling of unstructured targets, and low-power adaptive scheduling algorithms, aiming to build an intelligent water environment monitoring and management system with a closed loop of perception–decision–control.

(3) Reinforcement learning-driven intelligent regulation and policy optimization.

In complex water environment scenarios—such as reservoir scheduling, pollutant tracking, and dynamic area coverage for water quality monitoring—coordinated deployment of heterogeneous platforms (e.g., UAVs, surface patrol boats, and underwater sensor networks) is often required. To enhance system responsiveness and task execution efficiency under environmental uncertainties, researchers have increasingly introduced artificial intelligence technologies into studies of path optimization and multi-platform cooperative control.

Conventional methods typically rely on environment perception data collected via optical or LiDAR sensors, which may include visual imagery, laser scans, and infrared signals. These are commonly coupled with artificial potential fields or rule-based path planning methods for area coverage and obstacle avoidance [73,74]. However, such approaches often struggle to simultaneously meet the requirements of real-time performance, feasible path planning in dynamic environments, and constraints on energy consumption and navigation stability imposed by complex aquatic structures.

To address these challenges, deep reinforcement learning (DRL) has emerged as an adaptive strategy optimization tool, enabling systems to iteratively interact with the environment and learn optimal control policies through trial and error. In water environment tasks, DRL can be applied to solve multi-objective scheduling problems such as path generation, task area selection, and energy-aware control. When deployed on mobile sensing platforms, the learned outputs can be directly mapped to path decisions or operational commands.

Given the high cost and risk associated with real-world trial-and-error training in aquatic environments, Yan et al. proposed a DDPG (Deep Deterministic Policy Gradient)-based path optimization method. This approach utilizes disparity angles between the monitoring platform and surrounding obstacles as input and incorporates a cost function to generate energy-efficient and communication-aware obstacle avoidance paths, illustrated in Figure 8 [75].

Moreover, information coordination and task allocation among heterogeneous platforms are essential for enhancing overall system effectiveness. Hu et al. proposed a Distributed Guiding-Vector Field (DGVF) control architecture, establishing a perception network composed of aerial drones and surface vessels. The system employs a hierarchical control structure to decouple task allocation from command tracking [76].

Under more complex marine–atmospheric interference conditions, Sana et al. reviewed motion control techniques for hybrid-powered, high-mobility platforms operating in uncertain environments. Their work provides multi-dimensional dynamic modeling support for intelligent cruise monitoring in aquatic systems [77].

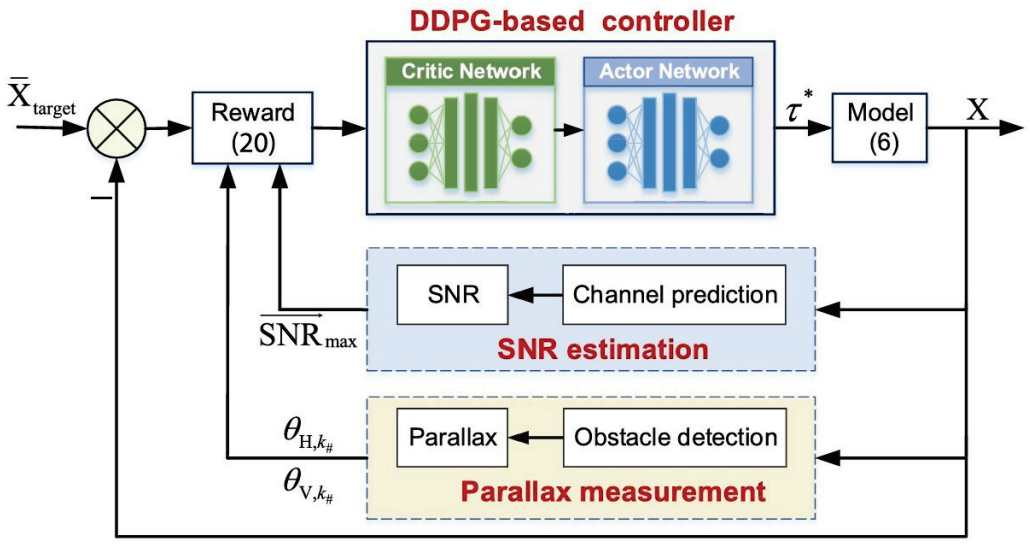


Figure 8. Illustration of a DDPG-based obstacle avoidance strategy using parallax and channel state estimation.

4.4. Towards Application-Oriented Deployment of Cross-Domain Control Architectures

Building upon the systematic review of key technologies such as task allocation mechanisms, cross-medium communication architectures, and environment-aware control strategies, it is evident that water environment systems are gradually evolving from localized autonomous control toward large-scale collaborative intelligence. Through hybrid centralized–distributed deployment, air–ground–water coordinated communication schemes, and graph-structured path planning algorithms empowered by reinforcement learning, multi-agent systems have demonstrated promising capabilities in cross-domain coordination and adaptive regulation.

However, most of these control architectures and mechanisms remain in the theoretical or experimental validation phase. Their effectiveness in complex real-world systems requires further evaluation through scenario-based systematic testing. For example, urban drainage systems demand rapid regulation in response to heavy rainfall and pollutant surges; water supply systems aim for energy efficiency and pressure stability during peak demand periods; hydropower systems must balance water availability, electricity generation, and ecological sustainability; and multi-platform water quality monitoring faces challenges such as sensor heterogeneity and unreliable communication.

Therefore, the subsequent chapters will focus on representative water system scenarios to explore the deployment pathways and application effectiveness of multi-agent reinforcement learning (MARL) in real-world water environment control. By engaging in task-specific modeling practices and comparative performance analysis, the next section aims to delineate the applicability boundaries, latent challenges, and evolutionary trends of existing methods—ultimately laying both theoretical and engineering foundations for a more robust and generalizable intelligent control framework in water environment systems.

5. Application Practices of Multi-Agent Reinforcement Learning in Water Environment Systems

5.1. Urban Drainage and Emergency Pollution Control

5.1.1. Intelligent Regulation Requirements and Reinforcement Learning

Under the dual pressures of rapid urbanization and increasing frequency of extreme rainfall events, urban drainage systems face escalating challenges, including combined sewer overflows (CSOs) and urban flooding, which pose serious threats to urban water

safety and aquatic ecosystems [78]. As a key component of smart infrastructure, real-time control (RTC) has been widely adopted to dynamically adjust hydraulic loads based on sensor and actuator feedback, enabling proactive intervention in the drainage process.

In recent years, reinforcement learning (RL) has emerged as a powerful control method in urban drainage systems (UDSs), offering the ability to learn optimal control policies through environmental interaction without requiring precise physical models. Compared to traditional model predictive control (MPC) approaches, RL demonstrates lower computational cost and greater adaptability to environmental variability [79]. Studies have shown that RL algorithms can effectively reduce CSO occurrences, mitigate flooding risks, and limit pollutant discharges. However, RTC systems are often susceptible to real-world uncertainties such as actuator failures, sensor inaccuracies, and communication breakdowns, posing challenges to the robustness and safe deployment of RL-based controllers [80].

HEO proposed an explainable MARL-driven flexible SBR system, which enables coordinated control of DO and EC regulators during the SBR process. The approach achieved an average energy consumption reduction of 4.93% while maintaining effluent quality standards under varying influent conditions [81]. Huang employed a multi-agent deep reinforcement learning (DRL) approach to evaluate the effectiveness of real-time control (RTC) strategies and developed an integrated assessment framework comprising multiple quantitative metrics, including control objectives, decision-making latency, robustness, and adaptability. A total of 31 historical rainfall events were analyzed. The results demonstrated that compared with conventional RTC methods, the DRL-based approach reduced flood and overflow risks by an average of 15.1% to 43.5% [82].

5.1.2. Robustness Assessment and Fault Tolerance Under Communication Failures

Sensors and actuators in urban drainage networks are typically spatially distributed and rely on wireless communication to connect with centralized controllers or intelligent agents. Thus, communication failures or signal delays can significantly impair control performance. To evaluate system resilience under such conditions, researchers constructed two simulated fault scenarios: (1) sensor data loss leading to observation gaps; and (2) delayed control signals preventing timely actuation.

In centralized DQN systems, control signal failures can result in total loss of functionality. However, decentralized architectures such as IQL and VDN can incorporate default fallback rules or historical behavior references to retain partial autonomy in the absence of communication. To further enhance fault tolerance, an integrated control architecture was proposed, where a central agent governs remote control under normal conditions, while local agents autonomously take over during communication interruptions, improving emergency response stability.

Simulation experiments under varying communication failure rates revealed that decentralized MARL structures demonstrate superior robustness and performance retention, particularly in CSO volume reduction and response time, as illustrated in Figure 9. These findings underscore the practical advantages of MARL-based architectures in urban water management, supporting the development of resilient, communication-tolerant, and intelligent drainage systems.

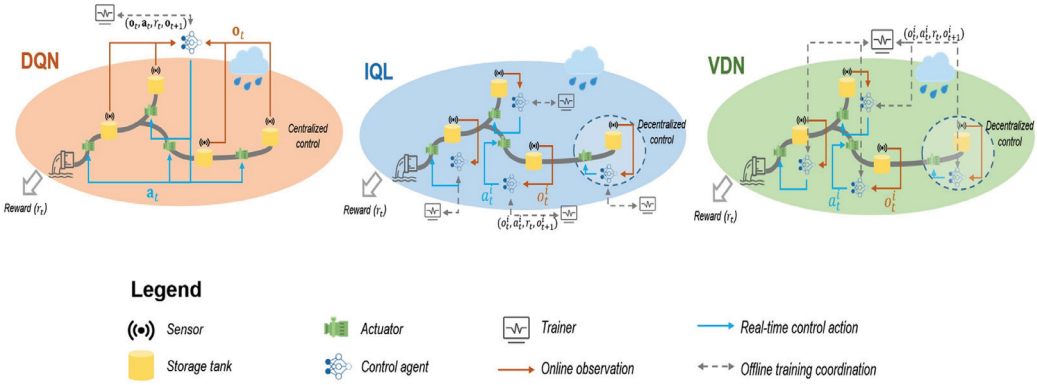


Figure 9. Integrated control architecture combining centralized, decentralized, and default fallback strategies.

5.2. Multi-Agent Intelligent Control for Urban Water Distribution Systems

In the context of smart urban water management, the water distribution network (WDN)—as the core infrastructure of urban supply systems—has emerged as a critical focus for enhancing resource utilization, operational efficiency, and water security. Unlike drainage systems that primarily manage gravity-based flow and stormwater routing, supply systems must address challenges such as pressure regulation, dynamic demand response, and coordinated pump–valve control [83,84].

5.2.1. Intelligent Control Requirements in Water Supply Systems

Urban water supply systems must maintain service reliability while minimizing energy consumption, ensuring pressure balance, and prolonging equipment lifespan. These systems are structurally complex, typically comprising multiple sets of pumps, tanks, valves, pipelines, and user demand nodes. Given the spatial heterogeneity and temporal variability of demand, scheduling decisions must strike a balance between real-time responsiveness, cost-efficiency, and system safety. Currently, the main control problems in WDNs are related to pump scheduling and valve operation, where pump activity directly influences energy usage and carbon emissions, while valves govern pressure distribution and flow equilibrium. These control processes are highly nonlinear and strongly interconnected, rendering conventional rule-based or optimization methods inadequate in the face of abrupt demand fluctuations, equipment malfunctions, or uncertain disturbances. Hence, there is a pressing need for intelligent control strategies capable of environmental perception, temporal memory, and multi-objective coordination.

5.2.2. Multi-Agent Modeling and Task Decomposition for Water Supply Networks

A water distribution system can be abstracted as a graph-based structure, where nodes represent reservoirs, tanks, consumers, or junctions, while edges represent physical components such as pipes, pumps, or valves. The control objective is typically to maintain acceptable tank levels while achieving network-wide pressure minimization under minimum pressure constraints. In a MAS framework, key control components—such as pump stations, main valves, and tanks—are modeled as autonomous agents. Each agent independently perceives its local state (e.g., pressure, tank level, demand flow) and selects actions (e.g., adjusting pump speed or valve position) accordingly. System-wide optimization is achieved through indirect coordination via environmental dynamics. Each agent is defined by four core functions: perception, memory, decision-making, and execution. Agents operate in a distributed manner without requiring explicit inter-agent communication, enhancing system scalability, robustness, and fault tolerance.

5.2.3. Applications of Deep Reinforcement Learning in Water Supply Control

To enhance autonomous decision-making, deep reinforcement learning (DRL) techniques have been widely applied to pump scheduling and valve operation, as illustrated in Figure 10. Representative algorithms such as Proximal Policy Optimization (PPO) and Deep Q-Networks (DQN) are trained through state–action–reward interactions in simulated environments to derive long-term optimal policies.

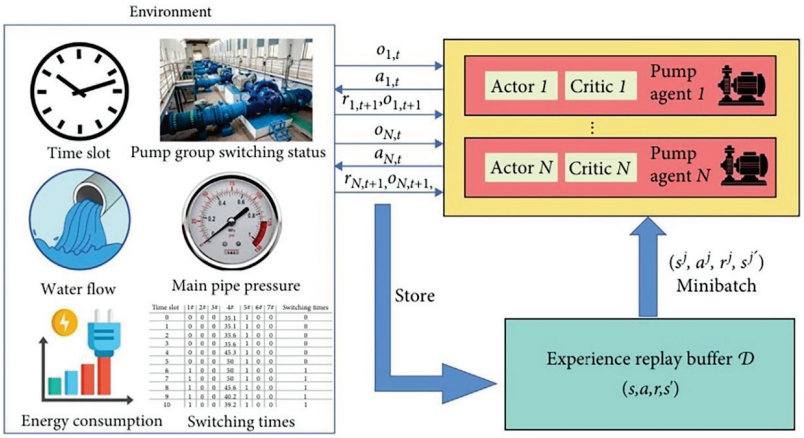


Figure 10. Real-time pump scheduling framework in a water distribution network.

PPO, a stable policy gradient method, clips the objective function to prevent excessively large updates. Its loss function is defined in Equation (12):

$$L(\theta) = \mathbb{E}_t\left[\log \pi_\theta(a_t|s_t) \cdot \hat{A}_t\right] \tag{12}$$

where $\hat{A}_t$ is the advantage function, measuring the deviation of the action from the average policy performance. PPO often uses a two-layer neural network to output parameters α and β of a Beta distribution for the action policy, defined in Equation (13):

$$h(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot x^{\alpha-1}(1-x)^{\beta-1} \tag{13}$$

To ensure uni-modal concavity, the outputs are transformed via Equation (14):

$$\alpha = \text{LeakyReLU}(x) + 1, \quad \beta = \text{Softplus}(x) + 1 = \log(1+\exp(x)) + 1 \tag{14}$$

In practical applications, PPO agents are trained with an hourly time step over a 24 h operational cycle. The state inputs include pressure, tank level, and time, while the actions represent valve positions or pump speed settings. EPANET is used for hydraulic simulation and reward computation.

To reflect real-world demand variability, Zaman et al. applied a DQN strategy to optimize pump operating periods, reducing energy consumption while maintaining supply reliability [85]. Candelieri et al. used Q-learning for pump on/off control, demonstrating scalability and tolerance to demand uncertainty in multi-pump systems [86]. Xu et al. integrated knowledge assistance using maximum state values from historical data to accelerate PPO convergence under unstable conditions [87]. Vercouter further enhanced DQN with offline learning and supervised pretraining (e.g., k-NN) for improved policy performance. Hung used Q-learning to allocate water quantities in non-stationary supply environments [88].

Hung employed Q-learning to allocate water resources under non-stationary supply conditions, effectively addressing dynamically changing water availability [87]. Overall,

deep reinforcement learning (DRL) algorithms—particularly PPO, DQN, MADDPG, and Q-learning—have demonstrated promising applicability in urban water supply systems, especially under multi-scenario, multi-objective, and high-disturbance settings [89]. Future research may further enhance the generalization capabilities and cross-regional deployment efficiency of multi-agent policies by integrating meta-learning, transfer learning, and federated learning techniques.

Wang applied a data-driven personalized federated multi-agent attention-based deep reinforcement learning (PFL-MAADRL) algorithm to address the intake scheduling problem of three water abstraction pump stations in an urban water treatment system. The model simultaneously optimized energy consumption, reservoir water levels, and trunk line pressure. Under uncertainty, the proposed method demonstrated robust performance and achieved up to a 10.6% reduction in maximum energy consumption compared to benchmark approaches [90].

Jiang proposed a coupled framework that integrates a comprehensive water quality model with a deep reinforcement learning (DRL) algorithm and applied it to the eutrophic Lake Dianchi—the largest freshwater lake in China—for validation. The results showed that compared to previous operational strategies, the proposed approach significantly reduced total nitrogen and total phosphorus concentrations in the lake by 7% and 6%, respectively [7].

5.3. Multi-Agent Regulation in Hydropower Energy Systems

Hydropower systems, functioning as integrated water infrastructure that combines water supply, flood control, power generation, and ecological regulation, involve highly complex and nonlinear processes. Particularly in scenarios such as multi-reservoir joint scheduling, real-time flood discharge, and energy conversion efficiency optimization, the system is characterized by significant uncertainty and conflicting multi-objective demands, which pose challenges for traditional optimization approaches [91]. In recent years, reinforcement learning (RL) and deep reinforcement learning (DRL) have been increasingly applied to the intelligent operation of hydropower systems, demonstrating superior adaptability and generalization capabilities. In contexts where accurate environmental modeling is infeasible or traditional methods fail to yield timely solutions, multi-agent systems (MASs) offer a scalable paradigm for addressing high-dimensional and dynamic control problems collaboratively, as illustrated in in Figure 11.

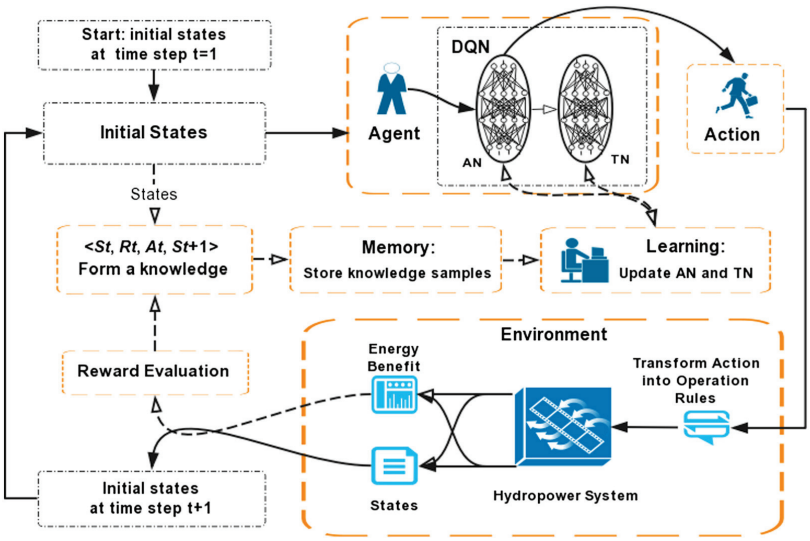


Figure 11. Deep reinforcement learning framework for hydropower system control.

5.3.1. Hydropower Scheduling Problems and DRL-Based Hydropower System Control

Hydropower operation problems can be formulated as Markov Decision Processes (MDPs), where the state space typically includes reservoir levels, inflow rates, and inflow forecasts, the action space includes turbine flow or spill discharge, and the reward function integrates electricity revenue, water utilization efficiency, and risk mitigation. Hu et al. were among the first to apply the Deep Q-Network (DQN) to real-time multi-reservoir scheduling, using two separate neural networks to estimate action values and policy functions. This approach effectively overcomes the "curse of dimensionality" faced by conventional RL in high-dimensional state spaces and outperforms both dynamic programming and decision tree-based strategies, particularly in balancing hydropower benefits and interpretability [92].

Luo et al. further enhanced DQN performance by integrating dueling networks, Double DQN (DDQN), and prioritized experience replay, which improved learning stability and sample efficiency [93]. Jiang et al. extended DQN to a multi-energy system by incorporating solar and wind power forecasts, achieving significant improvements in overall system profit through cross-energy coordination [94]. Wu et al. developed a multi-objective RL framework to optimize water supply, ecological preservation, and power generation simultaneously. By introducing weighted combinations into the reward function and evaluating multiple policy strategies, the study demonstrated superior performance in multi-objective trade-offs compared to traditional baselines [95].

Mitjana et al. addressed constraint scheduling under reservoir level uncertainty using the REINFORCE algorithm. By incorporating chance constraints and fallback mechanisms, they enhanced system safety and operational robustness [96]. Riemer-Sorensen employed the Soft Actor–Critic (SAC) algorithm to optimize weekly hydropower dispatch, leveraging historical electricity price and inflow data to reduce overflow loss and increase generation profit [97].

5.3.2. Cutting-Edge Applications of Multi-Agent DRL in Hydropower Systems

Multi-agent architectures are particularly well-suited for coordinated scheduling tasks in multi-reservoir and multi-hydropower systems. Hu et al. developed a collaborative model based on Multi-Agent Deep Deterministic Policy Gradient (MADDPG) for hydropower dispatching, achieving significant improvements in both energy consumption control and electricity generation efficiency [98].

MADDPG adopts a centralized training and decentralized execution (CTDE) paradigm, wherein each agent (e.g., reservoir unit) makes decisions independently using local observations, while policy updates are performed using global information. This enhances the generalization ability and robustness of the overall system.

Xi et al. systematically evaluated the sensitivity of scheduling performance with respect to key hyperparameters, including scheduling time step, state space representation, and policy update frequency. To alleviate the training burden caused by high-dimensional inputs, state space compression and feature construction techniques were applied to reduce input dimensionality [99].

To improve modeling under high-dimensional state inputs, researchers have incorporated factors such as hydraulic interdependencies among reservoirs, seasonal forecasts, and stochastic inflow disturbances. These enhancements enable DRL models to better adapt to abrupt hydrological changes across river basins. Moreover, the design of DRL decision networks is evolving toward deeper and more modular architectures. By introducing residual connections, attention mechanisms, and contrastive learning components, the models exhibit enhanced capabilities in extracting and leveraging temporal features for robust and adaptive scheduling.

5.4. Intelligent Task and Path Control in Multi-Platform Autonomous Monitoring Systems

In response to the growing demand for dynamic sensing, real-time response, and efficient spatial coverage in water environmental systems, multi-platform autonomous monitoring systems have demonstrated substantial value in tasks such as pollution tracking, water quality assessment, and ecological early warning. Approaches based on multi-agent systems (MASs)—covering task allocation, path planning, and water body monitoring—have emerged as key technologies driving intelligent regulation in aquatic environments.

5.4.1. Multi-Agent Task Allocation and Coordination Mechanisms

A key challenge lies in optimizing task allocation and execution coordination among multiple AUV/SV platforms. To address issues such as task priority, platform heterogeneity, and communication constraints, researchers have proposed various distributed strategies—such as auction-based mechanisms, market-driven games, and graph-theoretic modeling—for multi-objective task planning. Wu et al. proposed an improved consensus algorithm based on bundle assignment for dynamic task allocation among heterogeneous AUV platforms [100].

Furthermore, Nguyen developed a distributed coordination architecture for underwater multi-robot systems with limited communication bandwidth, which improved task coverage while maintaining manageable communication loads [101]. Lin et al. introduced a software-defined networking framework for centralized management of AUV networks. By integrating graph-based optimization strategies with reinforcement learning, they proposed a graph-structured SAC algorithm for optimal path scheduling and system input selection [102]. Their system further incorporates a three-tier network architecture—SDNA-ITS—comprising Detection AUVs at the network layer, Local Control AUVs at the intermediate layer, and a USV-based Master Controller (USV-MC) at the global coordination layer, enabling robust collaborative scheduling for complex pollution detection tasks.

5.4.2. Autonomous Path Planning and Dynamic Obstacle Avoidance Strategies

In intricate and obstacle-dense underwater environments, effective inspection of submerged structures relies on advanced path planning systems. These systems must not only address energy efficiency and obstacle avoidance but also dynamically adapt to ocean currents, physical obstructions, and sensor feedback. A hierarchical MER-SAC framework for AUV path planning is illustrated in Figure 12.

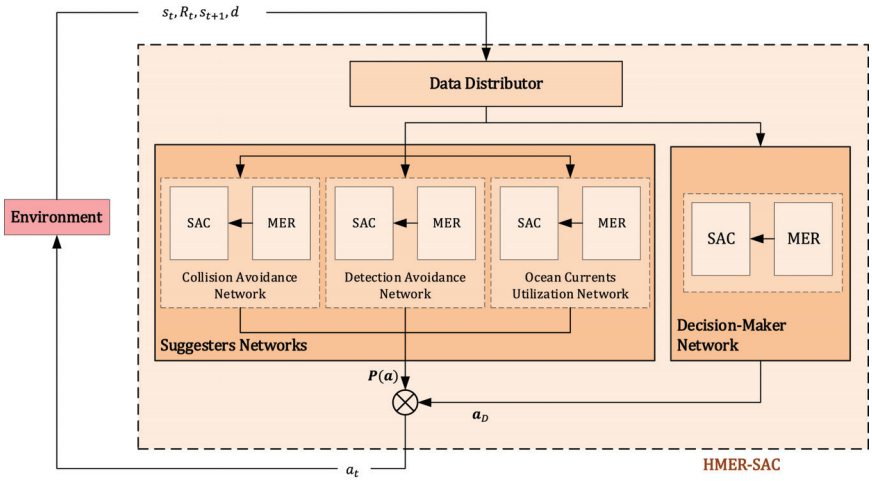


Figure 12. Schematic of the hierarchical MER-SAC architecture for AUV path planning.

Studies have shown that traditional methods perform well in static environments, but their effectiveness is limited in complex underwater scenarios [103]. In the realm of

intelligent path optimization, bio-inspired algorithms—including Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), and Cuckoo Search—have been widely applied to path generation for underwater missions [104]. However, their robustness is limited when facing abrupt environmental changes or unstructured scenarios. In recent years, deep reinforcement learning (DRL) has demonstrated significant advantages in underwater path planning, offering improved adaptability to uncertain environments and enhancing the overall intelligence of navigation systems [105]).

In addition, underwater defect detection tasks based on optical imaging impose higher requirements on path planning. Underwater image quality is often degraded by lighting variability, color attenuation, and turbidity. Therefore, path planning systems must simultaneously optimize imaging position, illumination angle, and obstacle avoidance strategy. To address the complexity and adaptivity required for underwater structure inspection, Tao proposed multi-objective DRL-based path planning frameworks [106]. These models, incorporating algorithms such as SAC or DDPG, enable dynamic responsiveness and real-time policy adaptation during the exploration of targeted regions.

5.4.3. Multi-Source Data Fusion and Real-Time Control in Water Body Monitoring Tasks

In water body monitoring tasks, multi-platform systems must ensure high spatiotemporal resolution in data acquisition over target regions and provide rapid responses to abnormal events such as sudden pollution incidents or algal bloom outbreaks. To achieve this, constructing a perception and control architecture based on multi-source heterogeneous data fusion has become a critical enabler, as illustrated in Figure 13.

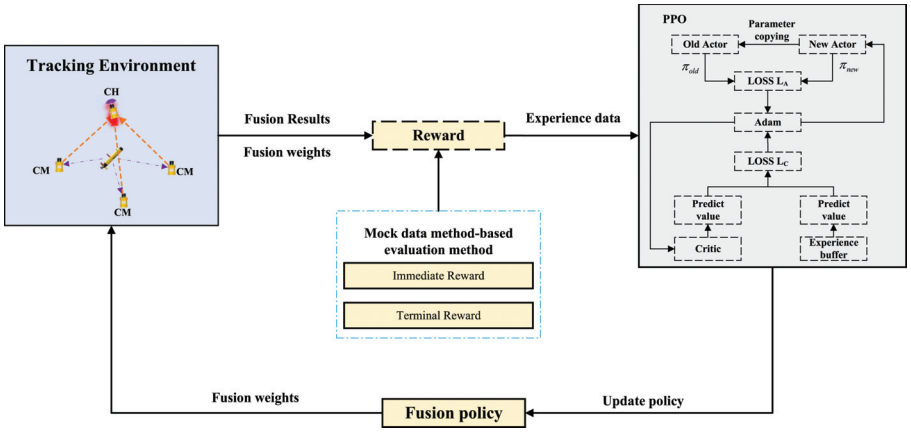


Figure 13. Architecture of a deep reinforcement learning-based framework for multi-sensor data fusion.

Punit proposed a water quality monitoring framework that integrates data from multiple sensors, enabling real-time perception of key environmental parameters and improving accuracy through error correction techniques [107]. Joana et al. further utilized a collaborative autonomous underwater vehicle (AUV) sensing mechanism to significantly enhance response times during algal bloom events [108].

Mancy proposed the *FL-MAPPO* model, which promotes decentralized and privacy-preserving decision-making to minimize latency and eliminate single points of failure. The framework adopts a multi-layer architecture encompassing data acquisition, distributed learning, and real-time execution. Experimental results showed that the proposed method outperformed existing large-scale real-time flood forecasting systems, achieving a mean squared error (MSE) of 0.112, R-squared (R^2) of 0.953, and mean absolute error (MAE) of 0.207 [109].

Building upon edge intelligence principles, Ren et al. developed a water environment monitoring system that integrates computation, analytics, and feedback mechanisms into a unified edge architecture. The real-time regulation system leverages wireless communication and edge computing to upload local sensing results to the control center, enabling rapid command dissemination and adaptive task updating [110].

Future advancements in multi-platform collaborative monitoring are expected to focus on intelligent formation control, energy-efficient coordination, and fine-grained target recognition. These developments will further enhance the system's comprehensive capabilities in emergency response, pollutant diffusion tracking, and ecological risk assessment.

6. Cross-Domain Generalization and Digital Twin Integration in Multi-Agent Systems

6.1. Policy Transfer and Robustness in Heterogeneous Scenarios

The deployment of multi-agent systems (MASs) in complex water environment scenarios faces substantial challenges, including diverse task types, heterogeneous geographic conditions, sensor inconsistency, and communication instability. To enhance the generalization and adaptability of learned policies across such non-stationary conditions, researchers have explored *transfer learning* and *meta-reinforcement learning (meta-RL)* techniques [111].

Yang et al. investigated the issue of suboptimal renewable energy utilization in integrated power systems (IPS) combining hydropower, photovoltaics, and pumped storage. They proposed a parallel multi-agent dynamic scheduling framework that integrates spatiotemporal information, enabling more efficient coordination and adaptation in dynamic, multi-modal energy environments [112].

Other studies have employed adversarial perturbation training and parameter regularization to improve the robustness of RL agents to sensor drift and input disturbances [113]. Under conditions of inconsistent or missing multi-source data, Bayesian policy optimization and distributed collaborative training were introduced to stabilize cross-regional deployments [114]. In heterogeneous systems such as urban water networks and mountainous reservoirs, knowledge distillation and agent weight-sharing strategies have been applied to transfer experiential knowledge across domains, significantly reducing the cost of retraining.

6.2. Integration Potential of Digital Twins in Intelligent Water System Control

Digital twins (DTs), as a key technology linking physical systems with virtual models, are increasingly being integrated into water system control. By fusing sensor data, simulation models, and control logic, DTs enable real-time mapping, prediction, and optimization of MAS operations. Heo et al. implemented a DT system in a wastewater treatment plant for dynamic aeration control using reinforcement learning, enabling feedback from online policy updates to physical processes and thus forming a closed-loop optimization cycle [115].

In urban water supply systems, Salvatore et al. developed a multi-scale DT framework that integrates LSTM-based prediction modules with MAS-based scheduling to achieve coordinated optimization of tank levels and pump energy consumption [116]. In multi-AUV systems, DTs are used to replicate individual agent states, supporting dynamic simulations of task planning, path optimization, and energy consumption while enabling real-time policy adjustment [117]. Looking ahead, DT-enabled intelligent water systems are expected to integrate complex system modeling, real-time simulation, and policy iteration, serving as foundational platforms for building high-reliability intelligent control systems.

Despite the strong theoretical synergy between digital twin (DT) technologies and reinforcement learning (RL), their integration in practical engineering systems still faces several critical challenges.

First, the simulation-to-reality gap (sim-to-real gap) remains a major bottleneck limiting the generalizability of trained models. In simulated environments, it is difficult to fully replicate real-world sensor precision, control dynamics, and external disturbances. As a result, policies trained within DT environments may experience performance degradation and lack robustness when deployed in real systems.

Second, model fidelity is a fundamental prerequisite for constructing effective digital twins. Any deviation or simplification in the physical modeling used for policy learning can impair the multi-agent system's perception capabilities, thereby undermining the stability and interpretability of the training process.

Third, system latency is a key factor influencing real-time control performance. In scenarios involving multi-source sensing, data transmission, model inference, and control execution, excessive system response time relative to the dynamics of the target process may result in ineffective decisions or system instability. This issue is particularly critical in high-frequency control tasks such as pump regulation and aeration control.

Therefore, the future integration of DT and multi-agent reinforcement learning (MARL) requires in-depth research into cross-modal data alignment, enhanced environmental modeling fidelity, and low-latency architecture design in order to build intelligent water system control frameworks that are stable, trustworthy, and practically deployable.

6.3. Edge Computing and Control Latency Optimization

In multi-agent systems deployed in real-world water infrastructure, limited wireless communication capacity and the spatial distribution of agents impose significant constraints on centralized coordination. As a result, both control commands and observation data must be processed at edge nodes near the source to mitigate communication latency.

Zhang proposed a distributed control strategy based on multi-agent reinforcement learning (MARL) to enhance system-level coordination and robustness under unstable communication conditions [118]. This approach enables agents to make decentralized decisions locally, while still contributing to global system objectives.

Applied to real-time control (RTC) in urban drainage systems, the method not only improves regulation efficiency but also addresses potential risks arising from hardware malfunctions in sensors and actuators or communication failures. The framework supports resilient operations in cyber–physical water infrastructure by leveraging edge intelligence for timely and adaptive control.

In large-scale AUV clusters, Zhang et al. deployed lightweight inference models at each edge node and used information compression mechanisms to upload only key control parameters to the cloud, maintaining global system consistency while minimizing network load [119]. For high-frequency control devices such as pumps, gates, and monitoring units, Qian et al. introduced edge-side control caches and sliding window prediction mechanisms, improving fault tolerance and latency robustness to achieve microsecond-level response times [120]. By integrating mobile edge computing (MEC), 5G network slicing, and asynchronous event-driven control architectures, future edge-intelligent water systems are expected to feature modular design, high-frequency responsiveness, and adaptive fault tolerance, driving the evolution toward low-latency and highly robust intelligent control.

7. Research Challenges and Future Directions

7.1. Challenges and Technical Limitations

Despite the promising applications of multi-agent reinforcement learning (MARL), digital twins, and edge intelligence in water environment systems, several challenges remain in their real-world deployment and adaptation to complex scenarios. These challenges can be broadly categorized into four major aspects:

(1) Limited policy generalization and poor system transferability.

Current MARL frameworks face significant difficulties in cross-regional deployment and heterogeneous scenario adaptation. Most policy transfer and meta-learning approaches rely heavily on structural similarities between source and target tasks, which are often lacking in complex water systems due to variations in geographic features, task objectives, and sensor configurations. While adversarial training and parameter regularization have been proposed to improve robustness, these methods still struggle to generalize under unpredictable disturbances and multi-source uncertainties encountered in real-world settings.

(2) Incomplete system-level co-modeling and control architectures.

Water environment systems are inherently composed of multiple interconnected subsystems characterized by multi-scale, multi-objective, and multi-constraint dynamics. Most existing methods remain focused on localized optimization, lacking a unified architecture for system-wide coordinated control. In large-scale, heterogeneous MARL settings, issues such as role assignment, communication protocols, and policy synchronization among agents are yet to be fully addressed. These limitations hinder task integration, reduce policy efficiency, and make it difficult to achieve real-time optimization under complex and dynamic conditions.

(3) Complexity in twin and edge system construction with limited real-time responsiveness.

Digital twin systems require high-fidelity physical modeling and high-quality real-time data, leading to high implementation costs and poor scalability. They are currently applicable mainly to isolated or small-scale systems. While edge computing can reduce control latency, challenges remain in terms of information sharing, global consistency maintenance, and resilience to communication failures. Moreover, limitations in computing power and sensing range at edge nodes hinder their scalability and adaptability in large-scale deployments.

(4) Limited adaptability and safety robustness under extreme conditions.

Most existing intelligent control models are primarily trained under stable operating conditions, lacking sufficient capacity to model system behavior in the face of nonlinear environmental responses, data incompleteness, or equipment malfunctions encountered during extreme events. This results in poor dynamic adaptability and inadequate coordination across multiple data and control sources during emergencies. In scenarios such as sudden flooding, contamination accidents, or system failures, agents typically lack risk-awareness mechanisms and fault-tolerant strategies, making it difficult to achieve resilient, safe, and continuous control—key objectives in the management of critical water infrastructure systems.

7.2. Future Research Directions

7.2.1. Building a Data Ecosystem and AI Integration for Intelligent Water Systems

Given the complexity of pollutant compositions, limitations in detection techniques, and severe fragmentation of data across water systems, constructing a comprehensive data ecosystem integrated with AI and machine learning has become a crucial direction for enhancing intelligent water system regulation. In recent years, researchers have proposed

using advanced spectroscopic detection, two-dimensional chromatography, and multi-channel online sensors to collect high-dimensional and structured environmental data. These data are then processed using AI models to identify key pollutants, enabling precise water quality risk warnings and real-time optimization.

Ding et al. developed an AI-assisted framework that integrates detection, analysis, and control optimization for identifying disinfection by-product (DBP) precursors. This framework separates dissolved organic matter (DOM) using two-dimensional chromatography, converts it into a digital matrix via UV/fluorescence detectors, and applies AI/ML methods to model the relationship between DOM composition and DBP formation potential [121]. The resulting predictive model informs the adjustment of disinfection parameters, thereby improving both the timeliness and accuracy of pollution detection and supporting dynamic optimization of water treatment processes.

Furthermore, integrating such approaches with digital twin systems enables real-time virtual–physical synchronization, dynamic water quality mapping, and strategy co-optimization. Future research should focus on multi-source data fusion modeling, the development of standardized and shareable data platforms, and the evolution of AI algorithms from black-box models to explainable frameworks, all aiming to enhance the reliability and autonomy of intelligent water environment management.

7.2.2. Perception-Driven Communication Mechanisms

In intelligent aquatic environments, clusters of distributed agents—such as water quality sensors, storage control units, and mobile sampling devices—must collaboratively perform complex tasks under conditions of partial observability. Given the constraints imposed by limited communication bandwidth, restricted sensing range, and dynamic task environments, the design of efficient systems has emerged as a critical enabler for intelligent system coordination.

Regarding communication timing, non-continuous (event-triggered) strategies have received considerable attention. In this paradigm, agents initiate communication only when significant state changes occur or when interaction is deemed essential. This approach effectively reduces communication overhead and is particularly suitable for scenarios that demand both responsiveness and resource efficiency, such as dynamic hydrological response, emergency pollution mitigation, or stormwater regulation.

In terms of interaction targets, group-based communication mechanisms offer a hybrid approach by integrating global and local communication strategies. Agents dynamically select communication partners based on current task demands, enabling broader coordination coverage while minimizing redundant exchanges. This approach is highly effective for cross-regional water resource allocation and multi-site water quality monitoring, facilitating task decomposition and goal partitioning among agents.

The encoding and compression of communication content are also critical for improving system scalability. Current research emphasizes techniques such as localized state sharing and neighbor-based compressed representations to reduce computational and memory burdens, making them well-suited for large-scale distributed deployments. These methods are applicable to scenarios including large-scale water network simulation and regional decision integration in urban drainage systems.

From the perspective of information processing, the adoption of mutual information optimization and attention-based mechanisms enhances the ability to extract salient signals from noisy observations. Attention-driven communication protocols dynamically assess the relevance of peer-provided information, allowing agents to focus on task-critical exchanges. Such mechanisms exhibit strong adaptability in applications such as pollution plume tracking and strategic protection of critical water sources [122].

In summary, a perception-driven communication framework structured around four core dimensions—communication timing, interaction targeting, message encoding, and information processing—lays a theoretical and technical foundation for effective coordination in distributed intelligent systems. Future research directions include

- Developing task-sensitive communication triggers tailored to non-stationary environments;
- Designing multi-scale communication structures and hierarchical architectures to accommodate system heterogeneity;
- Incorporating graph-based attention and multi-modal fusion for efficient heterogeneous information integration;
- Exploring co-optimization frameworks that jointly address communication and control strategies.

Together, these strategies can drive the evolution of intelligent water systems from "multi-point coordination" toward "ubiquitous intelligence," enabling advanced management and decision optimization across domains such as smart water engineering, real-time operations, and ecological protection.

7.2.3. Self-Organizing Multi-Agent Collaboration

Self-organizing multi-agent collaboration aims to develop autonomous agents capable of general collaboration, even in the absence of predefined protocols or prior knowledge. This approach is highly relevant to heterogeneous and decentralized water systems, especially in scenarios such as cross-basin water allocation, multi-agency pollution prevention, and emergency flood response, where agents must coordinate with unfamiliar or dynamically changing partners.

Early studies assumed static and observable agent behaviors, analogous to predefined operations in hydraulic facilities. However, real-world actors often exhibit behavioral uncertainty. Subsequent work has relaxed these assumptions by enabling agents to dynamically model their teammates through interaction learning. For example, agents can observe the behaviors of others to infer strategies and adapt accordingly—critical in systems where autonomous water quality sampling robots adjust their paths based on peer activity to enhance monitoring efficiency.

Other studies enhanced inter-agent communication through robust, low-bandwidth protocols to support temporary cooperation. These approaches are particularly applicable to distributed hydrological sensor networks and real-time coordination in urban drainage systems under low-latency constraints.

Despite progress, many approaches remain confined to closed-world settings with fixed agent types. In contrast, open-world water systems—comprising mobile sensors, heterogeneous infrastructure, and human interventions—require agents to adapt as their population evolves. Emerging methods use graph neural networks (GNNs), such as the GPL framework, to model evolving interaction structures, enabling effective coordination in scenarios like cross-watershed dispatching.

Further research has also considered adversarial robustness—such as handling malicious pollution sources or unreliable nodes—and few-shot collaborative learning for rapid adaptation in new or emergency tasks. In flood events, for example, agents must quickly learn effective strategies from minimal experience, while strategy coverage methods can help build more generalized agent libraries to support scalable simulation and training.

7.2.4. Collaborative Control Architecture for Complex Water Systems

Water environment systems exhibit a high degree of complexity and multi-scale coupling, encompassing a wide range of interconnected subsystems such as urban drainage, water resource scheduling, water quality management, and ecological conservation. These

subsystems are characterized by strong dynamic dependencies and spatiotemporal interactions, necessitating coordinated control at the system level to ensure operational safety, efficient resource allocation, and long-term ecological sustainability.

However, most reinforcement learning (RL)-based approaches to date are constrained to isolated subsystems or static scenarios, making them insufficient for capturing the inter-module dependencies and asynchronous dynamics inherent in such complex systems. Particularly under conditions involving multi-objective trade-offs and complex operational constraints, conventional RL methods often suffer from limited learning efficiency, poor scalability, and lack of coordination. These limitations highlight the urgent need for a novel control architecture that supports modularity, hierarchy, and collaboration.

Hierarchical control architectures (HCAs) provide a systematic framework for decomposing complex control problems into multiple layers, each responsible for perception, planning, and execution at different levels of granularity. This structure enhances both the manageability and learning efficiency of the system. Representative modeling methods include the following:

1. **Task decomposition and role assignment:** The global control objective is decomposed into multiple sub-goals, each corresponding to a control layer or agent role. For instance, in water resource management, the upper layer may handle high-level scheduling tasks such as water allocation and inter-regional distribution, while the lower layer focuses on local actuation tasks like pump control and feedback-based flow adjustments.
2. **Hierarchical policy learning:** The high-level controller (meta-controller) learns to assign sub-tasks or intermediate goals to lower-level controllers, while the low-level controllers optimize their actions based on local observations to fulfill the high-level directives. This structure can be implemented using the *options framework* or *hierarchical reinforcement learning (HRL)*, which improves policy interpretability and enhances transferability across tasks.
3. **State abstraction and coordination mechanisms:** Each layer can define its own state–action space abstraction to reduce learning complexity. Communication between layers or across subsystems can be facilitated through intermediate variables, shared representations, or attention-based mechanisms, enabling inter-layer coordination and policy alignment.
4. **Multi-agent cooperation within hierarchical structures:** Each sub-task can be further handled by one or more agents operating under a hierarchical framework. Mechanisms such as *centralized training with decentralized execution (CTDE)*, *heterogeneous reward design*, and *graph neural networks (GNNs)* are employed to promote agent collaboration and information sharing. For example, in wastewater treatment, individual agents can control flow regulation, chemical dosing, and energy optimization, while in watershed-scale management, agents can coordinate water storage operations, pollutant dispersion control, and ecological flow management.

Such a hierarchical multi-agent architecture inherently supports scalability, mitigates the curse of dimensionality in large state spaces, and addresses asynchronous dynamics and local conflicts in distributed environments. Moreover, state abstraction and inter-layer communication are key enablers for architectural efficiency. In practice, high-level policies operate on low-dimensional abstract states representing system trends or progress, while low-level agents respond to raw sensor inputs in real time. To ensure coherence between local actions and global objectives, communication protocols based on GNNs, attention mechanisms, or policy embeddings can be employed to construct shared state representations.

Future research should further explore the integration of *physics-informed modeling*, *multi-source heterogeneous data fusion*, and *cross-scale process coupling* into hierarchical multi-

agent frameworks. Embedding such architectures within *digital twin platforms* will enable closed-loop optimization from sensing and perception to cognitive decision-making. This integrated approach is expected to significantly enhance the intelligence, adaptability, and robustness of water environment systems under uncertainty, paving the way for a new generation of resilient and sustainable water governance systems.

7.2.5. Modeling Large-Scale Heterogeneous Multi-Agent Systems

Most reinforcement learning (RL) applications in intelligent water management primarily focus on small- to medium-scale homogeneous environments, where agents typically share similar architectures, objectives, and operational tasks. However, real-world water systems exhibit significantly greater complexity, encompassing functionally heterogeneous subsystems—such as urban drainage management, watershed dispatching, water treatment operations, and ecological preservation—along with multi-objective missions and diverse, multi-modal data sources. This paradigm shift calls for control models that can effectively manage large-scale, heterogeneous perception processes.

A comprehensive water management task may involve diverse agent types: sluice gate controllers responsible for hydraulic regulation in river networks, industrial supervisory agents focused on pollution mitigation, process-level agents optimizing wastewater treatment operations, and ecological agents tasked with maintaining environmental flows. These agents differ in their structural designs, functional objectives, underlying control mechanisms, and operational spatiotemporal scales. When operating within shared and dynamic environments, they must achieve strategic coordination across multiple layers to ensure system stability and task diversity fulfillment.

Currently, there is a notable absence of scalable RL frameworks specifically designed for heterogeneous multi-agent systems in complex water environments. This limits their deployment in real-world settings where variability, conflict, and cooperation are prevalent. Future research efforts should prioritize the following directions:

- **Modeling heterogeneous tasks and role-based structures:** Develop agent modeling frameworks that accommodate varying structures, roles, capabilities, and objectives, including clear task decomposition and role assignments;
- **Multi-objective game theory and cooperative strategy learning:** Employ game-theoretic models and multi-objective optimization techniques to support robust strategy formation under both cooperative and adversarial conditions;
- **Cross-scale coordination and communication protocol design:** Design efficient and scalable communication protocols to resolve asynchrony and coordination issues arising from distributed, multi-scale decision-making processes;
- **Generalization and scalability:** Construct RL training architectures that offer high generalizability and scalability, thereby improving performance, adaptability, and robustness in complex, uncertain real-world environments.

Addressing these challenges will be essential for transitioning from isolated RL deployments to fully integrated, large-scale intelligent water management systems capable of supporting long-term sustainability, operational resilience, and adaptive governance.

7.2.6. Adaptability and Robustness of Intelligent Models Under Extreme Conditions

The ability to optimize control strategies under extreme conditions is essential for developing resilient and secure control architectures in water environment systems. In practice, such systems are often subjected to abrupt disturbances, including floods, droughts, pollution incidents, equipment malfunctions, and extreme weather events. These anomalies may induce partial system failures, abrupt changes in resource availability, or severe sensor

outages, all of which pose considerable challenges to control models trained under nominal operational conditions.

Most existing reinforcement learning (RL) approaches were developed with a focus on optimal control in stable and routine scenarios, frequently overlooking the inherent uncertainties and nonlinear dynamics associated with extreme events. However, in such critical scenarios, both the system's safety boundaries and its operational objectives may undergo significant shifts.

Safety-constrained reinforcement learning algorithms have introduced distributed risk-aware critic mechanisms to enhance the adaptability of agents in high-risk scenarios. By incorporating distributed risk evaluation modules and context-aware policy optimization strategies, these frameworks enable the construction of intelligent control architectures capable of dynamically adjusting objective weights and strategy priorities under varying extreme conditions. Future research should prioritize the following directions:

- **Extreme scenario modeling and synthetic environment construction:** Develop high-fidelity simulation environments that emulate events such as flooding, pollution surges, and water shortages to support adversarial training and generalization of agent behavior;
- **Safety-constrained reinforcement learning:** Introduce risk-sensitivity metrics, dynamic safety threshold functions, and multi-objective constraint modeling to improve system boundary awareness and ensure secure policy execution;
- **Uncertainty modeling and robust optimization:** Apply Bayesian reinforcement learning and distributed decision-making frameworks to explicitly represent environmental and prediction uncertainties, thereby improving model robustness under disruptive conditions;
- **Integration of fault tolerance and emergency response:** Incorporate fault diagnosis, contingency planning, and multi-source coordination mechanisms to combine RL-based control with traditional emergency management systems for enhanced responsiveness during critical events.

By embedding these capabilities, future intelligent water systems will be more capable of handling rare yet high-impact disruptions, thereby contributing to the realization of robust, adaptive, and secure environmental management infrastructures.

8. Conclusions

Multi-agent reinforcement learning (MARL) is emerging as a foundational methodology for intelligent control in water environment systems, enabling a paradigm shift from centralized, rule-based operations toward distributed, cooperative, and adaptive decision-making frameworks. This review has systematically traced the evolution of MARL applications in the domain, highlighting its capabilities in improving water resource allocation, enhancing pollution response strategies, reducing operational energy consumption, and reinforcing system resilience under complex and dynamic conditions.

Despite recent advances, deploying MARL in real-world water systems remains a substantial challenge. Key limitations stem from agent heterogeneity, cross-domain task coupling, communication constraints, and the inherent non-stationarity of environmental dynamics. These challenges hinder the scalability, reliability, and interpretability of existing MARL-based control frameworks in large-scale operational settings.

To address these gaps, future research must focus on several critical directions: (i) developing hierarchical and cross-layer modeling frameworks that support system-level abstraction and inter-agent coordination; (ii) enhancing generalization and transfer capabilities across spatiotemporally variable tasks with limited training samples; (iii) designing interpretable and adaptive control architectures that remain robust under uncertain, multi-

domain operating conditions; and (iv) establishing unified benchmarks and semi-realistic testbeds that integrate hydrodynamic variability, sensing faults, communication latency, and external disturbances for the systematic validation of MARL algorithms.

In summary, MARL holds strong potential to become a core enabling technology for the next generation of intelligent, self-organizing, and sustainable water environment systems. Advancing this vision will require closer integration across control theory, machine learning, and environmental systems engineering, as well as a shift toward system-level co-design, multi-scale optimization, and field-level deployment to fully bridge the gap between algorithmic innovation and practical impact.

Author Contributions: Supervision, Y.P. and L.J.; investigation, L.J.; resources, L.J. and Y.P.; visualization, L.J.; writing—original draft, L.J.; writing—review & editing, Y.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, F.; Sewlia, M.; Dimarogonas, D.V. Cooperative control of heterogeneous multi-agent systems under spatiotemporal constraints. *Annu. Rev. Control* **2024**, *57*, 100946. [CrossRef]
2. Li, W.; Shi, F.; Li, W.; Yang, S.; Wang, Z.; Li, J. Review on Cooperative Control of Multi-Agent Systems. *Int. J. Appl. Math. Control Eng.* **2024**, *7*, 10–17.
3. Jain, G.; Kumar, A.; Bhat, S.A. Recent developments of game theory and reinforcement learning approaches: A systematic review. *IEEE Access* **2024**, *12*, 9999–10011. [CrossRef]
4. Zhu, R.; Liu, L.; Li, P.; Chen, N.; Feng, L.; Yang, Q. DC-MAC: A delay-aware and collision-free MAC protocol based on game theory for underwater wireless sensor networks. *IEEE Sens. J.* **2024**, *24*, 6930–6941. [CrossRef]
5. He, X.; Hu, Z.; Yang, H.; Lv, C. Personalized robotic control via constrained multi-objective reinforcement learning. *Neurocomputing* **2024**, *565*, 126986. [CrossRef]
6. Cheng, J.; Cheng, M.; Liu, Y.; Wu, J.; Li, W.; Frangopol, D.M. Knowledge transfer for adaptive maintenance policy optimization in engineering fleets based on meta-reinforcement learning. *Reliab. Eng. Syst. Saf.* **2024**, *247*, 110127. [CrossRef]
7. Jiang, Q.; Li, J.; Sun, Y.; Huang, J.; Zou, R.; Ma, W.; Guo, H.; Wang, Z.; Liu, Y. Deep-reinforcement-learning-based water diversion strategy. *Environ. Sci. Ecotechnol.* **2024**, *17*, 100298. [CrossRef]
8. Zuccotto, M.; Castellini, A.; Torre, D.L.; Mola, L.; Farinelli, A. Reinforcement learning applications in environmental sustainability: A review. *Artif. Intell. Rev.* **2024**, *57*, 88. [CrossRef]
9. Jiao, P.; Ye, X.; Zhang, C.; Li, W.; Wang, H. Vision-based real-time marine and offshore structural health monitoring system using underwater robots. *Comput.-Aided Civ. Infrastruct. Eng.* **2024**, *39*, 281–299. [CrossRef]
10. Kartal, S.K.; Cantekin, R.F. Autonomous Underwater Pipe Damage Detection Positioning and Pipe Line Tracking Experiment with Unmanned Underwater Vehicle. *J. Mar. Sci. Eng.* **2024**, *12*, 2002. [CrossRef]
11. Ravier, R.; Garagić, D.; Galoppo, T.; Rhodes, B.J.; Zulch, P. Multiagent Reinforcement Learning and Game-Theoretic Optimization for Autonomous Sensor Control. In Proceedings of the 2024 IEEE Aerospace Conference, Big Sky, MT, USA, 2–9 March 2024; pp. 1–12. [CrossRef]
12. Shi, H.; Li, J.; Mao, J.; Hwang, K.S. Lateral Transfer Learning for Multiagent Reinforcement Learning. *IEEE Trans. Cybern.* **2023**, *53*, 1699–1711. [CrossRef]
13. Han, S.; Dastani, M.; Wang, S. Sparse communication in multi-agent deep reinforcement learning. *Neurocomputing* **2025**, *625*, 129344. [CrossRef]
14. Luo, J.; Zhang, W.; Jiongming, S.; Yuan, W.; Chen, J. Research Progress of Multi-Agent Game Theoretic Learning. *J. Syst. Eng. Electron.* **2022**. [CrossRef]
15. Rutherford, A.; Ellis, B.; Gallici, M.; Cook, J.; Lupu, A.; Ingvarsson Juto, G.; Willi, T.; Hammond, R.; Khan, A.; Schroeder de Witt, C.; et al. Jaxmarl: Multi-agent rl environments and algorithms in jax. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 50925–50951.
16. Nagoev, Z.; Bzhikhatlov, K.; Pshenokova, I.; Unagasov, A. *Algorithms and Software for Simulation of Intelligent Systems of Autonomous Robots Based on Multi-Agent Neurocognitive Architectures*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 381–391.
17. Wu, P.; Guan, Y. Multi-agent deep reinforcement learning for computation offloading in cooperative edge network. *J. Intell. Inf. Syst.* **2025**, *63*, 567–591. [CrossRef]

18. Milani, S.; Topin, N.; Veloso, M.; Fang, F. Explainable reinforcement learning: A survey and comparative review. *ACM Comput. Surv.* **2024**, *56*, 1–36. [CrossRef]
19. Ernst, D.; Louette, A. *Introduction to Reinforcement Learning*; Feuerriegel, S., Hartmann, J., Janiesch, C., Zschech, P., Eds.; Springer: Singapore, 2024; pp. 111–126.
20. Tam, P.; Ros, S.; Song, I.; Kang, S.; Kim, S. A survey of intelligent end-to-end networking solutions: Integrating graph neural networks and deep reinforcement learning approaches. *Electronics* **2024**, *13*, 994. [CrossRef]
21. Boyajian, W.L.; Clausen, J.; Trenkwalder, L.M.; Dunjko, V.; Briegel, H.J. On the convergence of projective-simulation–based reinforcement learning in Markov decision processes. *Quantum Mach. Intell.* **2020**, *2*, 13. [CrossRef] [PubMed]
22. Albrecht, S.V.; Christianos, F.; Schäfer, L. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*; MIT Press: Cambridge, MA, USA, 2024.
23. Li, Y.; Tsang, Y.P.; Wu, C.H.; Lee, C.K.M. A multi-agent digital twin–enabled decision support system for sustainable and resilient supplier management. *Comput. Ind. Eng.* **2024**, *187*, 109838. [CrossRef]
24. Tong, Y.; Fei, S. Research on Multi object Occlusion Tracking and Trajectory Prediction Method Based on Deep Reinforcement Learning. In Proceedings of the 2024 6th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI), Nanjing, China, 6–8 December 2024; pp. 656–659.
25. Palmer, G.; Tuyls, K.; Bloembergen, D.; Savani, R. Lenient multi-agent deep reinforcement learning. *arXiv* **2017**, arXiv:1707.04402.
26. Zheng, Y.; Meng, Z.; Hao, J.; Zhang, Z. Weighted double deep multiagent reinforcement learning in stochastic cooperative environments. In *Pacific Rim International Conference on Artificial Intelligence*; Springer: Cham, Switzerland, 2018; pp. 421–429.
27. Micieli, M.; Botter, G.; Mendicino, G.; Senatore, A. UAV thermal images for water presence detection in a Mediterranean headwater catchment. *Remote Sens.* **2021**, *14*, 108. [CrossRef]
28. Zhang, C.; Wu, Z.; Li, Z.; Xu, H.; Xue, Z.; Qian, R. Multi-agent Reinforcement Learning-Based UAV Swarm Confrontation: Integrating QMIX Algorithm with Artificial Potential Field Method. In Proceedings of the 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Kuching, Malaysia, 6–10 October 2024; pp. 161–166.
29. Romero, A.; Song, Y.; Scaramuzza, D. Actor-critic model predictive control. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024; pp. 14777–14784.
30. Zhang, J.; Han, S.; Xiong, X.; Zhu, S.; Lü, S. Explorer-actor-critic: Better actors for deep reinforcement learning. *Inf. Sci.* **2024**, *662*, 120255. [CrossRef]
31. Zhao, E.; Zhou, N.; Liu, C.; Su, H.; Liu, Y.; Cong, J. Time-aware MADDPG with LSTM for multi-agent obstacle avoidance: A comparative study. *Complex Intell. Syst.* **2024**, *10*, 4141–4155. [CrossRef]
32. Kuba, J.G.; Chen, R.; Wen, M.; Wen, Y.; Sun, F.; Wang, J.; Yang, Y. Trust region policy optimisation in multi-agent reinforcement learning. *arXiv* **2021**, arXiv:2109.11251.
33. Harris, A.; Liu, S. Maidrl: Semi-centralized multi-agent reinforcement learning using agent influence. In Proceedings of the 2021 IEEE Conference on Games (CoG), Copenhagen, Denmark, 17–20 August 2021; pp. 01–08.
34. Ke, C.; Chen, H. Cooperative path planning for air–sea heterogeneous unmanned vehicles using search-and-tracking mission. *Ocean Eng.* **2022**, *262*, 112020. [CrossRef]
35. Zhang, Y.; Chen, X.; Mao, Y.; Shuai, C.; Jiao, L.; Wu, Y. Analysis of resource allocation and $PM_{2.5}$ pollution control efficiency: Evidence from 112 Chinese cities. *Ecol. Indic.* **2021**, *127*, 107705. [CrossRef]
36. Zhang, H.; Zhao, H.; Liu, R.; Kaushik, A.; Gao, X.; Xu, S. Collaborative task offloading optimization for satellite mobile edge computing using multi-agent deep reinforcement learning. *IEEE Trans. Veh. Technol.* **2024**, *73*, 15483–15498. [CrossRef]
37. Jouini, O.; Sethom, K.; Namoun, A.; Aljohani, N.; Alanazi, M.H.; Alanazi, M.N. A survey of machine learning in edge computing: Techniques, frameworks, applications, issues, and research directions. *Technologies* **2024**, *12*, 81. [CrossRef]
38. Han, S.; Zhang, T.; Li, X.; Yu, J.; Zhang, T.; Liu, Z. The Unified Task Assignment for Underwater Data Collection with Multi-AUV System: A Reinforced Self-Organizing Mapping Approach. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 1833–1846. [CrossRef]
39. Davis, A.; Wills, P.S.; Garvey, J.E.; Fairman, W.; Karim, M.A.; Ouyang, B. Developing and field testing path planning for robotic aquaculture water quality monitoring. *Appl. Sci.* **2023**, *13*, 2805. [CrossRef]
40. Ullah, R.; Abbas, A.W.; Ullah, M.; Khan, R.U.; Khan, I.U.; Aslam, N.; Aljameel, S.S. EEWMP: An IoT-Based Energy-Efficient Water Management Platform for Smart Irrigation. *Sci. Program.* **2021**, *2021*, 5536884. [CrossRef]
41. Lenczner, G.; Chan-Hon-Tong, A.; Le Saux, B.; Luminari, N.; Le Besnerais, G. DIAL: Deep interactive and active learning for semantic segmentation in remote sensing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 3376–3389. [CrossRef]
42. Moubayed, A.; Sharif, M.; Luccini, M.; Primak, S.; Shami, A. Water leak detection survey: Challenges & research opportunities using data fusion & federated learning. *IEEE Access* **2021**, *9*, 40595–40611.
43. Xue, D.; Yuan, L.; Zhang, Z.; Yu, Y. Efficient Multi-Agent Communication via Shapley Message Value. In *IJCAI*; Cambridge University Press: Cambridge, UK, 2022; pp. 578–584. [CrossRef]
44. Abadal, S.; Jain, A.; Guirado, R.; López-Alonso, J.; Alarcón, E. Computing graph neural networks: A survey from algorithms to accelerators. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–38. [CrossRef]

45. Balasubramanian, E.; Elangovan, E.; Tamilarasan, P.; Kanagachidambaresan, G.; Chutia, D. Optimal energy efficient path planning of UAV using hybrid MACO-MEA* algorithm: Theoretical and experimental approach. *J. Ambient Intell. Humaniz. Comput.* **2023**, *14*, 13847–13867. [CrossRef] [PubMed]
46. Li, S.; Gupta, J.K.; Morales, P.; Allen, R.; Kochenderfer, M.J. Deep implicit coordination graphs for multi-agent reinforcement learning. *arXiv* **2020**, arXiv:2006.11438.
47. Ou, M.; Xu, S.; Luo, B.; Zhou, H.; Zhang, M.; Xu, P.; Zhu, H. 3D Ocean Temperature Prediction via Graph Neural Network with Optimized Attention Mechanisms. *IEEE Geosci. Remote. Sens. Lett.* **2024**, *21*, 1–5. [CrossRef]
48. Cao, W.; Yan, J.; Yang, X.; Luo, X.; Guan, X. Communication-Aware Formation Control of AUVs with Model Uncertainty and Fading Channel via Integral Reinforcement Learning. *IEEE/CAA J. Autom. Sin.* **2023**, *10*, 159–176. [CrossRef]
49. Yan, M.; Wang, Z. Water Quality Prediction Method Based on Reinforcement Learning Graph Neural Network. *IEEE Access* **2024**, *12*, 184421–184430. [CrossRef]
50. Fan, X.; Zhang, X.; Yu, X. A graph convolution network-deep reinforcement learning model for resilient water distribution network repair decisions. *Comput.-Aided Civ. Infrastruct. Eng.* **2022**, *37*, 1547–1565. [CrossRef]
51. Zhu, M.; Wen, Y.Q. Design and Analysis of Collaborative Unmanned Surface-Aerial Vehicle Cruise Systems. *J. Adv. Transp.* **2019**, *2019*, 1323105. [CrossRef]
52. Chen, J.; Wang, Z.; Srivastava, G.; Alghamdi, T.A.; Khan, F.; Kumari, S.; Xiong, H. Industrial blockchain threshold signatures in federated learning for unified space-air-ground-sea model training. *J. Ind. Inf. Integr.* **2024**, *39*, 100593. [CrossRef]
53. Du, Y.; Fu, H.; Wang, S.; Sun, Z. A Real-Time Collaborative Mapping Framework using UGVs and UAVs. In Proceedings of the 2024 IEEE International Conference on Unmanned Systems (ICUS), Nanjing, China, 18–20 October 2024; pp. 584–590.
54. Samadzadegan, F.; Toosi, A.; Dadrass Javan, F. A critical review on multi-sensor and multi-platform remote sensing data fusion approaches: Current status and prospects. *Int. J. Remote Sens.* **2025**, *46*, 1327–1402. [CrossRef]
55. Matsuki, H.; Scona, R.; Czarnowski, J.; Davison, A.J. CodeMapping: Real-Time Dense Mapping for Sparse SLAM using Compact Scene Representations. *arXiv* **2021**, arXiv:2107.08994. [CrossRef]
56. Ghasemieh, A.; Kashef, R. Towards explainable artificial intelligence in deep vision-based odometry. *Comput. Electr. Eng.* **2024**, *115*, 109127. [CrossRef]
57. Zhong, X.; Pan, Y.; Behley, J.; Stachniss, C. Shine-mapping: Large-scale 3d mapping using sparse hierarchical implicit neural representations. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; pp. 8371–8377.
58. Billings, G.; Camilli, R.; Johnson-Roberson, M. Hybrid visual SLAM for underwater vehicle manipulator systems. *IEEE Robot. Autom. Lett.* **2022**, *7*, 6798–6805. [CrossRef]
59. Brecko, A.; Kajati, E.; Koziorek, J.; Zolotova, I. Federated Learning for Edge Computing: A Survey. *Appl. Sci.* **2022**, *12*, 9124. [CrossRef]
60. Chen, D.; Deng, T.; Jia, J.; Feng, S.; Yuan, D. Mobility-aware decentralized federated learning with joint optimization of local iteration and leader selection for vehicular networks. *Comput. Netw.* **2025**, *263*, 111232. [CrossRef]
61. Lindsay, J.; Ross, J.; Seto, M.L.; Gregson, E.; Moore, A.; Patel, J.; Bauer, R. Collaboration of Heterogeneous Marine Robots Toward Multidomain Sensing and Situational Awareness on Partially Submerged Targets. *IEEE J. Ocean. Eng.* **2022**, *47*, 880–894. [CrossRef]
62. Braca, P.; Willett, P.; LePage, K.; Marano, S.; Matta, V. Bayesian Tracking in Underwater Wireless Sensor Networks with Port-Starboard Ambiguity. *IEEE Trans. Signal Process.* **2014**, *62*, 1864–1878. [CrossRef]
63. Chen, J.; Wu, H.T.; Lu, L.; Luo, X.; Hu, J. Single underwater image haze removal with a learning-based approach to blurriness estimation. *J. Vis. Commun. Image Represent.* **2022**, *89*, 103656. [CrossRef]
64. Wang, L.; Xu, L.; Tian, W.; Zhang, Y.; Feng, H.; Chen, Z. Underwater image super-resolution and enhancement via progressive frequency-interleaved network. *J. Vis. Commun. Image Represent.* **2022**, *86*, 103545. [CrossRef]
65. Guo, Y.; Li, Y.; Tan, H.; Zhang, Z.; Ye, J.; Ren, C. Research on Target Tracking Simulation System Framework for Multi-Static Sonar Buoys. *J. Phys. Conf. Ser.* **2023**, *2486*, 012097. [CrossRef]
66. Xuan, W.; Jian-She, G.; Bo-Jie, H.; Zong-Shan, W.; Hong-Wei, D.; Jie, W. A lightweight modified YOLOX network using coordinate attention mechanism for PCB surface defect detection. *IEEE Sens. J.* **2022**, *22*, 20910–20920. [CrossRef]
67. Yan, J.; Yi, M.; Yang, X.; Luo, X.; Guan, X. Broad-Learning-Based Localization for Underwater Sensor Networks with Stratification Compensation. *IEEE Internet Things J.* **2023**, *10*, 13123–13137. [CrossRef]
68. Han, L.; Tang, G.; Cheng, M.; Huang, H.; Xie, D. Adaptive Nonsingular Fast Terminal Sliding Mode Tracking Control for an Underwater Vehicle-Manipulator System with Extended State Observer. *J. Mar. Sci. Eng.* **2021**, *9*, 501. [CrossRef]
69. Motoi, N.; Hirayama, D.; Yoshimura, F.; Sabra, A.; Fung, W.k. Sliding Mode Control with Disturbance Estimation for Underwater Robot. In Proceedings of the 2022 IEEE 17th International Conference on Advanced Motion Control (AMC), Padova, Italy, 18–20 February 2022; pp. 317–322. [CrossRef]

70. Chen, B.; Hu, J.; Zhao, Y.; Ghosh, B.K. Finite-time observer based tracking control of uncertain heterogeneous underwater vehicles using adaptive sliding mode approach. *Neurocomputing* **2022**, *481*, 322–332. [CrossRef]
71. Soriano, T.; Pham, H.A.; Gies, V. Experimental investigation of relative localization estimation in a coordinated formation control of low-cost underwater drones. *Sensors* **2023**, *23*, 3028. [CrossRef]
72. Zeng, J.; Wan, L.; Li, Y.; Zhang, Z.; Xu, Y.; Li, G. Robust composite neural dynamic surface control for the path following of unmanned marine surface vessels with unknown disturbances. *Int. J. Adv. Robot. Syst.* **2018**, *15*, 1729881418786646. [CrossRef]
73. Luo, G.; Shen, Y. A study on path optimization of construction machinery by fusing ant colony optimization and artificial potential field. *Adv. Control Appl. Eng. Ind. Syst.* **2024**, *6*, e125. [CrossRef]
74. Tang, J.; Pan, Q.; Chen, Z.; Liu, G.; Yang, G.; Zhu, F.; Lao, S. An improved artificial electric field algorithm for robot path planning. *IEEE Trans. Aerosp. Electron. Syst.* **2024**, *60*, 2292–2304. [CrossRef]
75. Yan, J.; Zhang, L.; Yang, X.; Chen, C.; Guan, X. Communication-Aware Motion Planning of AUV in Obstacle-Dense Environment: A Binocular Vision-Based Deep Learning Method. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 14927–14943. [CrossRef]
76. Hu, B.B.; Zhang, H.T.; Liu, B.; Ding, J.; Xu, Y.; Luo, C.; Cao, H. Coordinated Navigation Control of Cross-Domain Unmanned Systems via Guiding Vector Fields. *IEEE Trans. Control Syst. Technol.* **2024**, *32*, 550–563. [CrossRef]
77. Sharif, S.; Zeadally, S.; Ejaz, W. Space-aerial-ground-sea integrated networks: Resource optimization and challenges in 6G. *J. Netw. Comput. Appl.* **2023**, *215*, 103647. [CrossRef]
78. Garofalo, G.; Giordano, A.; Piro, P.; Spezzano, G.; Vinci, A. A distributed real-time approach for mitigating CSO and flooding in urban drainage systems. *J. Netw. Comput. Appl.* **2017**, *78*, 30–42. [CrossRef]
79. Tian, W.; Fu, G.; Xin, K.; Zhang, Z.; Liao, Z. Improving the interpretability of deep reinforcement learning in urban drainage system operation. *Water Res.* **2024**, *249*, 120912. [CrossRef]
80. Zhang, M.; Xu, Z.; Wang, Y.; Zeng, S.; Dong, X. Evaluation of uncertain signals' impact on deep reinforcement learning-based real-time control strategy of urban drainage systems. *J. Environ. Manag.* **2022**, *324*, 116448. [CrossRef]
81. Heo, S.; Nam, K.; Kim, S.; Yoo, C. XRL-FlexSBR: Multi-agent reinforcement learning-driven flexible SBR control with explainable performance guarantee under diverse influent conditions. *J. Water Process Eng.* **2024**, *66*, 105991. [CrossRef]
82. Huang, Z.; Wang, Y.; Dong, X. Dimensions of superiority: How deep reinforcement learning excels in urban drainage system real-time control. *Water Res. X* **2025**, *28*, 100313. [CrossRef]
83. Mendoza, E.; Andramuño, J.; Núñez, J.; Córdova, L. Intelligent multi-agent architecture for a supervisor of a water treatment plant. *J. Phys. Conf. Ser.* **2021**, *2090*, 012124. [CrossRef]
84. Jiménez, A.F.; Cárdenas, P.F.; Jiménez, F. Intelligent IoT-multiagent precision irrigation approach for improving water use efficiency in irrigation systems at farm and district scales. *Comput. Electron. Agric.* **2022**, *192*, 106635. [CrossRef]
85. Zaman, M.; Tantawy, A.; Abdelwahed, S. Optimizing Smart City Water Distribution Systems Using Deep Reinforcement Learning. In Proceedings of the 2023 IEEE 20th International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET), Boca Raton, FL, USA, 4–6 December 2023; pp. 228–233.
86. Candelieri, A.; Perego, R.; Archetti, F. Bayesian optimization of pump operations in water distribution systems. *J. Glob. Optim.* **2018**, *71*, 213–235. [CrossRef]
87. Xu, J.; Wang, H.; Rao, J.; Wang, J. Zone scheduling optimization of pumps in water distribution networks with deep reinforcement learning and knowledge-assisted learning. *Soft Comput.* **2021**, *25*, 14757–14767. [CrossRef]
88. Donâncio, H.; Vercouter, L.; Roclawski, H. The Pump Scheduling Problem: A Real-World Scenario for Reinforcement Learning. *arXiv* **2022**, arXiv:2210.11111.
89. Hung, F.; Yang, Y.E. Assessing adaptive irrigation impacts on water scarcity in nonstationary environments—A multi-agent reinforcement learning approach. *Water Resour. Res.* **2021**, *57*, e2020WR029262. [CrossRef]
90. Wang, D.; Li, A.; Yuan, Y.; Zhang, T.; Yu, L.; Tan, C. Energy-saving scheduling for multiple water intake pumping stations in water treatment plants based on personalized federated deep reinforcement learning. *Environ. Sci. Water Res. Technol.* **2025**, *11*, 1260–1270. [CrossRef]
91. Quan, Y.; Xi, L. Smart generation system: A decentralized multi-agent control architecture based on improved consensus algorithm for generation command dispatch of sustainable energy systems. *Appl. Energy* **2024**, *365*, 123209. [CrossRef]
92. Hu, S.; Gao, J.; Zhong, D.; Wu, R.; Liu, L. Real-Time Scheduling of Pumps in Water Distribution Systems Based on Exploration-Enhanced Deep Reinforcement Learning. *Systems* **2023**, *11*, 56. [CrossRef]
93. Luo, W.; Wang, C.; Zhang, Y.; Zhao, J.; Huang, Z.; Wang, J.; Zhang, C. A deep reinforcement learning approach for joint scheduling of cascade reservoir system. *J. Hydrol.* **2025**, *651*, 132515. [CrossRef]
94. Jiang, W.; Liu, Y.; Fang, G.; Ding, Z. Research on short-term optimal scheduling of hydro-wind-solar multi-energy power system based on deep reinforcement learning. *J. Clean. Prod.* **2022**, *385*, 135704. [CrossRef]
95. Wu, R.; Wang, R.; Hao, J.; Wu, Q.; Wang, P. Multiobjective multihydropower reservoir operation optimization with transformer-based deep reinforcement learning. *J. Hydrol.* **2024**, *632*, 130904. [CrossRef]

96. Mitjana, F.; Denault, M.; Demeester, K. Managing chance-constrained hydropower with reinforcement learning and backoffs. *Adv. Water Resour.* **2022**, *169*, 104308. [CrossRef]
97. Riemer-Sørensen, S.; Rosenlund, G.H. Deep Reinforcement Learning for Long Term Hydropower Production Scheduling. In Proceedings of the 2020 International Conference on Smart Energy Systems and Technologies (SEST), Istanbul, Turkey, 7–9 September 2020; pp. 1–6. [CrossRef]
98. Hu, S.; Gao, J.; Zhong, D. Multi-agent reinforcement learning framework for real-time scheduling of pump and valve in water distribution networks. *Water Supply* **2023**, *23*, 2833–2846. [CrossRef]
99. Zhang, X.; Wang, Q.; Yu, J.; Sun, Q.; Hu, H.; Liu, X. A multi-agent deep-reinforcement-learning-based strategy for safe distributed energy resource scheduling in energy hubs. *Electronics* **2023**, *12*, 4763. [CrossRef]
100. Wu, X.; Gao, Z.; Yuan, S.; Hu, Q.; Dang, Z. A dynamic task allocation algorithm for heterogeneous UUV swarms. *Sensors* **2022**, *22*, 2122. [CrossRef]
101. Nguyen, T.; La, H.M.; Le, T.D.; Jafari, M. Formation Control and Obstacle Avoidance of Multiple Rectangular Agents with Limited Communication Ranges. *IEEE Trans. Control Netw. Syst.* **2017**, *4*, 680–691. [CrossRef]
102. Lin, C.; Han, G.; Zhang, T.; Shah, S.B.H.; Peng, Y. Smart Underwater Pollution Detection Based on Graph-Based Multi-Agent Reinforcement Learning Towards AUV-Based Network ITS. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 7494–7505. [CrossRef]
103. Zhang, C.; Cheng, P.; Du, B.; Dong, B.; Zhang, W. AUV path tracking with real-time obstacle avoidance via reinforcement learning under adaptive constraints. *Ocean Eng.* **2022**, *256*, 111453. [CrossRef]
104. Guo, Y.; Liu, H.; Fan, X.; Lyu, W. Research progress of path planning methods for autonomous underwater vehicle. *Math. Probl. Eng.* **2021**, *2021*, 8847863. [CrossRef]
105. Chu, Z.; Wang, F.; Lei, T.; Luo, C. Path planning based on deep reinforcement learning for autonomous underwater vehicles under ocean current disturbance. *IEEE Trans. Intell. Veh.* **2022**, *8*, 108–120. [CrossRef]
106. Tao, M.; Li, Q.; Yu, J. Multi-Objective Dynamic Path Planning with Multi-Agent Deep Reinforcement Learning. *J. Mar. Sci. Eng.* **2024**, *13*, 20. [CrossRef]
107. Khatri, P.; Gupta, K.K.; Gupta, R.K. Raspberry Pi-based smart sensing platform for drinking-water quality monitoring system: A Python framework approach. *Drink. Water Eng. Sci.* **2019**, *12*, 31–37. [CrossRef]
108. Fonseca, J.; Bhat, S.; Lock, M.; Stenius, I.; Johansson, K.H. Adaptive sampling of algal blooms using autonomous underwater vehicle and satellite imagery: Experimental validation in the baltic sea. *arXiv* **2023**, arXiv:2305.00774.
109. Mancy, H.; Ghannam, N.E.; Abozeid, A.; Taloba, A.I. Decentralized multi-agent federated and reinforcement learning for smart water management and disaster response. *Alex. Eng. J.* **2025**, *126*, 8–29. [CrossRef]
110. Ren, J.; Zhu, Q.; Wang, C. Edge Computing for Water Quality Monitoring Systems. *Mob. Inf. Syst.* **2022**, *2022*, 5056606. [CrossRef]
111. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. *Int. Conf. Mach. Learn.* **2017**, *70*, 1126–1135.
112. Yang, J.; Liu, J.; Qiu, G.; Liu, J.; Jawad, S.; Zhang, S. A spatio-temporality-enabled parallel multi-agent-based real-time dynamic dispatch for hydro-PV-PHS integrated power system. *Energy* **2023**, *278*, 127915. [CrossRef]
113. Li, Y.; Wu, B.; Feng, Y.; Fan, Y.; Jiang, Y.; Li, Z.; Xia, S.T. Semi-supervised robust training with generalized perturbed neighborhood. *Pattern Recognit.* **2022**, *124*, 108472. [CrossRef]
114. Yang, Z.; Jin, H.; Tang, Y.; Fan, G. Risk-Aware Constrained Reinforcement Learning with Non-Stationary Policies. In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, Auckland, New Zealand, 6–10 May 2024; pp. 2029–2037.
115. Heo, S.; Oh, T.; Woo, T.; Kim, S.; Choi, Y.; Park, M.; Kim, J.; Yoo, C. Real-scale demonstration of digital twins-based aeration control policy optimization in partial nitritation/Anammox process: Policy iterative dynamic programming approach. *Desalination* **2025**, *593*, 118235. [CrossRef]
116. Cavalieri, S.; Gambadoro, S. Digital twin of a water supply system using the asset administration shell. *Sensors* **2024**, *24*, 1360. [CrossRef]
117. Yang, J.; Xi, M.; Wen, J.; Li, Y.; Song, H.H. A digital twins enabled underwater intelligent internet vehicle path planning system via reinforcement learning and edge computing. *Digit. Commun. Networks* **2024**, *10*, 282–291. [CrossRef]
118. Zhang, Z.; Tian, W.; Liao, Z. Towards coordinated and robust real-time control: A decentralized approach for combined sewer overflow and urban flooding reduction based on multi-agent reinforcement learning. *Water Res.* **2023**, *229*, 119498. [CrossRef] [PubMed]
119. Li, M.; Zhang, X.; Guo, J.; Li, F. Cloud–edge collaborative inference with network pruning. *Electronics* **2023**, *12*, 3598. [CrossRef]
120. Qian, P.; Liu, G. Application of High-Frequency Intelligent Sensing Network in Monitoring and Early Warning of Water Quality Dynamic Change. *Int. J. Comput. Intell. Syst.* **2024**, *17*, 195. [CrossRef]

121. Ding, Y.; Sun, Q.; Lin, Y.; Ping, Q.; Peng, N.; Wang, L.; Li, Y. Application of artificial intelligence in (waste)water disinfection: Emphasizing the regulation of disinfection by-products formation and residues prediction. *Water Res.* **2024**, *253*, 121267. [CrossRef]
122. Jiang, J.; Lu, Z. Learning attentional communication for multi-agent cooperation. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*; MIT Press: Cambridge, MA, USA, 2018; pp. 7265–7275.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Design and Development of a Flexible Manufacturing Cell Controller Using an Open-Source Communication Protocol for Interoperability

Evangelos Tzimas *, George Papazetis, Panorios Benardos * and George-Christopher Vosniakos

Manufacturing Technology Laboratory, School of Mechanical Engineering, National Technical University of Athens, Heroon Polytechniou 9, GR15772 Athens, Greece; gpapazetis@gmail.com (G.P.); vosniak@mail.ntua.gr (G.-C.V.)

* Correspondence: v_tzimas@mail.ntua.gr (E.T.); pbenard@mail.ntua.gr (P.B.); Tel.: +30-210-772-1799 (P.B.)

Abstract: Flexible manufacturing cells provide significant advantages in low-volume mass-customization production but also induce added complexity and technical challenges in terms of integration, control, and extensibility. The variety of closed-source industrial protocols, the heterogeneous equipment, and the product's manufacturing specifications are main points of consideration in the development of such a system. This study aims to describe the approach, from concept to implementation, for the development of the controller for a flexible manufacturing cell consisting of heterogeneous equipment in terms of functions and communication interfaces. Emphasis is put on the considerations and challenges for effective integration, extensibility, and interoperability. Scheduling and monitoring performed by the developed controller are demonstrated for a manufacturing cell producing microfluidic devices (bioMEMS) that consists of six workstations and a robot-based handling system. Communication between the system controller and the workstations was based on open-source technologies instead of proprietary software and protocols, to support interoperability and, to a considerable extent, code reusability.

Keywords: manufacturing system integration; manufacturing controller; production monitoring; communication protocols; remote procedure call; production scheduling; interoperability

1. Introduction

The prevailing pattern of market demand for a large variety of customized products is pushing industry towards flexible and adaptive workflow models maintaining reasonable product prices to achieve high competitiveness [1]. Flexible manufacturing cells (FMC) provide a solution that bridges the gap between flexibility and automation [2].

An FMC may typically incorporate several processing and tending stations with widely varying functionality, such that it may not be possible to apply a uniform approach for directly commanding them to perform an action. The same heterogeneity is also encountered in terms of each machine's numerical control and communication support, especially in cases where legacy equipment is upgraded to serve as a component of an FMC. Furthermore, the specialized nature of each FMC usually leads to the necessity of creating customized software solutions for its high-level control to meet application specific requirements [3].

The purpose of this study is to develop a cell controller software (CCS), from a generic concept attempting to address common issues encountered in software integration and control, to the full implementation and deployment for an FMC. The CCS is demonstrated on an FMC that is tasked with the production of microfluidic devices (bioMEMS) consisting of six manufacturing workstations, a robot-based material handling and storage system, and a central control PC that runs the CCS. The heterogeneity of the equipment in terms of operation and communication renders this FMC a representative case study, depicting

the challenges that arise during the development of the CCS along with the solutions that were applied.

The current work aims to present the development of a manufacturing cell control software that is based on open-source technologies to address operational flexibility, data interoperability, and equipment heterogeneity challenges. The main novelty lies both at the conceptual stage, which introduces a modular software architecture made up of components that can be continuously updated to reflect any changes in the manufacturing system assets or their behavior and representation over time, as well as at the communication level between the software and the relevant equipment, which employs an abstraction layer where each cell equipment is represented by a dedicated client module that exposes the available set of methods and states to the rest of the control software. The latter enables a robust implementation since a variety of communication protocols can still be used without problem. The proposed architecture can be used as a basis to develop applications for FMC process control and monitoring tasks, which emphasize ease of adoption and customization while being easily extendable and able to integrate with larger applications. Since this architecture is based on open-source technologies, it can serve as an alternative to existing commercial platforms. The paper covers all stages from conceptual design to implementation, while explaining the specific requirements and constraints that are involved in each individual stage along with the provided solutions that were developed.

The next section provides a short account of the state-of-the-art in flexible manufacturing controller architectures in light of Industry 4.0. Section 3 describes the proposed CCS concept. Section 4 provides implementation details of this concept using a particular real-life FMC as a demonstration basis. Section 5 reports on tests and sample results of deploying the CCS. Section 6 summarizes findings and provides an outlook.

2. Previous Work

FMCs have been being developed since the late 1970s [4], yet the digitalization of the manufacturing workflow and the interconnection of physical assets and processes have become increasingly important with the emergence of industry 4.0 and the market demand for personalized production, leading to the adoption of agile manufacturing practices [5]. Industry 4.0 technologies offer numerous possibilities, ranging from cloud-based monitoring [6] and cloud-based [7] applications to the deployment of fully integrated systems from the production process level to the business level [2] with enhanced performance in terms of downtime and quick response to market changes [8].

Highly digitalized, integrated manufacturing systems can bring competitive advantages, yet their development is a non-standard, often challenging process [9]. System reference architectures, their semantics and ontologies [10], connectivity and interoperability of systems, and the standardization of Industrial Internet of Things (IIoT) applications, currently comprise domains of interest widely studied yet still evolving.

Reference architectures provide classifications, entity hierarchies, and viewpoints that aim to facilitate the development of industrial systems that conform to specific standards. The Reference Architecture Model Industry 4.0 (RAMI 4.0) is the dominant framework of IIoT applications in the manufacturing domain [11]. RAMI 4.0 expands on the Smart Grid Architecture Model (SGAM) and introduces a three-dimensional representation of the industry 4.0 space along three directions: (a) six layers corresponding to various IT perspectives of the system (e.g., functional descriptions, communications behavior, hardware/assets, etc.); (b) the product lifecycle and value stream based on the IEC 62890 standard; and (c) the hierarchical levels of assets within the plant (product, field device, control device, station, etc.) according to standards IEC 62264 and IEC 6151 [12]. The Industrial Internet Reference Architecture (IIRA) [13] based on the ISO "ISO/IEC/IEEE 42010:2011" architecture concepts is another multi-layer proposal on modeling IIoT systems, but it is not specific to manufacturing. It considers four layers (viewpoints), namely the business, usage, functional, and implementation, each one having a distinct scope and objectives. In essence, following this model provides a methodology for the stakeholders to identify the

vision for their organization and how to benefit from a pertinent IIoT system. While IIRA and RAMI 4.0 have been developed independently with different objectives, scopes, and approaches, they can be considered complementary rather than conflicting. Furthermore, joint efforts from their creators are applied to bridge possible differences and eventually improve interoperability among systems [14].

Admittedly, reference architectures are still quite abstract, and as a result, adhering to one could yield a range of significantly divergent end results. This becomes particularly evident when delving into the implementation of specific services relevant to a manufacturing system, such as the control service of an FMC integrating a variety of heterogeneous equipment, the real-time monitoring service of the equipment, and the predictive maintenance services, to name but a few. Alternative architecture proposals that are closer to particular systems or even to particular asset types naturally tend to be more concrete [15,16].

Interoperability of systems and versatility in the use of equipment in the context of smart factories and smart manufacturing is heavily affected by the communication protocols and interfaces that are applied across them. The Open Platform Communications Unified Architecture (OPC UA) provides a well-established standard to facilitate the exchange of information [17]. It is proposed as the communication standard to be used for developing RAMI 4.0 applications. Semantics and vocabulary for the manufacturing domain comprise the MTConnect standard, which in combination with the OPC UA can provide a solution guaranteeing interoperability among equipment that adopts these standards [18]. Alternative popular communication protocols are Message Queuing Telemetry Transport (MQTT), Constrained Application Protocol (CoAP), and Data Distribution Service (DDS) [19]. However, despite the fact that these protocols and standards have been available for almost two decades, there are still significant challenges. One such challenge is how to extend the capabilities of the already operational hardware infrastructure so that it can take advantage of modern ways of information management and communication technologies [20].

Flexibility, maintainability, and interoperability are fostered by service-oriented architectures (SOA) in the development of supporting software for manufacturing systems [21–23]. One of the most common applications for the developed software concerns solving the scheduling problem where evolutionary optimization approaches are trending [24,25]. The variety of technologies and different standards in Industry 4.0, along with the co-existence of modern and legacy equipment in industrial sites, induces challenges in performing their digital transformation [26]. SOA allows an incremental transition to the smart digital plant, while also facilitating extension of the functional capability of the system [27]. Software integration of equipment that conforms to different standards is another common issue that can be facilitated by middleware services/applications that act as an adapter, converting an arbitrary interface to the desired one [28–30].

Open-source technologies, in a broader sense, are also trending as practical solutions for enhancing the robustness and scalability of manufacturing systems, addressing the challenges of operational flexibility and interoperability presented in Industry 4.0. Shifting to open-source technology provides low-cost solutions and reduces dependency on proprietary systems, thus fostering innovation and collaboration across different sectors [31–33]. Especially for small to medium enterprises (SMEs), this approach can be a significant enabler by lowering the entry-level investment barrier and offering enhanced customization options. Furthermore, the use of open-source technologies in IoT and IIoT applications ensures greater security and reliability, with continuous updates and community-driven support that keeps the systems up to date with the latest security standards and technological advancements.

The development of the CCS reported next has been inspired by the concepts provided in this section. Even though it is not a typical large-scale industrial application and thus not all concepts presented are applicable, it is also built over a layered architecture and promotes SOA, too. The communication between the system components is realized mainly with the use of gRPC [34], a modern open-source communication protocol.

3. CCS Concept

An FMC is, in general, a complex system with numerous possible configurations and processes that often forwards this complexity to be handled at the CCS level. Nevertheless, apart from the differences across FMCs, there are also components that, at least at an abstract level, can be identified in most, if not all, such systems. Aiming to simplify the design of an FMC controller software, this section initially attempts a grouping of the essential common elements for the operation of any FMC into a set of component categories. These are then used in the proposed generic CCS software architecture that serves as a template, handling basic functionality, while allowing customizability according to application-specific requirements.

3.1. Common/Shared FMC Components

3.1.1. Physical Asset Components

The first class of components that should be included in the shared set are those directly recognized as physical assets of FMC. These are typically workstations and tending stations. For each of these categories, the creation of an interface is proposed. The methods exposed can vary depending on specific FMC setups, but a generic rule would be to include two subsets of methods, one to target the state monitoring of an asset (machine) and the other to target relevant actions (see Table 1 for an example).

Table 1. Physical asset components—interface example.

Method Categories	WorkStation	TendingStation
State monitoring	getProcessState getDoorState getCondition	getState
Action execution	startProcess stopProcess pauseProcess	movePiece

In essence, the use of interfaces for workstations and tending machines is intended to create an abstraction layer over the heterogeneity of the equipment and the rest of the application modules and to also simplify addressing a machine. Each workstation extends the relevant interface with a concrete implementation, hiding complicated logic, communication protocols used, and additional details, which are not useful beyond this layer.

3.1.2. Behavioral Components

Moving from physical assets, we focus our attention on FMC standard behaviors. The selected behaviors that were regarded as "standard" across FMCs are (i) process monitoring, (ii) task scheduling, and (iii) high-level control of machines. These operations are interdependent in a cyclic manner, namely, performing a scheduled task requires control of machines, which will subsequently trigger system state changes that need to be monitored. The new state is then used as an input to the scheduling algorithm that will generate new tasks. Naturally, the workflow is also directly affected by the human operator's decisions, e.g., selection of product mix for manufacturing, enabling/disabling of workstations, etc. The proposed software components to support this operation are summarized in Table 2 along with a short description of their responsibilities. The behavioral components are described at an abstract level since the implementation strategies could vary widely, without affecting the structure of the application.

The main concept is to use the FMC_controller as the high-level component that provides the endpoints of a service and is composed of the Supervisor, Scheduler, and Machine Operator components. The Supervisor, Scheduler, and Machine Operator components are coupled to interact through the state of the system and provide the monitoring, scheduling, and control functionality of the service.

Table 2. Behavioral components and responsibilities.

Supervisor
- Maintains system state - Notifies subscribers for state updates
Scheduler
- Runs scheduling algorithm to find the next FMC task/job
Machine Operator
- Executes tasks/jobs utilizing workstation/tending station components
FMC_controller
- Provides the endpoints for communication with other services - Maintains the cell's physical assets configuration - Allows the interaction of the Scheduler, Supervisor and Machine Operator components - Performs high-level operations, e.g., places new job orders, enables/disables scheduling.

3.1.3. System State Representation

It is obvious that the system and workstations' state play a central role in the interaction of the components. Even though data structures that are sufficient for describing the system state and maintaining the maximum capability for the subsystems cannot be standardized, there is a minimum requirement that needs to be met to effectively develop a CCS application.

Thus, the data structures need to represent unambiguously two state categories, namely, (i) the physical assets state and (ii) the FMC process state. Regarding the physical assets, simple enumerations could be sufficient. The values of the enumerations would then be the responses provided by invoking the state monitoring methods of the physical asset components. For example, invoking the getDoorState method of a workstation could produce one of the following values: "door open", "door closed", "door moving", and "door error".

However, a more elaborate data structure needs to be created for describing the process state of the FMC, as it needs to contain information about jobs being performed by the cell, routings, available resources, and more. An example of such a data structure is provided in Section 4.2.1.

3.2. Main CCS Service

The software components presented, are the main building blocks for a minimal CCS application. Figure 1 visualizes the hierarchical structure and the relationships of the components to form a framework for an application/service that supports process monitoring, task scheduling, and system setting. The main drive for using this design is building a modular main CCS application with interchangeable components that could easily be integrated as a service in a larger application.

The introduction of software entities, i.e., physical assets and behavioral components, along with their relationships and respective interfaces, effectively forms the CCS service. This structure is well suited to both service-oriented and traditional object-oriented approaches for their concrete implementation. This generic template is not limited to specific technologies and has provided the basis for developing the FMC controller application that is the subject of the following section.

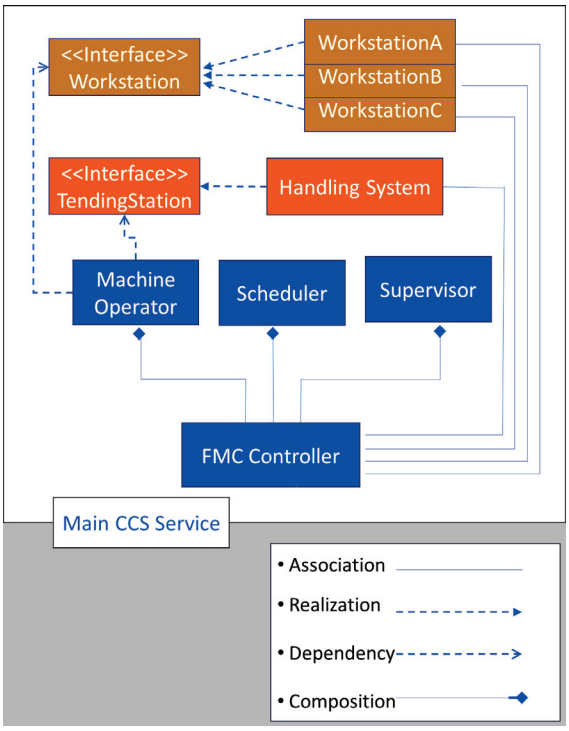


Figure 1. Components overview for the generic cell controller software service.

4. Implementation Example

A CCS application has been developed based on the concept and design proposed in Section 3. The application was demonstrated on an FMC dedicated to the production of microfluidic lab-on-chip devices (bioMEMS).

4.1. CCS and Subsystem Requirements

The functional requirements for the FMC operation were defined at an abstract level from the early design phase of the project. The primary purpose of the CCS is to integrate the various standalone machines, i.e., workstations and robot, into a system that effectively operates as an FMC. The software should facilitate the phased development and integration of workstations into the FMC, allowing them to be taken offline while the rest of the system remains fully operational. A graphical user interface should be included to allow for FMC process and system monitoring, as well as high-level control.

Each workstation is required to be addressable over the LAN and expose an API that enables the CCS to communicate with it, exchange information, and require specific actions. The CCS should be capable of flexibly supporting the integration of machines that meet this requirement but may expose a variety of interfaces and communication protocols.

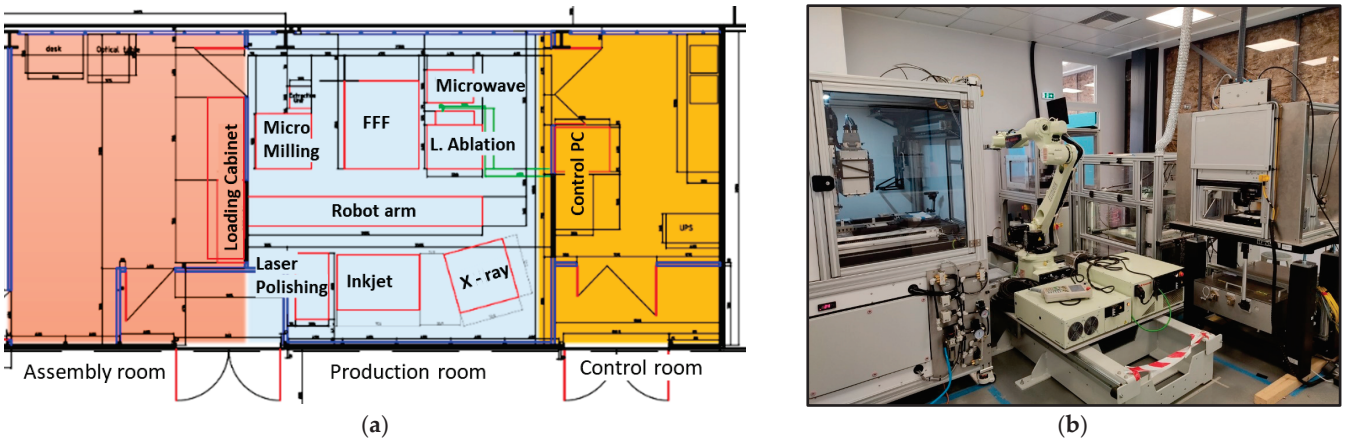
The FMC should be operated and monitored from the control room via the CCS software, which is hosted on the control PC (see Appendix C for specifications), and for safety purposes exclusively under trained personnel supervision. Consequently, the CCS is designed as a desktop application, and the control PC is not accessible remotely.

4.2. Flexible Manufacturing Cell Description

4.2.1. Layout

The manufacturing site is divided into three areas, namely, (i) the production room, (ii) the control room, and (iii) the assembly room (see Figure 2). The production room is the central area of the site, containing 5 processing stations, one metrology station, and the robot-based material handling system. Across the center of the room, a linear rail has been installed, allowing the industrial robot to move along it, tending the workstations. The latter are laid out on both sides of the rail, thus maximizing the reach and maneuvering

capability of the robot. An electrically operated gripper is the end-effector of the robot, capable of transferring specially designed carriers [35]. The carriers are the platform for transferring semi-finished products across the workstations and the storage cabinet. Each workstation as well as the storage cabinet are equipped with spigots to accurately position and securely hold the carriers.



(**a**) (**b**)

Figure 2. FMC layout (**a**) drawing; (**b**) physical.

The storage cabinet located in the assembly room is placed so that it can be directly accessible from both the robot operating in the production room and the human operator working in the assembly room. In this way, the human operator can collect from the cabinet the carriers with the finished products and replace them with empty ones to allow undisrupted operation of the FMC. The control room accommodates the control PC and auxiliary equipment. Using the CCS's user interface, the FMC operator monitors the process and system state and controls the overall production progress.

4.2.2. Workstations—Handling System

To perform the manufacturing of the lab-on-chip devices, a set of workstations and the handling system have been specifically developed or upgraded to meet the requirements of the FMC. Each subsystem enables interaction with the CCS by providing an appropriate service. The CCS sends requests or remote procedure calls, either to get a subsystem's state or to demand actions, and once the subsystem's software handles them, an appropriate response is returned. All involved systems communicate using the local area network of the production site.

A fused filament fabrication (FFF) 3D printer, capable of hosting and processing (mirror printing) three carriers simultaneously, is the first stage of the manufacturing process. The printing table has been modified appropriately to permanently host three spigots as placeholders for the received carriers. An enclosure has been added to the workstation to enable the automated operation of a custom loading/unloading door. The FFF–enclosure system is equipped with a Raspberry Pi 4TM microcomputer (see Appendix C for specifications), used primarily to provide control and monitoring capabilities of the workstation to an external application. This task is achieved by running a RepetierTM [34] server instance for the 3D printing process, leveraging the relevant REST API, and a gRPC server application developed to manage the enclosure's door subsystem.

The next stage of the manufacturing process involves micro-milling channels on the 3D-printed substrates. This is taken care of by the relevant workstation that has been specifically designed to be used as a component of the FMC. The micro-milling station is capable of processing one carrier at a time, and on top of its control software, it is running a gRPC server application to allow integration with the CCS by providing the necessary state monitoring and control interface.

Depending on the exact design of the bioMEMS device, the next steps involve a combination of laser processing and fluid dispensing processes. Custom processing workstations that support laser ablation, laser polishing, and drop-on-demand fluid dispensing of various materials have been developed and integrated into the FMC. The last step of the manufacturing process concerns the quality inspection of the product and is performed in a high-resolution X-ray metrology workstation. The mentioned workstations also provide a communication interface via the gRPC protocol for integration with the CCS application.

A crucial part of the FMC is the robot-based handling system, responsible for tending all the subsystems [35]. The industrial robot has been programmed to allow communication with the CCS software using the TCP/IP communication protocol, accepting carrier transfer commands, and providing status updates.

Each custom workstation developed for the FMC is equipped with a dedicated PC that hosts all necessary software for numerical control (NC), programming, and communication with the CCS application. Process data specific to each machine are stored in the respective PC. The FFF workstation provides the same functionality through the Raspberry Pi 4TM microcomputer. These systems can be accessed from the control PC via VNC software (version 7.1.0). The network configuration is illustrated in Figure 3.

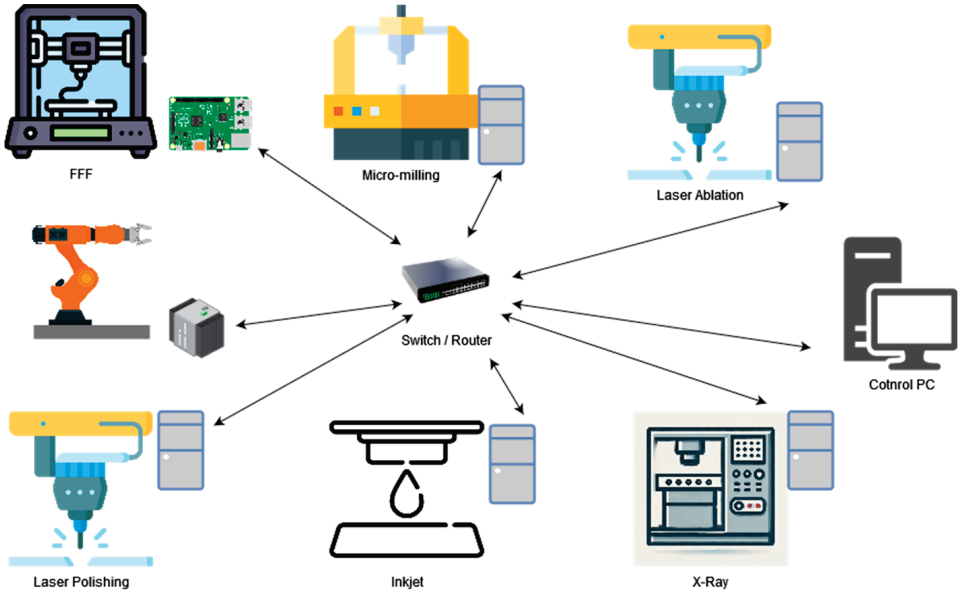


Figure 3. Communication network.

4.2.3. CCS Main Operation

The CCS enables the operator to issue production orders by creating and assigning process plans (or "recipes") to carriers. The CCS automatically instructs the workstations and the handling system to perform appropriate tasks based on its scheduling algorithm output and the subsystem's state that is acquired by frequent polling. Figure 4 documents this workflow through the execution of a basic one-step process plan involving the FFF workstation.

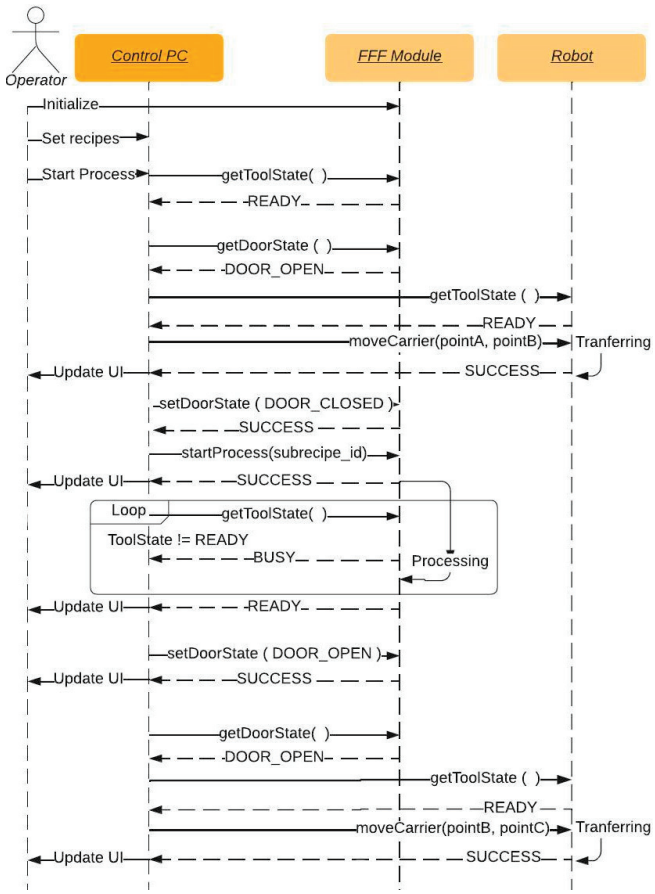


Figure 4. Process sequence diagram.

4.2.4. CCS Design

The CCS as an application to be executed in the control computer of the production area provides the following functionality:

1. handling communication with workstations
2. scheduling workstation and handling system tasks to complete the production process
3. monitoring production process and subsystems' status
4. supporting multi-product production process
5. providing GUI to facilitate operator in
    a. creating and assigning process plans
    b. controlling and monitoring automated production process.

High network input-output performance is crucial for the CCS, as it handles a large number of requests to monitor the state of workstations and issue commands accordingly, while not being involved in process-intensive tasks. To address this need, a single-threaded asynchronous approach has been adopted for both the main process and the graphical user interface workflow. In this case, JavaScript was used as the main programming language, facilitating an event-driven approach using event-listeners and attaching handlers [36]. A sequential-like workflow would be difficult to maintain since there are many disruptive events that can occur and would need to be handled as exceptions to this workflow.

The CCS structure should facilitate the phased development and integration of workstations into the FMC, allowing them to be taken offline while the rest of the system remains fully operational. The CCS application is composed of two processes, e.g., (i) the main process and (ii) the front-end, that complement each other. The main process is designed in accordance with the main CCS template architecture presented in Figure 1, and it is responsible for handling communication with the workstations, process scheduling, system monitoring, file/operating system manipulation, and serves the main window of the

application. The front-end part of the application consists of components that form the user interface. They define the way in which the information is displayed and the controls provided for the user to interact with the application. Figure 5 provides an overview of the modules and the flow of information between them.

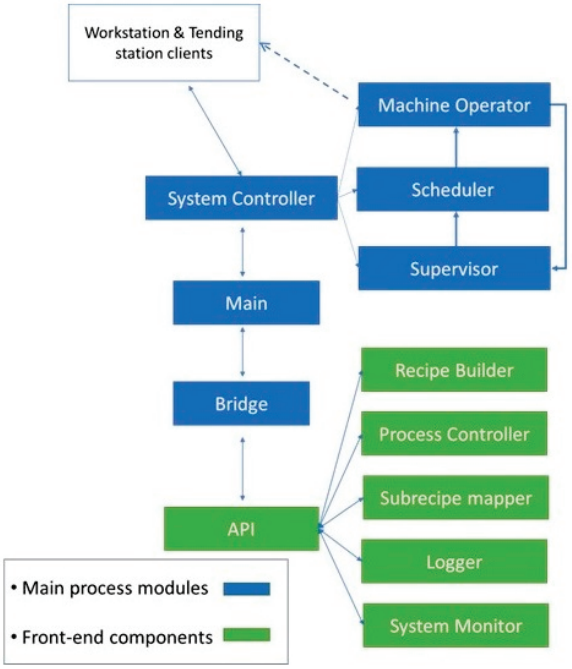


Figure 5. CCS module and component overview.

The communication framework between CCS and the workstations is required to be open, proven, flexible, and easy to implement. For the presented FMC use case, gRPC [34] was selected, as it is a robust, open-source Remote Procedure Call (RPC) framework that favors interoperability by supporting the use of a large variety of programming languages and environments, while allowing data structures and communication interfaces to be organized clearly and concisely with the use of special files (proto files) devoted to this purpose. The industrial robot and the FFF are commercial systems providing the means to communicate with the CCS over TCP/IP and HTTP, respectively. Note that the CCS is not limited to integration only with systems conforming to the above communication protocols.

The main process is an extended version of the proposed main CCS template (see Figure 1). It is developed using Electron [37], an open-source JavaScript framework for building cross-platform desktop applications using web technologies. The functionality of the main process is provided through the interaction of its various modules (see Figure 6). For the implementation of the CCS in this specific use case, the lack of common interfaces among the various workstations necessitated the development of dedicated software adapters, thus precluding the possibility of service reuse, which is one of the main advantages of SOA. This constraint, along with the increased complexity and overhead for service governance, dictated the choice of a monolithic design.

The main module is responsible for serving the application window to be populated by the front-end components and handles interactions with the operating and file systems. Along with the bridge module, they provide the interface for the bi-directional communication between the main process and the front-end. This has been facilitated by the utilities provided by the Electron framework. Event-listeners have been paired with suitable event-handlers, enabling the application to react to various triggers and update front-end and main process states accordingly. Certain types of events, such as completion of job orders, are triggered by the main process forcing the front-end to update, while others, such as click-events, are triggered by the front-end and handled by the main process.

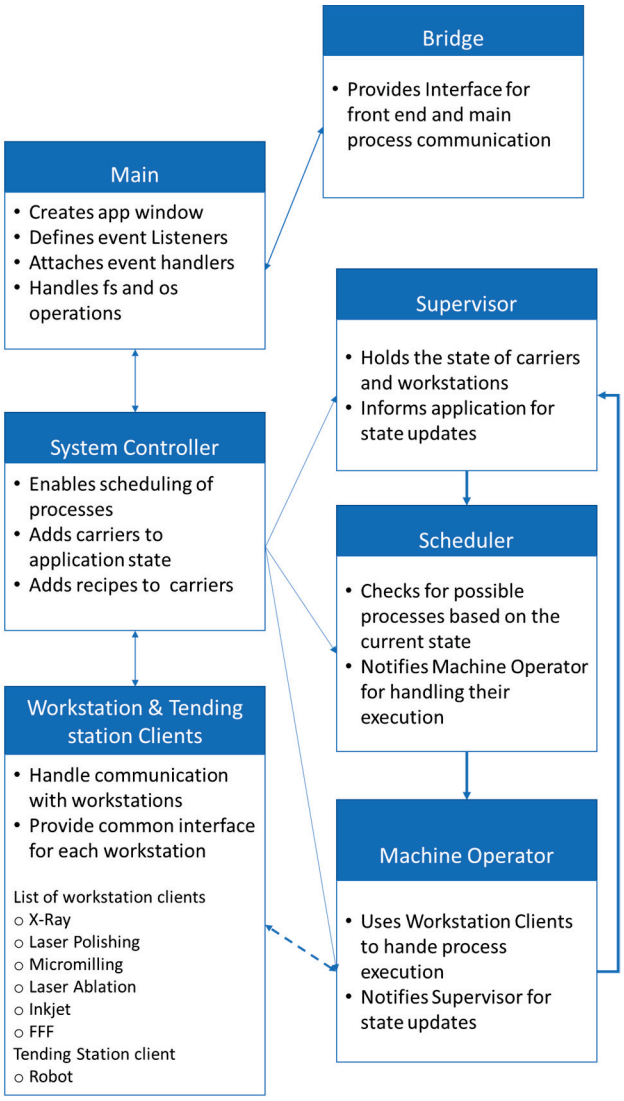


Figure 6. Main process modules.

The rest of the modules follow the architecture of the main CCS template. For each workstation and the handling system, there is a dedicated client module exposing a common set of methods to be used by the rest of the application (implementing either the workstation or the tending station interface presented in Section 3.1.1). The client modules effectively create an abstraction layer that highly promotes interoperability and flexibility. It has allowed the application to interact with workstations in a uniform manner, while the underlying communication has been implemented by a variety of technologies and protocols:

- gRPC services—micro-milling, laser ablation, laser polishing, inkjet, X-ray workstations,
- RESTful API service—FFF workstation
- TCP/IP service—industrial robot of the handling system

The system controller module is a high-level component, its main purpose being to hold the information regarding the physical setup of the FMC (e.g., the machines comprising it) and update the behavior of the application at runtime by providing handlers for user interactions, such as adding job orders, disabling scheduling, etc. It is composed of the Scheduler, Supervisor, and Machine Operator objects and defines the relationship among them.

The Supervisor, Machine Operator, and Scheduler objects belong to the same hierarchical level; their interaction effectively provides the desired control, monitoring, and scheduling capabilities of the application. They all depend on or manipulate the system state, and there is a cyclic dependency among them, where the output of one is input for the next, as illustrated in the following block of pseudocode.

```
while carrier processing is enabled:
    task = Scheduler(state)
    carrier_state = MachineOperator.perform(task)
    state = Supervisor.updateState(carrier_state)
```

The Supervisor is responsible for holding the carrier and system states and the Scheduler on updates. The Machine Operator performs the various tasks using the appropriate workstation or handling system clients and notifies the Supervisor to update the state accordingly. The Scheduler, based on the system and carriers' states, searches for possible actions and instructs the Machine Operator to handle their execution. An action can either be a carrier transfer or a manufacturing process at a workstation.

A greedy algorithm has been applied for scheduling the asynchronous execution of actions. It is worth mentioning that when a multi-carrier process takes place, there is the possibility for scheduling deadlocks to occur. When the current location occupied by one carrier obstructs the production sequence of another, it effectively prevents the latter from moving to its intended location. If the second carrier, in turn, obstructs yet another carrier, and this situation repeats cyclically, eventually blocking the first one from reaching its intended next location, a deadlock arises. The Scheduler module is responsible for detecting and resolving deadlocks by moving one of the involved carriers back to the loading cabinet that serves the purpose of a buffer in such a situation. Figure 7 provides a Gantt chart with the results on a multi-carrier, multiple-process-plan production, while Figure 8 displays the basic scheduling workflow.

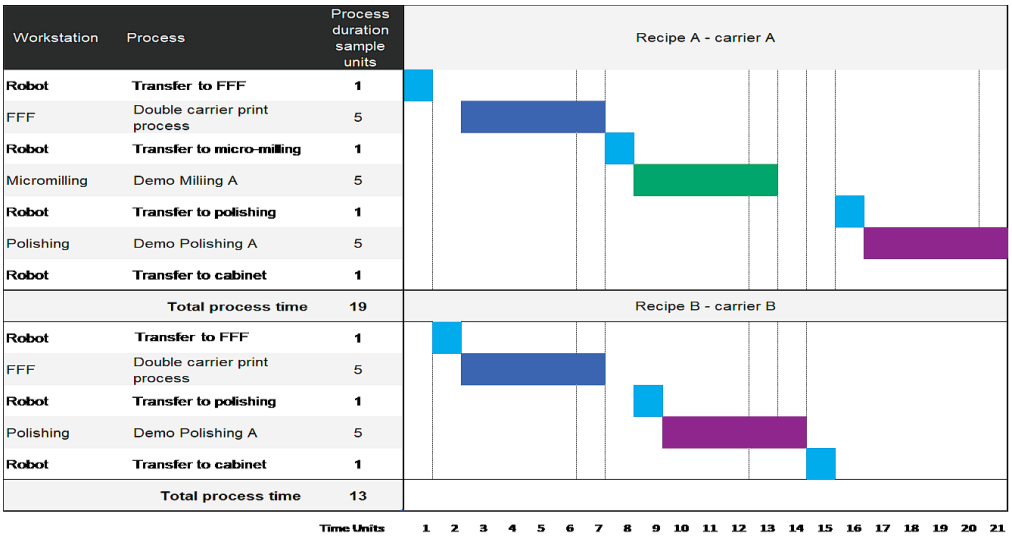


Figure 7. Multi-carrier multi-recipe Gantt chart.

The data structures utilized to encapsulate the information for the system and manufacturing process states are very important elements for the application. Job orders along with the accompanying process plans are organized in a data structure named "recipe" composed by an array of the pending steps required for the job order's completion and an array containing the completed steps. A recipe step is defined as a combination of a manufacturing process and a workstation name. The physical carriers simultaneously

serve as both the platform for transferring products and the base plate for executing all manufacturing processes. As a result, in-process items are modeled through the carrier data structure. Each digital carrier is defined by its current location, the linked process plan, an ID that also corresponds to a unique position in the loading cabinet, and additional attributes that describe its activity status (e.g., is enabled in the process queue, or is currently being processed). The number of digital carriers that can be concurrently handled by the CCS application for manufacturing products of the bioMEMS parts family matches the physical capacity of the cell. Appendix A provides an example, beginning with the definition of a process plan and proceeding with the carrier state array.

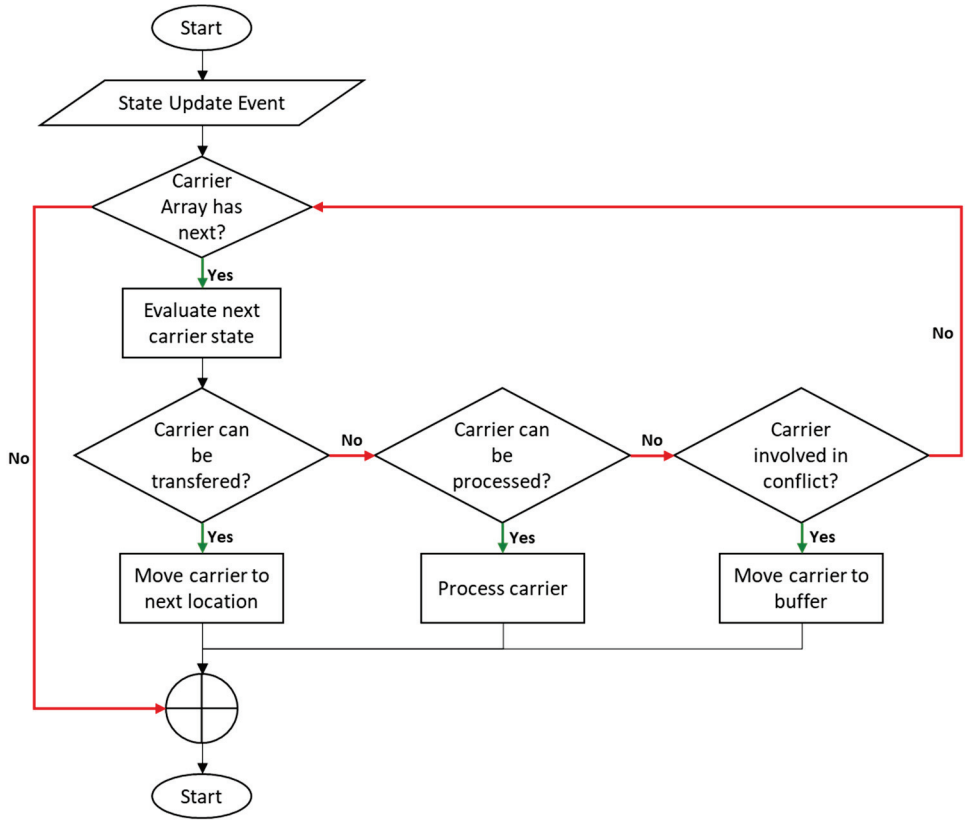


Figure 8. Task scheduling workflow.

4.2.5. Front-End

The graphical user interface of the application is equally important to the main process since it provides all the controls and display capabilities to the operator of the FMC. The front-end has been developed using React^{TM} [38], a JavaScript library for building user interfaces with reusable high-level components in a declarative style. Figure 9 shows the functionality of the individual components and the structure of the application.

A typical snapshot of the CCS application GUI is shown in Figure 10. The normal workflow for the operator of the FMC after initializing the workstations and loading the cabinet with empty carriers would be to get to the control PC and perform the following steps using the GUI:

1. load "digital carriers" corresponding to the physical carriers in the loading cabinet
2. assign process plans to them
3. activate scheduling
4. monitor production progress and system status
5. [optionally] create/update/delete recipes (process plans)
6. [optionally] update list of supported sub-recipes for a given workstation

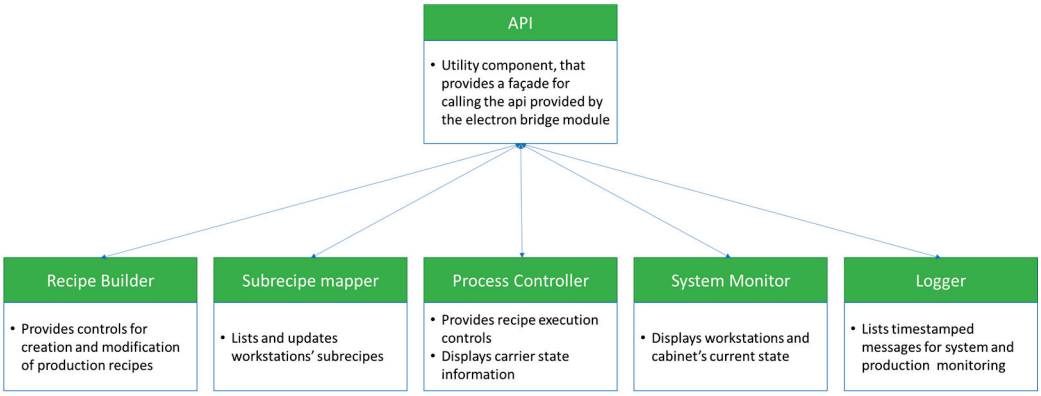


Figure 9. Front-end components.

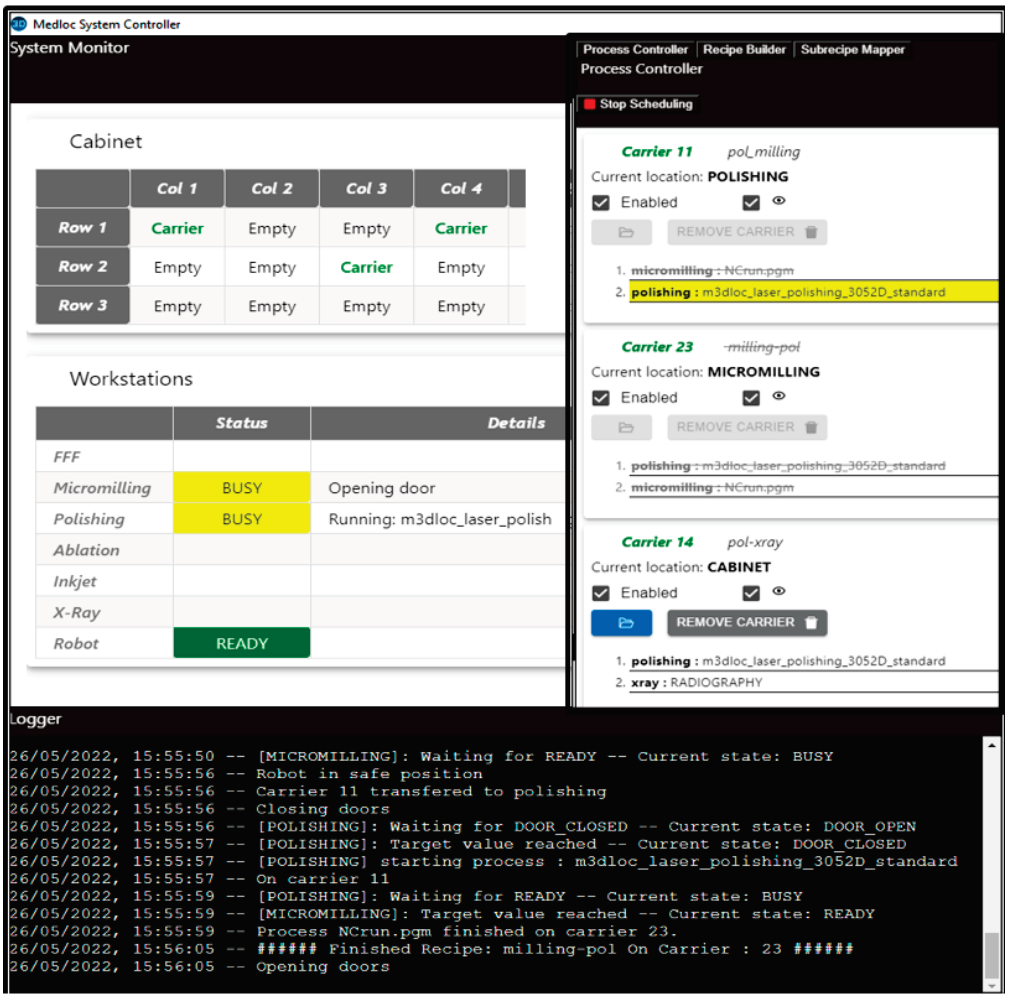


Figure 10. CCS user interface.

The front-end components provide the operator with the means to perform the mentioned activities. The system monitor plays a dual role, as it displays information regarding the state of the workstations and of the robot along with the available and occupied positions of the loading cabinet. The interactive digital cabinet is used by the operator of the FMC to insert or remove carriers and attach process plans to them.

The Recipe Builder component facilitates the creation and modification of process plans. Processes grouped per workstation can be added as steps in the production sequence via a dropdown menu containing the valid options. A closely related component is the

Sub-recipe Mapper, which allows the operator to update the list of supported processes per workstation.

The Process Controller component provides the means for the user to manipulate the production workflow. Specifically, the user can enable or disable the FMC operation and scheduling activities and even selectively deactivate carriers from taking part in the production process.

The Logger provides timestamped messages concerning the system state. The log messages are also saved locally in the control PC and are available for postprocessing or troubleshooting.

5. Implementation and Testing

The integration of the workstations and subsystems has been performed in phases and in parallel with the development of the CCS application. Creating different modules for each workstation that would be added to the cell simplified CCS development and deployment, while a strict procedure for testing had been enforced.

Software services for simulating communication and interaction with each workstation were developed and made available on the CCS development computer, allowing offline testing for each new feature of the CCS. The main functionality of these services is to act as finite state-machines implementing an identical interface with the corresponding workstations in terms of the supported communication protocol, the available methods, and the request-response messages. Process mechanics simulation is not supported; however, when these standalone services are deployed as a group, they effectively provide an FMC mock-up that can be used for offline testing new CCS features. This testing capability has been extensively exploited when CCS modules were added to support integration of a new workstation as well as for the development and evaluation of the scheduling strategy.

In the following section, the results for three scheduling strategy variations are demonstrated, essentially constituting a further testing process following basic functionality testing. Each variation is tested using the FMC mock-up to simulate four production scenarios.

Scenarios 1–3 simulate the production of identical items (15 items of a 5-step process plan) with varying probability of deadlock occurrence. Scenario 4 simulates the concurrent execution of 3 different process plans, each consisting of 5 items assigned to corresponding carriers (see Table 3).

Table 3. Simulation scenarios definition.

Scenario	Deadlock Probability	Items	Workstation Sequence	Single Item Production Time (s)	Sequential Production Time (s)
1	Low	15	A, I, M, P, X	335	5025
2	Moderate	15	P, A, X, A, M	330	4950
3	High	15	P, X, P, X, A	270	4050
4	Recipe 1: Low	5	A, I, M, P, X	335	5000
	Recipe 2: Moderate	5	P, A, X, A, M	330	
	Recipe 3: Low	5	X, P, M, I, A	335	

Note that for the purpose of this simulation, each process that is executed on a workstation has a known and fixed duration and so has every carrier transfer (see Table 4). Different production sequences have been used in the testing scenarios, corresponding to different product types in the FMC.

The following normalized key performance indicators were introduced to evaluate scheduling strategies:

- Production time saving = 1 − (Batch completion time/Sequential production time)
- WS utilization = Total time loaded with carrier/total processing time
- Robot workload = 1 − (Actual count/minimum number of carrier transfers for batch completion)

Table 4. Scenario shared elements.

Processing	
Workstation	**Process Time (s)**
Ablation (A)	60
Inkjet (I)	65
Micro-milling (M)	90
Polishing (P)	50
X-ray (X)	40
Transferring	
Min count of transfer carriers	90
Min total transfer duration (s)	450

The fundamental scheduling workflow is common to all the variations compared. Essentially, all schedulers, upon invocation, search the carrier state for (1) possible transfer operation, (2) possible process operation, and (3) conflict condition. If there is a carrier matching the search criteria, then the scheduler notifies the machine operator to perform the pertinent task (i.e., a carrier transfer or a process operation at a workstation).

The first scheduling variation considers as possible transfer operations only those that lead to advancing the progress of a process plan. To decide whether a transfer operation will be requested, the algorithm iterates over the carrier state array, and for each carrier, it checks if the workstation indicated by the next step of the process plan is available. If the workstation is available or there are no steps left in the process plan, a transfer operation is requested to either move the carrier to the next workstation or to the cabinet as a finished product. If the robot is not available, the scheduler will skip the search for transfer operations. A workstation process operation is considered to be possible when the current location of a carrier matches the next step of the process plan and the workstation is ready to perform the process. A conflict condition arises when a carrier is involved in a deadlock as described in Section 4.2.4 and it is resolved by moving the carrier to the buffer.

The second scheduling strategy differs from the first only in searching for possible carrier transfers. In this case, the carrier state array is scanned, and a possible transfer operation is requested if the carrier is currently located in an idle workstation (having finished the process operation). If the workstation indicated by the next step of the carrier's recipe is available, then the carrier will be moved there; otherwise, it will be moved to the cabinet. The purpose of this variation is to improve the availability of workstations by freeing them from carriers unnecessarily occupying their production area. However, this increases the robot workload since additional transfer motions are requested to move carriers back and forth with respect to the cabinet.

The third scheduler is similar to the first one in terms of searching for possible carrier transfers and possible process operations, but it takes an alternative approach in considering conflict conditions. For each carrier, it searches the rest of the carrier state array to find if the location of the current carrier is the next recipe location for any of the others, essentially blocking the latter's production progress. If at the same time the carrier is also blocked from moving to the workstation indicated by its next recipe step, then the carrier is involved in a conflict condition that is resolved by moving it back to the cabinet.

Numerical results of simulation runs for the three schedulers are shown in Appendix B. Characteristic performance metrics for the three schedulers in all four scenarios are shown in Figure 11.

The primary performance indicator for the schedulers is the production time-saving index since it correlates with throughput. Schedulers B and C outperform scheduler A in scenarios 2 and 4. Scheduler B yields the best results in terms of workstation availability, having the lowest workstation utilization index but achieving the highest use of the robotic transport system. Scheduler A achieves minimal use of the transport system. Based on the simulation results, scheduler C has been chosen for the application deployment in

the FMC due to its promising performance without imposing excessive workload on the transport system.

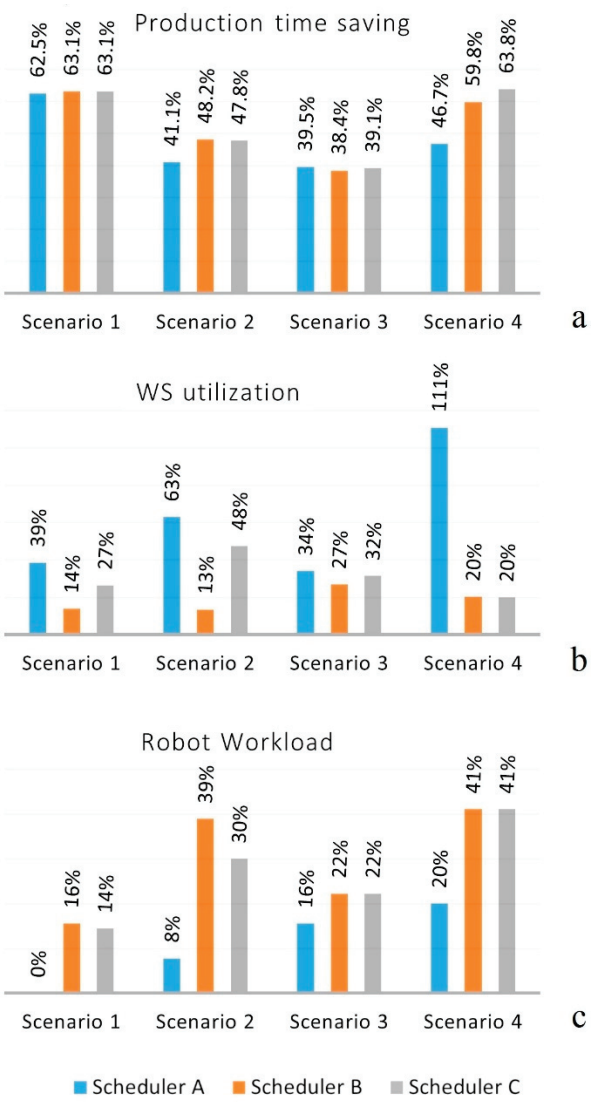


Figure 11. Performance metrics comparison of the schedulers in all scenarios. (**a**) Production time saving. (**b**) Workstation utilization. (**c**) Robot workload.

The offline testing capability was extensively utilized to validate the correct sequence of RPCs between the application and mock-up FMC, taking into account the states of the workstations at any point in time. After completion of offline testing and on-site deployment of the application, an additional procedure was put in place to facilitate the integration of workstations. In particular, after the initial commissioning of a workstation that would ensure that the system is functional in stand-alone mode and in compliance with the safety and operational requirements of the FMC, the integration phase should begin.

Initially, the RPC methods exposed by the workstation for its automatic operation should be individually tested by an external application. The subsequent test should involve the execution of demo process plans, involving only the workstation and the transport system, to uncover any unforeseen errors under real operating conditions. The final test entails the execution of a process plan involving all the workstations of the FMC. Upon the successful completion of these three tests, the workstation can be considered an integral part of the FMC.

6. Conclusions and Future Work

FMC setup variations render the development of CCS applications a non-standard, complicated procedure. Using the proposed architecture as a guide for implementing critical parts of the application has significantly simplified development. The generic interfaces for the workstations and the handling system have successfully allowed addressing in a uniform manner systems that are characterized by very different principles of operation and functionality.

The structure of the controller, consisting of distinct Scheduler, Supervisor, and Machine Operator modules, enabled testing various implementation strategies for each one of them. This is especially important for the Scheduler, as it is a key element in the overall performance of the manufacturing cell.

The client modules effectively create an abstraction layer that highly promotes interoperability and flexibility. The underlying communication has been implemented by a variety of technologies and protocols, mainly gRPC, RESTful API, and TCP/IP service.

Offline testing using mock workstation server applications proved to be of major importance. It significantly accelerated integration and testing of the various workstations upon their commissioning and reduced probable software-related failures at early stages. Nonetheless, extensive testing under realistic conditions remained essential, as it is impossible to anticipate all potential failure causes in complex systems consisting mainly of heterogeneous equipment.

The proposed cell controller architecture offers a template for FMC process control and monitoring applications. However, each implementation should consider the particular characteristics of the use case and could involve additional effort and modifications. Furthermore, considerable technical expertise will be required for efficient implementation as well as maintaining high integration flexibility standards. Finally, there are some limitations regarding the use of open-source technologies in terms of long-term support and code maintenance.

The next step in further development is to exploit a microservices-style architecture so that the main control process and the workstation clients comprise distinct services. This decoupling will enable workstation services to communicate with other services, freeing them from reliance on the main control process. Additionally, services for system diagnostics or predictive maintenance can be developed and integrated in a modular fashion. The ultimate objective is to provide a structured solution for building highly customizable, performant, and scalable applications to support FMC operations.

Author Contributions: Conceptualization, E.T., G.P., P.B. and G.-C.V.; methodology, E.T. and G.-C.V.; software, E.T.; validation, E.T. and G.P.; writing—original draft preparation, E.T.; writing—review and editing, E.T., P.B. and G.-C.V.; supervision, P.B. and G.-C.V.; funding acquisition, G.-C.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by European Union's Horizon 2020 research and innovation programme, M3DLoC (Additive Manufacturing of 3D Microfluidic MEMS for Lab-on-a-Chip applications) under grant Agreement No 760662.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

```
// Recipe step
recipe_step = {
    workstation: "micromilling",
    subrecipe_name: "drilling_operation_1"
}

// Recipe composed from multiple steps
recipe = {
    name: "biomems_device_1",
    steps: [
        {
            workstation: "micromilling",
            subrecipe_name: "drilling_pr_1"
        }, ...
    ],
    completed_steps: [
        {
            workstation: "fff",
            subrecipe_name: "base_printjob"
        }
    ]
}

// Carrier state array
carrier_state = [{
    id: 1,
    current_location: "micro-milling"
    recipe: {
        name: "device_1",
        completed_steps: [...]
        remaining_steps: [next_step, ...]
    },
    is_enabled: true,
    is_being_processed: false
}, ...]
```

Figure A1. System state example.

Appendix B

Table B1. Net process time for batch production in seconds.

	Scenario			
	1	**2**	**3**	**4**
Workstation				
A	900	1800	900	1200
I	975			650
M	1350	1350		1350
P	750	750	1500	750
X	600	600	1200	600

Table B2. Occupied by carrier time (s) results for simulation scenarios 1 to 4.

Scenario	**1**			**2**			**3**			**4**		
Scheduler	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
Workstation												
A	1026	1017	982	2333	2038	2044	996	1137	995	2225	1423	1366
I	1091	1107	1075							2175	746	747
M	1466	1440	1447	1475	1480	1441				1533	1456	1472
P	1395	903	969	2391	884	2052	1923	1984	1823	2132	1087	973
X	1395	745	1315	1139	704	1105	1907	1447	1916	1525	762	900
Carrier Transfers	90	104	103	97	125	117	104	110	110	108	127	127
Completion time (s)	1884	1856	**1852**	2918	**2566**	2583	**2452**	2495	2466	2663	2011	**1812**

The bold corresponds to the minimum time for each of the Scenarios.

Appendix C

Table C1. Control PC specifications.

CPU	Ryzen 5 5600X (AM4/3.7 GHz/32 MB)
RAM	32 GB 3200 MHz DDR4
GPU	GeForce GT 1030 2 GB GDDR5
OS	Windows 10 Pro 64-bit

Table C2. Raspberry Pi specifications.

CPU	Quad core Cortex-A72 (ARM v8) 64-bit SoC @ 1.8 GHz
RAM	8 GB LPDDR4-3200 SDRAM
OS	Raspberry Pi OS (former Raspbian)

References

1. Wang, Y.; Ma, H.-S.; Yang, J.-H.; Wang, K.-S. Industry 4.0: A way from mass customization to mass personalization production. *Adv. Manuf.* **2017**, *5*, 311–320. [CrossRef]
2. Esmaeilian, B.; Behdad, S.; Wang, B. The evolution and future of manufacturing: A review. *J. Manuf. Syst.* **2016**, *39*, 79–100. [CrossRef]
3. Zeid, A.; Sundaram, S.; Moghaddam, M.; Kamarthi, S.; Marion, T. Interoperability in Smart Manufacturing: Research Challenges. *Machines* **2019**, *7*, 21. [CrossRef]
4. Qin, J.; Liu, Y.; Grosvenor, R. A Categorical Framework of Manufacturing for Industry 4.0 and Beyond. *Procedia CIRP* **2016**, *52*, 173–178. [CrossRef]
5. Potdar, P.K.; Routroy, S.; Behera, A. Agile manufacturing: A systematic review of literature and implications for future research. *Benchmarking Int. J.* **2017**, *24*, 2022–2048. [CrossRef]
6. Toro, R.; Correa, J.E.; Ferreira, P.M. A Cloud-Monitoring Service for Manufacturing Environments. *Procedia Manuf.* **2018**, *26*, 1330–1339. [CrossRef]
7. Hu, L.; Nguyen, N.-T.; Tao, W.; Leu, M.C.; Liu, X.F.; Shahriar, R.; Al Sunny, S.M.N. Modeling of Cloud-Based Digital Twins for Smart Manufacturing with MT Connect. *Procedia Manuf.* **2018**, *26*, 1193–1203. [CrossRef]
8. Javaid, M.; Haleem, A.; Singh, R.P.; Suman, R. Enabling flexible manufacturing system (FMS) through the applications of industry 4.0 technologies. *Internet Things Cyber-Physical Syst.* **2022**, *2*, 49–62. [CrossRef]
9. Sisinni, E.; Saifullah, A.; Han, S.; Jennehag, U.; Gidlund, M. Industrial Internet of Things: Challenges, Opportunities, and Directions. *IEEE Trans. Ind. Informatics* **2018**, *14*, 4724–4734. [CrossRef]
10. Ameri, F.; Sormaz, D.; Psarommatis, F.; Kiritsis, D. Industrial ontologies for interoperability in agile and resilient manufacturing. *Int. J. Prod. Res.* **2021**, *60*, 420–441. [CrossRef]
11. Weyrich, M.; Ebert, C. Reference Architectures for the Internet of Things. *IEEE Softw.* **2015**, *33*, 112–116. [CrossRef]
12. VDI/VDE-Gesellschaft (2015) Reference Architecture Model Industrie 4.0 (RAMI4.0)—Status Report. Available online: https://www.zvei.org/fileadmin/user_upload/Presse_und_Medien/Publikationen/2016/januar/GMA_Status_Report__Reference_Archtitecture_Model_Industrie_4.0__RAMI_4.0_/GMA-Status-Report-RAMI-40-July-2015.pdf (accessed on 19 December 2023).
13. Industry IoT Consortium. The Industrial Internet Reference Architecture. 2022. Available online: https://www.iiconsortium.org/wp-content/uploads/sites/2/2022/11/IIRA-v1.10.pdf (accessed on 19 December 2023).
14. Sino-German Industrie 4.0/Intelligent Manufacturing Standardisation Sub-Working Group. Alignment Report for Reference Architectural Model for Industrie 4.0/Intelligent Manufacturing System Architecture. Berlin. 2018. Available online: https://www.plattform-i40.de/IP/Redaktion/EN/Downloads/Publikation/hm-2018-manufacturing.html (accessed on 19 December 2023).
15. Pisching, M.A.; Pessoa, M.A.O.; Junqueira, F.; dos Santos Filho, D.J.; Miyagi, P.E. An architecture based on RAMI 4.0 to discover equipment to process operations required by products. *Comput. Ind. Eng.* **2018**, *125*, 574–591. [CrossRef]
16. Baptista, L.F.; Barata, J. Piloting Industry 4.0 in SMEs with RAMI 4.0: An enterprise architecture approach. *Procedia Comput. Sci.* **2021**, *192*, 2826–2835. [CrossRef]
17. Drahos, P.; Kucera, E.; Haffner, O.; Klimo, I. Trends in industrial communication and OPC UA. In Proceedings of the 29th International Conference on Cybernetics & Informatics (K&I), Lazy pod Makytou, Slovakia, 31 January–3 February 2018; pp. 1–5. [CrossRef]
18. MTConnect Institute (2023) MTConnect ® Standard Part 1.0-Fundamentals Version 2.1.0. Available online: https://docs.mtconnect.org/MBSD_MTConnect_Part_1_2-1-0.pdf (accessed on 19 December 2023).

19. Lu, Y.; Witherell, P.; Jones, A. Standard connections for IIoT empowered smart manufacturing. *Manuf. Lett.* **2020**, *26*, 17–20. [CrossRef] [PubMed]
20. González, I.; Calderón, A.J.; Figueiredo, J.; Sousa, J.M.C. A Literature Survey on Open Platform Communications (OPC) Applied to Advanced Industrial Environments. *Electronics* **2019**, *8*, 510. [CrossRef]
21. Weckenborg, C.; Schumacher, P.; Thies, C.; Spengler, T.S. Flexibility in manufacturing system design: A review of recent approaches from Operations Research. *Eur. J. Oper. Res.* **2024**, *315*, 413–441. [CrossRef]
22. Ashraf, M.; Hasan, F. Configuration selection for a reconfigurable manufacturing flow line involving part production with operation constraints. *Int. J. Adv. Manuf. Technol.* **2018**, *98*, 2137–2156. [CrossRef]
23. Giret, A.; Garcia, E.; Botti, V. An engineering framework for Service-Oriented Intelligent Manufacturing Systems. *Comput. Ind.* **2016**, *81*, 116–127. [CrossRef]
24. Elgendy, A.; Yan, J.; Zhang, M. A Parallel distributed genetic algorithm using Apache Spark for flexible scheduling of multitasks in a cloud manufacturing environment. *Int. J. Comput. Integr. Manuf.* **2023**, *37*, 652–667. [CrossRef]
25. An, Y.; Chen, X.; Li, Y.; Zhang, J.; Jiang, J. Flexible job-shop scheduling and heterogeneous repairman assignment with maintenance time window and employee timetable constraints. *Expert Syst. Appl.* **2021**, *186*, 115693. [CrossRef]
26. Zambetti, M.; Khan, M.A.; Pinto, R.; Wuest, T. Enabling servitization by retrofitting legacy equipment for Industry 4.0 applications: Benefits and barriers for OEMs. *Procedia Manuf.* **2020**, *48*, 1047–1053. [CrossRef]
27. Al-Jaroodi, J.; Mohamed, N.; Jawhar, I. A service-oriented middleware framework for manufacturing industry 4.0. *SIGBED* **2018**, *15*, 29–36. [CrossRef]
28. Kannoth, S.; Schnicke, F.; Antonino, P.O. Enabling Industry 4.0 Communication Protocol Interoperability: An OPC UA Case Study. In Proceedings of the ECBS 2021: 7th Conference on the Engineering of Computer Based Systems, Novi Sad, Serbia, 26–27 May 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 1–9.
29. Chiraga, N.; Walker, A.; Bright, G. Middleware communication system for fixture based flexible manufacturing systems. In Proceedings of the 13th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2016), Lisbon, Portugal, 29–31 July 2016; SciTePress: Lisbon, Portugal, 2016; pp. 146–152.
30. Melo, P.F.S.; Godoy, E.P.; Ferrari, P.; Sisinni, E. Open Source Control Device for Industry 4.0 Based on RAMI 4.0. *Electronics* **2021**, *10*, 869. [CrossRef]
31. Ayele, E.D.; Gavriel, S.; Gonzalez, J.F.; Teeuw, W.B.; Philimis, P.; Gillani, G. Emerging Industrial Internet of Things Open-Source Platforms and Applications in Diverse Sectors. *Telecom* **2024**, *5*, 369–399. [CrossRef]
32. Omidi, S.A.; Baig, M.J.A.; Iqbal, M.T. Design and Implementation of Node-Red Based Open-Source SCADA Architecture for a Hybrid Power System. *Energies* **2023**, *16*, 2092. [CrossRef]
33. Folgado, F.J.; Calderón, D.; González, I.; Calderón, A.J. Review of Industry 4.0 from the Perspective of Automation and Supervision Systems: Definitions, Architectures and Recent Trends. *Electronics* **2024**, *13*, 782. [CrossRef]
34. Introduction to gRPC. Available online: https://grpc.io/docs/what-is-grpc/introduction/ (accessed on 19 December 2023).
35. Papazetis, G.; Tzimas, E.; Benardos, P.; Vosniakos, G.-C. Development of a robot-based handling system for a high precision manufacturing cell. In Proceedings of the 32nd International Conference on Flexible Automation and Intelligent Manufacturing (FAIM2023), Porto, Portugal, 18–22 June 2023; Lecture Notes in Mechanical Engineering Series. Springer: Porto, Portugal, 2023; pp. 29–36.
36. JavaScript—The Event Loop. Available online: https://developer.mozilla.org/en-US/docs/Web/JavaScript/EventLoop (accessed on 19 December 2023).
37. Electron Docs: What Is Electron. Available online: https://www.electronjs.org/docs/latest/ (accessed on 19 December 2023).
38. React: The Library for Web and Native User Interfaces. Available online: https://react.dev/ (accessed on 19 December 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Design of Connector Assembly Equipment for the Automotive Industry

Pedro M. P. Curralo [1], Raul D. S. G. Campilho [1,2,*], Joaquim A. P. Pereira [1] and Francisco J. G. Silva [1,2]

[1] CIDEM, ISEP—School of Engineering, Polytechnic of Porto, R. Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal; 1160889@isep.ipp.pt (P.M.P.C.); joaquim.pinto@wrk.pt (J.A.P.P.); fgs@isep.ipp.pt (F.J.G.S.)

[2] INEGI—Institute of Science and Innovation in Mechanical and Industrial Engineering, Pólo FEUP, Rua Dr. Roberto Frias, 400, 4200-465 Porto, Portugal

* Correspondence: raulcampilho@gmail.com; Tel.: +351-939526892

Abstract: The automotive industry is one of the most demanding sectors of all manufacturing industries due to its competitiveness. It is necessary to innovate through the implementation of automated and robotic equipment, leading to cycle time and labor cost reduction. This work aims to design semi-automatic equipment to assemble cabling connectors used in the automotive sector, replacing a manual process currently taking place in an automotive components company. In the proposed equipment, the operator places a connector in the equipment, and the components (pins and seals) are automatically inserted. A vision sensor with artificial intelligence then confirms the correct application. The equipment operation defined as Finite Element Method (FEM) was applied for structural verification; the materials and fabrication processes were detailed; the associated costs were calculated, and the equipment subsets were validated. The design was successfully accomplished, and the imposed requirements were fulfilled, with significant advantages over the current process, providing new knowledge on how semi-automatic systems can be deployed to enhance the productivity and quality of manufacturing processes. The design principles and insights gained from this work can be applied to other automation challenges, particularly where manual processes need to be replaced by more efficient semi-automatic or automatic systems. The modularity of the overall solution and the design concepts of the component inserter, component feeder, and assembly process allow for its use in different assembly scenarios beyond the automotive sector, such as electronics or aerospace, providing a contribution to increased competitiveness and survival in the global market.

Keywords: automotive industry; mechanical design; process automation; connector assembly; feeder system; Finite Element Method

1. Introduction

The automotive industry plays a leading role in the world economy. This industry presents itself in the market as one of the most demanding sectors of all manufacturing industries due to its competitiveness, being constantly pressured to meet tight deadlines and high-quality requirements [1]. Thus, it is necessary to upgrade processes through the implementation of automated and robotic equipment [2–4]. With automation, it is possible to obtain less waste and more quality and productivity [5], leading to shorter cycle times and smaller labour costs. Currently, the priority of companies is the development or acquisition of machines that use as few operators as possible, and there is a tendency for full automation in the future [6]. In automated equipment, the processes are managed automatically and without the need for significant human intervention, which increases efficiency [7,8]. Thus, fatigue and occupational diseases tend to be excluded as they are factors that affect productivity [9]. With automation, the setup time is also reduced, and

assembly accuracy is guaranteed. Automation can be divided into three types: fixed; flexible; and programmable [10]. Fixed automation consists of using a machine that repeats the same automatic process without any flexibility [10,11]. In flexible automation, a computer controls product combinations and their reconfiguration in short adjustment times [10,12]. Programmable automation is flexible in the assembly or manufacturing process, but it is necessary to reprogram the automated action for new configurations [10,13]. In recent years, the introduction of more advanced automation-related technology led to a further reduction in labour and an increase in productivity [14]. With intelligent feeder systems, the automotive industry can group different vehicle models on the same production line. This principle can also be used on assembly lines for small automotive components. On the other hand, the demand for robots has significantly increased because they are robust, fast, flexible, and highly accurate [15]. These are more economical than human labour as they are produced in greater quantity and with higher quality. Nowadays, robotics is applied all over the world [16]. Although robots are expensive, they have a lower cost per component, high efficiency, and high production [10]. Nowadays, most manufacturing processes require robotised operations, and the automotive industry is currently the most robotised industry. More and more manufacturing processes are carried out using automated machines, and robotic solutions are progressively used to increase production and assist the operator [17,18].

In the field of engineering design, the systematic development of products and equipment follows structured methodologies aimed at optimising functionality, efficiency, and cost-effectiveness [19]. According to Pahl et al. [20], developers of the Systematic Design Approach, engineering design processes typically involve stages such as conceptual design, embodiment design, and detailed design, each providing a framework for turning abstract ideas into practical solutions. This approach is used broadly in mechanical engineering and product design, especially for complex, multi-component systems. One key area of focus in modern engineering design is the integration of automation to enhance production efficiency and reduce human error, which constitutes the Product Development Process developed by Ulrich and Eppinger [21]. This approach includes the steps of planning, concept development, system-level design, detail design, testing and refinement, and production ramp-up. The V-Model (Verification and Validation) was developed by Forsberg et al. [22] and applied to the context of systems engineering, providing a detailed view of verification and validation processes. This is a model primarily used in systems engineering, where each design phase is validated and verified in a parallel step. It is often used in highly regulated industries like aerospace and automotive engineering. Stuart Pugh's Total Design approach [23] integrates the design process with business and market requirements. It uses methods like the Pugh Decision Matrix for concept evaluation, helping designers systematically compare multiple design solutions. It is used in product engineering, particularly in industries where both functional and market requirements must be balanced. The Theory of Inventive Problem Solving (TRIZ), developed by Altshuller [24], is a problem-solving methodology that focuses on resolving contradictions within the design by using inventive principles derived from patent analysis. It encourages innovation by highlighting ways to overcome common design challenges. It is especially useful in industries that require high innovation (e.g., electronics, automotive, and aerospace). The Axiomatic Design [25] focuses on creating systems where the design parameters (physical solutions) map clearly and independently to the functional requirements, reducing design complexity. This approach is used in manufacturing systems, software development, and product design. The Design Science Research (DSR) methodology is another structured methodology that, though it originates primarily from the fields of information systems and management science [26], has been increasingly adopted in engineering design, especially in research contexts [27]. This methodology focuses on creating and evaluating artefacts (like models, systems, or processes) to solve identified problems. It combines scientific rigour (research-based problem solving) with practical relevance (ensuring solutions have real-world applicability). Another influential approach is the Design Research Methodology (DRM) proposed by

Blessing and Chakrabarti [28]. DRM offers a comprehensive framework specifically tailored for design research aimed at understanding and improving design processes and outcomes. Unlike other methodologies, DRM places a strong emphasis on the research dimension of design projects, particularly the systematic investigation of design processes to generate knowledge that can improve both design practice and theory. By systematically guiding the research through descriptive and prescriptive phases, DRM ensures that the outcomes of the design are both scientifically validated and practically applicable. In addition to these established methodologies, several more recent developments have gained prominence in the field of engineering design. These approaches address modern challenges in product development, such as reducing complexity, improving reliability, and enabling automation, including the following:

- **Design for Manufacturing and Assembly (DfMA)** is a concept that integrates design and manufacturing considerations early in the design process to simplify production and reduce costs. The main goal of DfMA is to design products that are easier and more cost-effective to manufacture and assemble, minimising parts and simplifying assembly operations [29]. This approach ensures that the design is optimised not only for functionality but also for the efficient production;
- **Graph-based design languages** are an approach to formalising design processes using mathematical graph structures to represent the relationships between components, systems, and functions. This method provides a visual and algorithmic means to explore design alternatives and optimise complex systems, enabling designers to handle intricate dependencies and constraints more effectively [30];
- **Fault-tolerant design** is a critical approach used to ensure that systems continue to operate reliably even when some of their components fail. In the context of automated systems, this approach is essential for maintaining high levels of uptime and minimising the risk of costly production delays. Fault-tolerant design involves identifying potential failure points and incorporating redundancy, error detection, and recovery mechanisms into the system [31].

Automated and robotised systems have become increasingly complex, and both implementations and use of protocols are described in the scientific literature [32–35], eventually applying a specific engineering design approach. Costa et al. [36] solved quality problems in the transmission system of automotive windshield wipers using a design approach resembling the DSR methodology in the applied steps. The previous system was semi-automatic and could not detect and reject components during the process prior to assembly, leading to the production of defective parts. For optimisation, the equipment and the product to be manufactured were analysed. Subsequently, the best solutions for the detected faults were discussed and chosen. A fully automated system capable of detecting faulty components was then implemented. The equipment now has a 100% quality control, i.e., it no longer manufactures defective products. Furthermore, production rates remained within automotive industry standards, and productivity increased by 19%. Figueiredo et al. [37] added an automatic coating stripping operation onto an existing machine to produce control cables for the automotive industry, using three simple steps as a design approach: a study of the manual stripping system, a study of the machine cycle, and a redesign of the mechanisms to comply with the specification. The new stripping operation removes the coating from the cable edges, allowing the injection of zamak terminals without die-casting defects. The equipment's production cycle was analysed, and an automated solution based on multiple blades did not affect productivity. The proposed device produces cables up to 2.5 m long, stripped on both sides, with a cycle time of 7 s. Santos et al. [38] presented a new solution to transport conduits used to protect control cables for the automotive industry along an automatic production line in which other operations were applied to the control cables, such as cutting, deburring, and punching of terminals at the ends. This work was divided into three steps: establishing the main requirements for the new concept; finding a solution for every identified problem; and combining all the solutions found into one system that surpassed the identified problems. As a result, a simple and economical solution to be

implemented in the line was studied and developed, which had a much higher reliability than the current mechanism. After implementation, the setup time was reduced by 97%, and system stoppages due to problems dramatically decreased. In the scope of robotics, Pereira et al. [39] designed an automatic robotic cell to insert clips into two different car parts after their fabrication by an injection process. A self-developed design approach was described, including elements of both DSR and V-Model approaches, as follows: analysis of requirements; information gathering; equipment division in sub-assemblies; definition of principles for each sub-assembly; preliminary design with possible solution and solution selection using a Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis; design using Computer-Aided Design (CAD) and the FEM; and validation by prototype construction and workflow simulation. Resulting from this work, a six-axis robot with a special tool to hold the clips was implemented. An automatic clip-loading system was designed, which loaded the clips in the tool and delivered them to the robot while the empty tool was left by the robot to charge with clips. The equipment was fabricated, validated in a real scenario, including meeting the cycle time imposed by the client, and delivered for factory implementation. Michalos et al. [40] addressed the implementation of a robotic system for a collaborative human–robot assembly process. An active workflow design methodology was proposed by the authors, starting with assembly task identification, followed by operation sequence definition, resource alignment, layout design, collaborative workflow generation, safety concept definition, engineering system development, and finally, risk assessment. The steps were not tight since feedback in each stage could lead to modifications in a previous one and recalculation. The proposed solution aimed to combine heavy-duty industrial robots with human intervention in a fenceless workspace. Augmented reality glasses and smartwatches enabled the closing of the robot–human communication loop. The system was validated by applying the concept to a case study in the automotive industry. A safety analysis was supported by Safety Related Parts/Control Systems (SRC/CS) design approach, further corroborating the possibilities of human–robot collaborations.

This work aims to design semi-automatic equipment to assemble cabling connectors used in the automotive sector, replacing a currently manual process currently taking place in an automotive components company. In the proposed equipment, the operator places a connector in the equipment, and the components (pins and seals) are automatically inserted. A vision sensor with artificial intelligence then confirms the correct application. The equipment operation was defined; FEM was applied for structural verification; the materials and fabrication processes were detailed; the associated costs were calculated, and the equipment subsets were validated.

2. Methods

2.1. Design Approach

The DSR design methodology focuses on the creation and evaluation of artefacts, such as models, processes, and systems, to address real-world problems while simultaneously contributing to academic knowledge [27]. The DSR approach follows a structured process in which researchers and practitioners collaborate to define the problem, design the artefact, and evaluate its effectiveness. This dual emphasis on problem-solving and knowledge generation makes DSR especially relevant for projects involving innovation and the development of new technologies, such as the semi-automatic assembly equipment presented in this work. The key stages of the DSR process applied to the design and development of the equipment are as follows:

- **Problem description**—Examination of the existing system or process, identifying inefficiencies, limitations, or areas where improvements can be made. This stage involves describing specific problems that the new design will address;
- **Goals and requirements**—Establishment of clear goals and requirements for the new solution. This includes both functional objectives, such as operational efficiency or reliability, and any constraints related to cost, materials, or technology;

- **Design and development**—Creation of a new design or concept that meets the defined objectives. This stage involves iterative design processes, where multiple concepts may be generated and refined;
- **Demonstration**—Materialisation of the developed concept by constructing a prototype or implementing a functional CAD model. This concept should be capable of demonstrating the core functions of the equipment;
- **Evaluation**—Assessing the performance of the developed solution through virtual or physical testing. This stage involves verifying whether the equipment complies with the established goals and requirements;
- **Conclusions**—Consist of the final assessment comparing the results of the equipment with the goals and requirements and drawing conclusions about the overall success of the new design.

2.2. Problem Description

The current project arises from a challenge posed by the automotive industry. A specific company needs equipment that can perform the assembly process of automotive cabling connectors in a semi-automatic way, which are currently assembled manually and pose a significant cost and human resource allocation to the company, serving as the process bottleneck for a subsystem being delivered for a vehicle manufacturer. Thus, specific equipment is necessary to overcome these limitations and increase the company's competitiveness. The mentioned connectors are part of the wiring harnesses that interconnect all the electrical parts of the vehicle. Some connector channels are selected to be connected to the ends of electrical cables, which may or may not belong to the same harness, and the ones chosen are not always the same because they depend on the function to be performed in a certain part of the vehicle. The connector assembly equipment should work semi-automatically to insert components (pins and seals) into various connectors. The function of the components is to seal/close the free channels of the connector, i.e., the channels that are not connected to the ends of electrical cables. Figure 1a shows the pins and the seal that the equipment inserts. Figure 1b shows the connector being worked on. Figure 2 shows the pins and seal inserted into the connector to assemble.

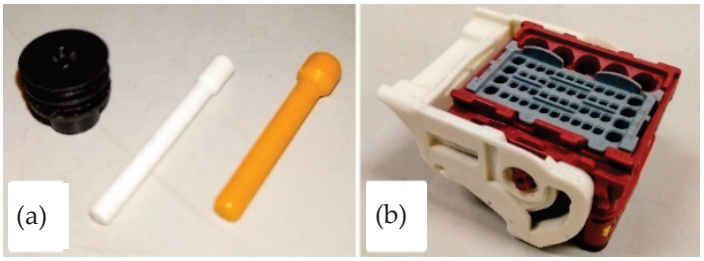


Figure 1. Seals and pins (**a**) and connector to assemble (**b**).

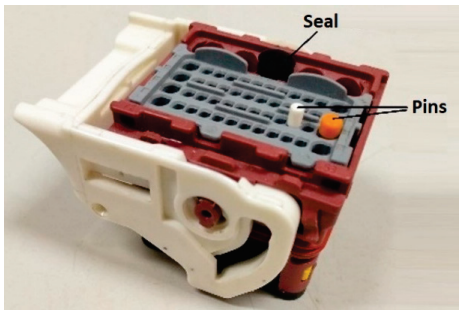


Figure 2. Pins and seal inserted into the connector to assemble.

2.3. Goals and Requirements

The main goal of this work involves designing equipment capable of performing the semi-automatic assembly of cabling connectors used in vehicles. The assembly equipment to be designed is a semi-automatic system intended for the insertion of components (pins and seals) into various types of connectors. The function of these components is to seal or close the unused channels of the connector, meaning the channels that are not connected to the ends of electrical cables.

The requirements are divided into project requirements, pin and seal assembly in the plastic connector, and pin and seal detection in the plastic connector. The project requirements, pre-defined by the client company, aim to ensure the good functionality of the equipment (Table 1). For the pin and seal assembly in the plastic connector, several assembly considerations were considered (Table 2). Pin and seal detection in the plastic connector is carried out to ensure correct insertion, which can be achieved by various methods and equipment (Table 3 shows the main considerations to be taken into account when selecting the type of detection to be implemented).

Table 1. Project requirements.

Structure and shielding	The equipment should be compact and prepared to be placed on a bench; The main structure shall have a technical aluminium profile with square section of 30 × 30 mm^2; The door of the equipment must have an aluminium profile with a square section of 30 × 30 mm^2; The structure of the door must allow for visual access to the work area. Therefore, they must be isolated with crystal compact polycarbonate.
Dimensions	The equipment must have a maximum width of 600 mm and a maximum length of 1100 mm.
Workboard	Connector placement must be performed by the operator.
Component feeding	The pins and seals must be deposited in the equipment in bulk; The equipment must have an opening to the outside that allows for the deposition of components; The feeder system must be covered by a protection that prevents the entry of unwanted parts.
Quality	The correct assembly must be verified by a vision sensor.
Autonomy	The equipment should operate continuously, with no component supply, for 2 h.
Market components	Pneumatics: SMC; sensors and vision sensor: Keyence or Contrinex; linear movement and technical profile: Bosch Rexroth; control/interface: Esaware.
Treatment	Parts susceptible to wear must receive the appropriate thermal and chemical treatment.
Electric panel	As safety margin, the electrical panel should have at least 25% free space.
Fixation	The equipment must be fixed by four support feet.
Operating sequence	The operator places a connector in the equipment, and the equipment automatically assembles the pins and seals. Finally, the equipment verifies the correct assembly of these components and releases the connector. Subsequently, the operator removes the connector, and the cycle repeats.

Table 2. Pin and seal assembly requirements.

Connector grasping	A positioning and clamping system ensures that the connector always maintains its position. This system guarantees the correct assembly of the pins/seals.
Assembly force	The assembly force depends on the type of pin/seal to be inserted and the connector geometry. This force should be based on the maximum force stated on the data sheet of the insertion cylinder used, and it should be assumed that it will not damage the pin/seal or the connector.
Insertion system	For the automatic assembly of pins and seals, an insertion system should be designed to suit their geometry and material behaviour.

Table 3. Pin and seal detection requirements.

Detection type	The choice of detection type considers the size, shape, colour, and material of the pins and seals.
Position	For reliable detection, the pins and seals should not move after assembly.
Detection spot	To ensure the correct assembly of the pins and seals, a key detection point is selected.
Detection distance	The detection distance needs to be analysed. This distance must consider a balance between the characteristics of the detector, component, and its assembly.
Safety	The sensors must be protected from impacts and dust.

3. Results

3.1. Final Equipment and Operation Sequence

The proposed equipment is a prototype designed for series production (Figure 3). The design has been developed to facilitate future modification of the workboard, which is necessary to assemble different connectors. The equipment allows for the simple replacement of the feeder system and the inserter for the assembly of different components. Additionally, the equipment is assembled using mechanically fastened connections. This equipment is semi-closed regarding the exterior, having only one door and an opening on its top part. The door allows the operator to place the workpiece on the workboard, while the opening enables the deposition of components in the feeder. With this isolation, the safety of the operators is ensured.

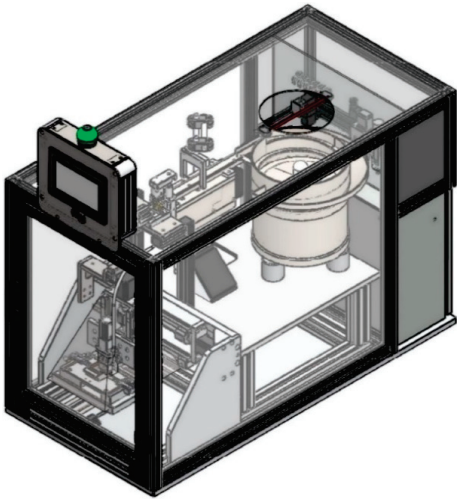


Figure 3. Isometric view of the equipment.

The operation of the final equipment is described as follows. First, the operator deposits the components in bulk through an opening in the equipment. Then, the type of connector to work on is selected using the console. The operator puts the connector into the nest of the workboard and closes the door of the equipment. Next, the inserter automatically inserts the components, which include pins and seals, into the connector according to the pre-set programming. Afterwards, the artificial vision system checks to ensure that the components are correctly inserted into the connector. Then, the equipment informs the operator whether the insertion of the components is correct and releases the connector. Finally, the operator opens the door and removes the connector from the equipment. The equipment is composed of subsets that work together to ensure optimal performance. The arrangement of these subsets was carefully evaluated for space, minimising the overall dimensions. These subsets consist of the following (Figure 4):

- Structure/shields: block access to the mechanisms to ensure the operator's safety during operation;

- Workboard: has a nest where the operator positions the connector, which blocks it in the *x* and *y*-axes, while the workboard blocks it in the *z*-axis. It detects the presence of the connector;
- Inserter: inserts the pins/seals into the connector. This system receives the pins and seals from the feeder system and inserts them into the workpiece by application of a vertical force;
- Axle systems: responsible for positioning the inserter at the insertion points of the components. The guided movements of the axles are driven by electric motors;
- Feeder system: to provide the correct orientation and separate the components (deposited in bulk);
- Artificial vision system: checks correct assembly by analysing the entire insertion area (upper face) of the connector and verifying if the pins and seals are in the pre-defined locations.

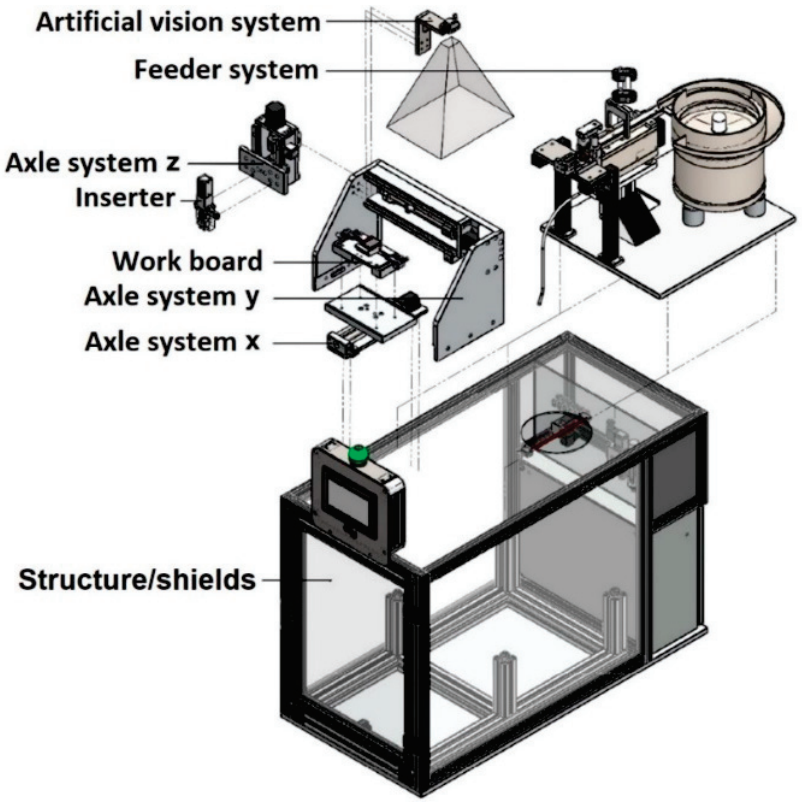


Figure 4. Layout of the equipment.

3.2. Subset Detail

The various subsets that make up the equipment are described next. These are the structure/shields, workboard, inserter, axle systems, feeder system, and artificial vision system.

3.2.1. Structure/Shields

The structure/shields of the equipment are shown in Figure 5, providing operator safety during operation. The equipment dimensions are 1080 × 700 × 530 mm^3. Isolated mechanisms with polycarbonate and Dibond® materials provide safety, with the polycarbonate enabling visibility inside the equipment. The magnetic lock and plate (Figure 6a) mechanically secure the door through magnetic forces.

A magnetic safety switch depicted in Figure 6b verifies whether the door is properly closed and transmits this information to the automation system. If the door is not properly closed, the equipment will not operate. The equipment is supported by four feet, such as

those shown in Figure 6c, to maintain stability and levelness on the workbench. A pneumatic platform provides pressurised air to power the equipment. The electric panel incorporates the necessary electrical components, and its dimensions were selected based on the size of the equipment and the expected number of electrical elements. The programmable interface device provides information to the operator on the status of the equipment and the steps to be executed, with a warning light of different colours to indicate the equipment's status. The structure supports all loads exerted by the subset and is made of technical aluminium profiles, allowing for connection flexibility for a wide variety of connectors and components. The aluminium profile door gives the operator access to the inside.

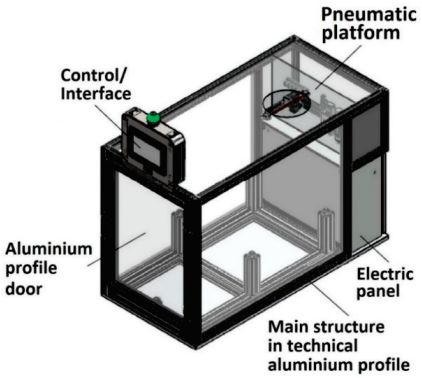
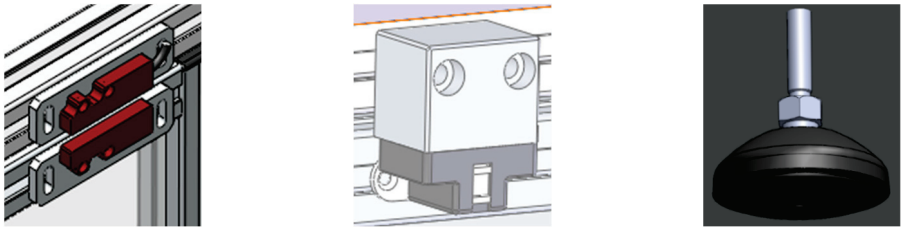


Figure 5. Layout of the structure/shields.

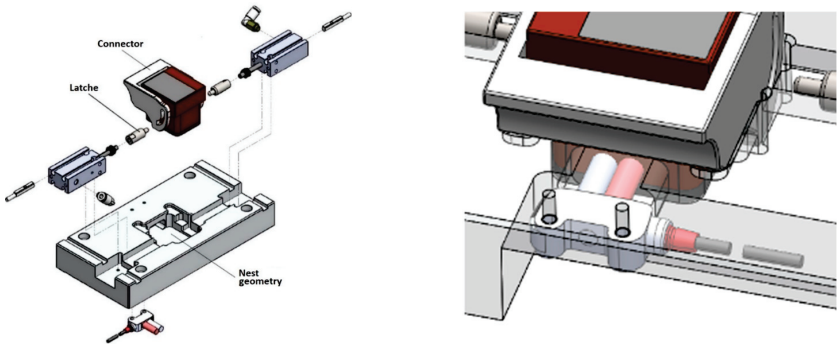


(**a**) (**b**) (**c**)

Figure 6. Magnetic lock and plate (**a**), magnetic safety switch (**b**), and support feet (**c**).

3.2.2. Workboard

The operator places the connector to assemble in a nest on the workboard (Figure 7a), which secures the connector in the x and y-axes. The upper part of the connector is left free for component insertion (pins and seals). The latches lock the connector in position on the z-axis until the insertion of components is complete. After insertion, the connector is unlocked, and the operator removes it from the nest. The short-range photoelectric sensor shown in Figure 7b verifies the presence of the connector on the workboard.



(**a**) (**b**)

Figure 7. Workboard layout (**a**) and connector detection via short-range photoelectric sensor (**b**).

3.2.3. Inserter

The inserter operates through the four steps represented in Figure 8:

1. The component is received by the inserter from the blow system, located in the feeder system;
2. The equipment places the inserter on the insertion point;
3. The grip opens, and the component descends to the insertion area (upper face) of the connector;
4. As soon as the component descends, the insertion cylinder advances, exerting a vertical force on the component and inserting it into the connector.

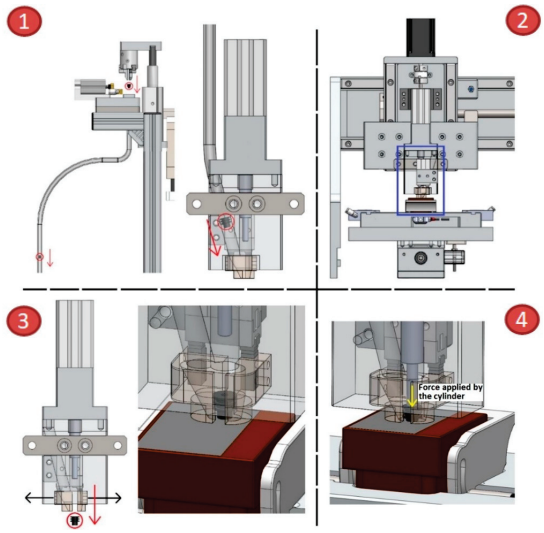


Figure 8. Stages of operation of the inserter.

The inserter, depicted in detail in Figure 9, is responsible for receiving components, such as pins or seals, from the feeder system and inserting them into the connector. The insertion process involves applying a vertical force on the component to fit it into the connector.

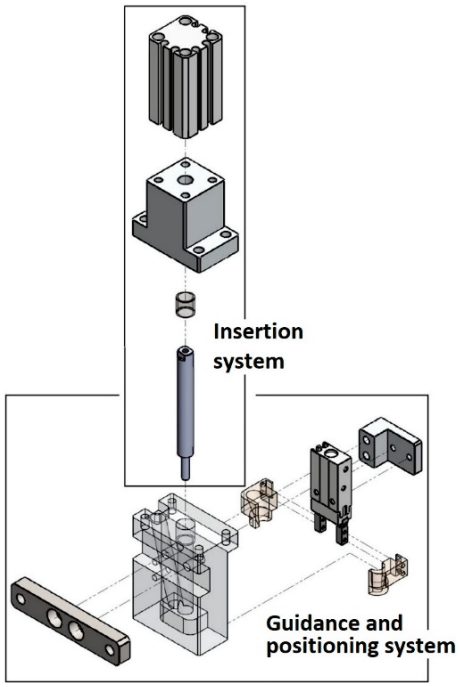


Figure 9. Layout of the inserter.

3.2.4. Axle Systems

The position of the inserter at the component insertion points is determined by the axis systems presented in Figures 10–12, which are equipped with electrical drives and guided movements. The linear movement in each axis is carried out by electric motors such as the one depicted in Figure 13.

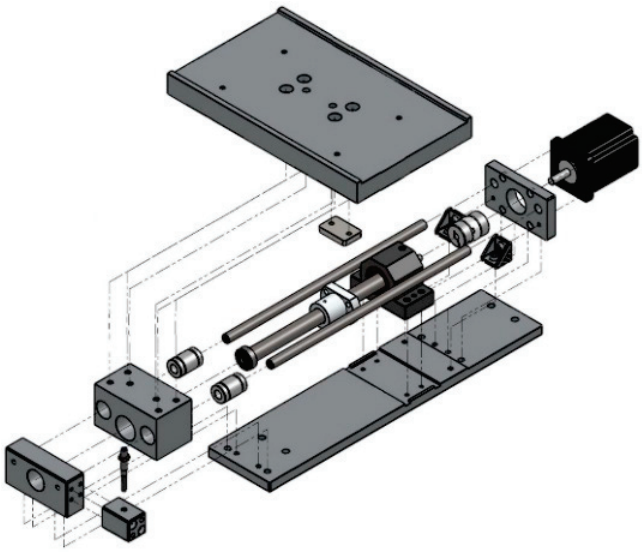


Figure 10. Layout of the *x*-axis system.

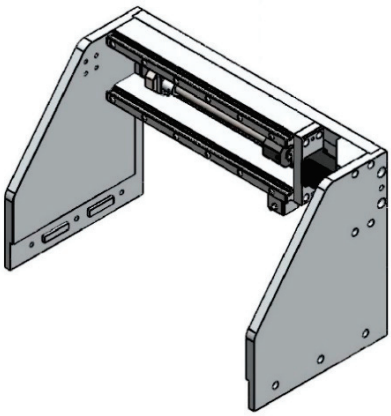


Figure 11. Layout of the *y*-axis system.

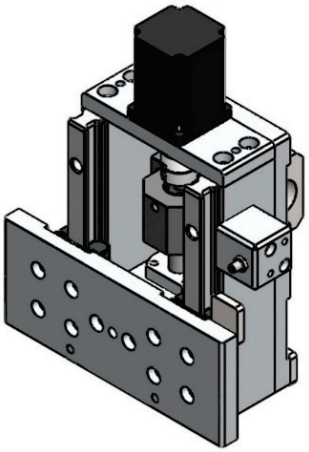


Figure 12. Layout of the *z*-axis system.

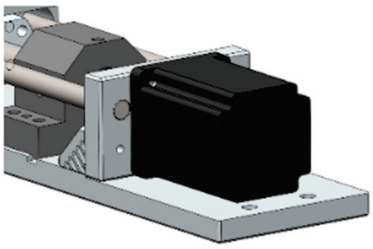
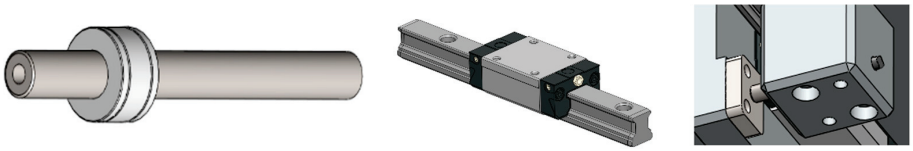


Figure 13. Electric motor.

The x-axis system, as shown in Figure 14a, is guided by precision shafts and linear bearings, while the y-axis and z-axis systems, as shown in Figure 14b, are guided by linear guides in the form of skids and rails. The position of the "home" sensor beam, as shown in Figure 14c, represents the initial travel limit for each axis system.

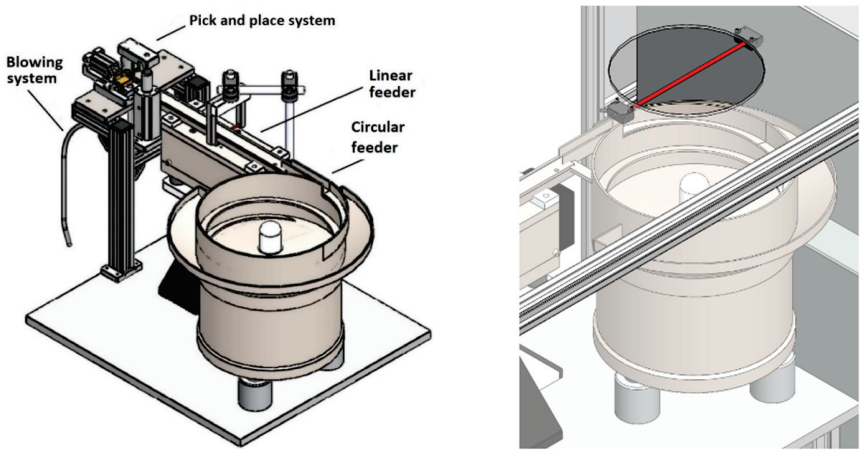


(**a**) (**b**) (**c**)

Figure 14. Precision shaft and linear bearing (**a**), linear guide (skid and rail) (**b**), and "home" sensor beam position (**c**).

3.2.5. Feeder System

The feeder system depicted in Figure 15a operates in multiple stages. The operator first deposits the components (pins or seals) randomly and in bulk through the supply hole. Then, the circular feeder guides and positions the components correctly before they move to the linear feeder, which vibrates them to the loading point of the pick and place system, ensuring the required output rate. The pick and place system then transports the components from the linear feeder to the blowing system, which finally delivers them to the inserter. In Figure 15b, a polycarbonate plate is installed above the feeder system to prevent any foreign materials from falling into the feeders. Moreover, a supply hole and a safety sensor are in place for additional safety measures.



(**a**) (**b**)

Figure 15. Layout of the feeder system for the components (**a**) and polycarbonate plate, supply hole, and security sensor (**b**).

3.2.6. Artificial Vision System

The artificial vision system represented in Figure 16 verifies the insertion of components (pins and seals) in the connector after assembly according to pre-defined programming. The inspection involves a thorough examination of the connector surface, with correct assembly determined by the components' placement in pre-defined locations. The system signals conformity (OK) if the assembly is correct and non-conformity (NOK) if it is not.

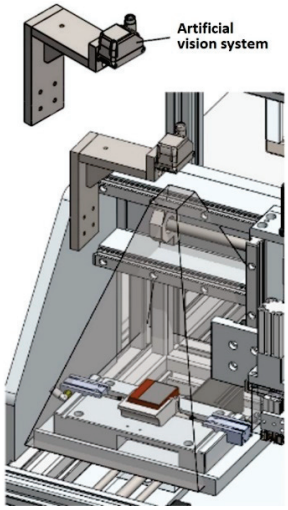
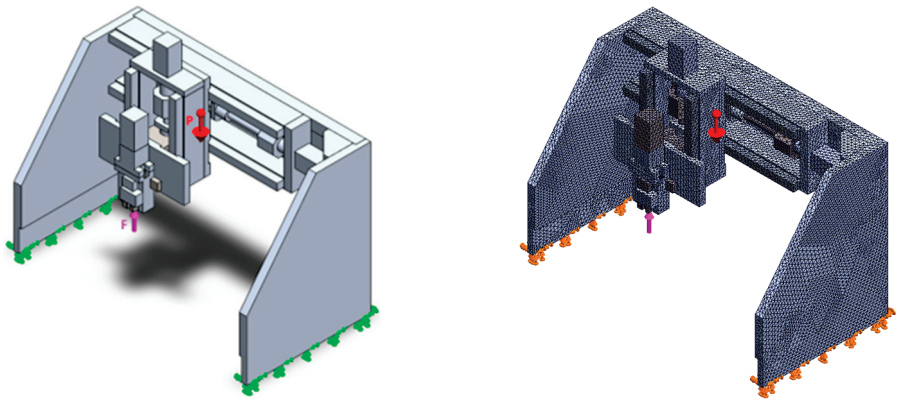


Figure 16. Artificial vision system.

3.3. Structural Verification

To validate the structure of the *y*-axis system, which corresponds to the main axle structure and also involving the *z*-axis system and requires stiffness and strength verification, a linear static analysis was conducted using the Simulation (FEM) module of Solidworks^{TM}. The analysis considered the influence of the different materials used in the parts. As shown in Figure 17a, the analysis considers the impact of the load exerted by the insertion cylinder and the weight of the *z*-axis system with the inserter applied on the *y*-axis system. As boundary conditions, the bases of both side plates were clamped, which is a reasonable approximation to the real behaviour due to the mechanical fastening of this set with the structure. The reaction on the cylinder was defined as 2000 N (defined by the chosen cylinder) and applied upwards.



(**a**) (**b**)

Figure 17. Boundary conditions and loads (**a**) and selected mesh (**b**).

A mesh convergence study was carried out to ensure the existence of accurately predicted stresses and strains. For the analyses performed, quadratic tetrahedral elements of 10

nodes were used, and a mesh based on curvature was considered. The convergence study was executed with gradually increasing mesh sizes at three points in the model. The final mesh selected for the analysis is shown in Figure 17b, including a total of 211,733 elements. Figure 18a presents the resultant displacements (URES) in the model, emphasising that the point where the cylinder exerts force shows the maximum URES (0.404 mm). This value is considered small enough not to affect the proper functioning of the equipment or the insertion process of the component due to the existence of self-centring faces between components to be inserted and the respective housings and not affect the required insertion length. Thus, in terms of displacements, the system is considered validated. Figure 18b presents von Mises equivalent stresses with the scale limited to 40 MPa, which accounts for most regions in the model. Nonetheless, localised stress concentrations were detected near the component insertion tip, where the cylinder exerts force up to 188.3 MPa. This analysis is complemented by an iso-clipping visualisation of the factor of safety (FOS), shown in Figure 19, by filtering the data for a coefficient lower than 20 while the scale ranges from 0 to 100. The highlighted areas represent the regions of the highest stress concentration, which are typically located at the corners between the components and the component application tip. In view of the attained stress results, the *y*-axis design is considered validated.

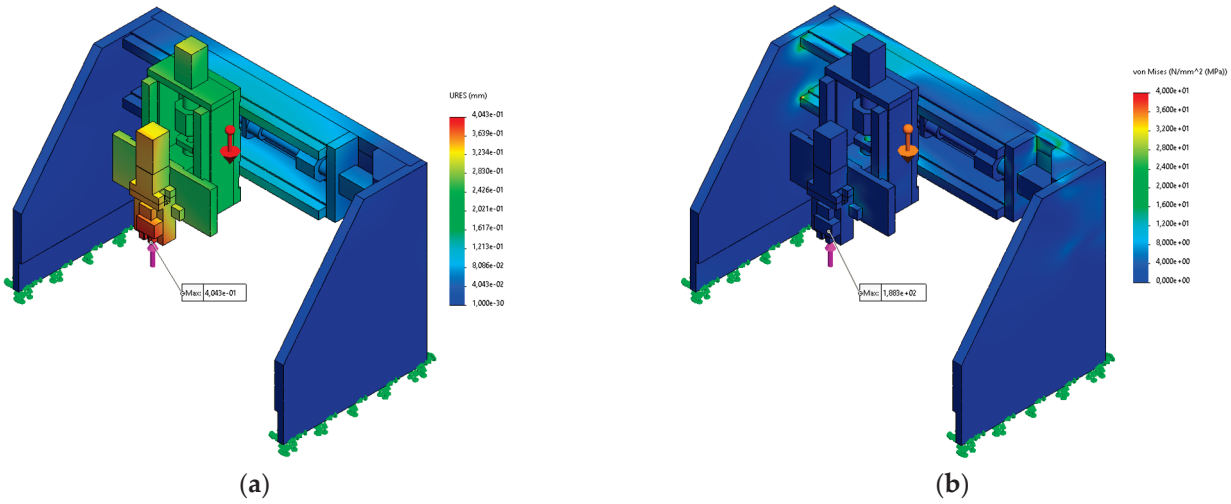


Figure 18. URES results (**a**) and von Mises stresses (**b**) in the *y*-axis system.

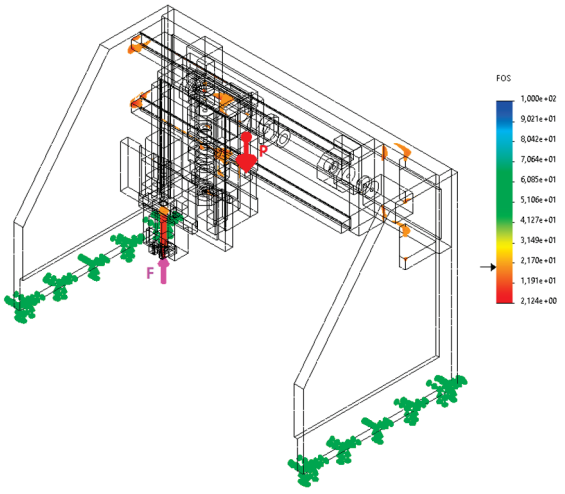


Figure 19. FOS analysis (critical areas) for the *y*-axis system.

3.4. Materials and Processes

The most suitable material was selected for each component to be manufactured based on its function in the equipment and the applied loads. In summary, the selected materials for the equipment were as follows:

- Aluminium alloy 3.3547 (DIN EN-AW 5083) is the most used in the equipment due to easy machining (leading to short fabrication times), reduced weight, and acceptable strength;
- Steel 1.1191 (DIN Ck45) was chosen for parts with higher wear and loads due to its strength and hardness requirements. Some treatments are applied, including quenching, which increases resistance to wear and fatigue, nitriding, which hardens the component surface, and oxidation, which creates a protective layer to increase corrosion resistance and durability;
- Stainless steel 1.4301 (DIN X 5 CrNi 18-10) was chosen for the equipment sheets since untreated steel parts are susceptible to corrosion. Therefore, stainless steel was chosen for some parts as it does not require treatment to resist corrosion;
- Dibond (aluminium-polyethylene composite) and polycarbonate: Dibond is an opaque material purchased in sheets, with a white finish on one side and silver on the other. Polycarbonate is bought in sheets with its final dimensions, chosen to enclose the sides and top of the equipment.

The main manufacturing processes used in the production of the equipment's parts were as follows:

- Milling: process in which material removal is accomplished by feeding the part against a rotating cutter, with relative movement in at least three axes, giving the part the desired shape and finish;
- Turning: in this machining process, a lathe is used to rotate the part while a cutting tool moves in length and depth to remove metal along the diameter, creating a revolution-based shape;
- Laser cutting: thermal process in which a focused laser beam is used to melt material, used to cut sheets and tubes with high quality and precision;
- Bending: it plastically deforms the sheet and enables the production of cylindrical, conical, and prismatic surfaces. This process is utilised for small-batch production;
- Electro-erosion: machining process that involves the destruction of metal particles through wire-induced electric discharges. This method is ideal for cutting intricate contours or complex cavities.

3.5. Working Principle

From the time the connector arrives at the equipment until it leaves, the equipment performs a series of functions responsible for assembling the components and obtaining the final product. The equipment's operating sequence is defined in the diagram of Figure 20. As this is semi-automatic equipment, a few steps in the operation sequence require the operator's intervention, namely, connector type selection via the equipment's console, which indicates the references that can be assembled, followed by manual connector insertion into the workboard and door lock.

To ensure that the operator's tasks are carried out correctly, detection systems are available that eliminate faults, such as the connector's presence and door lock detections. The connector's presence in the workboard is detected, and then it is locked, which guarantees its stability throughout the assembly process. During assembly, the equipment moves its axes, placing the inserter on the different component assembly points. When the inserter is positioned over one insertion point, its operating sequence begins, namely, checking for the presence of a component to be inserted, followed by opening the grip, and finally advancing the insertion cylinder. Next, these steps are reverted in preparation for the next insertion. At the same time, the feeder system remains in operation, ensuring that there is no component shortage in the insertion zone. The power supply system is

in operation throughout the entire equipment's operating period. During the equipment start-up phase, the operator is asked to supply the circular feeder with components. At this stage, the operator puts as many components as possible into the system, ensuring production for a few hours. Inside the circular feeder, the components are orientated and positioned so that they can then be sent to the insertion area. Once the equipment has inserted all the components, the workboard is positioned so that the vision system can check that the components have been correctly assembled, thus starting the process of extracting the connector from the equipment. Once the assembly has been successfully completed, the equipment moves the workboard back to its initial position and releases the connector. At this point, the operator opens the equipment door, removes the finished connector, and begins assembling the next connector.

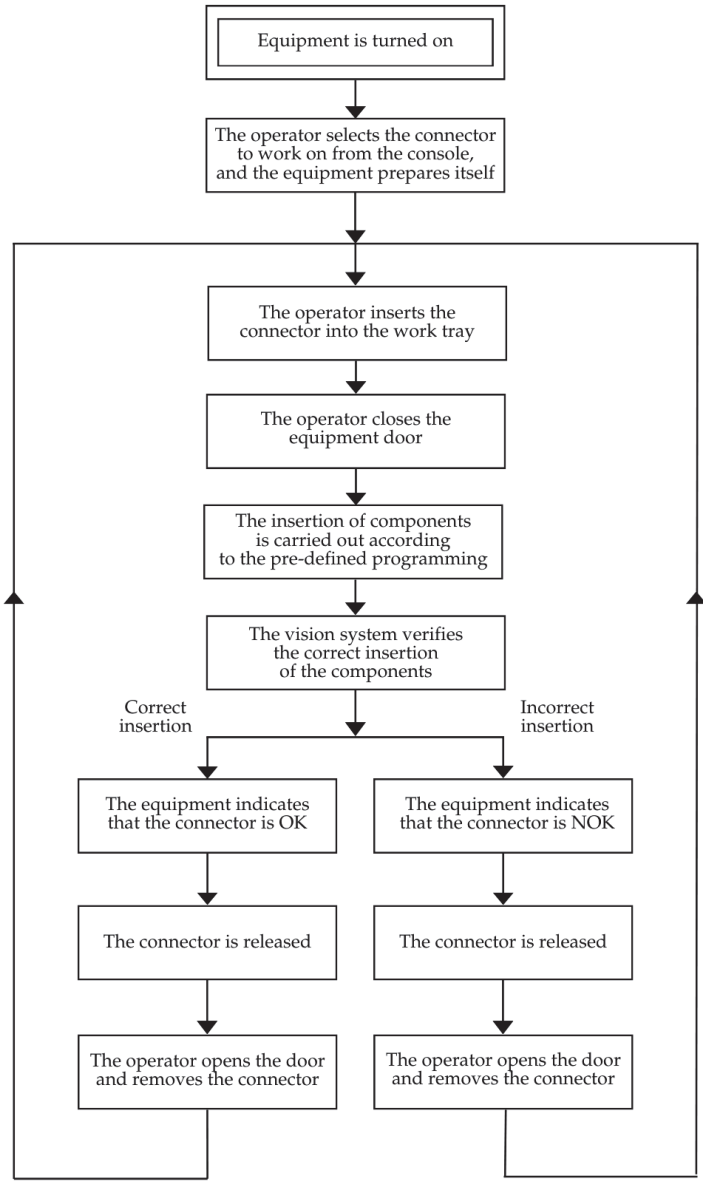


Figure 20. Equipment's operating sequence.

3.6. Cost Analysis

The final purpose of the proposed equipment is to commercialise it, which makes it appropriate to carry out an analysis of associated costs. Table 4 lists all the points considered in the cost evaluation phase of the equipment, costs that range from the design to the profit forecast associated with its commercialisation.

Table 4. Costs associated with the equipment.

Mechanical design
Materials procurement
Machining hours
Mechanical and electrical assembly
Automation design
Profit margin

Labour hours and the profit margin are only defined at the commercialisation stage, at which point an estimate is made depending on the complexity of the equipment and the customer who is going to buy it, so the following is simply an analysis of the costs associated with the purchase of materials, which is the main cost and directly influences the other costs. Table 5 shows approximate figures for the main costs associated with purchasing materials. There is a major expense associated with the acquisition of the main systems, which includes the power supply system. Then, the expenditure is distributed between structural materials, raw materials, pneumatic materials, and electrical materials, showing that the sum of the expenditure on the acquisition of materials is approximately EUR 18,000.

Table 5. Costs associated with the purchase of materials.

Acquisition of the main systems	8750 €
Structural material	2120 €
Raw materials	2400 €
Pneumatic material	665 €
Electrical material	3678 €
External services	80 €
	Aprox. EUR 18,000

3.7. Subsets Validation

This project was developed with the intention of responding to the needs of a particular client, but due to the negative phase the market has gone through in recent years, the equipment has not yet reached a production stage. However, all the main systems were developed using knowledge acquired by the design company on other projects.

Figure 21 shows application examples of the principles used in the different subsets of the equipment, including the workboard, the inserter, the feeder system, and the axle system. The fact that the project is based on systems that have already been implemented and tested gives confidence in its application, something that is sought after in all the special equipment developed. Although the equipment has not entered a production phase, it can be adapted to other processes in the future. The fact that the equipment has a three-axis system, an artificial vision system and that it is compact allows it to be adapted and transformed into a 3D printer or a laser engraving system, for example, with short design time.

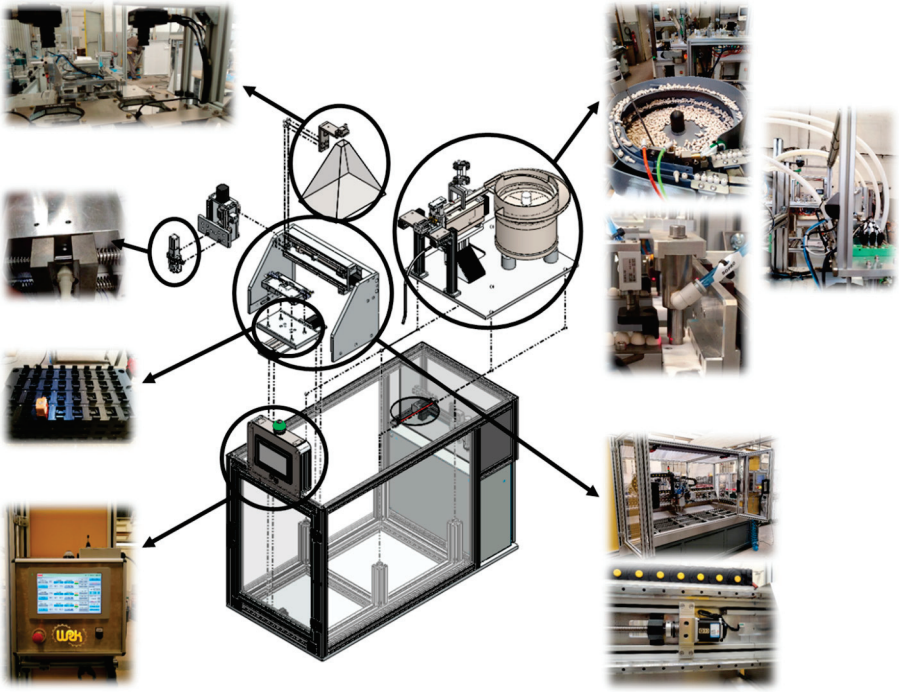


Figure 21. Examples of system applications.

4. Conclusions

The present work aimed to design semi-automatic equipment that could assemble cabling connectors used in the automotive sector as a replacement for a currently manual process that takes place in an automotive components company to increase its competitiveness. The design was based on principles such as automation, flexibility, high productivity, and competitive pricing. The pre-defined objectives and requirements for this project were fully met:

- Develop equipment that works semi-automatically and uses an inserter, two-component types and one connector type—the equipment works semi-automatically with two component types (pins and seals) and one connector type. There is only human intervention when depositing the components in the circular feeder and placing the connector in the nest of the workboard;
- Implement a component feeder system. The components must be deposited randomly by the operator—a feeder system has been implemented with a circular feeder ready to receive randomly deposited components via a supply hole;
- Design systems to transport components from the feeder system to the inserter—the pick and place and the blower transport the components from the feeder system to the inserter and have been developed taking into account the geometry of the component;
- Create positioning systems for the connector to secure it during the assembly of components (pins and seals)—a workboard was developed consisting of a nest that locks the connector in a given position on the x and y-axes. In addition, the workboard includes latches that lock the connector in a position on the z-axis;
- Apply a system capable of detecting the correct assembly of components in the connector—vision sensor with artificial intelligence was attached to the coordinate mechanism (y-axis);
- Carry out a linear static analysis using FEM—validation of the axle system (y-axis) was accomplished by means of a linear static analysis using the FEM;
- Estimate the costs associated with the equipment—the costs were estimated by categories (main systems, structural material, raw materials, pneumatic material, electrical material, and external services).

The equipment design was successfully accomplished, and the imposed requirements by the client were fulfilled, with significant advantages over the current manual process, such that an alternative to a less efficient manual process can be implemented. Overall, a novel solution was proposed combining automation, modular feeding, assembly mechanisms (which can be adapted for different components, making it flexible for other applications and industries), and a vision system with artificial intelligence. While designed for the automotive sector, the principles of this semi-automatic system, especially the feeder system and the artificial intelligence-driven quality control, can be transferred to other industries (such as electronics, aerospace, or consumer goods), where similar manual processes are still in place. By highlighting the design of a system that is both high-performing and cost-efficient, both the technical and economic challenges of automation in competitive industries like automotive manufacturing are addressed.

Author Contributions: Conceptualisation, P.M.P.C. and J.A.P.P.; methodology, R.D.S.G.C. and F.J.G.S.; software, P.M.P.C.; validation, P.M.P.C. and J.A.P.P.; formal analysis, R.D.S.G.C. and F.J.G.S.; investigation, P.M.P.C.; resources, R.D.S.G.C. and F.J.G.S.; writing—original draft preparation, P.M.P.C. and J.A.P.P.; writing—review and editing, R.D.S.G.C. and F.J.G.S.; supervision, R.D.S.G.C. and F.J.G.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

CAD	Computer-Aided Design
DfMA	Design for Manufacturing and Assembly
DRM	Design Research Methodology
DSR	Design Science Research
FEM	Finite Element Method
FOS	Factor of Safety
NOK	Indicator of Non-conformity
OK	Indicator of Conformity
SRC/CS	Safety Related Parts/Control Systems
SWOT	Strengths, Weaknesses, Opportunities, and Threats
TRIZ	Theory of Inventive Problem Solving
URES	Resultant Displacements
V-Model	Verification and Validation

References

1. Klink, G.; Mathur, M.; Kidambi, R.; Sen, K. Contribution of the automobile industry to technology and value creation. *Auto Tech Rev.* **2014**, *3*, 18–23. [CrossRef]
2. Benotsmane, R.; Dudás, L.; Kovács, G. Survey on new trends of robotic tools in the automotive industry. In Proceedings of the Vehicle and Automotive Engineering 3 (VAE 2020), Miskolc, Hungary, 25–26 November 2021; pp. 443–457.
3. Gabbar, H.A.; Idrees, M. ARSIP: Automated robotic system for industrial painting. *Technologies* **2024**, *12*, 27. [CrossRef]
4. Dulina, L.; Zuzik, J.; Furmannova, B.; Kukla, S. Improving material flows in an industrial enterprise: A comprehensive case study analysis. *Machines* **2024**, *12*, 308. [CrossRef]
5. Lo, W.; Yang, C.-M.; Zhang, Q.; Li, M. Increased productivity and reduced waste with robotic process automation and generative AI-powered IoE services. *J. Web Eng.* **2024**, *23*, 53–87. [CrossRef]
6. Konstantinidis, F.K.; Mouroutsos, S.G.; Gasteratos, A. The role of machine vision in industry 4.0: An automotive manufacturing perspective. In Proceedings of the 2021 IEEE International Conference on Imaging Systems and Techniques (IST), Kaohsiung, Taiwan, 24–26 August 2021; pp. 1–6.
7. Gajmal, S.S.; Bhatwadekar, S.G. Low cost automation (LCA): A case study. *Int. J. Adv. Eng. Glob. Tech.* **2014**, *2*, 172–181.
8. Salah, B.; Alsamhan, A.M.; Khan, S.; Ruzayqat, M. Designing and developing a smart yogurt filling machine in the Industry 4.0 era. *Machines* **2021**, *9*, 300. [CrossRef]

9. Magalhães, A.; Silva, F.; Campilho, R. A novel concept of bent wires sorting operation between workstations in the production of automotive parts. *J. Braz. Soc. Mech. Sci. Eng.* **2018**, *41*, 25. [CrossRef]
10. Groover, M.P. *Automation, Production Systems and Computer-Integrated Manufacturing*; Pearson: London, UK, 2016.
11. Dey, C.; Sen, S.K. *Industrial Automation Technologies*; CRC Press: Boca Raton, FL, USA, 2020.
12. Shaloo, M.; Princz, G.; Hörbe, R.; Erol, S. Flexible automation of quality inspection in parts assembly using CNN-based machine learning. *Procedia Comput. Sci.* **2024**, *232*, 2921–2932. [CrossRef]
13. Lowe, B.D.; Hayden, M.; Albers, J.; Naber, S. Case studies of robots and automation as health/safety interventions in small manufacturing enterprises. *Hum. Factors Ergon. Man.* **2023**, *33*, 69–103. [CrossRef]
14. Ford, M. *The Lights in the Tunnel: Automation, Accelerating Technology and the Economy of the Future*; Acculant Publishing: New York, NY, USA, 2009.
15. Zhou, Z.; Guo, J.; Zhu, Z.; Guo, H. Uncalibrated visual servoing based on Kalman filter and mixed-kernel online sequential extreme learning machine for robot manipulator. *Multimed. Tools Appl.* **2024**, *83*, 18853–18879. [CrossRef]
16. Ulanov, A. World robot market: Key characteristics and trends. In *Digital International Relations*; Baykov, A., Zinovieva, E., Eds.; Springer Nature: Singapore, 2023; pp. 235–246.
17. Kiran, D. *Production Planning and Control: A Comprehensive Approach*; Butterworth-Heinemann: Oxford, UK, 2019.
18. Ahmad, R.; Plapper, P. Safe and automated assembly process using vision assisted robot manipulator. *Procedia CIRP* **2016**, *41*, 771–776. [CrossRef]
19. Almoslehy, S.A.M. *Design for Profitability: Guidelines to Cost Effectively Manage the Development Process of Complex Products*; CRC Press: Boca Raton, FL, USA, 2018.
20. Pahl, G.; Beitz, W.; Feldhusen, J.; Grote, K.-H. *Engineering Design: A Systematic Approach*, 3rd ed.; Springer: Berlin, Germany, 2007.
21. Ulrich, K.T.; Eppinger, S. *Product Design and Development*, 6th ed.; McGraw-Hill Education: New York, NY, USA, 2016.
22. Forsberg, K.; Mooz, H.; Cotterman, H. *Visualizing Project Management: Models and Frameworks for Mastering Complex Systems*; John Wiley & Sons: Hoboken, NJ, USA, 2005.
23. Pugh, S. *Total Design: Integrated Methods for Successful Product Engineering*; Addison-Wesley: Boston, MA, USA, 1991.
24. Altshuller, G. *40 Principles: TRIZ Keys to Technical Innovation*; Technical Innovation Center: Worcester, MA, USA, 2002.
25. Suh, N.P. *Axiomatic Design: Advances and Applications*; Oxford University Press: New York, NY, USA, 2001.
26. Peffers, K.; Tuunanen, T.; Rothenberger, M.A.; Chatterjee, S. A design science research methodology for information systems research. *J. Manag. Inform. Syst.* **2007**, *24*, 45–77. [CrossRef]
27. Tojal, M.; Silva, F.J.; Campilho, R.; Pinto, A.; Pinto Ferreira, L. Case-based product development of a high-pressure die casting injection subset using design science research. *FME Trans.* **2022**, *50*, 32–45. [CrossRef]
28. Blessing, L.T.; Chakrabarti, A. *DRM: A Design Reseach Methodology*; Springer: Berlin/Heidelberg, Germany, 2009.
29. Boothroyd, G.; Dewhurst, P.; Knight, W.A. *Product Design for Manufacture and Assembly*; CRC Press: Boca Raton, FL, USA, 2010.
30. Ehrgott, M.; Gandibleux, X. Multiobjective combinatorial optimization—Theory, methodology, and applications. In *Multiple Criteria Optimization: State of the Art Annotated Bibliographic Surveys*; Springer: Boston, MA, USA, 2003.
31. Noura, H.; Theilliol, D.; Ponsart, J.-C.; Chamseddine, A. *Fault-Tolerant Control Systems: Design and Practical Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009.
32. Feldmann, K.; Müller, B.; Haselmann, T. Automated assembly of lightweight automotive components. *CIRP Ann.* **1999**, *48*, 9–12. [CrossRef]
33. Mo, F.; Rehman, H.U.; Monetti, F.M.; Chaplin, J.C.; Sanderson, D.; Popov, A.; Maffei, A.; Ratchev, S. A framework for manufacturing system reconfiguration and optimisation utilising digital twins and modular artificial intelligence. *Robot. Comput. Integr. Manuf.* **2023**, *82*, 102524. [CrossRef]
34. Mateu-Gomez, D.; Martínez-Peral, F.J.; Perez-Vidal, C. Multi-arm trajectory planning for optimal collision-free pick-and-place operations. *Technologies* **2024**, *12*, 12. [CrossRef]
35. Ling, L.; Zhang, X.; Hu, X.; Fu, Y.; Yang, D.; Liang, E.; Chen, Y. Research on spraying quality prediction algorithm for automated robot spraying based on KHPO-ELM neural network. *Machines* **2024**, *12*, 100. [CrossRef]
36. Costa, M.; Gouveia, R.; Silva, F.; Campilho, R. How to solve quality problems by advanced fully-automated manufacturing systems. *Int. J. Adv. Manuf. Technol.* **2018**, *94*, 3041–3063. [CrossRef]
37. Figueiredo, D.; Silva, F.; Campilho, R.; Silva, A.; Pimentel, C.; Matias, J. A new concept of automated manufacturing process for wire rope terminals. *Procedia Manuf.* **2021**, *51*, 431–437. [CrossRef]
38. Santos, P.R.; Silva, F.J.G.; Campilho, R.D.S.G.; Pinto, G.F.L.; Baptista, A. A novel concept of a conduit transport system. *Procedia Manuf.* **2019**, *38*, 848–857. [CrossRef]
39. Pereira, J.A.P.; Campilho, R.D.S.G.; Silva, F.J.G.; Sánchez-Arce, I.J. Robotized cell design for part assembly in the automotive industry. *Proc. Inst. Mech. Eng. Part C* **2022**, *236*, 8807–8822. [CrossRef]
40. Michalos, G.; Kousi, N.; Karagiannis, P.; Gkournelos, C.; Dimoulas, K.; Koukas, S.; Mparis, K.; Papavasileiou, A.; Makris, S. Seamless human robot collaborative assembly: An automotive case study. *Mechatronics* **2018**, *55*, 194–211. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Toward Inclusive Smart Cities: Sound-Based Vehicle Diagnostics, Emergency Signal Recognition, and Beyond

Amr Rashed [1], Yousry Abdulazeem [2], Tamer Ahmed Farrag [3,4], Amna Bamaqa [5,4], Malik Almaliki [6,4], Mahmoud Badawy [5,4,7,*] and Mostafa A. Elhosseini [6,4,7]

1 Department of Communications and Electronics Engineering, Faculty of Engineering, Mansoura University, Mansoura 35516, Egypt; amr_rashed2@hotmail.com
2 School of Computational Sciences and Artificial Intelligence (CSAI), Zewail City of Science and Technology, Giza 12578, Egypt; yousry@mans.edu.eg
3 Department of Electrical Engineering, College of Engineering, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia; t.farrag@tu.edu.sa
4 King Salman Center for Disability Research, Riyadh 11614, Saudi Arabia; abamaqa@taibahu.edu.sa (A.B.); mrmalki@taibahu.edu.sa (M.A.); melhosseini@mans.edu.eg (M.A.E.)
5 Computer Science and Information Department, Applied College, Taibah University, Medinah 41461, Saudi Arabia
6 Department of Computer Science, College of Computer Science and Engineering, Taibah University, Yanbu 46421, Saudi Arabia
7 Computers and Control Systems Engineering Department, Faculty of Engineering, Mansoura University, Mansoura 35516, Egypt
* Correspondence: engbadawy@mans.edu.eg

Abstract: Sound-based early fault detection for vehicles is a critical yet underexplored area, particularly within Intelligent Transportation Systems (ITSs) for smart cities. Despite the clear necessity for sound-based diagnostic systems, the scarcity of specialized publicly available datasets presents a major challenge. This study addresses this gap by contributing in multiple dimensions. Firstly, it emphasizes the significance of sound-based diagnostics for real-time detection of faults through analyzing sounds directly generated by vehicles, such as engine or brake noises, and the classification of external emergency sounds, like sirens, relevant to vehicle safety. Secondly, this paper introduces a novel dataset encompassing vehicle fault sounds, emergency sirens, and environmental noises specifically curated to address the absence of such specialized datasets. A comprehensive framework is proposed, combining audio preprocessing, feature extraction (via Mel Spectrograms, MFCCs, and Chromatograms), and classification using 11 models. Evaluations using both compact (52 features) and expanded (126 features) representations show that several classes (e.g., Engine Misfire, Fuel Pump Cartridge Fault, Radiator Fan Failure) achieve near-perfect accuracy, though acoustically similar classes like Universal Joint Failure, Knocking, and Pre-ignition Problem remain challenging. Logistic Regression yielded the highest accuracy of 86.5% for the vehicle fault dataset (DB1) using compact features, while neural networks performed best for datasets DB2 and DB3, achieving 88.4% and 85.5%, respectively. In the second scenario, a Bayesian-Optimized Weighted Soft Voting with Feature Selection (BOWSVFS) approach is proposed, significantly enhancing accuracy to 91.04% for DB1, 88.85% for DB2, and 86.85% for DB3. These results highlight the effectiveness of the proposed methods in addressing key ITS limitations and enhancing accessibility for individuals with disabilities through auditory-based vehicle diagnostics and emergency recognition systems.

Keywords: smart cities; intelligent transportation systems (ITSs); machine learning (ML); Bayesian optimization

1. Introduction

Intelligent Transportation Systems (ITSs) play a crucial role in developing smart cities through advanced technologies to enhance the efficiency and sustainability of transportation networks [1,2]. These systems integrate data from sensors, cameras, and GPS devices that provide real-time information on traffic flow, weather conditions, and other relevant factors. ITSs can also use this to adjust the traffic signals dynamically, manage toll road usage, and give drivers personalized route recommendations to reduce congestion and lower travel time. Additionally, ITSs can support autonomous vehicles and shared mobility services, further improving overall urban transportation system performance.

In smart cities, ITSs can also contribute to air quality and greenhouse gas emission reduction by making public transportation, cycling, and walking real options instead of private car travel [3]. This can be achieved through incentives for using sustainable modes of transportation by implementing smart parking systems and congestion pricing schemes, thus decreasing the overall demand for fossil fuel-powered vehicles. This would further allow easy intermodal connections, making urban mobility and the general urban environment even more accessible, sustainable, and inclusive.

The role of vision in systems has become an integral part of modern infrastructure, especially in ITSs, for furthering both safety and efficiency. However, these systems have a significant limitation: they cannot "hear" necessary auditory signals such as emergency sirens, mechanical faults, and environmental hazards [4]. There are various limitations of vision-based systems, such as the inability to detect auditory signals [5], environmental noise interference [6], mechanical fault detection [7]. Other than emergency signals and mechanical faults, environmental hazards such as construction noise, falling debris, or wildlife may create serious dangers. In this regard, the vision-based systems cannot detect such hazards until they become visually apparent, which may be too late to act effectively. Thus, sound-based diagnostics can provide early warnings against such dangers and enhance safety. Sound-based diagnostics can improve the reliability and efficiency of existing systems manifoldly. There are several reasons that sound-based diagnosis is highly respected, including enhanced situational awareness [8,9], real-time monitoring and alerts [10], cost-effectiveness [11], accessibility for hearing-impaired people [12], and improved emergency response [12].

ITSs are designed to improve the functionality and safety of transport systems using advanced technologies [13]. Whereas these systems are designed to streamline the process, people with disabilities face unique barriers to their movement and access to essential services in such contexts. These include, but are not limited to, information inaccessibility, lack of warnings in emergencies, navigation barriers, inadequate access to vehicles, and communication barriers. Sound-based systems are essential in offering complementarities and ensuring increased access within ITS environments based on alternative means of communication and information sharing. Some of the possible solutions include the following: audio cues for navigation, visual alerts for the hearing-impaired, sound detection for vehicle fault detection, emergency sirens and alerts, and improved communication systems [14]. ITS application covers a wide range of fields in implementing sound detection. For example, this will help in emergency response for quick and timely mitigation of hazards. In the field of public transportation, increased safety and efficiency for passengers may also be achieved. Smart cars use it to enhance driver awareness and the vehicle's behavior.

Recent advances in artificial intelligence (AI) have paved the way for sophisticated data analysis techniques that drive innovative applications across various fields, including sound-based diagnostics. Within this AI framework, machine learning (ML)—defined as using algorithms and statistical models that enable computer systems to learn from data and improve their performance on specific tasks without explicit programming—has emerged

as a key enabler. By leveraging ML, our study can analyze complex auditory signals from vehicle faults and environmental sounds, transforming raw data into actionable insights for Intelligent Transportation Systems (ITSs).

We begin our approach with foundational ML models that provide robust baseline performance. Techniques such as Logistic Regression (LR) and k-nearest neighbors (kNN) are utilized to establish initial classification capabilities, forming the groundwork for further enhancements. These basic models are critical for understanding the underlying patterns in the audio data and serve as benchmarks against which more sophisticated methods can be compared.

The study introduces three novel datasets that capture various sounds relevant to Intelligent Transportation Systems. The first dataset (DB1) consists of 27 distinct vehicle fault classes featuring critical sounds such as Engine Misfire, Fuel Pump Cartridge Fault, Radiator Fan Failure, and Strut Mount Failure, all directly generated by the vehicle. The second dataset (DB2) comprises 22 environmental sound classes, including emergency signals like sirens and various transportation-related and ambient environmental noises. These datasets provide a rich collection of auditory signals that form the basis for robust sound-based diagnostic systems. The third dataset (DB3) also merges DB1 and DB2 to create a comprehensive collection of 49 classes. This unified dataset enables the framework to classify any sound from vehicle faults or external environmental events into the correct category. The study lays the groundwork for advancing machine learning research in sound-based diagnostics by addressing the scarcity of specialized, publicly available auditory datasets. It contributes to the development of more inclusive and responsive ITS applications.

Building on these fundamentals, our framework incorporates advanced variants of ML to address the challenges of differentiating acoustic similar classes. Ensemble methods such as AdaBoost, Random Forest (RF), and Gradient Boosting (GB) enhance accuracy by combining the strengths of multiple weak learners. Additionally, Support Vector Machines (SVM) and Stochastic Gradient Descent (SGD) optimize decision boundaries in complex feature spaces, while Decision Trees (DTs) provide interpretable classification logic. Complementing these are the CN2 algorithm and Naive Bayes (NB), which handle complex rule-based classification and probabilistic inference. Together, these diverse ML techniques form a comprehensive diagnostic system capable of robust performance in real-world ITS applications.

Integrating auditory intelligence into ITSs addresses several key research challenges, notably the difficulty of distinguishing acoustically similar classes—such as Universal Joint Failure versus Bad CV Joint and Knocking versus Pre-ignition Problem. These challenges demand advanced feature extraction techniques and robust machine learning models that capture subtle differences in sound signatures. Additionally, the scarcity of specialized, publicly available auditory datasets has historically hindered progress in this area; this research overcomes that barrier by introducing a comprehensive, curated dataset that serves as a benchmark for future work.

This study addresses a critical gap in Intelligent Transportation Systems (ITSs) by explicitly defining its aim to detect faults directly from sounds generated by vehicles, such as engine or brake noises, and to classify external alert sounds, including emergency sirens. The intended applications of these predictive outputs are articulated, emphasizing their role in real-time diagnostics for smart vehicle systems and providing auditory-to-visual alert conversions to assist sound-impaired drivers. Additionally, the study highlights the potential of auditory capabilities to enhance vehicle fault detection and accessibility for individuals with disabilities while addressing the scarcity of specialized datasets in this domain. This is achieved through the following key contributions:

- Introducing a novel dataset comprising vehicle fault sounds, emergency sirens, and environmental noises filling a critical gap in publicly available resources.

- Developing a comprehensive methodology for audio preprocessing, including normalization, resampling, and segmentation.
- Proposing robust feature extraction techniques, such as Mel Spectrograms, MFCCs, and Chromatograms, enabling compact and expanded feature representations.
- Evaluating multiple ML models in the first scenario, including neural networks, Logistic Regression, and Random Forests.
- Proposing a Bayesian-Optimized Weighted Soft Voting with Feature Selection (BOWSVFS) approach in the second scenario, achieving a classification accuracy of 91.04% on the car fault dataset (DB1) and outperforming the first scenario's results.
- Demonstrating the relevance of sound-based ITSs in promoting accessibility by offering real-time alerts and auditory-to-visual conversion solutions for individuals with disabilities.
- Aligning sound-based diagnostics with broader smart city goals, contributing to the development of safer and more inclusive transportation systems.

This research establishes a strong foundation for integrating auditory intelligence into ITSs, with significant implications for safety, accessibility, and inclusivity in smart cities. The framework demonstrates strong performance overall, with several fault classes being recognized with near-perfect accuracy. For example, classes such as Engine Misfire, Fuel Pump Cartridge Fault, Radiator Fan Failure, Strut Mount Failure, and Suspension Arm Fault consistently achieve 100% accuracy in many cases. This indicates that the framework effectively captures the distinct acoustic signatures associated with these faults. However, challenges remain for acoustically similar classes. Universal Joint Failure, for instance, is occasionally misclassified—often confused with Bad CV Joint—while Bad Wheel Bearing also shows minor misclassifications. More notably, the Knocking and Pre-ignition Problem classes face significant difficulties, with Pre-ignition Problem instances frequently being predicted as Engine Misfire. These misclassifications highlight the areas where further refinement in feature extraction or model tuning may be necessary to better differentiate between closely related acoustic patterns.

The structure of this paper is as follows: A summary of the current literature is given in Section 2, along with potential directions for further research. Section 3 introduces the datasets. Section 4 focuses on materials and explains the proposed methodology. Section 5 provides an overview of the experiments, including the experimental setup, methods, and findings collected, focusing on the performance metrics attained. The overall discussion in Section 6 wraps up with conclusions and future work, summarizing the paper's key contributions and suggesting directions for subsequent research in this domain.

2. Literature Review

This section discusses earlier attempts at sound recognition and sound-based defect detection in machinery, vehicles, trains, and aircraft systems. The majority of the methods assessed were developed using ML methods. Some, meanwhile, are more recent and rely on deep learning or vision transformers.

Nasim et al. introduced a sound-based early fault detection system for vehicles utilizing ML technology [15]. This system is specifically designed to target the faults in vehicles at their initial stages by analyzing the sound emitted by the vehicle. The system starts working by binary classification, which can decide whether the vehicle is faulty or healthy. They utilized time domain, frequency domain, and time–frequency domain features to detect normal and abnormal vehicle conditions effectively. Additionally, they employed abnormal vehicle data to classify them into fifteen other typical vehicle issues. Through experimentation, the random forest algorithm yielded the best accuracy of 97% for fault detection and 92% for problem classification when utilizing time–frequency features.

Hamad et al. proposed a rule-based ML technique that automatically detects engine problems [16]. The generalizability of the system is considered by time domain, frequency domain, and time–frequency domain features. The robustness of the developed system is evaluated using noisy sound data collected under various normal and abnormal conditions. The experimental results demonstrated that the approach outperformed other techniques by 2.6−6.0% and yielded the highest performance accuracy of 98.6%. Yildirim et al. proposed a testing and evaluation procedure on the sound quality of two types of cars [17]. The sound quality is analyzed through the car's road running test on the provided ground with varying running speeds. They proposed a neural network predictor to model the system for possible experimental applications. In their experiments, only objective factors of loudness, sharpness, speech intelligibility, and sound pressure level are considered essential for sound quality. The computer simulations and experiments show evidence that the neural predictor algorithm provides reasonable accommodation in different cases and allows superior prediction in two-car sound analysis.

Mel-Frequency Cepstral Coefficients (MFCCs), DWT-based features, and the Extreme Learning Machine (ELM) classifier were employed in the vehicle problem diagnostic system that Akbalik et al. presented [18]. The proposed framework uses a big, diversified dataset that includes many vehicle models and real-world operating situations. The experiment results show that the MFCC-based features combined with the ELM classifier outperform the others in terms of accuracy, precision, recall, F1-score, macro F1-score, and weighted F1-score, which are 92.17%, 92.24%, 92.22%, 92.10%, and 92.06%, respectively. Murovec et al. created an acquisition system using the Zero-Crossing Signature (ZCS) technique [19]. To accomplish precise engine type classification, the study used a unique level-crossing (ZCS) feature that demonstrated excellent performance in differentiating engine sounds from surrounding noise. A dataset of 417 vehicle recordings was examined, and the classification performance of the ZCS was compared to the traditional Zero-Crossing (ZC) technique utilizing a Self-Organizing Map (SOM) with a 1D grid of nine neurons. Wang et al. proposed a method for diagnosing engine acoustic signal faults using multi-level supervised learning and time-frequency transformation [20]. First, it decomposes the fault diagnostic problem into feature augmentation, fault detection, and identification. Second, based on several time–frequency studies, it proposes an adaptive fault feature band extraction approach aimed at distinct features from different vehicle data. Finally, a frequency band attention module was designed to focus on the most meaningful frequency range to the characteristics of engine failure.

Boztas et al. proposed a learning model for improving machine fault classification using handcrafted attributes [21]. The approach utilized texture and statistical features in classifying faults with high performance. They developed a hybrid and multilevel feature extraction technique that maintains high efficiency while lowering the complexity associated with deep learning frameworks. Using a Chi2 feature selector to eliminate redundant features, the model focused on the most informative features throughout the classification step. In the MIMII (noisy) dataset, the proposed model effectively classified more than 90% of the five cases. A Variational Autoencoder/Convolutional Neural Network (VAE-CNN) was created by Wang et al. to diagnose rolling bearing faults [22]. The model was developed to extract complex vibration signal features to detect and categorize faults. While the CNN component increases the expressiveness of signal data and successfully handles issues like gradient vanishing and explosion, the VAE component improves noise robustness. The diagnostic accuracy of the VAE-CNN model for various fault types at varying rotational speeds typically exceeded 90%, yielding generally satisfactory diagnostic results. Xinwen Guo developed a defect diagnostic approach based on feature extraction and a word bag model using acoustics and vibration engineering science theories [23]. This approach mainly expands the three-layer structure of the word bag model and constructs

codebooks for each layer's feature vectors based on this model. Thereafter, it develops the failure detection system of a rolling bearing based on the adaptive extended word bag model. The findings revealed that the defect detection technique has excellent diagnostic accuracy and stability, offering dependable technical assistance for regular operation and safe mechanical equipment maintenance.

Li et al. developed a defect diagnosis system for railway turnout switch machines based on sound signals [24]. The method used Eigenmode Decomposition to improve the sound signal, reduce noise, and extract important statistical information from the time and frequency domains. The ReliefF algorithm is used for feature selection, dimension reduction, and fault classification with weighted parameters to address redundant information in high-dimensional features. The selected feature parameters are then utilized to train the Support Vector Machine. The results showed a defect diagnostic accuracy of 98% in the positioning work mode and 95.67% in the reversing work mode. Kreuzera et al. proved that diagnosing bearing defects in railway vehicles using aerial sound data is possible, even in complex real-world settings [25]. To that purpose, many characteristics are investigated, including Mel Frequency Cepstral Coefficients (MFCCs), which are best suited for diagnosing bearing problems by analyzing airborne sound. The MFCCs were utilized to train an MLP classifier. The suggested technique is assessed using real-world data from a cutting-edge commuter train car in a dedicated measurement effort. The classification results showed that the chosen MFCC features allowed for the reliable detection of bearing defects, including those not included in the training. Eunsun Yun and Minjoong Jeong proposed a feature extraction technique for fault sound identification in EPS motors [26]. This technique reduced the feature dimensionality while preserving the original raw waveform, which is crucial to maintaining the essential features in the waveform for anomaly detection. They combined DFMT with MFCC to optimize feature extraction. They applied LSTM-AE to classify data by segregating standard data from abnormal ones using reconstruction error metrics. The experimental results of the proposed method were proved efficient with an accuracy of 99.2%, recall of 94.0%, precision of 95.6%, and F1-score of 94.7%.

A sound-based engine classification was proposed by Shajie et al. to detect flaws in engine ball bearings [27]. They used sound-based component extraction techniques to find reoccurring patterns across time. They proposed modifications to the ResNet and hybrid CNN models based on the NASA-bearing dataset. To identify TIM-bearing faults, they employed time and frequency features that may be inferred from the signals and their spectra. The experiments considered realistic scenarios found in real-world industrial settings. They gained insights into the method's performance with reasonable accuracy rates. To improve industrial productivity and minimize machinery downtime, Khan et al. developed a technique for diagnosing robotic manipulator faults using motor sound analysis [28]. They investigate the efficiency of deep learning and conventional ML in detecting motor abnormalities using a dataset created with a specifically designed robotic manipulator. It obtained an F1-score of over 92%, outperforming the traditional methods significantly, hence proving the potential of sound analysis for automatic defect identification in robotic systems using the proposed custom CNN and 1D-CNN models. Kim et al. proposed a deep denoising autoencoder method to filter out various industrial noise levels from audio data [29]. They applied unsupervised learning models for rapid and accurate anomaly detection. They preprocessed audio data to adapt the denoising technique to the noise levels of different industrial contexts. Several experiments using different industrial equipment types demonstrated the proposed technique's effectiveness, efficiency, and processing speed. Senanayaka et al. diagnosed machinery defects by isolating audio sources from complex mixtures of sound waves [30]. First, they activated fault sound isolation and separated distinct fault noises from a complicated blend of

sound signals. Then, the isolated fault noises were passed through a 1D-CNN classifier to ensure correct classification. A machine fault simulator by Spectra Quest equipped with a condenser mic was employed to evaluate the proposed model. To improve early vehicle defect recognition, Hameed et al. investigated the application of ML for real-time engine knocking detection [31]. They analyzed several machine-learning techniques and retrieved frequency modulation amplitude demodulation features from engine sound data. With a classification accuracy of 66.01%, the coarse decision tree approach proved the most successful. The accuracy was then increased by employing deep learning models; a deep learning recurrent neural network (RNN) model in LSTM attained 90% accuracy.

Naryanto et al. developed a deep learning model to detect and classify damage or defects in diesel engines using artificial neural networks and convolutional neural networks [32]. They utilized the DEFault dataset, which has 3500 rows of data organized under four distinct labels. Results showed that ANN outperformed CNN for noisier datasets, but it outperformed for less noisy datasets. Yuan et al. proposed a defect detection approach for new energy vehicle engines using wavelet transforms and Support Vector Machines [33]. First, an abnormal noise signal identification model for vehicle engine faults is developed, and the time–frequency parameters of the basis function are adaptively changed. The engine surface radiation noise is then split into the inner mechanical and battery excitation components. The new energy vehicle engine failure signal was decomposed using feature decomposition and multiscale separation. Furthermore, fuzzy clustering and time–frequency analysis of fault signals in the fractional Fourier domain were used to detect faults in new energy vehicle engines. Chu et al. proposed an intelligent identification model for diesel engine faults based on mixed attention [34]. They proposed a multi-cylinder whole-machine fault diagnosis model that integrates 1D-CNN with self- and mutual attention mechanisms. Single-cylinder sensor data were integrated using self-attention in the model, and signal features of each cylinder were fused using the mutual attention mechanism. Simultaneously considering the mechanism knowledge of cylinder structural consistency and signal time delay similarity, this approach utilized single-cylinder fault data to develop a comprehensive fault recognition model for all cylinders. The average diagnosis accuracy reached 100% in known fault data and about 96.65% in unknown fault data.

Lee et al. proposed a bearing failure detection using an LSTM autoencoder with self-attention based on graph convolution networks [35]. Accordingly, they trained their model using data from the Fault Simulator Testbed and the Case Western Reserve University dataset. Results demonstrated that the proposed model attained an accuracy of 97.3% and 99.9%,, respectively, in the CWRU dataset and Fault Simulator Dataset. Using a single microphone and a data-driven approach, Spadini et al. developed a model for intelligent fault diagnosis in rotating equipment that successfully identified 42 classes of defect kinds and severities [36]. They considered reliable data from the unbalanced MaFaulDa dataset to balance high performance and minimal resource consumption. The model achieved remarkable performance in terms of the analysis by time, frequency, mel-frequency, and statistical parameters with an accuracy of 99.54% and F-Beta of 99.52%. Using sound samples, Gantert et al. proposed a multiclass method for identifying anomalous samples in industrial machinery [37]. Integrating binary models commonly found in the literature aims to improve the model's generality while decreasing the number of classifiers. Using MIMII and ToyADMOS, two industrial sound datasets, they compared the proposed multiclass models with the binary alternative. Experiments revealed that 98% of the Toy-ADMOS dataset and 93% of the MIMII dataset were correctly classified. Table 1 summarizes the papers mentioned in this study highlighting the main characteristics and problems of each one.

Table 1. Comparison between literature review papers.

Study	Domains of Usage	Underlying Methodologies	Dataset	Size	Type	Classification	Accuracy	Problems
[15]	Vehicles	ML	Recordings	351	Car sounds	15 classes	92%	Performance depends on the quality and representativeness of the training data.
[16]	Vehicles	ML	Recordings	555	Car sounds	15 classes	98.6%	Difficult to adapt to new or unexpected problems; require extensive manual tuning.
[17]	Vehicles	NN	Recordings	2	Two Cars	Binary	$R^2 = 0.99$	Limited to the specific sound quality factors considered; may not capture subjective aspects of sound quality; requires labeled data for training.
[18]	Vehicles	ML	Recordings	280	Car sounds	6 classes	92.17%	Sensitive to the choice of hidden nodes; require careful tuning; may not capture all relevant information.
[19]	Vehicles	NN	Recordings	417	Car sounds	Binary	$F1 = 0.86$	Sensitive to noise and may not generalize well to unseen engine types; require careful parameter tuning.
[20]	Engines	DL	Recordings	100	Vehicles/ Induction motors	6 classes	$F1 = 0.95$	Time-frequency transformation can be computationally expensive; performance depends on the effectiveness of the adaptive fault feature band extraction.
[21]	Industrial	ML	MIMII	5101	Machine sounds	5 classes	95%	Handcrafted features may not capture all relevant information; performance may be limited by the quality of the feature selector.
[22]	Rolling bearing	VAE-CNN	CWRU	2048	Rolling bearing	10 classes	96.62%	VAE-CNN models can be complex and computationally expensive to train; require a large amount of data to achieve good performance.
[24]	Railway	ML	Recordings	1600	Turnout switch machine	10 classes	98%	Eigenmode Decomposition can be computationally expensive; SVMs can be sensitive to the choice of kernel and parameters.
[25]	Railway	ML	Recordings	25,000	Commuter train dataset	Binary	97.04%	MFCCs may not capture all relevant information in the sound signal; MLP classifiers can be prone to overfitting.
[26]	EPS motors	LSTM-AE	Recordings	29,759	Rolling bearing	Binary	99.2%	LSTM-AE models can be complex and computationally expensive to train; performance depends on the quality of the reconstruction error metric.
[27]	Vehicles	CNN	NASA Bearing Dataset	9463	Engine Ball Bearing	Binary	91%	CNN models can be data-hungry and computationally expensive . to train; performance depends on the choice of network architecture and hyperparameters
[28]	Robotic manipulator	CNN	Recordings	181	Motors	Binary	92.34%	Custom CNN models may be difficult to generalize to other robotic manipulators; performance depends on the quality of the custom dataset.

Table 1. *Cont.*

Study	Domains of Usage	Underlying Methodologies	Dataset	Size	Type	Classification	Accuracy	Problems
[29]	Industrial	AE	MIMII	5101	Machine sounds	Binary	96.51%	Denoising autoencoders may require careful tuning of the noise level; performance depends on the characteristics of the industrial noise.
[30]	Industrial	CNN	Recordings	60 stem files	Engine Ball Bearing	4 classes	99.58%	Audio source isolation can be challenging in complex environments; 1D-CNN models may not capture all relevant spatial information in the sound field.
[31]	Vehicles	DL	Recordings	153	Engine Sounds Knocking	Binary	90%	LSTM models can be computationally expensive; performance depends on the quality of the frequency modulation amplitude demodulation features.
[32]	Diesel engines	CNN	DEFault	3500	Engine Sounds	4 classes	99.37%	ANN performance can be sensitive to parameter initialization and network architecture; CNN performance dependent on the noise level; dataset size.
[33]	New energy vehicles	ML	Recordings	N/A	Engine Sounds	Binary	90%	Wavelet transforms can be computationally expensive; SVMs can be sensitive to the choice of kernel and parameters.
[34]	Diesel engines	Mixed attention	N/A	N/A	Engine Sounds	Binary	98.17%	Complexity, computational cost, reliance on single-cylinder data for whole-machine diagnosis.
[35]	Rolling bearing	AE with Self-Attention	CWRU	N/A	Bearing fault	N/A	97.3%	Complexity, computational cost, potential overfitting, and reliance on specific datasets.
[36]	Industrial	ML	MaFaulDa	1951	Machine sounds	6 classes	99.54%	Complexity, computational cost, difficulty generalizing to new datasets.
[37]	Industrial	ML	MIMII	5101	Machine sounds	4 classes	93%	Possible overfitting for limited datasets, computational cost for multiclass integrations, and difficulty generalizing to new datasets.
			ToyADMOS	N/A		3 classes	98%	

Research gap: While Intelligent Transportation Systems have advanced significantly through vision-based technologies, a critical gap exists in integrating sound-based fault detection mechanisms. This gap is particularly evident in three areas: (1) the limited development of audio-based diagnostic systems that utilize real-time analysis of vehicle-generated sounds (e.g., engine or brake noises) and external emergency alert sounds (e.g., sirens), (2) the scarcity of comprehensive public datasets designed explicitly for vehicle sound analysis, and (3) insufficient attention to accessibility needs for individuals with disabilities within ITS frameworks. These limitations hinder the development of more inclusive and comprehensive transportation monitoring systems. To address these limitations, this study aims to develop a comprehensive dataset that serves as a "conscious ear" for intelligent systems in modern cities, vehicles, and transportation networks. This effort seeks to enhance the auditory capabilities of smart systems, enabling them to respond effectively to complex auditory scenarios, thereby enhancing safety and functionality in various applications.

3. Dataset Creation

The main problem with transportation sound-based fault diagnosis is the availability of datasets. Therefore, data from various sources are collected to create a tailored dataset for car sound analysis. Reliable audio samples are built by downloading videos from YouTube related to car faults, animal sounds, car crashes, siren sounds, etc. After this step, the model splits videos into segments and extracts those sections that may contain the target audio. Then, it converts the files into audio format. To expand the dataset, additional audio is supplemented from Kaggle datasets: FSC22 [38], Google AudioSet [39], Audio Classifier Dataset [40], Sound Classification of Animal Voice [41], and Vehicle Sounds Dataset [42]. Lastly, the model ensures that every sample is labeled and verified.

In this vein, the dataset was created and reviewed using a combination of publicly available datasets and real-world recordings, covering a wide range of vehicle faults, crashes, emergency sirens (police, ambulance, fire truck), wild animal sounds, car and truck horns, and other environmental road sounds. This approach ensures a diverse and realistic dataset that enhances model performance in detecting road-related events. The dataset creation procedure involves different stages as shown in Figure 1.

3.1. Data Collection and Annotation

We carefully selected publicly available datasets that include real traffic scenarios and vehicle fault cases. Additionally, we extracted relevant frames and sequences from YouTube videos, ensuring a diverse representation of traffic conditions and vehicle behaviors. Each data sample was manually labeled based on predefined criteria, focusing on vehicle states, traffic interactions, and specific fault conditions.

3.2. Expert Review and Validation

To enhance reliability, domain experts with extensive experience in automotive engineering and machine learning reviewed the dataset. The experts cross-checked and validated the labels to ensure accuracy and consistency with real-world vehicle behaviors.

3.3. Publicly Available Datasets Referenced

We utilized multiple datasets containing wild animal sounds, vehicle faults, and environmental noises to build a comprehensive dataset. The key datasets referenced include:

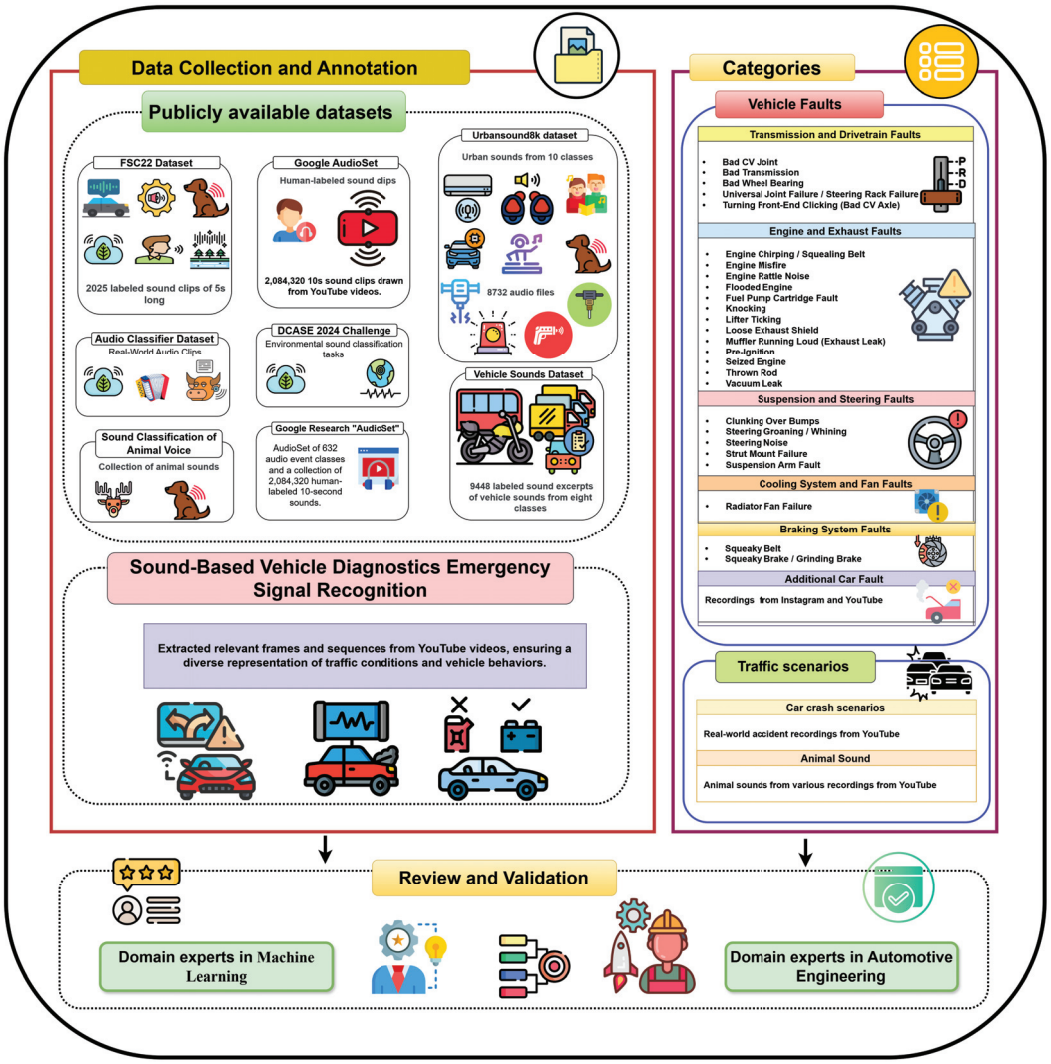


Figure 1. Dataset creation.

1. FSC22 Dataset: A collection focused on various sound categories, including vehicle sounds and environmental noises, useful for sound classification models [38].
2. Google AudioSet: A large-scale collection of audio data across thousands of categories, aimed at improving sound classification models [39].
3. Audio Data: A dataset containing diverse audio clips across various categories, useful for developing classification models [40].
4. Sound Classification of Animal Voice: A dataset containing sounds from different animals, useful for animal sound classification tasks [41].
5. DCASE 2024 Challenge: A dataset designed for the DCASE 2024 challenge, covering environmental sound classification tasks [43].
6. UrbanSound8K Dataset: Contains 8732 labeled sound excerpts from urban environments, categorized into 10 classes such as car horns and sirens [44].
7. AudioSet by Google Research: A vast dataset with over 2 million human-labeled 10 s sound clips spanning thousands of audio categories [45].
8. Vehicle Sounds Dataset: Contains various vehicle sounds useful for training models focused on transportation-related sound classification [42].

This dataset has been meticulously designed and validated to provide a diverse and realistic representation of road-related sounds, ensuring high-quality training data for machine learning models in the domain of automotive fault detection, traffic event classification, and environmental sound analysis. To standardize the process, files should

have the same duration; however, this is not the case. Preprocessing is performed as a solution to ensure samples of the same duration. Algorithm 1 contains the pseudo-code for this stage. The primary steps in the algorithm are:

1. Repeat the audio until it achieves the required duration.
2. Normalize the audio to standardize levels.
3. Resample the audio to a consistent sampling rate (e.g., 16 kHz or 44.1 kHz).
4. Segment lengthy recordings into shorter clips (e.g., 2–5 s each).

Algorithm 1: Pseudo-code for Audio Preprocessing Script

Input: Input audio file (F)
Output: Preprocessed Audio segments (F_P)
$rate \leftarrow$ Target sample rate = 16 kHz
$dur \leftarrow$ Audio segment duration = 2500 ms
$dur_{min} \leftarrow$ Audio file minimum duration = 10 s
if $(dur(F) < dur_{min})$ **then**
    $rep \leftarrow \lceil dur_{min}/dur(F) \rceil$ `// No. of repetitions`
    `// Extend Audio to Minimum Duration`
    for $i \leftarrow 0$ **to** rep **do**
        $F_{Ext} \leftarrow F_{Ext} + F$
    $F_{Ext} \leftarrow trim(F_{Ext}, dur)$
else
    $F_{Ext} \leftarrow F$
$F_{norm} \leftarrow normalize(F_{Ext})$ `// Normalize audio levels`
$F_{norm} \leftarrow resample(F_{norm}, rate)$ `// Resample audio to TARGET SAMPLE RATE`
if $F_{norm} > dur_{min}$ **then**
    $F_P \leftarrow segment(F_{norm}, dur_{min})$
else
    $F_P \leftarrow F_{norm}$
Return F_P

4. Methodology

This study aims to develop an advanced sound-based early diagnosis system to support Intelligent Transportation Systems (ITS) by enabling real-time detection of vehicle faults and identification of emergency sounds. The main steps of this system are illustrated in Figure 2. To achieve this objective, addressing the primary challenges encountered in this field is essential, starting with the absence of comprehensive public datasets specifically designed for vehicle sound analysis. The initial phase of the proposed model involves the creation of a dataset that contains recordings of car fault sounds, emergency sirens, and ambient noises. This process includes audio data collection and preprocessing. Subsequently, the most significant features are extracted in two versions: a compact version with 52 features and an expanded one with 126 features.

In the final step, both sets of extracted features are classified using 11 distinct ML models. Another phase of optimization is provided by the system to enhance the accuracy of classification. It utilizes the best ML models with the highest-ranked features to build an ensemble optimization model. The following subsections provide further details on each stage of this model.

The key steps in our audio preprocessing pipeline to provide further insight into how the system operates are as follows:

1. Fixed Time Windows for Feature Extraction:

- We preprocess the raw audio files by extending them to a minimum duration of 10 s (MIN_DURATION_MS = 10,000 ms), normalizing their levels, and resampling them to a target sample rate of 16 kHz (TARGET_SAMPLE_RATE = 16,000 Hz).
- The preprocessed audio is then segmented into fixed-length clips of 2.5 s (SEGMENT_DURATION_MS = 2500 ms).
- Only segments meeting the required length are retained for further processing.

2. Time Window Length Determination:
   - The nature of the vehicle fault guided the choice of segment duration sounds, which are typically periodic and repetitive over short durations.
   - To ensure consistency across all samples, we set a minimum duration of 10 s for all audio files. If an audio file is shorter than this threshold, it is repeated and trimmed to match the minimum duration.
   - The fixed 2.5 s time window used for feature extraction ensures that features such as MFCCs, Mel Spectrogram, and Chroma Features capture sufficient temporal and spectral characteristics of the sounds.
3. Sliding Factor Consideration:
   - A fixed windowing approach is used over overlapping sliding windows during segmentation. This ensures non-redundant segments while maintaining dataset balance.
   - However, future work could explore the impact of using overlapping windows to capture more temporal variations while controlling data redundancy.

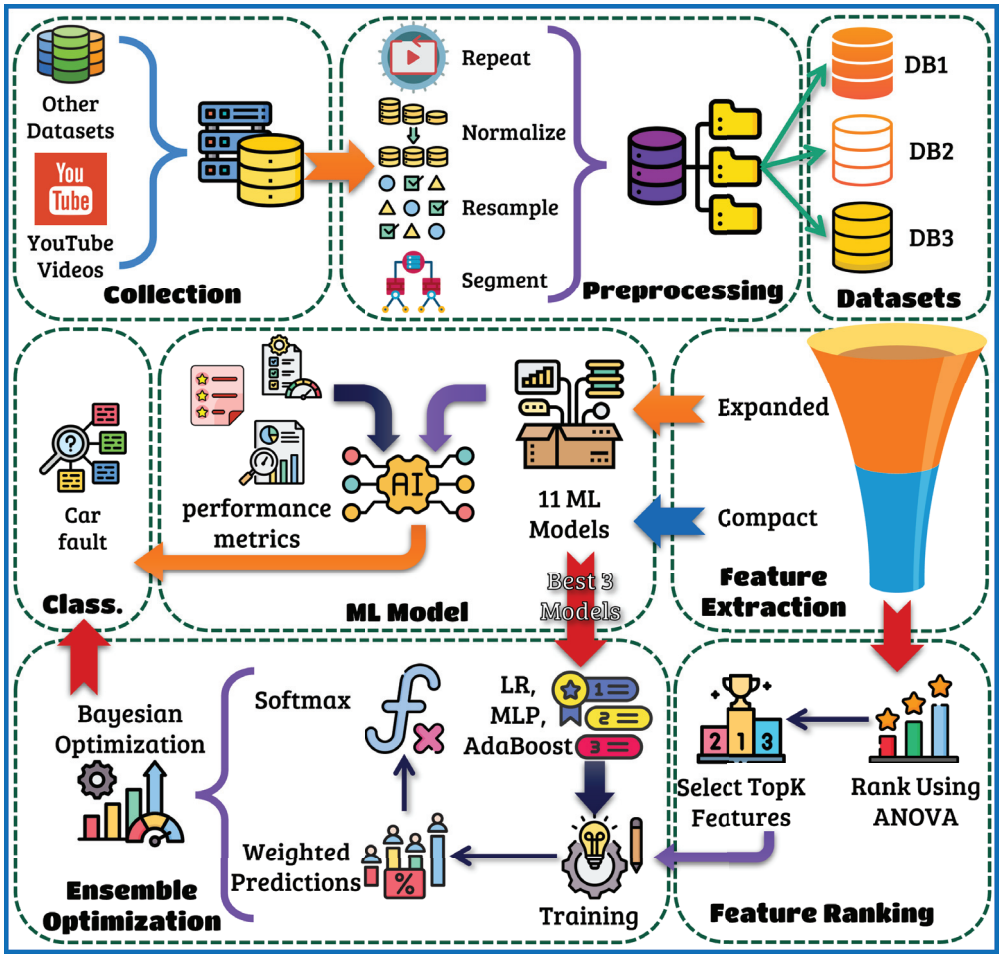


Figure 2. Main system phases and steps.

This structured approach ensures that the extracted sound features represent the fault categories well while maintaining computational efficiency.

By the end of this phase, audio files are sampled, labeled, and normalized to build the dataset. Three datasets are created: the first one contains car faults (DB1) with 133 audio files and 27 distinct classes, the second dataset contains other sounds (DB2) with 1031 audio files and 22 distinct classes, and the third dataset is a merged version between the latter two (DB3) with 1164 audio files and 49 different classes. Tables 2 and 3 show the labels and the corresponding file counts for DB1 and DB2, respectively.

Table 2. DB1 labels and file counts.

Labels	Count	Labels	Count
Bad Wheel Bearing	21	Squeaky Belt	4
Universal Joint Failure or Steering Rack Failure	10	Seized Engine	4
Knocking	5	Pre-Ignition	4
Wheel Bearing, Transmission Whining Noise and Catalytic Converter Issues	5	Bad Transmission	4
Bad CV Joint	4	Strut Mount Failure	4
Radiator Fan Failure	4	Lose Exhaust Shield	4
Turning Front End Clicking Bad CV Axle	4	Lifter Ticking	4
Steering Noise	4	Flooded Engin	4
Steering Groaning Whining Low Power Steering Fluid	4	Engine Rattle Noise	4
Squeaky Brake/Grinding Brake	4	Engine Misfire	4
Muffler Running Loud Exhaust Leak	4	Thrown Rod	4
Clunking Over Bumps Bad Stabilizer Link Noise	4	Suspension Arm Fault	4
Engine Chirping/Squealing Belt	4	Vacuum Leak	4
Fuel Pump Cartridge Fault	4	Total (27)	133

Table 3. DB2 labels and file counts.

Group	Labels	Count	Group	Labels	Count	Group	Labels	Count
Animals	Cats	200	Vehicles and Transportation	Car Crashes	103	Emergency Vehicles	Police Car Siren	41
	Sheep	80					Fire Truck Siren	37
	Bear	68		Car Horn	24		Ambulance Siren	30
	Dog	68		Motorcycle	20	Construction and Machinery	Drilling	24
	Monkey	60		Bus	20			
	Lions	48		Bike	20	Weapons and Explosions	Gunshot	14
	Wolf	47		Train	20			
	Horse	40		Truck	20	Total (22)		1031
	Mouse	28		Truck Horn	19			

4.1. Feature Extraction

After preparing the datasets, the next stage in the proposed system is feature extraction. Audio feature extraction is a significant task in processing an audio signal for the purpose of sound classification. From an audio signal, meaningful features can be extracted to analyze and understand the content of the audio. Figure 3 shows some key features commonly extracted from audio signals.

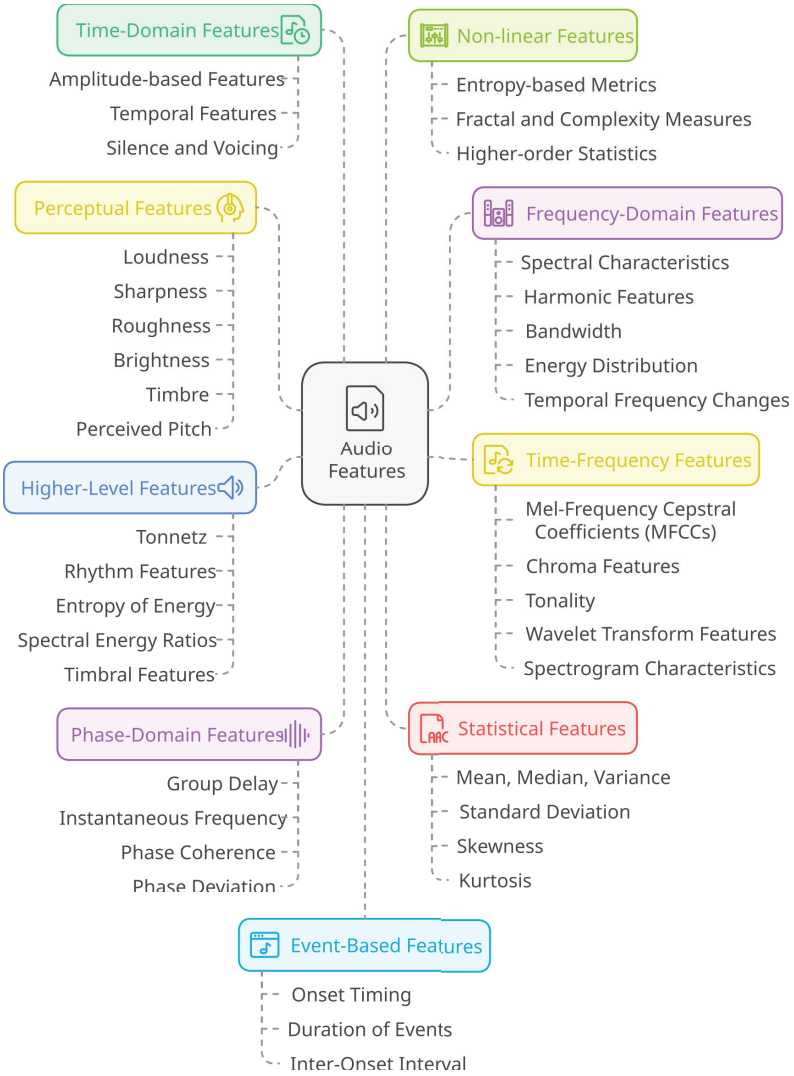


Figure 3. Common audio features.

The essential features in our study are extracted in two versions: a compact version with 52 features and an expanded one with 126 features. In the compact version, Mel Spectrogram [46], MFCCs [47], and Chroma Features [48] were used. Figure 4 shows an example of a Mel Spectrogram.

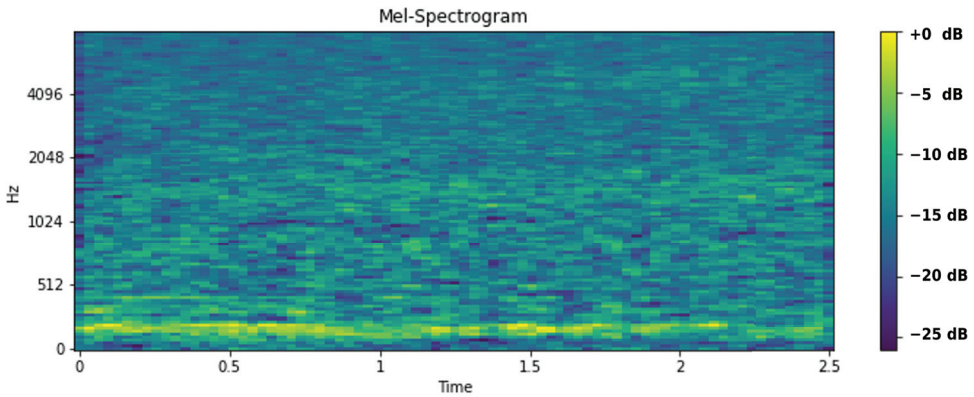


Figure 4. Example of a Mel Spectrogram of a bad CV joint.

For generating the expanded version, Spectral Features [49], Zero-Crossing Rate [19], Root Mean Square Energy (RMSE) [50], Chroma Features, MFCCs, and Extended MFCCs [51] were used. Table 4 defines these features, including their counts and the version(s), Compact (C) or Expanded (E), they appeared in. Figure 5 shows two-dimensional data projection DB1 compact features.

Table 4. Features extracted in the study.

Features Domain	Feature	Definition	Count	Usage
Time-Domain	Zero-Crossing Rate	How often the signal changes from positive to negative.	2	E
	RMSE	Measures the energy of the audio signal.	2	E
Frequency-Domain (Spectral Features)	Spectral Centroid	Indicates where the center of mass of the spectrum is located.	2	E
	Spectral Bandwidth	Measures the width of the spectrum around its centroid.	2	E
	Spectral Roll-off	The frequency below which 85% of the total spectral energy resides.	2	E
	Spectral Contrast	Refers to the difference in amplitude between peaks and valleys in the spectrum of an audio signal.	14	E
Time-Frequency	Mel Spectrogram	A visual representation of the frequency spectrum of an audio signal over time.	2	C
	Chroma Features	Represent the energy distribution across the 12 different pitch classes (notes) of the Western music scale.	24	C & E
	MFCCs	Capture spectral features related to the timbre of audio. They represent the short-term power spectrum of sound.	26	C & E
	MFCCs Delta		26	E
	MFCCs Delta2		26	E

The pseudo-code for extracting compact and expanded feature lists from audio files are listed in Algorithms 2 and 3, respectively.

Algorithm 2: Pseudo-code for extracting Compact feature list

Input: Input audio file (F)
Output: Compact features list ($Feat_{Comp}$)
$Feat_{Comp}, Features \leftarrow Emptylist$
$rate \leftarrow 16kHz$
$F \leftarrow resample(F, rate)$ `// Resample audio to Target sample rate`
$Mel \leftarrow melspectrogram(F)$
$Mel_{db} \leftarrow power_to_db(Mel)$
$MFCC \leftarrow mfcc(F, 13)$ `// 13 features of MFCC`
$Chroma \leftarrow chroma(F)$ `// 12 classes of Chroma`
$Features \leftarrow Mel_{db} + MFCC + Chroma$
foreach f *in* $Features$ **do**
    $f_{mean} \leftarrow mean(f)$
    $Feat_{Comp} \leftarrow Feat_{Comp} + f_{mean}$
    $f_{std} \leftarrow std(f)$
    $Feat_{Comp} \leftarrow Feat_{Comp} + f_{std}$
Return $Feat_{Comp}$

Algorithm 3: Pseudo-code for extracting Expanded feature list

Input: Input audio file (F)
Output: Expanded features list ($Feat_{Exp}$)
$Feat_{Exp}, Features \leftarrow Emptylist$
$rate \leftarrow 16kHz$
$F \leftarrow resample(F, rate)$ `// Resample audio to Target sample rate`
$ZCR \leftarrow zero_crossing_rate(F)$
$RMS \leftarrow root_mean_square(F)$
$S_{centroid} \leftarrow spectral_centroid(F)$
$S_{bandwidth} \leftarrow spectral_bandwidth(F)$
$S_{rolloff} \leftarrow spectral_rolloff(F)$
$S_{contrast} \leftarrow spectral_contrast(F)$ `// 7 values of contrast`
$Chroma \leftarrow chroma(F)$ `// 12 classes of Chroma`
$MFCC \leftarrow mfcc(F, 13)$ `// 13 features of MFCC`
$MFCC_{delta} \leftarrow mfcc_delta(MFCC)$
$MFCC_{delta_delta} \leftarrow mfcc_delta(MFCC_{delta})$
$Features \leftarrow ZCR + RMS + S_{centroid} + S_{bandwidth} + S_{rolloff} + S_{contrast} + Chroma + MFCC + MFCC_{delta} + MFCC_{delta_delta}$
foreach f *in* $Features$ **do**
    $f_{mean} \leftarrow mean(f)$
    $Feat_{Exp} \leftarrow Feat_{Exp} + f_{mean}$
    $f_{std} \leftarrow std(f)$
    $Feat_{Exp} \leftarrow Feat_{Exp} + f_{std}$
Return $Feat_{Exp}$



Figure 5. Two-dimensional data projection using t-SNE for compacted features extracted from DB1.

4.2. Classification

In this proposed system, the input audio is classified using ML techniques. The two versions of feature lists are used to test eleven different models on the three datasets created.

Neural Network (NN): A computational model consisting of interconnected neurons [52]. It is used for both regression and classification tasks. A neural network can be formulated by:

$$Y = f(WX + b)$$

where f is an activation function, W are weights, x is input, and b is bias.

Naive Bayes (NB): A probabilistic classifier based on Bayes' theorem, assuming independence among predictors [53]. The NB equation is given by:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

where C is the class and X is the feature vector.

Logistic Regression (LR): A statistical method for predicting binary classes [54]. The outcome is modeled using a logistic function, which outputs probabilities. Logistic Regression is formulated as follows:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)}}$$

Stochastic Gradient Descent (SGD): An iterative method for optimizing an objective function, commonly used in training ML models, particularly neural networks [55].

$$\theta = \theta - \eta \nabla J(\theta)$$

where η is the learning rate and $\nabla J(\theta)$ is the gradient of the loss function.

k-Nearest Neighbors (kNN): A non-parametric method used for classification and regression by finding the k nearest data points in the feature space [56].

$$\hat{y} = \text{mode}(y_i) \text{ for } i \in \text{k nearest neighbors}$$

Decision Tree (DT): A flowchart-like structure where each internal node represents a feature test, each branch represents an outcome, and each leaf node represents a class label [57].

$$\text{Class} = \text{leaf node based on features}$$

Random Forest (RF): An ensemble learning method that constructs multiple decision trees during training and outputs the mode of their predictions for classification tasks [58].

$$H(x) = \text{mode}(h_1(x), h_2(x), ..., h_T(x))$$

Support Vector Machine (SVM): A supervised learning model that finds the optimal hyperplane that best separates different classes in the feature space [59].

$$f(x) = \text{sign}(W \cdot X + b)$$

CN2 Rule Induction: An algorithm for inducing classification rules from examples [60]. It generates rules based on the attributes of the training data.

$$\text{Class} = \text{if } (A_1 \wedge A_2 \wedge ... \wedge A_n) \text{ then } C$$

where A_i are conditions based on attributes and C is the class label.

Adaptive Boosting (AdaBoost): An ensemble method that combines multiple weak classifiers to create a strong classifier by focusing on errors made by previous classifiers [61].

$$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$$

where α_t is the weight of the classifier h_t.

Gradient Boosting (GB): An ensemble technique that builds models sequentially, with each new model correcting errors made by the previous ones [62].

$$F(x) = F_{m-1}(x) + \nu h_m(x)$$

where ν is the learning rate and h_m is the new model.

4.3. Feature Ranking

Feature ranking is considered a very crucial step in machine learning and data analysis and can select relevant features that contribute to the predictive models [63]. Four important feature ranking methods have been applied to estimate the importance of the different features used in previous classification experiments, including information gain, analysis of variance, ReliefF, and fast correlation-based filters. **Information Gain (IG)** measures the reduction in entropy or uncertainty after splitting a dataset based on a feature [64]. IG calculates the difference between the entropy of the target variable and the conditional entropy given the feature. Features with higher IG values are more informative.

$$IG = H(Y) - H(Y|X)$$

where $H(Y)$ is the entropy of the target variable, and $H(Y|X)$ is the conditional entropy.

Analysis of Variance (ANOVA) measures the ratio of between-class variance to within-class variance for a feature [65]. Features with higher ANOVA values are more discriminative.

$$ANOVA = \frac{Variance_{between-classes}}{Variance_{within-class}}$$

ReliefF is an extension of the Relief algorithm, estimating feature relevance by measuring the difference between the feature's values for nearest neighbours from different classes [66].

$$W(F) = W(F) + \frac{1}{k}\sum_{j=1}^{k}(\Delta(F, H_j) - \Delta(F, M_j))$$

where $W(F)$ is the current weight of feature F, k is the number of nearest neighbors considered, H_j are the nearest neighbors from the same class (hits), and M_j are the nearest neighbors from different classes (misses).

Fast Correlation-Based Filter (FCBF) evaluates feature relevance using correlation and redundancy [67]. It selects features with a high correlation to the target variable and low redundancy.

$$FCBF = Correlation(X, Y) - Redundancy(X)$$

4.4. Bayesian-Optimized Weighted Soft Voting with Feature Selection (BOWSVFS)

Ensemble learning is currently one of the most powerful methods in machine learning, combining many models to produce predicted performance better to that of standalone models [68]. Optimal combination in model weight determination remains an optimization challenge. To address this issue, the proposed approach employs Bayesian Optimization combined with Weighted Soft Voting.

In this model, WSV is the central process, with a few classifiers voting on a final prediction and assigning weights to each. Soft voting, unlike hard voting, does not involve direct class prediction; instead, it uses cross-class probability distributions [69]. Each prediction is weighted based on the perceived importance of each classifier in the entire ensemble, and the weighted probabilities are combined to obtain the final prediction.

The weights assigned to each classifier play a crucial role in the ensemble's performance. Traditional approaches often use equal weights or weights determined through grid search. However, these methods can be computationally expensive and may not find the optimal weight configuration, especially when dealing with multiple classifiers and features. Algorithm 4 depicts the procedures for calculating WSV. Figure 6 shows the steps of Bayesian-Optimized Weighted Soft Voting procedure.

Algorithm 4: Pseudo-code for calculating WSV

Input: Dataset (D), weights (W), No. of features (F)
Output: Mean Accuracy (ACC_{mean})
$FOLDS_{num} \leftarrow 10$
$SEED_{rand} \leftarrow 42$
$X, l \leftarrow SeparateFeaturesAndLabels(D)$
$encoder \leftarrow LabelEncoder()$ `// Label Encoding`
$L_{encoded} \leftarrow encoder.Fit_Transform(l)$
`// Feature Ranking`
$selector \leftarrow SelectKBest(rankfunc, F)$
$F_{selected} \leftarrow selector.Fit_Transform(X, L_{encoded})$
`// Feature Scaling`
$scaler \leftarrow StandardScaler()$ $F_{scaled} \leftarrow scaler.Fit_Transform(F_{selected})$
`// Initialize Classifiers`
$classifiers \leftarrow [LR(SEED_{rand}), MLP(SEED_{rand}), AdaBoost(SEED_{rand})]$
`// Initialize Cross-Validation`
$cv_splits \leftarrow CreateKFold(FOLDS_{num}, shuffle, SEED_{rand})$
$ACC_{scores} \leftarrow []$
`// Perform Cross-Validation`
foreach i_{train}, i_{test} *in* cv_splits **do**
    $X_{train} \leftarrow F_{scaled}[i_{train}]$
    $l_{train} \leftarrow L_{encoded}[i_{train}]$
    $X_{test} \leftarrow F_{scaled}[i_{test}]$
    $l_{test} \leftarrow L_{encoded}[i_{test}]$
    $Pred_{probability} \leftarrow []$
    `// Train each classifier and get predictions`
    foreach *classifier in classifiers* **do**
        $classifier.Fit(X_{train}, l_{train})$
        $probs \leftarrow classifier.PredictProbabilities(X_{test})$
        $Pred_{probability}.append(probs)$
    `// Initialize weighted sum matrix with zeros`
    $Sum_{weighted} \leftarrow ZeroMatrix(LENGTH(Pred_{probability}[0]), CLASSES)$
    `// Combine predictions with weights`
    for $i \leftarrow 0$ *to length(W)* **do**
        $Sum_{weighted} \leftarrow Sum_{weighted} + W[i] \times Pred_{probability}[i]$
    $Prob \leftarrow Softmax(Sum_{weighted})$ `// Apply softmax activation`
    $class_{predicted} \leftarrow ArgMax(Prob)$ `// Get class with highest probability`
    $ACC \leftarrow CalculateAccuracy(l_{test}, class_{predicted})$ `// Calculate accuracy`
    $ACC_{scores}.append(ACC)$
$ACC_{mean} \leftarrow mean(ACC_{scores})$ `// Calculate mean accuracy`
Return ACC_{mean}

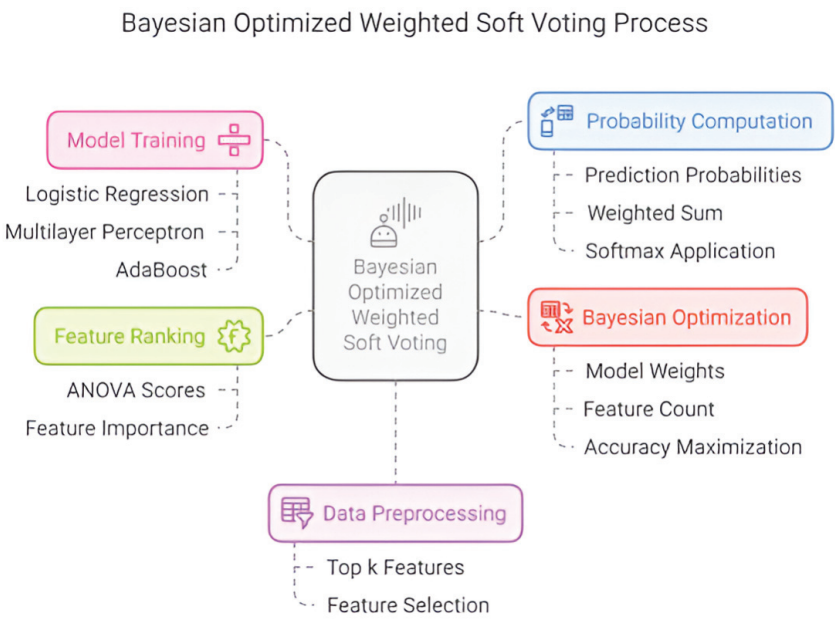


Figure 6. Steps for Bayesian-Optimized Weighted Soft Voting.

Bayesian optimization provides a more systematic and efficient technique to determining optimal weights [70]. Bayesian optimization uses a probabilistic model to predict the link between hyperparameters, weight, and feature counts, as well as model performance. This model is typically a Gaussian Process variation. This strategy is particularly useful since it swiftly explores the hyperparameter space by creating a proxy model of the objective function. It strikes a compromise between exploring unknown regions and exploiting known favorable locations, requiring fewer iterations than grid or random search algorithms. The steps for implementing the proposed Bayesian Optimization for Weighted Soft Voting are given in Algorithm 5, and these steps are:

1. **Feature Ranking Using ANOVA**: Rank features based on ANOVA (analysis of variance) scores.
2. **Data Preprocessing**: Select the top-k features based on ANOVA ranking using the following equation.

$$F_k = \text{SelectTopKFeatures}(F, \text{ANOVA}, k)$$

3. **Training Three Models**: Train Logistic Regression (LR), Multilayer Perceptron (MLP), and AdaBoost using the selected features with 10-fold cross-validation to ensure robustness using the following equation.

$$p^{(j)}(x) = f_j(x|F_k), \quad j \in \{LR, MLP, AB\}$$

4. **Computing Weighted Probabilities**:
   - Calculate the prediction probabilities for each model.
   - Compute the weighted sum of these probabilities using the following equation.

$$P_{sum}(c|x) = \sum_{j \in \{LR,MLP,AB\}} w_j p^{(j)}(x)_c$$

   - Apply softmax to the final weighted sum of probabilities using the following equation.

$$P(c|x) = \frac{\exp(P_{sum}(c|x))}{\sum_{c'} \exp(P_{sum}(c'|x))}$$

5. **Bayesian Optimization**: Optimize the model weights (w_1, w_2, w_3) and the feature count (k) using Bayesian optimization techniques to maximize accuracy using the following equation.

$$\{w_1^*, w_2^*, w_3^*, k^*\} = \arg \max_{w_1, w_2, w_3, k} \text{Accuracy}(y, \hat{y}(x))$$

6. **Final Predictions Using Optimized Parameters**: Use the optimized weights and features for the final soft voting decision using the following equation.

$$\hat{y}(x) = \arg\max_c P(c|x)$$

Algorithm 5: Pseudo-code for Bayesian Optimization Weighted Soft Voting

Input: Dataset (D)
Output: Optimal Weights (W_{opt}), Optimal Features (F_{opt}), Best Accuracy (ACC_{best})
$PARAM_BOUNDS \leftarrow [w_1, w_2, w_3 \in [0,1], F \in [20,30]]$ `// Parameter space bounds`
$MAX_ITERATIONS \leftarrow 20$
$INITIAL_POINTS \leftarrow 10$
$SEED_{rand} \leftarrow 42$
`// Initialize Bayesian Optimization`
$optimizer \leftarrow BayesianOptimizer(ObjectiveFunction, PARAM_BOUNDS, SEED_{rand})$
`// Perform optimization`
for $i \leftarrow 1$ *to* $INITIAL_POINTS$ **do**
    $optimizer.SampleRandomPoint()$
for $iteration \leftarrow 1$ *to* $MAX_ITERATIONS$ **do**
    `// Sample next point based on acquisition function`
    $point_{Next} \leftarrow optimizer.SuggestNextPoint()$
    `// Normalize weights`
    $W_{sum} \leftarrow point_{Next}.w_1 + point_{Next}.w_2 + point_{Next}.w_3$
    $W_{norm} \leftarrow [point_{Next}.w_1 / W_{sum}, point_{Next}.w_2 / W_{sum}, point_{Next}.w_3 / W_{sum}]$
    `// Perform weighted soft voting with the given parameters`
    $ACC \leftarrow CalculateWSVA(D, W_{norm}, point_{Next}.F)$
    `// Update optimizer with results`
    $optimizer.UpdateModel(point_{Next}, ACC)$
`// Extract best parameters`
$P_{best} \leftarrow optimizer.GetBestParameters()$
$ACC_{best} \leftarrow optimizer.GetBestScore()$
`// Normalize final weights`
$W_{sum} \leftarrow P_{best}.w_1 + P_{best}.w_2 + P_{best}.w_3$
$W_{opt} \leftarrow [P_{best}.w_1 / W_{sum}, P_{best}.w_2 / W_{sum}, P_{best}.w_3 / W_{sum}]$
$F_{opt} \leftarrow P_{best}.F$
Return $W_{opt}, F_{opt}, ACC_{best}$

Implementing the proposed approach in practice can be performed by detecting the onset of an alert or emergency sound through a preprocessing step such as a sound event or voice activity detection (VAD). By distinguishing between background noise and relevant sound events, these techniques can help the system identify when a sound starts, even in noisy environments. Additionally, microphones or sensors can be strategically placed in or around the vehicle to capture sound more accurately, such as in isolated engine compartments with noise-canceling technology to improve sound capture quality. Also, the system can be integrated with existing vehicle monitoring systems to automatically trigger sound detection when certain conditions are met, such as abnormal engine behavior,

sudden changes in vehicle speed, or other sensor data that might indicate a fault or emergency event.

5. Experiments and Discussion

In this study, we evaluated the performance of eleven distinct machine learning models on three datasets, utilizing two versions of feature lists: a compact version comprising 52 features and an expanded version consisting of 126 features. The models were assessed based on several performance metrics, including Area Under the Curve (AUC), Classification Accuracy (CA), F1-score (F1), Precision (Prec), Recall, Matthews Correlation Coefficient (MCC), Specificity (Spec), and Logarithmic Loss (LogLoss).

Data acquisition involved online tools for downloading YouTube videos, while segmenting and audio extraction utilized FFmpeg (v6.0) and Veed.io. Audio conversion was performed using Online Audio Converter. Audio processing, feature extraction, and analysis were conducted using Python (v3.12) (Jupyter Notebook and Spyder IDE) on a computer equipped with an Intel Core i7 processor and 16 GB RAM. Key libraries employed include Librosa (v0.10.1) and Pydub (v0.25.1). All processes were completed using standard software tools. To ensure transparency and reproducibility, all datasets and codes are publicly available in our GitHub (v3.15) repository, along with comprehensive documentation and 72 references for dataset collection, including publicly available sources and YouTube audio samples.

5.1. Performance Metrics

This study used several performance indicators to analyze the efficiency of the models under evaluation. One significant metric utilized is the Area Under the Curve (**AUC**), which measures a model's ability to distinguish between positive and negative classes by computing the Area under the Receiver Operating Characteristic (ROC) curve. The AUC is determined using the following formula:

$$\text{AUC} = \int_0^1 \text{ROC}(x)dx$$

where ROC is the true positive rate plotted against the false positive rate at various threshold settings.

Another important metric is **Classification Accuracy (CA)**, calculated as the ratio of correctly predicted instances to the total number of instances in a dataset. The formula runs as follows:

$$\text{CA} = \frac{TP + TN}{TP + TN + FP + FN}$$

where: TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives. **Precision** is the ratio of true positive predictions the model provides to total positive predictions, indicating how precise the model is when producing positive predictions. Precision can be calculated as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall, also known as sensitivity, is the ratio of true positive predictions compared to all actual positive instances. It is one of the key metrics for evaluating the performance of a predictive model by its ability to identify positive instances correctly. It is calculated using the following formula:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Specificity, on the other hand, measures the proportion of true negative predictions among all actual negative instances and is calculated as:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

The **F1-score** comprehensively evaluates a model's performance by calculating the harmonic mean of Precision and recall. It's calculated with the following formula:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{ Recall}}{\text{Precision} + \text{ Recall}}$$

The Matthews Correlation Coefficient (**MCC**) is a well-balanced measure that considers all four categories of the confusion matrix, providing a more comprehensive metric for binary classification. The MCC is calculated as follows:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Logarithmic Loss (**LogLoss**) is a metric used to evaluate the performance of a classification model. It measures the accuracy of the probabilities assigned to each class. The LogLoss is calculated as:

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1 - y_i)\log(1 - p_i))$$

where N = total number of instances, y_i = actual label (0 or 1), p_i = predicted probability of the positive class.

5.2. Hyperparameters

Hyperparameter tuning is one of the essential steps in any machine learning classification process [71,72]. It basically involves the selection of the best hyperparameters that control the learning process of a model. These hyperparameters are fixed before training since they are not learned from the data during training. Well-set hyperparameters can significantly improve the accuracy of a model and its generalization on unseen data. They form the basis for finding a good trade-off between bias and variance. Appropriate hyperparameters can speed up the training process, which means faster model development and deployment. Efficient hyperparameter settings can optimize resource utilization, thus reducing training time and costs.

Table 5 lists the configurations used for each model. The following are some common hyperparameters and their effect:

- **Learning Rate**: This hyperparameter controls the step size during gradient descent when moving toward the minimum. A very high learning rate results in instability, while a very low one slows down training.
- **Number of Trees/Estimators**: The number of trees in ensembling techniques like Random Forest and Gradient Boosting. More trees provide higher accuracy, but training a model takes longer.
- **Tree Depth**: The hyperparameter for tree-based models defines each tree's maximum depth. Deep trees can easily capture complex patterns but tend to overfit much more.
- **Regularization**: Methods such as L1 and L2 regularization prevent overfitting by penalizing large weights. The strength of regularization is a hyperparameter that needs tuning.
- **Number of Hidden Layers and Neurons**: This governs the model's architecture in neural networks.

Cross-validation is one of the best methods for hyperparameter tuning, and it was employed in this study. It evaluates model performance using techniques such as k-fold cross-validation to obtain a more accurate assessment of its performance. The sampling type used was a 10-fold cross-validation.

Table 5. Hyperparameters for each model and their values.

Model	Hyperparameter	Value	Model	Hyperparameter	Value
SGD	Algorithm		CN2 Rule Inducer	Rule ordering	Ordered
	Classification loss function	Hinge		Covering algorithm	Exclusive
	Regression loss function	Squared Loss		Rule search	
	Regularization method	Ridge (L2)		Evaluation measure	Entropy
	Regularization strength (α)	0.00001		Beam width	5
	Learning parameters			Rule filtering	
	Learning rate	Constant		Min. rule coverage	1
	Initial learning rate	0.01		Max. rule length	5
Neural Network	No. of hidden neurons	100	SVM	Cost	1
	Activation	ReLU		Regression loss (ϵ)	0.1
	Solver	SGD		Kernel	RBF
	Regularization (α)	0.0001		Gamma (γ)	Auto
	No. of iterations	1000		Numerical tolerance	0.001
	Training	Replicable		Iteration limit	100
Decision Tree	Type	Induce binary tree	AdaBoost	Base estimator	Tree
	Min. instances in leaves	2		No. of estimators	50
	Split subsets	>5		Learning rate	1
	Max. tree depth	100		Classification algorithm	SAMME
	Stop when majority	95%		Regression loss function	Exponential
Random Forest	No. of trees	10	Logistic Regression	Regularization type	Lasso (L1)
	Split subsets	>5		Strength	1
kNN	No. of neighbors	5	Gradient Boosting	No. of trees	100
	Metric	Euclidean		Learning rate	0.1
	Weight	Uniform		Tree depth	3
Naive Bayes	Default Settings			Training instances	1

5.3. Car Faults DB1 Evaluation

Various ML models were tested on the car faults dataset (DB1), analyzing two versions of feature lists: compact and expanded. Tables 6 and 7 display the measured performance metrics in both cases.

Table 6. Performance metrics of DB1 with compacted features.

Model	Train	Test	AUC	CA	F1	Prec	Recall	MCC	Spec	LogLoss
LR	13.724	0.06	0.965	**0.865**	0.863	0.877	0.865	0.859	0.995	0.844
NN	5.466	0.129	0.975	0.842	0.833	0.854	0.842	0.834	0.989	0.852
SGD	0.387	0.118	0.89	0.789	0.783	0.792	0.789	0.779	0.988	7.271
AdaBoost	4.102	0.164	0.912	0.722	0.706	0.711	0.722	0.707	0.984	3.834
kNN	0.097	0.243	0.964	0.714	0.694	0.753	0.714	0.704	0.988	1.715
RF	0.255	0.06	0.946	0.684	0.671	0.707	0.684	0.668	0.983	2.047
SVM	0.468	0.136	0.92	0.639	0.6	0.663	0.639	0.628	0.952	1.909
NB	0.168	0.104	0.975	0.639	0.571	0.576	0.639	0.64	0.989	5.678
DT	0.637	0	0.834	0.624	0.618	0.673	0.624	0.606	0.98	10.742
GB	36.681	0.081	0.834	0.504	0.509	0.612	0.504	0.472	0.951	5.033
CN2	71.023	0.064	0.68	0.406	0.402	0.435	0.406	0.376	0.968	2.848

Table 7. Performance metrics of DB1 with expanded features.

Model	Train	Test	AUC	CA	F1	Prec	Recall	MCC	Spec	LogLoss
LR	131.735	0.151	0.926	**0.692**	0.671	0.669	0.692	0.676	0.98	1.837
SGD	0.6	0.299	0.812	0.647	0.638	0.653	0.647	0.626	0.974	12.205
NN	5.768	0.316	0.932	0.624	0.614	0.625	0.624	0.603	0.977	1.519
AdaBoost	13.698	0.259	0.874	0.624	0.611	0.655	0.624	0.602	0.971	4.449
DT	1.311	0	0.827	0.609	0.606	0.659	0.609	0.592	0.983	11.262
NB	0.413	0.157	0.965	0.594	0.52	0.48	0.594	0.594	0.987	10.384
RF	0.415	0.14	0.915	0.586	0.568	0.606	0.586	0.561	0.967	2.75
kNN	0.217	0.32	0.9	0.459	0.421	0.411	0.459	0.431	0.965	4.502
GB	73.66	0.17	0.823	0.436	0.428	0.477	0.436	0.398	0.945	5.396
CN2	160.273	0.124	0.683	0.406	0.414	0.471	0.406	0.372	0.963	2.856
SVM	0.834	0.306	0.811	0.323	0.224	0.214	0.323	0.311	0.883	2.534

It was shown in the results of testing DB1 by the compact version of the extracted feature list that the Logistic Regression has the highest classification accuracy with the lowest Log Loss value among all evaluated models. Logistic Regression had the best classification accuracy but the second lowest Log Loss value when testing DB1 with the expanded version of the list of extracted features.

5.4. Other Sounds DB2 Evaluation

ML models were tested on other sound datasets (DB2), analyzing two versions of feature lists: compact and expanded. Tables 8 and 9 display the measured performance metrics in both cases.

Table 8. Performance metrics of DB2 with compact features.

Model	Train	Test	AUC	CA	F1	Prec	Recall	MCC	Spec	LogLoss
NN	25.526	0.143	0.99	**0.884**	0.884	0.886	0.884	0.874	0.992	0.462
LR	132.968	0.071	0.979	0.847	0.844	0.845	0.847	0.834	0.989	0.71
AdaBoost	27.621	0.236	0.966	0.845	0.843	0.844	0.845	0.831	0.989	2.206
SVM	3.519	0.336	0.987	0.831	0.827	0.844	0.831	0.817	0.985	0.636
RF	0.659	0.079	0.975	0.831	0.828	0.836	0.831	0.817	0.987	1.371
SGD	0.745	0.169	0.894	0.807	0.798	0.803	0.807	0.791	0.986	6.667
kNN	0.146	0.356	0.966	0.806	0.802	0.811	0.806	0.79	0.989	2.09
GB	232.345	0.224	0.976	0.805	0.805	0.814	0.805	0.788	0.985	0.933
DT	2.452	0.002	0.876	0.742	0.742	0.747	0.742	0.72	0.983	7.533
CN2	666.881	0.102	0.887	0.682	0.68	0.686	0.682	0.654	0.977	1.818
NB	0.348	0.076	0.933	0.577	0.583	0.691	0.577	0.556	0.982	5.333

Based on the DB2 testing results of the compact version of the list of features extracted, the Neural Networks model has the highest accuracy classification with the lowest Log Loss value among all models evaluated. Using the expanded version of the list of extracted features, Neural Networks obtained the second-highest accuracy after AdaBoost.

Table 9. Performance metrics of DB2 with expanded features.

Model	Train	Test	AUC	CA	F1	Prec	Recall	MCC	Spec	LogLoss
AdaBoost	64.35	0.488	0.969	**0.844**	0.841	0.844	0.844	0.83	0.989	2.032
NN	32.243	0.435	0.983	0.84	0.839	0.842	0.84	0.826	0.989	0.7
GB	539.137	0.336	0.979	0.835	0.834	0.842	0.835	0.821	0.987	0.844
SGD	1.418	0.304	0.903	0.822	0.819	0.819	0.822	0.806	0.988	6.164
LR	191.823	0.198	0.98	0.815	0.811	0.817	0.815	0.799	0.987	0.707
RF	1.004	0.161	0.972	0.812	0.805	0.812	0.812	0.796	0.986	1.485
SVM	6.311	0.588	0.982	0.79	0.78	0.792	0.79	0.772	0.983	0.732
DT	5.391	0	0.871	0.726	0.726	0.732	0.726	0.702	0.982	7.822
kNN	0.287	0.446	0.941	0.696	0.681	0.68	0.696	0.67	0.979	3.288
CN2	1256.612	0.18	0.893	0.687	0.688	0.694	0.687	0.659	0.979	1.8
NB	0.62	0.186	0.95	0.637	0.646	0.729	0.637	0.618	0.985	6.159

5.5. DB3 Evaluation

ML models were tested on the merged dataset (DB3), analyzing two versions of feature lists: compact and expanded. Tables 10 and 11 display the measured performance metrics in both cases.

Table 10. Performance metrics of DB3 with compact features.

Model	Train	Test	AUC	CA	F1	Prec	Recall	MCC	Spec	LogLoss
NN	33.206	0.156	0.988	**0.855**	0.851	0.857	0.855	0.845	0.993	0.6
LR	498.042	0.132	0.977	0.81	0.805	0.81	0.81	0.797	0.99	0.879
AdaBoost	52.826	0.326	0.955	0.786	0.779	0.782	0.786	0.772	0.99	2.775
kNN	0.118	0.368	0.966	0.766	0.758	0.768	0.766	0.751	0.99	2.205
RF	0.93	0.081	0.962	0.761	0.749	0.758	0.761	0.745	0.986	2.198
SVM	5.95	0.513	0.985	0.759	0.734	0.732	0.759	0.744	0.984	1.006
GB	1989.676	0.219	0.94	0.737	0.735	0.754	0.737	0.72	0.984	2.304
SGD	0.944	0.129	0.858	0.735	0.72	0.724	0.735	0.717	0.988	9.169
DT	3.696	0	0.858	0.676	0.67	0.673	0.676	0.654	0.984	8.976
CN2	993.324	0.085	0.853	0.605	0.606	0.613	0.605	0.578	0.981	2.666
NB	0.255	0.057	0.937	0.206	0.195	0.359	0.206	0.207	0.996	14.929

Table 11. Performance metrics of DB3 with expanded features.

Model	Train	Test	AUC	CA	F1	Prec	Recall	MCC	Spec	LogLoss
AdaBoost	126.229	0.446	0.953	**0.799**	0.792	0.804	0.799	0.786	0.989	2.525
NN	38.35	0.334	0.981	0.78	0.773	0.784	0.78	0.765	0.988	0.96
RF	1.37	0.252	0.962	0.772	0.755	0.753	0.772	0.757	0.987	2.217
SGD	2.012	0.311	0.874	0.764	0.755	0.753	0.764	0.748	0.991	8.16
LR	401.936	0.167	0.972	0.741	0.727	0.73	0.741	0.723	0.986	1.024
GB	3372.378	0.305	0.929	0.717	0.717	0.741	0.717	0.699	0.984	2.674
SVM	10.61	0.806	0.976	0.703	0.67	0.67	0.703	0.684	0.979	1.196
DT	7.743	0.002	0.843	0.655	0.653	0.66	0.655	0.632	0.985	9.877
kNN	0.261	0.464	0.927	0.631	0.612	0.609	0.631	0.606	0.981	4.366
CN2	1853.444	0.183	0.864	0.621	0.619	0.628	0.621	0.596	0.983	2.614
NB	0.451	0.276	0.932	0.068	0.033	0.065	0.068	0.069	0.997	29.18

The DB3 test results for the compact form of the list of features extracted showed that the Neural Networks model presents the minimum Log Loss value and the maximum accuracy of classification compared with the other models. Using the expanded version of the list of extracted features, Neural Networks reached the second-highest accuracy after AdaBoost.

5.6. Feature Ranking

Feature ranking was performed using the compact feature list on the DB1 dataset. Table 12 shows the rankings of the 52 features of the compact list. The table's rankings demonstrate that the top features across approaches were MFCC features (mean_10, mean_3, mean_2, mean_4) and Mel Spectrogram features (mean). The dominance of MFCC methods is evident. MFCC mean features are statistically significant across all measures, consistently outperforming standard deviation features. Chromagram characteristics, notably standard deviations, have a lower overall relevance. However, a few exceptions, such as chromagram_mean_7, have moderate rankings.

Table 12. Feature ranking for compact feature list applied on DB1.

Features		IG	ANOVA	ReliefF	FCBF	Features		IG	ANOVA	ReliefF	FCBF
Type	**Name**					**Type**	**Name**				
mfcc	mean_3	1.5308	10.3679	0.1138	0.8767	chromagram	mean_2	1.0131	5.0941	0.0402	0.4475
mfcc	mean_2	1.4491	10.6247	0.1171	0.7928	chromagram	mean_4	1.0087	12.9994	0.0702	0.4447
mel_spectrogram	std	1.3937	1.1654	0.0103	0.0001	chromagram	mean_10	0.9987	6.0528	0.049	0.4384
mfcc	mean_4	1.3734	22.9451	0.107	0.7215	chromagram	std_10	0.9859	6.1512	0.047	0
mfcc	mean_1	1.3498	16.7839	0.1006	0.7004	chromagram	std_8	0.9857	7.7565	0.0578	0
mfcc	mean_6	1.3418	12.0147	0.0832	0.6934	chromagram	std_11	0.9716	5.3361	0.0417	0.4214
mfcc	mean_10	1.2915	25.8226	0.0737	0.6505	chromagram	std_6	0.9697	5.4978	0.0306	0.4203
mfcc	mean_7	1.2911	8.428	0.0588	0.6501	chromagram	std_4	0.9529	7.7805	0.0421	0.41
mfcc	mean_5	1.2857	17.4085	0.085	0.6457	chromagram	std_3	0.9522	4.0328	0.0216	0.4096
mfcc	mean_11	1.2359	17.177	0.0705	0.6056	mfcc	std_4	0.9462	5.2936	0.055	0.406
mel_spectrogram	mean	1.2194	17.6741	0.075	0.5926	chromagram	std_5	0.9289	4.2578	0.0342	0
mfcc	mean_8	1.2028	14.3887	0.0551	0.5799	chromagram	std_2	0.9077	7.5226	0.0397	0
mfcc	mean_0	1.2	2.7237	0.019	0.5778	mfcc	std_1	0.874	3.2042	0.0382	0.3637
chromagram	mean_7	1.1776	8.5138	0.0678	0.561	mfcc	std_3	0.8657	2.9155	0.0373	0.359
mfcc	mean_9	1.1709	11.2358	0.0494	0.556	chromagram	std_1	0.8654	6.8358	0.0418	0.3588
mfcc	mean_12	1.1682	18.0902	0.0914	0.554	mfcc	std_2	0.8516	2.9743	0.0403	0.3511
chromagram	mean_8	1.1526	10.7271	0.0623	0.5426	mfcc	std_7	0.8077	8.099	0.0446	0.3271
chromagram	mean_0	1.1297	9.6163	0.0597	0.5262	mfcc	std_9	0.7995	7.6018	0.0328	0.3227
chromagram	mean_5	1.1043	7.3469	0.0648	0.5083	chromagram	std_0	0.7954	6.62	0.0522	0.3205
chromagram	mean_6	1.0997	6.3285	0.0555	0.5051	mfcc	std_6	0.7791	4.7404	0.0413	0.3119
chromagram	mean_9	1.0812	7.6679	0.0542	0	mfcc	std_8	0.7719	5.5775	0.0459	0.3082
chromagram	mean_1	1.0713	8.7643	0.0605	0.4857	mfcc	std_5	0.736	4.8213	0.0417	0.2897
chromagram	mean_3	1.0665	5.8873	0.0425	0.4825	mfcc	std_12	0.669	2.8505	0.0464	0.2565
chromagram	std_7	1.0527	6.2475	0.0376	0.4733	mfcc	std_11	0.6665	1.725	0.0312	0.2553
chromagram	std_9	1.0188	8.5498	0.056	0.4512	mfcc	std_0	0.6607	0.4204	−0.0035	0.2525
chromagram	mean_11	1.0162	5.4859	0.061	0	mfcc	std_10	0.6427	2.6443	0.0556	0.244

5.7. Evaluation of BOWSV

To incorporate Bayesian Optimization and Weighted Soft Voting into the proposed model, the previously selected features are ranked using ANOVA F-scores to select the most relevant features. Standardization is performed to scale the variables to the same scale. Multiple classifiers with diverse bases are included. Optimization starts with defining the bounds for classifier weights and feature counts. Then, cross-validation is applied to the results to obtain robust performance estimates, and acquisition function guides the search for optimal parameters. The optimization objective function evaluates the ensemble's performance using cross-validation to ensure that the estimates of the generalization performance are reliable. It converges to the best answer by iteratively proposing different weight combinations and assessing how well they work. Table 13 shows the metrics of the three datasets, DB1, DB2, and DB3, after applying ensemble optimization. For each iteration, the weights (w1, w2, w3), the number of features used, and the achieved accuracy are depicted.

Table 13. Metrics of datasets after ensemble optimization.

iter	DB1 Dataset					DB2 Dataset					DB3 Dataset				
	W1	**W2**	**W3**	**Features**	**Accuracy**	**W1**	**W2**	**W3**	**Features**	**Accuracy**	**W1**	**W2**	**W3**	**Features**	**Accuracy**
1	0.95	0.73	0.60	24	0.8879	0.95	0.73	0.60	23.75	0.8788	0.95	0.73	0.60	27.49	0.8557
2	0.16	0.06	0.87	22	0.8429	0.16	0.06	0.87	21.56	0.8332	0.16	0.06	0.87	23.12	0.8058
3	0.71	0.02	0.97	26	0.8797	0.71	0.02	0.97	26.01	0.7934	0.71	0.02	0.97	32.02	0.7861
4	0.21	0.18	0.18	28	0.8797	0.21	0.18	0.18	28.32	0.8739	0.21	0.18	0.18	36.65	0.8668
5	0.52	0.43	0.29	23	0.8951	0.52	0.43	0.29	23.04	0.8778	0.52	0.43	0.29	26.08	0.8488
6	0.14	0.29	0.37	26	0.8802	0.14	0.29	0.37	26.12	0.8797	0.14	0.29	0.37	32.24	0.8685
7	0.79	0.20	0.51	25	0.8648	0.79	0.20	0.51	24.56	0.839	0.79	0.20	0.51	29.12	0.817
8	0.05	0.61	0.17	26	0.8797	**0.05**	**0.61**	**0.17**	**25.92**	**0.8885**	0.05	0.61	0.17	31.85	0.8556
9	0.95	0.97	0.81	21	0.8813	0.95	0.97	0.81	20.65	0.87	0.95	0.97	0.81	21.3	0.8419
10	**0.10**	**0.68**	**0.44**	**23**	**0.9104**	0.10	0.68	0.44	23.05	0.8807	0.10	0.68	0.44	26.09	0.8556
11	0.01	0.96	0.14	23	0.9033	0.93	0.68	0.62	23.81	0.8798	0.17	0.05	0.19	36.68	0.8341
12	0.05	0.85	0.89	23	0.9033	0.01	0.91	0.18	26.57	0.8807	**0.02**	**0.60**	**0.72**	**20.98**	**0.8453**
13	0.02	0.99	0.72	23	0.8731	0.02	0.97	0.97	23.87	0.8768	0.04	0.47	0.35	20.09	0.8496
14	**0.13**	**0.69**	**0.48**	**23**	**0.9104**	0.15	0.04	0.02	26.95	0.8312	0.71	0.99	0.37	39.42	0.8668
15	0.02	0.29	0.33	23	0.9033	0.28	1.00	0.02	23.64	0.8817	0.16	0.80	0.10	29.43	0.8591
16	0.15	0.07	0.98	23	0.8275	0.01	0.01	0.13	23.54	0.8254	0.71	0.45	1.00	38.77	0.8496
17	0.25	0.88	0.41	24	0.9027	0.59	0.86	0.99	23.08	0.8846	0.88	0.23	0.37	35.5	0.8281
18	0.22	0.50	0.05	23	0.9027	0.87	0.99	0.27	22.6	0.8807	0.62	0.25	0.77	23.87	0.8109
19	0.14	0.99	1.00	24	0.8577	0.26	0.33	0.38	29.4	0.8778	0.46	0.98	0.53	20.13	0.8487
20	0.62	0.99	0.15	23	0.9027	0.21	0.99	0.89	28.78	0.8768	0.53	0.30	0.02	34.85	0.853
21	0.99	0.99	0.11	23	0.8648	0.98	0.93	0.15	28.95	0.8729	0.53	0.26	0.53	37.74	0.8513
22	0.97	0.01	0.07	22	0.85	0.97	0.01	0.07	22.06	0.7847	0.97	0.01	0.07	24.12	0.7294
23	0.09	0.05	0.84	24	0.8126	0.84	0.05	0.98	28.97	0.8138	0.09	0.05	0.84	28.97	0.8445
24	0.96	0.97	0.30	27	0.8731	0.02	0.89	0.01	29.14	0.8768	0.80	0.92	0.97	21.66	0.847
25	0.85	0.41	0.89	30	0.872	0.02	0.99	0.63	29.95	0.8768	0.29	0.69	0.67	27.29	0.8582
26	0.02	0.86	0.08	20	0.8885	0.70	0.76	0.01	29.94	0.8788	0.61	0.90	0.76	39.74	0.8685
27	0.85	0.03	0.24	20	0.8505	0.01	0.98	0.29	28.04	0.8768	0.75	0.04	0.40	34.28	0.8032
28	0.10	0.04	0.94	27	0.8044	0.23	0.95	0.26	20	0.8768	0.72	0.30	0.54	36.91	0.8419
29	0.12	0.96	0.08	29	0.8725	0.19	0.13	0.00	29.99	0.873	0.81	0.69	0.61	21.69	0.8402
30	0.02	0.94	0.94	20	0.8885	0.29	0.02	1.00	20.01	0.7682	0.97	0.73	0.04	35.85	0.859

Due to the large number of classes (27 for car faults, 22 for environmental sounds, and 49 for the merged dataset), a full confusion matrix would be impractical. Instead, key examples of classification performance have been summarized. Several classes—including Engine Misfire, Fuel Pump Cartridge Fault, Radiator Fan Failure, Strut Mount Failure, Suspension Arm Fault, and others—are classified with 100% accuracy, and the Bad CV Joint class achieves around 75% accuracy. Furthermore, the Bayesian-Optimized Weighted Soft Voting with Feature Selection (BOWSVFS) approach demonstrates the robustness of the model by achieving an overall accuracy of 91.04% on the car fault dataset (DB1).

However, some classes present challenges. For instance, the Universal Joint Failure or Steering class has an 80% correct classification rate, with misclassifications primarily as engine rattling noise. The Knocking class, in particular, exhibits significant difficulty, with only 40% of instances correctly classified and misclassifications distributed across categories such as bad wheel bearing, squeaky, and squeaky brake (or grinding brake). These examples highlight the strengths and areas for improvement within the proposed framework.

5.8. Outlook and Future Perspectives

The practical implications of this research are far-reaching. The framework enhances overall transportation safety through timely interventions and improved emergency response by enabling early fault detection and real-time classification of vehicle and environmental sounds. Furthermore, the ability to accurately interpret auditory cues supports the development of more accessible and inclusive ITSs. For instance, auditory alerts can be transformed into visual or haptic signals, thereby assisting individuals with disabilities and ensuring that critical safety information is disseminated effectively. These advances pave the way for smarter, more responsive urban transportation systems that improve efficiency and significantly elevate the safety and quality of life in smart cities.

Although the proposed classification methods can achieve a high degree of accuracy in sound-based early fault detection for vehicles in ITSs, there remains potential for further enhancement through the incorporation of explicit user feedback, such as ratings of the classification results. The efficacy of machine learning systems can be notably augmented by fostering a collaborative relationship with users, improving the system's accuracy and enhancing user understanding and trust in the system [73–75]. Users can contribute to the classification model by providing explicit collective feedback regarding its classification accuracy and the early detection of faults for vehicles in ITSs. This feedback can subsequently be utilized to refine the overall accuracy of the classification model. For instance, users might assign scores or ratings to the accuracy of detected faults. Nonetheless, sustaining user motivation for continuous feedback poses a challenge, as many users exhibit limited interest in participating in such evaluations [76].

The gamification concept is employed as a behavioral change strategy to enhance user motivation toward engaging in desired behaviours, such as providing feedback on the classification accuracy of detected faults for vehicles in ITS [77,78]. A prevalent application of gamification involves incorporating elements of video games, such as points and levels, into non-gaming contexts, such as educational settings [79]. Gamification has demonstrated successful implementation across various domains, including the promotion of healthy lifestyle choices [80], the enhancement of student engagement in academic courses [81], and the improvement of quality and productivity within business environments [82]. There are four primary elements of gamification commonly utilized in non-gaming contexts [83]:

1. Points: Many gamification strategies rely on point systems, which may include features such as levels and leaderboards. The classification accuracy of detected faults can be quantified through user ratings regarding the quality of fault detection for vehicles in ITS. Points accumulated or lost will subsequently inform the classification model's training to enhance its ability to detect sound-based faults for vehicles in ITSs early. Nevertheless, points should be integrated with other gamification elements to effectively motivate users [83].
2. Digital Badges: Users may receive digital badges as recognition for acquiring specific skills, knowledge, or achievements, thereby showcasing their accomplishments [84]. These badges are typically awarded based on predefined criteria [85–87]. For example, users might earn digital badges by reaching a specified number of points corresponding to their ratings on the classification accuracy of early detected sound-based faults for vehicles in ITSs.
3. Levels: Users must accumulate points to advance to higher levels. Upon reaching a predetermined point threshold, they can level up, thereby unlocking additional features within the system [88].
4. Leaderboards: Users can establish leaderboards to reflect their achievements or points earned or to track progress toward specific goals [86].

A recent study [89] identifies several factors that affect users' perceptions and responses to gamification elements utilized for feedback collection, revealing diverse preferences in this context. This underscores the necessity of systematically gathering users' explicit and collective feedback, which can be instrumental in optimizing our proposed classification model to align with user preferences. Neglecting this aspect could result in overseeing critical factors that enhance classification accuracy. To address this, one can utilize the application-independent conceptual framework proposed by [89], which can be adapted to gamify the feedback collection process regarding the accuracy of our sound-based early detection of faults for vehicles in the ITS system. This framework articulates the variations in user perceptions and needs concerning gamification elements, aiming to motivate users to provide high-quality feedback on the classification accuracy of our proposed system. It serves as a guiding resource for software engineers in encouraging users to offer their explicit and collective feedback, thereby facilitating further training of our classification model and potentially improving its early fault detection accuracy for vehicles in the ITS. Additionally, a category representing normal operational conditions or safe car sounds can be included to better differentiate faults from irrelevant auditory data.

6. Conclusions and Future Work

This study highlights the critical significance of sound-based diagnostics in improving Intelligent Transportation Systems (ITSs) in smart cities. The potential for real-time vehicle fault identification and enhanced accessibility for people with disabilities was demonstrated by creating a new dataset of automotive malfunction sounds and combining audio processing techniques with ML. The high accuracy rates attained by various ML models demonstrate the efficacy of sound-based techniques as a complement to classic vision-based systems. Finally, this study leads to a more inclusive and responsive transportation infrastructure, which aligns with the overall goals of smart city development.

Future research includes (i) expanding the dataset to cover diverse vehicles, faults, and real-world scenarios through collaborations with the automotive and public transport sectors; (ii) integrating sound data with visual and environmental sensors, which can enhance system robustness; (iii) developing real-time sound-based detection systems for urban applications and exploring advanced machine learning methods, such as deep learning and transfer learning, which will improve accuracy; (iv) incorporating the "no sound"

class in implementing the proposed approach in practice. (v) investigating generative model-based data augmentation strategies to boost dataset diversity and model resilience; and (vi) exploring domain adaptation techniques, few-shot learning, or data augmentation strategies to enhance generalization across a wider range of vehicles.

Author Contributions: A.R.: Conceptualization, Methodology, Software, Writing—Original Draft. Y.A.: Data Curation, Methodology, Investigation, Writing—Original Draft. T.A.F.: Visualization, Software. M.A.: Conceptualization, Methodology, Writing—Original Draft. A.B.: Data Curation, Investigation, Writing—Original Draft. M.B.: Methodology, Writing—Review and Editing, Supervision. M.A.E.: Supervision, Methodology, Writing—Review and Editing, Project Administration. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by funds from the King Salman Centre for Disability Research (Group no.: KSRG-2024-240).

Data Availability Statement: In this study, data from various sources are collected to create a tailored dataset for car sound analysis: Reliable audio samples are built by downloading videos from YouTube related to car faults, animal sounds, car crashes, siren sounds, etc. Additional audio is supplemented from Kaggle datasets: FSC22 [38], Google AudioSet [39], Audio Classifier Dataset [40], Sound classification of animal voice [41], and Vehicle Sounds dataset [42]. The data collected by the authors are available at: https://github.com/amrrashed/Sound-Based-Vehicle-Diagnostics-Emergency-Signal-Recognition/tree/main (accessed on 31 January 2025). Code availability: The code used is available at: https://github.com/amrrashed/Sound-Based-Vehicle-Diagnostics-Emergency-Signal-Recognition/tree/main/codes (accessed on 31 January 2025).

Acknowledgments: The authors extend their appreciation to the King Salman Centre for Disability Research for funding this work through Research Group No. KSRG-2024-240.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gong, T.; Zhu, L.; Yu, F.R.; Tang, T. Edge Intelligence in Intelligent Transportation Systems: A Survey. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 8919–8944. [CrossRef]
2. Khalil, R.A.; Safelnasr, Z.; Yemane, N.; Kedir, M.; Shafiqurrahman, A.; Saeed, N. Advanced Learning Technologies for Intelligent Transportation Systems: Prospects and Challenges. *IEEE Open J. Veh. Technol.* **2024**, *5*, 397–427. [CrossRef]
3. Sarwatt, D.S.; Lin, Y.; Ding, J.; Sun, Y.; Ning, H. Metaverse for Intelligent Transportation Systems (ITS): A Comprehensive Review of Technologies, Applications, Implications, Challenges and Future Directions. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 6290–6308. [CrossRef]
4. Wang, B.; Li, Q.; Mao, Q.; Wang, J.; Chen, C.L.P.; Shangguan, A.; Zhang, H. A Survey on Vision-Based Anti Unmanned Aerial Vehicles Methods. *Drones* **2024**, *8*, 518. [CrossRef]
5. Masal, K.M.; Bhatlawande, S.; Shingade, S.D. Development of a visual to audio and tactile substitution system for mobility and orientation of visually impaired people: A review. *Multimed. Tools Appl.* **2024**, *83*, 20387–20427. [CrossRef]
6. Liu, F.; Lu, Z.; Lin, X. Vision-based environmental perception for autonomous driving. *Proc. Inst. Mech. Eng. Part D J. Automob. Eng.* **2023**, *239*, 39–69. [CrossRef]
7. Kiranyaz, S.; Can Devecioglu, O.; Alhams, A.; Sassi, S.; Ince, T.; Avci, O.; Gabbouj, M. Exploring Sound Versus Vibration for Robust Fault Detection on Rotating Machinery. *IEEE Sens. J.* **2024**, *24*, 23255–23264. [CrossRef]
8. Alqudaihi, K.S.; Aslam, N.; Khan, I.U.; Almuhaideb, A.M.; Alsunaidi, S.J.; Ibrahim, N.M.A.R.; Alhaidari, F.A.; Shaikh, F.S.; Alsenbel, Y.M.; Alalharith, D.M.; et al. Cough Sound Detection and Diagnosis Using Artificial Intelligence Techniques: Challenges and Opportunities. *IEEE Access* **2021**, *9*, 102327–102344. [CrossRef]
9. Vranken, E.; Mounir, M.; Norton, T., Sound-Based Monitoring of Livestock. In *Encyclopedia of Smart Agriculture Technologies*; Springer International Publishing: Berlin/Heidelberg, Germany, 2023; pp. 1–12. . [CrossRef]
10. Pervez, F.; Shoukat, M.; Suresh, V.; Farooq, M.U.B.; Sandhu, M.; Qayyum, A.; Usama, M.; Girardi, A.; Latif, S.; Qadir, J. Medicine's New Rhythm: Harnessing Acoustic Sensing via the Internet of Audio Things for Healthcare. *IEEE Open J. Comput. Soc.* **2024**, *5*, 491–510. [CrossRef]
11. Kim, J.; Kim, J.; Kim, H. A Study on Gear Defect Detection via Frequency Analysis Based on DNN. *Machines* **2022**, *10*, 659. [CrossRef]

12. Koh, P.; Kim, S. Designing a Augmented Reality Auditory Training Game for in-situ training and diagnostic tool for the hearing impaired. In Proceedings of the Audio Engineering Society Conference: AES 2024 International Audio for Games Conference, Tokyo, Japan, 27–29 April 2024.
13. Oladimeji, D.; Gupta, K.; Kose, N.A.; Gundogan, K.; Ge, L.; Liang, F. Smart Transportation: An Overview of Technologies and Applications. *Sensors* **2023**, *23*, 3880. [CrossRef] [PubMed]
14. Mohammed, H.B.M.; Cavus, N. Utilization of Detection of Non-Speech Sound for Sustainable Quality of Life for Deaf and Hearing-Impaired People: A Systematic Literature Review. *Sustainability* **2024**, *16*, 8976. [CrossRef]
15. Nasim, F.; Masood, S.; Jaffar, A.; Ahmad, U.; Rashid, M. Intelligent Sound-Based Early Fault Detection System for Vehicles. *Comput. Syst. Sci. Eng.* **2023**, *46*, 3175–3190. [CrossRef]
16. Hamad, A.A.; Nasim, M.F.; Jaffar, A.; Khalaf, O.I.; Ouahada, K.; Hamam, H.; Akram, S.; Siddique, A. Cognitive Inspired Sound-Based Automobile Problem Detection: A Step Toward Xai. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4814232 (accessed on 25 January 2025) [CrossRef]
17. Yildirim, S.; Bingol, M.S. Design of a Proposed Neural Network for Sound Quality Analysis of Different Types for Car Systems. *Int. J. Mechatron. Appl. Mech.* **2024**, *16*, 76–81.
18. Akbalık, F.; Yıldız, A.; Ertuğrul, Ö.F.; Zan, H. Engine Fault Detection by Sound Analysis and Machine Learning. *Appl. Sci.* **2024**, *14*, 6532. [CrossRef]
19. Murovec, J.; Prezelj, J.; Ćirić, D.; Milivojčević, M. Zero Crossing Signature: A Time-Domain Method Applied to Diesel and Gasoline Vehicle Classification. *IEEE Sens. J.* **2024**, *25*, 3. [CrossRef]
20. Wang, S.; Xu, Q.; Zhu, S.; Wang, B. Making transformer hear better: Adaptive feature enhancement based multi-level supervised acoustic signal fault diagnosis. *Expert Syst. Appl.* **2025**, *264*, 125736. [CrossRef]
21. Boztas, G.; Tuncer, T.; Aydogmus, O.; Yildirim, M. A DCSLBP based intelligent machine malfunction detection model using sound signals for industrial automation systems. *Comput. Electr. Eng.* **2024**, *119*, 109541. [CrossRef]
22. Wang, Y.; Li, D.; Li, L.; Sun, R.; Wang, S. A novel deep learning framework for rolling bearing fault diagnosis enhancement using VAE-augmented CNN model. *Heliyon* **2024**, *10*, e35407. [CrossRef]
23. Guo, X. Fault Diagnosis of Rolling Bearings Based on Acoustics and Vibration Engineering. *IEEE Access* **2024**, *12*, 139632–139648. [CrossRef]
24. Li, Y.; Tao, X.; Sun, Y. A Fault Diagnosis Method for Turnout Switch Machines Based on Sound Signals. *Electronics* **2024**, *13*, 4839. [CrossRef]
25. Kreuzer, M.; Schmidt, D.; Wokusch, S.; Kellermann, W. Real-World Airborne Sound Analysis for Health Monitoring of Bearings in Railway Vehicles. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4923626 (accessed on 19 January 2025). [CrossRef]
26. Yun, E.; Jeong, M. Acoustic Feature Extraction and Classification Techniques for Anomaly Sound Detection in the Electronic Motor of Automotive EPS. *IEEE Access* **2024**, *12*, 149288–149307. [CrossRef]
27. Shajie, D.; Juliet, S.; Ezra, K.; Annie Flora, J.B. Diagnostic Sonance: Sound-Based Approach to Assess Engine Ball Bearing Health in Automobiles. *Prz. Elektrotechniczny* **2024**, *1*, 74–78. [CrossRef]
28. Khan, F.A.; Jamil, A.; Khan, S.A.; Hameed, A.A. Enhancing robotic manipulator fault detection with advanced machine learning techniques. *Eng. Res. Express* **2024**, *6*, 025204. [CrossRef]
29. Kim, S.M.; Soo Kim, Y. Enhancing Sound-Based Anomaly Detection Using Deep Denoising Autoencoder. *IEEE Access* **2024**, *12*, 84323–84332. [CrossRef]
30. Senanayaka, A.; Lee, P.; Lee, N.; Dickerson, C.; Netchaev, A.; Mun, S. Enhancing the Accuracy of Machinery Fault Diagnosis through Fault Source Isolation of Complex Mixture of Industrial Sound Signals. *Int. J. Adv. Manuf. Technol.* **2024**, *133*, 5627–5642. [CrossRef]
31. Hameed, U.; Masood, S.; Nasim, F.; Jaffar, A.; Ahmed, Z.; Khan, R.; Hussain, A.; Ali, S.; Mehmood, A.; Shah, R. Exploring the Accuracy of Machine Learning and Deep Learning in Engine Knock Detection. *Bull. Bus. Econ.* **2024**, *13*, 203–210. [CrossRef]
32. Naryanto, R.F.; Delimayanti, M.K.; Naryaningsih, A.; Adi, R.; Setiawan, B.A. Fault Detection in Diesel Engines using Artificial Neural Networks and Convolutional Neural Networks. *J. Theor. Appl. Inf. Technol.* **2024**, *102*, 683–690.
33. Yuan, G.; Yang, Y. Fault detection method of new energy vehicle engine based on wavelet transform and support vector machine. *Int. J. Knowl. Based Intell. Eng. Syst.* **2024**, *28*, 718–731. [CrossRef]
34. Chu, S.; Zhang, J.; Liu, F.; Kong, X.; Jiang, Z.; Mao, Z. Fault identification model of diesel engine based on mixed attention: Single-cylinder fault data driven whole-cylinder diagnosis. *Expert Syst. Appl.* **2024**, *255*, 124769. [CrossRef]
35. Lee, D.; Choo, H.; Jeong, J. GCN-Based LSTM Autoencoder with Self-Attention for Bearing Fault Diagnosis. *Sensors* **2024**, *24*, 4855. [CrossRef]
36. Spadini, T.; Nose-Filho, K.; Suyama, R. Intelligent Fault Diagnosis of Type and Severity in Low-Frequency, Low Bit-Depth Signals. *arXiv* **2024**, arXiv:2411.06299. [CrossRef]

37. Gantert, L.; Zeffiro, T.; Sammarco, M.; Campista, M.E.M. Multiclass classification of faulty industrial machinery using sound samples. *Eng. Appl. Artif. Intell.* **2024**, *136*, 108943. [CrossRef]
38. Bandara, M.; Jayasundara, R.; Ariyarathne, I.; Meedeniya, D.; Perera, C. FSC22 Dataset. 2022. Available online: https://www.kaggle.com/datasets/irmiot22/fsc22-dataset (accessed on 19 January 2025).
39. Gemmeke, J.F.; Ellis, D.P.W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio Set: An ontology and human-labeled dataset for audio events. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LO, USA, 5–9 March 2017. [CrossRef]
40. Jacob, I. Audio Classifier Dataset. 2025. Available online: https://www.kaggle.com/datasets/aklimarimi/audio-classifier-dataset (accessed on 19 January 2025)
41. Putthewad, R.B. Sound Classification of Animal Voice. 2025. Available online: https://www.kaggle.com/datasets/rushibalajiputthewad/sound-classification-of-animal-voice (accessed on 19 January 2025).
42. Abderrahim, J. Vehicle Sounds Dataset. 2025. Available online: https://www.kaggle.com/datasets/janboubiabderrahim/vehicle-sounds-dataset (accessed on 19 January 2025).
43. Community, D. DCASE 2024 Challenge. 2024. Available online: https://dcase.community/challenge2024/index (accessed on 19 January 2025).
44. Community, D. UrbanSound8K Dataset. DCASE 2024 Challenge. 2024. Available online: https://www.kaggle.com/code/prabhavsingh/urbansound8k-classification (accessed on 19 January 2025).
45. Research, G. AudioSet. Available online: https://www.kaggle.com/datasets/akela91/google-audioset (accessed on 19 January 2025).
46. Li, H.; Wang, Z. Anomaly identification of wind turbine blades based on Mel-Spectrogram Difference feature of aerodynamic noise. *Measurement* **2025**, *240*, 115428. [CrossRef]
47. Lakdari, M.W.; Ahmad, A.H.; Sethi, S.; Bohn, G.A.; Clink, D.J. Mel-frequency cepstral coefficients outperform embeddings from pre-trained convolutional neural networks under noisy conditions for discrimination tasks of individual gibbons. *Ecol. Inform.* **2024**, *80*, 102457. [CrossRef]
48. Pandeya, Y.R.; Lee, J. GlocalEmoNet: An optimized neural network for music emotion classification and segmentation using timbre and chroma features. *Multimed. Tools Appl.* **2024**, *83*, 74141–74158. [CrossRef]
49. Constantinescu, C.; Brad, R. An Overview on Sound Features in Time and Frequency Domain. *Int. J. Adv. Stat. IT&C Econ. Life Sci.* **2023**, *13*, 45–58. [CrossRef]
50. Balingbing, C.; Kirchner, S.; Siebald, H.; Van Hung, N.; Hensel, O. Determining the sound signatures of insect pests in stored rice grain using an inexpensive acoustic system. *Food Secur.* **2024**, *16*, 1529–1538. [CrossRef]
51. Sanchez-Morillo, D.; Sales-Lerida, D.; Priego-Torres, B.; León-Jiménez, A. Cough Detection Using Acceleration Signals and Deep Learning Techniques. *Electronics* **2024**, *13*, 2410. [CrossRef]
52. Rizvi, S.; Pettee, M.; Nachman, B. Learning likelihood ratios with neural network classifiers. *J. High Energy Phys.* **2024**, *2024*, 1–41. [CrossRef]
53. Peretz, O.; Koren, M.; Koren, O. Naive Bayes classifier—An ensemble procedure for recall and precision enrichment. *Eng. Appl. Artif. Intell.* **2024**, *136*, 108972. [CrossRef]
54. Khashei, M.; Etemadi, S.; Bakhtiarvand, N. A New Discrete Learning-Based Logistic Regression Classifier for Bankruptcy Prediction. *Wirel. Pers. Commun.* **2024**, *134*, 1075–1092. [CrossRef]
55. Azimjonov, J.; Kim, T. Stochastic gradient descent classifier-based lightweight intrusion detection systems using the efficient feature subsets of datasets. *Expert Syst. Appl.* **2024**, *237*, 121493. [CrossRef]
56. Sun, Y.; Liu, Q. Collaborative filtering recommendation based on K-nearest neighbor and non-negative matrix factorization algorithm. *J. Supercomput.* **2025**, *81*, 79. [CrossRef]
57. Larisa, L. Optimized Composition of Business Process Web Services via QoS-Based Categorization Using Decision Tree Classifier and Knowledge-Based Decision Support. *Am. J. Bus. Oper. Res.* **2025**, *12*, 1–14. [CrossRef]
58. Bouke, M.A.; Alramli, O.I.; Abdullah, A. XAIRF-WFP: A novel XAI-based random forest classifier for advanced email spam detection. *Int. J. Inf. Secur.* **2025**, *24*, 5. [CrossRef]
59. Li, Y.; Xie, X. Two novel deep multi-view support vector machines for multiclass classification. *Appl. Intell.* **2025**, *55*, 1–17. [CrossRef]
60. Maszczyk, C.; Sikora, M.; Wróbel, Ł. Classification, Regression, and Survival Rule Induction with Complex and M-of-N Elementary Conditions. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 554–579. [CrossRef]
61. Kumpf, K.; Protic, M.; Jovanovic, L.; Cajic, M.; Zivkovic, M.; Bacanin, N. Insider Threat Detection Using Bidirectional Encoder Representations From Transformers and Optimized AdaBoost Classifier. In Proceedings of the 2024 International Conference on Circuit, Systems and Communication (ICCSC), Fez, Morocco, 28–29 June 2024; pp. 1–6. [CrossRef]
62. Theerthagiri, P. Liver disease classification using histogram-based gradient boosting classification tree with feature selection algorithm. *Biomed. Signal Process. Control* **2025**, *100*, 107102. [CrossRef]

63. Aljohani, M.; AbdulAzeem, Y.; Balaha, H.M.; Badawy, M.; Elhosseini, M.A. Advancing feature ranking with hybrid feature ranking weighted majority model: A weighted majority voting strategy enhanced by the Harris hawks optimizer. *J. Comput. Des. Eng.* **2024**, *11*, 308–325. [CrossRef]
64. Gao, J.; Wang, Z.; Jin, T.; Cheng, J.; Lei, Z.; Gao, S. Information gain ratio-based subfeature grouping empowers particle swarm optimization for feature selection. *Knowl. Based Syst.* **2024**, *286*, 111380. [CrossRef]
65. Jamil, M.A.; Khanam, S. Influence of One-Way ANOVA and Kruskal–Wallis Based Feature Ranking on the Performance of ML Classifiers for Bearing Fault Diagnosis. *J. Vib. Eng. Technol.* **2024**, *12*, 3101–3132. [CrossRef]
66. Yan, M.; Deng, J.; Zhang, S.; Chen, P. Feature Selection Method Based on Improved Differential Evolution and ReliefF. In Proceedings of the 2024 Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence, DEAI 2024, Dongguan, China, 19–21 January 2024; pp. 539–543. [CrossRef]
67. Zhang, S.; Wang, T.; Worden, K.; Sun, L.; Cross, E.J. Canonical-correlation-based fast feature selection for structural health monitoring. *Mech. Syst. Signal Process.* **2025**, *223*, 111895. [CrossRef]
68. Liu, Z. Ensemble Learning. In *Artificial Intelligence for Engineers*; Springer Nature: Cham, Switzerland, 2025; pp. 221–242. [CrossRef]
69. Chhillar, I.; Singh, A. An improved soft voting-based machine learning technique to detect breast cancer utilizing effective feature selection and SMOTE-ENN class balancing. *Discov. Artif. Intell.* **2025**, *5*, 4. [CrossRef]
70. Mahboubi, N.; Xie, J.; Huang, B. Point-by-point transfer learning for Bayesian optimization: An accelerated search strategy. *Comput. Chem. Eng.* **2025**, *194*, 108952. [CrossRef]
71. Iturbe-Araya, J.I.; Rifà-Pous, H. Enhancing unsupervised anomaly-based cyberattacks detection in smart homes through hyperparameter optimization. *Int. J. Inf. Secur.* **2025**, *24*, 45. [CrossRef]
72. Widardo, F.; Chowanda, A. Hyperparameter tuning for deep learning model used in multimodal emotion recognition data. *Bull. Electr. Eng. Inform.* **2025**, *14*, 261–267. [CrossRef]
73. Stumpf, S.; Rajaram, V.; Li, L.; Burnett, M.; Dietterich, T.; Sullivan, E.; Drummond, R.; Herlocker, J. Toward harnessing user feedback for machine learning. In Proceedings of the 12th International Conference on Intelligent User Interfaces, IUI07, Honolulu, HI, USA, 28–31 January 2007; pp. 82–91. [CrossRef]
74. Lee, T.Y.; Smith, A.; Seppi, K.; Elmqvist, N.; Boyd-Graber, J.; Findlater, L. The human touch: How non-expert users perceive, interpret, and fix topic models. *Int. J. Hum. Comput. Stud.* **2017**, *105*, 28–42. [CrossRef]
75. Liao, Q.V.; Gruen, D.; Miller, S. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, Honolulu, HI, USA, 25–30 April 2020; pp. 1–15. [CrossRef]
76. Almaliki, M.; Ali, R. Persuasive and Culture-Aware Feedback Acquisition. In *Persuasive Technology*; Springer International Publishing:Berlin/Heidelberg, Germany, 2016; pp. 27–38. . [CrossRef]
77. Deterding, S.; Dixon, D.; Khaled, R.; Nacke, L. From game design elements to gamefulness: Defining "gamification". In Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, MindTrek '11, Tampere, Finland, 28–30 September 2011; pp. 9–15. [CrossRef]
78. Herzig, P.; Ameling, M.; Schill, A. A Generic Platform for Enterprise Gamification. In Proceedings of the 2012 Joint Working IEEE/IFIP Conference on Software Architecture and European Conference on Software Architecture, Helsinki, Finland, 20–24 August 2012; pp. 219–223. [CrossRef]
79. Nicholson, S. A RECIPE for Meaningful Gamification. In *Gamification in Education and Business*; Reiners, T., Torstenand Wood, L.C., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 1–20. [CrossRef]
80. Johnson, D.; Deterding, S.; Kuhn, K.A.; Staneva, A.; Stoyanov, S.; Hides, L. Gamification for health and wellbeing: A systematic review of the literature. *Internet Interv.* **2016**, *6*, 89–106. [CrossRef]
81. Pløhn, T.; Aalberg, T. Using gamification to motivate smoking cessation. In Proceedings of the European Conference on Games Based Learning, Steinkjer, Norway, 8–9 October 2015; p. 431.
82. Simões, J.; Redondo, R.D.; Vilas, A.F. A social gamification framework for a K-6 learning platform. *Comput. Hum. Behav.* **2013**, *29*, 345–353. [CrossRef]
83. Lister, M. Gamification: The effect on student motivation and performance at the post-secondary level. *Issues Trends Educ. Technol.* **2015**, *3*, 2. [CrossRef]
84. Abramovich, S.; Schunn, C.; Higashi, R.M. Are badges useful in education?: It depends upon the type of badge and expertise of learner. *Educ. Technol. Res. Dev.* **2013**, *61*, 217–232. [CrossRef]
85. Ahn, J.; Pellicone, A.; Butler, B.S. Open badges for education: What are the implications at the intersection of open systems and badging? *Res. Learn. Technol.* **2014**, *22*, 563. [CrossRef]
86. Domínguez, A.; Saenz-de Navarrete, J.; De-Marcos, L.; Fernández-Sanz, L.; Pagés, C.; Martínez-Herráiz, J.J. Gamifying learning experiences: Practical implications and outcomes. *Comput. Educ.* **2013**, *63*, 380–392. [CrossRef]

87. Hanus, M.D.; Fox, J. Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Comput. Educ.* **2015**, *80*, 152–161. [CrossRef]
88. Goehle, G. Gamification and Web-based Homework. *PRIMUS* **2013**, *23*, 234–246. [CrossRef]
89. Almaliki, M. Misinformation-Aware Social Media: A Software Engineering Perspective. *IEEE Access* **2019**, *7*, 182451–182458. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

A Unified Framework for Automated Testing of Robotic Process Automation Workflows Using Symbolic and Concolic Analysis †

Ciprian Paduraru *,‡, Marina Cernat ‡ and Adelina-Nicoleta Staicu ‡

Computer Science Department, University of Bucharest, 050663 Bucharest, Romania; marina.cernat@unibuc.ro (M.C.); nicoleta.staicu@unibuc.ro (A.-N.S.)

* Correspondence: ciprian.paduraru@unibuc.ro

† This paper is an extended version of our paper published in Paduraru, C.; Cernat, M.; Staicu, A. N. Concolic execution for RPA testing. In Proceedings of the 2023 27th International Conference on Engineering of Complex Computer Systems (ICECCS), Toulouse, France, 14–16 June 2023.

‡ These authors contributed equally to this work.

Abstract: Robotic Process Automation is a technology that replicates human interactions with user interfaces across various applications. However, testing Robotic Process Automation implementations remains challenging due to the dynamic nature of workflows. This paper presents a novel testing framework that first integrates symbolic execution and concolic testing strategies to enhance Robotic Process Automation workflow validation. Building on insights from these methods, we introduce a hybrid approach that optimizes test coverage and efficiency in specific cases. Our open-source implementation demonstrates that automated testing in the Robotic Process Automation domain significantly improves coverage, reduces manual effort, and enhances reliability. Furthermore, the proposed solution supports multiple Robotic Process Automation platforms and aligns with industry best practices for user interface automation testing. Experimental evaluation, conducted in collaboration with industry, validates the effectiveness of our approach.

Keywords: robotic process automation; symbolic execution; concolic testing; workflow testing; automated user interface testing; robotic process automation debugging; software quality assurance

1. Introduction

Robotic Process Automation (RPA) is a technology that can replicate human interactions at the user interface (UI) level when using various applications. It is one of the newest and fastest growing segments in the software market [1]. Based on the size and value of the company, market share, scope and depth of the solution, the UiPath platform is probably the leading provider of RPA tools [2] and, according to our analysis, offers the most sophisticated support for RPA testing. In addition to UiPath, there are other RPA vendors on the market, such as Automation Anywhere, Blue Prism, Datamatics, Microsoft, Automation Edge, and many more for which the framework described in the following chapters can also be used, but for this article UiPath has been chosen as an example. However, despite its widespread adoption, the methods used to verify and validate RPA workflows have not evolved at the same pace. In practice, testing remains largely manual or relies on fragile black-box techniques, resulting in limited test coverage, high maintenance effort, and a higher risk of undetected errors in production systems. This research addresses the lack of effective, automated white-box testing techniques tailored specifically for RPA workflows. Existing approaches typically do not analyze the internal control-flow structure

of workflows, nor do they leverage formal methods to improve test generation or reliability. Against this backdrop, our work is guided by the following research questions:

- How can symbolic and concolic execution techniques be adapted and applied to RPA workflows for white-box testing?
- Can a hybrid strategy improve test coverage and fault detection compared to existing black-box or ad hoc methods?
- What are the practical implications and benefits of applying such a framework in real-world RPA development environments?

To investigate these questions, we propose the following hypothesis:

Integrating symbolic and concolic execution techniques into an RPA testing framework can lead to improved test coverage and fault detection, while reducing manual testing effort and maintenance costs.

To test this hypothesis, we developed an open-source testing framework combining symbolic and concolic analysis for RPA workflows. The framework was evaluated on real-world use cases provided by industry partners, allowing us to assess both its effectiveness and practicality in real development settings.

The main contribution of this article can be summarized as follows:

- We gather various methods of testing RPA systems and we implement them in a single open-source package at https://github.com/unibuc-cs/rpa-testing (accessed on 6 June 2025). While the evaluation and results presented are based on RPA processes developed using the UiPath Studio tool, the proposed architecture—designed with a clear separation of concerns—allows for the integration of other similar tools with minimal effort.
- A novel greedy strategy, referred to as All states ones, is introduced by this paper in Section 4.3.
- We validate our ideas and prototypes in collaboration with industry partners. In particular, in our work with our industry partners at UiPath, and based on experiments, we identified a suite of best practices to be used for testing RPA scenarios. These practices incorporate adapted methodologies from the software testing literature, such as mocking, human-in-the-loop testing, and white-box/black-box fuzzing. The evaluation is supported by prototype implementations across a variety of industrial RPA use cases.

The proposed framework is based on a few assumptions that support its effectiveness in structured automation environments. It assumes the availability of control-flow graphs and basic annotations, which are commonly present in modern RPA development platforms and help guide the path exploration process. The workflows considered in this study come from real-world domains such as banking and healthcare, but are of moderate complexity; while they reflect practical scenarios, larger or more dynamic systems may present additional challenges. The selection between the execution strategies is currently performed manually, which allows flexibility depending on coverage goals and time constraints.

The remainder of the article is organized as follows. The next section reviews prior research on RPA, as well as broader work on software testing techniques. Section 3 provides an overview of RPA concepts, tools, and the motivation for automating the testing of UI application using RPA technologies, including illustrative examples. Section 4 details the testing of RPA workflows using the All states once coverage strategy and further elaborates on workflow testing through symbolic and concolic execution methodologies, respectively. The evaluation of these methods on real-world applications, along with a discussion of best practices, is extended in Sections 5 and 6. Finally, the last section, Section 7, summarizes the main findings of the research, discusses its limitations, and suggests possible directions

for future investigations, including enhancements to the proposed methodologies and their applications in different industrial contexts.

2. Related Work

After studying the literature on RPA and testing, we found that there is no prior work on white-box testing for RPA workflows. Ref. [3] describes the current level of software robot testing and provides ideas for automation using model-based testing. However, they do not provide further evaluation or implementation proposals. Ref. [4] created a testing environment by observing user interactions with application UI interfaces and documenting them with logs and photographs. This method falls under "black-box testing", as it does not analyze an RPA workflow. Although fascinating, it has shortcomings when compared to our proposed methods. (a) It is more challenging to bring user information into the test field, and (b) in contrast to white-box testing of a given graph, recording actions solely using computer vision techniques might severely restrict the coverage of the test state space. Second, graphical user interfaces were tested using RPA in the papers by [3,5]. Furthermore, ref. [6] employed RPA for regression testing and automation for more typical information system components, such as desktop, mobile, or web applications, or even for large-scale software's API functionality testing. In the studies [5,7], graphical user interfaces were tested using RPA. Furthermore, ref. [6] employed RPA for regression testing and automation for more common information system components, such as desktop, mobile, or web applications, or even for API functionality testing in large-scale software.

Our testing methodology aligns with contemporary advancements in white-box testing, particularly emphasizing symbolic execution techniques. Recent studies have highlighted the integration of large language models (LLMs) in white-box testing to enhance test case generation and optimize testing processes. For instance, the development of WhiteFox demonstrates the application of LLMs in compiler fuzzing, leading to improved detection of compiler defects [8]. In the realm of symbolic execution, tools like FuSeBMC have been developed to combine fuzzing and symbolic execution, applying bounded model checking to identify security vulnerabilities in C programs [9]. Additionally, approaches such as TracerX [10] enhance dynamic symbolic execution by incorporating interpolation methods to manage path explosion and improve verification efficiency. A recent survey by [11] provides an overview of new research trends in symbolic execution, with particular emphasis on applications to test generation. These advancements have inspired our approach to repurpose cutting-edge techniques and contribute to the field through research. We assess and offer open-source versions of the suggested modifications to make them suitable for the Robotic Process Automation (RPA) testing domain. By examining and adapting earlier research focused on source code, Low Level Virtual Machine (LLVM), or binary testing, we aim to develop effective testing methods for RPA processes. In particular, the concept of symbolic execution has been applied to various levels of program representation, including source code and LLVM. For instance, tools like Concolic Replay and Testing (CRETE) [12] operate at the LLVM representation level, providing insights that have influenced our annotation strategies, which are discussed in subsequent sections. By leveraging these contemporary tools and methodologies, we aim to enhance the robustness and reliability of RPA testing processes.

In our study, ref. [13] delves into the potential of Robotic Process Automation (RPA) within the gaming industry, illustrating how automation technologies can address operational inefficiencies and enhance productivity. The authors highlight RPA's ability to streamline repetitive and error-prone processes, such as software testing, quality assurance, and data validation, which are integral to game development and maintenance. Further-

more, the study emphasizes RPA's role in augmenting user satisfaction by ensuring faster issue resolution and improved in-game experience.

Key areas of focus include the scalability of automation solutions in dynamic gaming environments and their contribution to cost reduction and operational resilience. Ref. [13] also outlines practical implementation frameworks and strategies for integrating RPA into existing workflows, providing insights for both industry professionals and academic researchers. By leveraging RPA, the gaming sector stands to benefit from optimized processes, reduced time-to-market for games, and a more adaptive approach to emerging challenges.

This work serves as a foundational reference for exploring automation in gaming and underscores the strategic importance of RPA in an increasingly competitive and technology-driven market. While prior work has explored testing in the context of RPA, we observed that the vast majority of these efforts focus on black-box methods or use RPA tools to test external systems. To our knowledge, no existing research has addressed white-box testing approaches that are specifically tailored to RPA workflows. In particular, the internal control-flow structure of workflows remains largely unexplored in test generation strategies. This lack of attention to white-box techniques in RPA represents a clear research gap. Our work aims to fill this gap by adapting symbolic and concolic execution methods to the specifics of RPA, offering a new direction for improving test coverage and reliability in this domain.

3. An Introduction to RPA and Test Automation

UiPath Studio (https://docs.uipath.com/studio/standalone/latest, accessed on 6 June 2025) is the environment in which the robots are visually designed and provides three primary workflow types that can be used to visualize the order in which a robot performs tasks: Sequence, Flowchart and State Machine. Examples of these can be found in this document in Figures 6–8. In the sequence, the activities are presented in a linear fashion. The flowcharts, on the other hand, provide greater flexibility with multiple decision points and arrows that connect any two actions. They are comparable to BPMN (Business Process Model and Notation) models, a graphical standard for modeling business processes that visually represents workflows using symbols such as events, activities, and gateways, as well as Unified Modeling Language (UML) activity diagrams. The state machines offer even more expressive power than flowcharts, as they support conditional transitions and internal states for nodes, making them similar to traditional UML state machines.

Data can also be incorporated into the aforementioned forms as local variables or workflow arguments. Additionally, each workflow can be hierarchically embedded within another workflow, allowing for greater modularity and reusability. For example, a workflow can contain a series of flowcharts with local state machines in certain nodes. When creating the workflow, existing basic activities, libraries implemented in source code (e.g., `C#` code fragments) or other workflows are inserted using drag and drop. Basic functions include reading and writing data to and from a variety of formats (PDFs, Excel, Word, common databases, desktop or web-based applications), working with all kinds of variables (since UiPath Studio is based on the .NET framework), creating reports, managing mouse clicks and keystrokes, and even using optical character recognition (OCR) to convert handwritten or printed text, as well as text contained in images, into machine-readable text. To create workflows, the user has the option to create them manually in the UiPath Studio interface, use the recording function or choose a hybrid of both. UiPath recording is a feature that captures user interactions with applications, such as mouse clicks, keyboard inputs, and UI navigation, to create automation workflows. It simplifies the process of building automation by generating sequences of activities that mimic human actions.

UiPath offers different recording types, including Basic, Desktop, Web, Image, and Native Citrix, each optimized for different environments. Once recorded, the automation can be modified, enhanced with logic, and executed by robots to streamline repetitive tasks efficiently.

We chose UiPath as the RPA provider because it offers a robust and flexible environment suited for implementing symbolic and concolic execution techniques tailored to RPA workflows. The selection criteria included the platform's support for workflow customization, accessibility of internal workflow structures, compatibility with our testing framework, and its widespread use in industry, which ensures practical relevance. While other RPA platforms could potentially be adapted to our approach, UiPath provided the best balance of technical features and ease of integration for this study.

3.1. Improving UI Test Automation with RPA

Automated tests for Robotic Process Automation (RPA) can significantly improve user interface (UI) testing. UI testing is often considered difficult due to various factors. Modern applications are often constantly updated to incorporate new features and functionalities, making comprehensive UI testing a challenge [14]. Furthermore, the increasing complexity of applications with features such as complex flowcharts, maps, embedded frames and diagrams further complicates UI testing [15]. Moreover, the dynamic nature of UI elements, which change frequently, leads to difficulties in maintaining UI test scripts [16]. When multiple or remote applications are involved in UI testing, additional challenges arise, such as limited control over graphical user interface (GUI) elements compared to testing a single web application with tools such as Selenium [17]. RPA can solve some of these problems by working effectively at the user interface level in heterogeneous environments. It often works without scripts, making it more accessible, and uses advanced technologies such as computer vision to improve stability and performance [18].

RPA has both opportunities and challenges, with many initiatives failing despite its potential benefits. Specifically, RPA's development life cycle is still in its early stages since it is a new paradigm. A key component is quality assurance, a laborious process that is essential to a project's success. The most popular method for ensuring that the deployed robots are operating as intended and with few faults is testing. Nevertheless, the majority of RPA testing is now carried out manually, and there are not many tools available to help with it. Based on the following examples, a discussion can be opened on the topic of RPA testing and, more precisely, how the state of the practice can be improved with state-of-the-art research in testing. In this way the study of automation in RPA testing can provide new research topics.

Based on the following examples, a discussion can be initiated on the topic of RPA testing, specifically focusing on how the current state of the practice can be enhanced by state-of-the-art research in testing.

3.2. Motivating Examples and Current Practice

3.2.1. Testing a Calculator Using RPA

Among the various available options, test automation can now also be achieved through RPA. A small example of this implementation can be found in the article "Improving UI Test Automation using Robotic Process Automation" [7], where two calculator apps were tested—one installed locally and the other accessible online. For this case, a robot was developed to test both calculators in parallel using test data stored in an Excel spreadsheet. The robot automatically interacts with the interfaces of both calculators, implementing a data-driven testing approach, extracts the results, and compiles them into a file for test reporting. Unlike traditional UI web testing frameworks, such as Selenium, which are

limited to working with browser applications, the robot is capable of switching between all three applications (the web calculator, the desktop calculator, and the Excel spreadsheet).

3.2.2. Testing the Back-Office of a Banking Application Using RPA

Another example of end-to-end testing using an unattended RPA robot can be found in the back-office operations of a banking application. Initially, the goal is to verify if the application functions correctly across multiple web browsers. The robot should be able to log in successfully regardless of the browser being used. Once the robot evaluates the login functionality, it proceeds to test the internal money transfer feature. The robot verifies whether the transferred amount reaches the intended account. If the transfer is successful, the test passes, and the outcome is saved. After the payment is completed, both the bank clerk and the robot should be able to view and store the payment confirmation in PDF format. To test this feature, the robot navigates to the menu, locates the payment confirmations, and saves the most recent file. It then opens the PDF file and checks the relevant fields to ensure it includes the transfer amount and the recipient account. If these fields are accurate and present, the test is deemed successful.

3.2.3. An Application for a Bank Loan

Consider a simple procedure (adapted from UiPath's documentation) that determines the type of bank loan based on its amount and duration.

Figure 1 shows the flowchart of the Create_Loan process, which contains three decisions (represented by a '?' and a condition). The first two relate to the loan amount and change the flow direction in a branch based on the loan category (low, medium or high). The last decision is about how long the loan can run for. The loan is considered short-term if it has a term of less than, say, five years; otherwise, it is considered long-term.

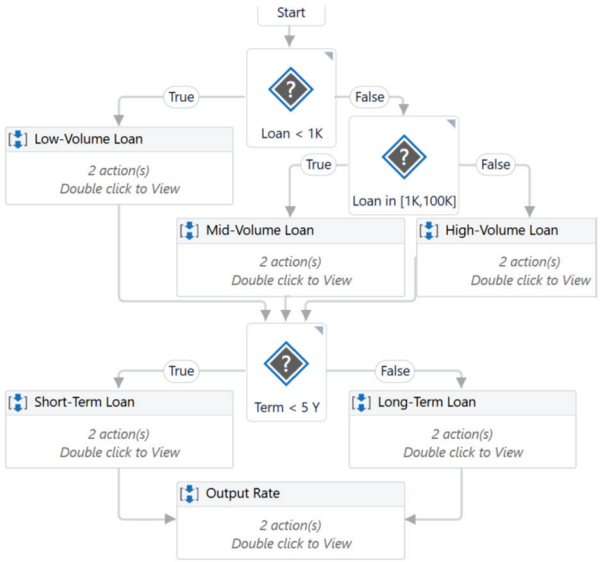


Figure 1. The main workflow of the BankLoan application.

Another, more complex example involves rule-based, high-volume processes in the retail industry. Rule-based, high-volume procedures are commonly automated in the retail sector, with refunding returned orders being a typical B2C (business-to-consumer) operation. This process often requires significant human resources, particularly during peak sales periods. By automating refund tasks with robots, businesses can save time, improve compliance, and optimize resource allocation. The robot's input is a web-based report containing details such as sales order numbers, customer account data, return dates,

and more. After logging into a customer relationship management (CRM) system, the robot retrieves each sales order number from the input report. It then identifies which items are eligible for a refund and calculates the refund amount for each request. If the customer has a voucher, the robot applies any relevant discounts. Once the total refund amount is determined, the robot accesses a second application, a payment platform, to transfer the full refund amount to the customer. The robot generates an output Excel file containing the sales order numbers, refunded amounts, and the status of each activity.

Current RPA testing techniques either rely on black-box UI observation or unstructured script-based testing. There is no existing symbolic or concolic execution engine tailored for workflow-based RPA environments with visual activity models. Our work fills this gap by adapting symbolic [19] and concolic analysis techniques to the structure of RPA workflows and implementing a scalable, open-source testing engine based on SMT solvers such as Z3 [20]. Symbolic execution is a program analysis technique that explores program behavior by determining the input values required to execute different code paths. While it theoretically offers high code coverage, practical adoption is often limited by the path explosion problem, where the number of feasible execution paths grows exponentially with program complexity.

Two principal variants of symbolic execution are employed in software testing: online symbolic execution and offline symbolic execution, also known as concolic execution.

In concolic execution, the program is executed with concrete inputs while symbolically tracking path constraints encountered at conditional branches. These constraints, representing logical conditions on symbolic inputs, are recorded as execution traces. At the end of each run, these traces are analyzed offline to generate new input values by solving the collected path constraints, enabling systematic exploration of alternative execution paths. This decoupling of path exploration and execution makes concolic execution more scalable and better suited for large and complex software systems compared to pure online symbolic execution.

These challenges underscore the need for more systematic, scalable, and automated testing methodologies in the RPA domain. While current RPA testing is largely manual and tool-limited, the dynamic and heterogeneous nature of UI workflows demands techniques that can adapt and reason about workflows structurally. This motivates our use of white-box testing methods, particularly symbolic execution and concolic analysis, to explore the workflow's internal logic and generate test cases with high coverage and minimal manual intervention. In the remainder of this paper, we formalize this approach and validate it through open-source implementation and empirical evaluation.

4. Methods

In this section, the first two subsections provide an overview of the known concolic and symbolic testing methods applied to the RPA domain, summarizing the current state of the art. We then present the technical details and insights gained during the implementation of these methods in the proposed open-source package, which constitutes the primary contribution of this paper. In Section 4.3, we introduce a novel technique, a hybrid of the previous two methods.

4.1. RPA Testing Using Symbolic Execution

The methods described below assume that the RPA workflow is represented as a directed graph G, where each action is modeled as a node. This graph may include cycles, with the starting action (i.e., node) designated as Initial(G). The decision factor for generating and executing test cases is primarily based on branch points, which represent activities that vary depending on specific conditions.

Algorithm 1 illustrates the main loop of the implementation pseudocode within our framework. The process begins at the initial activity node of the graph G (Line 2), inserts it with a single-node into a worklist W, extracts a promising path at each step of the loop (with customizable strategies for evaluating path priorities), and moves forward to Line 7. The worklist W is a priority queue sorted by priority of the element (path). The default priority calculation is determined based on the path's depth in the exploration tree of G. The algorithm prioritizes paths at the top of the exploration tree for quicker traversal, rather than focusing on long paths with frequent activity nodes. Our framework allows users to alter default behavior with function hooks according to specific testing goals.

Algorithm 1 The main loop of Symbolic Execution using DFS strategy.

1: `// Initialize the starting exploration path with the entry node of the workflow graph`
2: $start_path \leftarrow \text{SMTPath}([G.entry_node], \text{priority} = \text{MAX_PRIORITY})$
3: `// Add the initial path to the worklist`
4: $W.\text{add}(start_path)$
5: **while** $W \neq \varnothing$ **do**
6: $\quad currPath \leftarrow W.\text{extract}()$
7: $\quad$ EXECUTEPATH($currPath$, W)
8: **end while**

New entries in the work queue W are prioritized based on their depth index in the exploration tree, where the branching constraint is reversed. When sorting the priority queue W by priority, the default behavior favors modifications to paths that diverge closer to the root of previously explored paths. This approach is also employed in [21] and promotes broader early exploration. However, the user can override this behavior by providing a custom hook function if different prioritization is desired.

Algorithm 2 presents the code responsible for continuing the execution of a given path. The SMTPath class abstracts the execution path in this symbolic form. Line 3 adds an initial set of restrictions on variables in the DataStore (D), such as those propagated from annotations, to the SMT solver's state. The implementation parses the current node to make the continuation decision (Line 4).

When a branch node occurs (Lines 8–31), the current path is cloned and split into two; one takes the True branch and the other takes the False branch. To create a future worklist, we prioritize possible paths and assign the next node based on the branch under consideration. Line 27 shows a redundant DataStore (D) object for one path. This guarantees that each uses separate data context values. This enables concurrent execution of worklist branches. The implementation follows a depth-first search (DFS) approach, with the currently active path taking precedence over the other. The latter path is added to the worklist and may be run later (Line 29). Some implementations use a breadth-first-search (BFS) method, where both branches are added to the worklist based on priority, and a new path is explored at each step. Figure 2 illustrates the difference between the two.

Algorithm 2 Continuation of execution for a given path.

```
1: procedure EXECUTEPATH(path : SMTPath, W : Worklist)
2:     currNode ← path.nextNodeToExecute ▷ Each path stores the next node to execute (initial is entry node)
3:     path.conditions_smt ← DataStoreTemplate.getVariablesAssertions() ▷ Add constraints from DataStore annotations
4:     while currNode ≠ None do
5:         if currNode.type ≠ BRANCH then
6:             WorkflowExecutor.Execute(currNode)
7:             currNode ← currNode.next()
8:         else
9:             if not HasSymbolicVariables(currNode.condition) then
10:                Result ← WorkflowExecutor.Execute(currNode)
11:                currNode.advance(Result)
12:                continue
13:            end if
14:            newPath_onTrue ← SMTPath(
       nodes = currPath.nodes + currNode.nextNode(True),
       conditions_smt = currPath.conditions_smt + currNode.condition
15:            )
16:            newPath_onFalse ← SMTPath(
       nodes = currPath.nodes + currNode.nextNode(False),
       conditions_smt = currPath.conditions_smt + Not(currNode.condition)
17:            )
18:            solvableNewPaths ← []
19:            if SMTSolver(newPath_onTrue) has solution then
20:                solvableNewPaths.add(newPath_onTrue)
21:            end if
22:            if SMTSolver(newPath_onFalse) has solution then
23:                solvableNewPaths.add(newPath_onFalse)
24:            end if
25:            for all newPath in solvableNewPaths do
26:                newPath.scorePath()
27:                newPath.data_store ← currPath.data_store.clone()
28:                newPath.nextNodeToExecute ← currNode.nextNode(True if newPath = newPath_onTrue else False)
29:                W.add(newPath_onFalse)
30:            end for
31:            break
32:        end if
33:        StreamOut(currPath)
34:    end while
35: end procedure
```

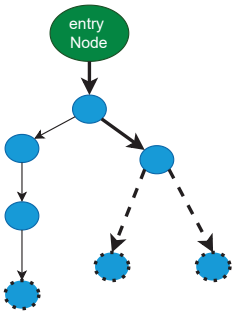


Figure 2. The distinction between the DFS and BFS execution types is depicted in the figure. Broken circles are used to symbolize frontier nodes, and visited nodes are those that remain. The route that is emphasized is the one that is currently being followed. The next iteration of BFS-style execution chooses the top continuation node (by priority), which may even be the unexplored node on the left side of the workflow graph. Both new boundary nodes are added to the priority queue worklist. One of the two new nodes is added to the work queue in the DFS manner, and additional execution proceeds on it.

4.2. Concolic Execution for RPA Testing

Our review of the literature revealed a notable gap in testing methodologies tailored to RPA workflows of varying complexity. To the best of our knowledge, no prior work has thoroughly addressed this issue. While the article "Towards Automated Testing of RPA Implementations" and the preceding section offer initial suggestions for automating testing through model-based techniques, they lack further analysis, reusable implementations, and technical depth. Our approach builds on whitebox testing principles. We explore, evaluate, and provide open-source implementations of the proposed adaptations to ensure their applicability to the RPA domain, leveraging state-of-the-art techniques. Specifically, we revisit and tailor prior research originally designed for source code, LLVM, or binary testing, adapting it to meet the unique needs of RPA workflows.

4.2.1. Overview of the Testing Pipeline

The RPA process specification undergoes a series of transformations before being handed off to specialized fuzzing techniques tailored to our use cases. The entire workflow is illustrated in Figure 3. Initially, the user builds an RPA workflow (W0) and an annotation set (S) with a variety of parameters. These inputs are then bundled and passed to a process called WorkflowParser, which parses W0 and S to produce a format that is easy to grasp for the subsequent fuzzing operations carried out by the WorkflowExecuter process. At the end of this process, a test suite is returned, which the tested robot can subsequently run. It is important to differentiate that the WorkflowExecuter simulates the state space environment to efficiently generate test cases, while the RPA Robot agent represents the actual machine or process in production. It is also feasible to generate test cases in real time on the robot side, but this would often involve a large latency and is therefore less efficient.

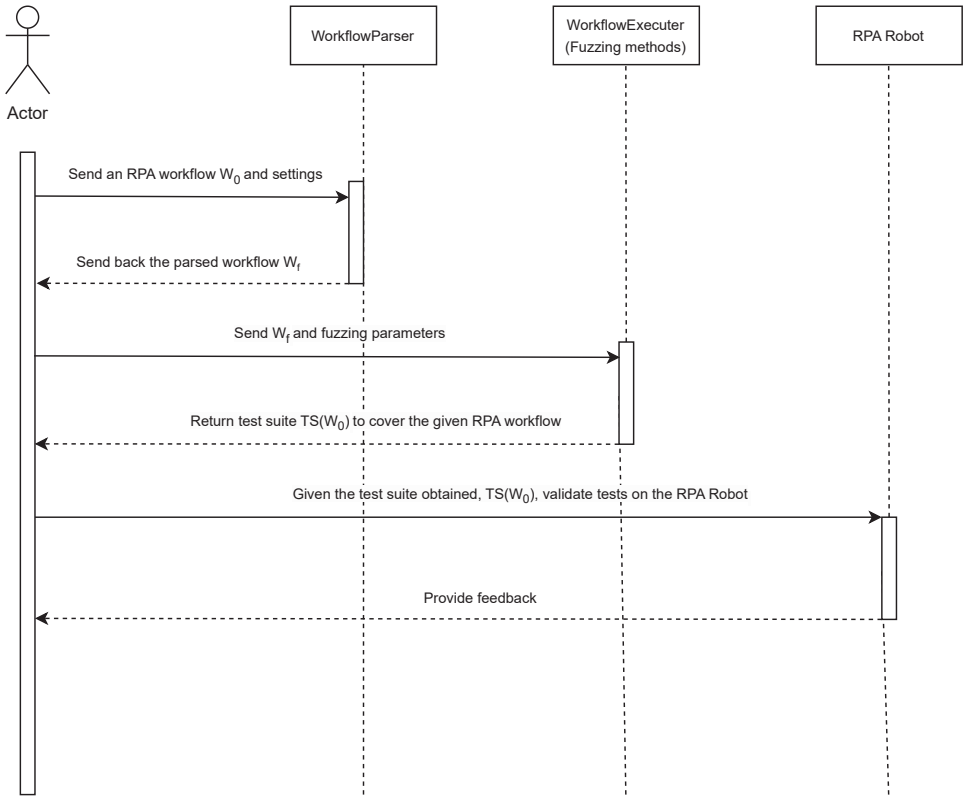


Figure 3. The overview of the full testing process, starting from the RPA workflow W0 given by the user to obtain a set of test cases and validate them on the robot.

Figure 3 provides a comprehensive overview of the full testing process, illustrating the roles and interactions of the various components described in this section.

4.2.2. Parsing Operation

The WorkflowParser procedure takes an RPA workflow W0 (Figure 2) that is generated in a text or visual application and converts it into a ContextFlowGraph before exporting it physically as a JavaScript Object Notation (JSON) file (Wf). The exported format represents each activity in the workflow as a hypernode in the graph G. If there is a decision branch in the activity, then the related nodes in G are linked appropriately. For example, if the activity involves querying databases and setting variables based on the findings, the hypernode is made up of multiple interconnected nodes that build an abstract syntax tree (AST) with low-level operations. Figure 4 illustrates the hypernode formed by an activity with a source code, as shown in Algorithm 1. Our environment supports several types of AST nodes by default, including logical and mathematical expressions, function calls, array and data table access, and variable reading and writing utilizing a DataStore, as described in Figure 5. Our collaboration with industry partners highlighted the importance of allowing users to contribute their own nodes to the project, as different functionalities may be required depending on the workflows being tested. Our research indicates that annotation support is crucial for high-quality testing. Annotation involves adding additional information to activity nodes, such as hints. Annotations can specify limits and patterns, saving testing resources from unnecessary pathways. The human-in-the-loop approach allows users to offer valuable information for fuzzing methods. This reduces computation and creates test cases that are understandable to humans. To ensure visibility for all stakeholders, we recommend setting annotations directly in the RPA workflow editor (see Algorithm 3).

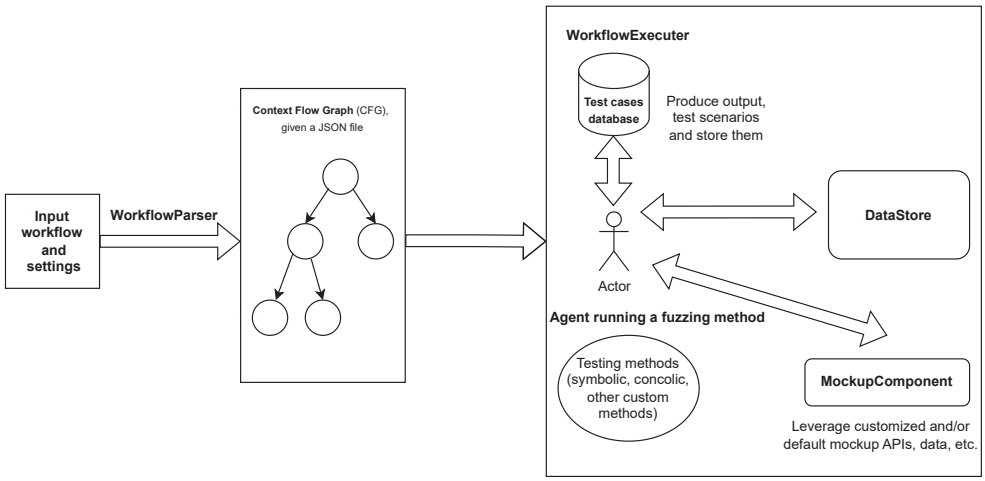


Figure 4. The detailed overview of the steps and components involved in the testing process.

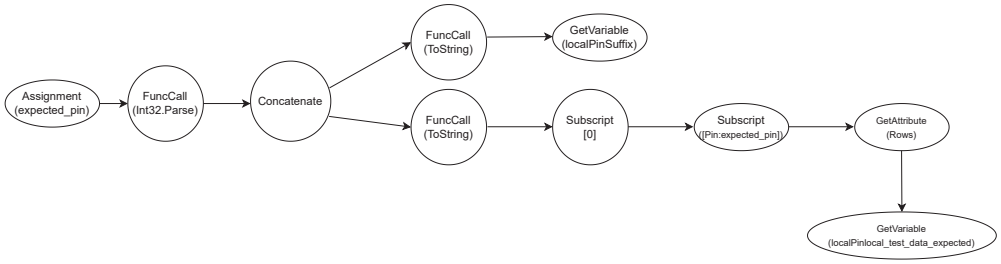


Figure 5. The AST generated internally by the WorkflowParser after evaluating the source code presented in Algoritm 1.

Algorithm 3 Example of an expression contained in an activity node.

```
// Input:  local_test_data_expected, localPinSuffix
// Output:  expected_pin
expected_pin = Int32.Parse(
   local_test_data_expected.Rows[0]["Pin:expected_pin"].ToString()
   + localPinSuffix.ToString())
```

4.2.3. Annotations

In the proposed testing framework, annotations are used to define boundaries and patterns that help the engine avoid executing irrelevant or invalid paths. Annotation refers to the insertion of additional data—such as hints or metadata—into activity nodes. For instance, by applying a human-in-the-loop approach, users can contribute valuable domain knowledge to guide fuzzing techniques. This collaboration not only reduces computational overhead but also leads to more human-readable and meaningful test scenarios. We found that making annotations directly visible in the RPA workflow editor is a practical design choice, helping both developers and testers understand how test data is controlled and interpreted. Currently, our system supports two main categories of annotations:

- Variable-level annotations are defined as additional information associated with variables, such as min/max ranges, pattern restrictions, or user-specified input values. This metadata is used to reduce the search space during test data generation and ensure realistic distributions. It also allows the user to insert CSV mock files to simulate input values for specific variables. For example, the user can set a variable named actual_pin_values to accept only numbers between 0 and 9999 and also declare it as a

user input. A mock data file can then be used to populate the value of this variable during test execution. Algorithm 4 provides an example.

- Source-code-level annotations are used to insert small code snippets that perform additional operations before or after a node is executed. This is mainly used to define assertions or apply mock behavior. Algorithm 5 shows a simple expression annotation that is evaluated as an assertion to verify whether the value of the pin at a certain index equals the expected value. These expressions are written using a domain-specific expression language and are evaluated during test execution.

Algorithm 4 Example: Various support for annotating the variables in the activity.

```
@Variables
{
  "local_number_retries":{"min": 0, "max": 3},
  "local_test_data_expected":{"path": "pin_mocked_data.csv"},
  "local_test_data_actual":{"path": "pin_real_data.csv"},
  "actual_pin_values":{
    "bounds": 10,
    "min": 0,
    "max": 9999,
    "userInput": "True"
  }
}
```

Algorithm 5 Example of additional code sanity check for a PIN number added as annotation to an activity in the workflow.

```
@Expression
{
  "expression":
  "actual_pin_values.GetElementAt(local_number_retries) ==
expected_pin"
}
```

Annotations are processed by the testing component, which takes as input the exported Control Flow Graph (CFG) and annotation metadata in JSON format, both generated by the WorkflowParser component (Figure 3).

A central DataStore (D) holds both concrete values and symbolic representations of variables influenced by user input. In the literature, this influence is referred to as tainting—describing how external inputs (like user actions or network data) can impact variables. The initial set of tainted variables, SymbolicVars (D), is either user-supplied through annotations or automatically detected as external input. As the workflow executes, other related variables become symbolic as well. The DataStore contains a structure Values (D), which maintains the current concrete values for all variables—symbolic or not—throughout evaluation. then executes the CFG node by node, following branches and links. At each step, the DataStore updates both the concrete and symbolic/tainted values to reflect current context and branching conditions. When a node is executed, its functionality—defined by the workflow—is evaluated. These evaluations may have visible side effects in the DataStore. In the abstract syntax tree (AST) of a node (see Figure 4), a leaf might represent a function like GetVariable, which fetches data from the DataStore.

The expression is then propagated to the root of the tree and evaluated. The evaluation of node-defined functionality can cause apparent side effects in the DataStore. Understanding the function and significance of real-time node evaluation is critical. In the matching node AST (Figure 4), a leaf has a GetVariable function call that fetches data from the *DataStore*. The expression is then moved to the top of the tree.

$$\begin{aligned} expected_pin &= Eval(node.expression) \\ Values(D) &= Values(D) \cup expected_pin \end{aligned} \tag{1}$$

The WorkflowExecutor is capable of executing the graph with different strategies around decision points or branches.

4.2.4. Mocking

Mocking refers to the practice of replacing real data sources, library calls, or system components with controlled substitutes. This can help reduce test complexity, avoid reliance on external systems, and focus testing on specific components or logic paths. The proposed framework supports three types of mocking:

- Data sources mocking is useful when the workflow relies on remote databases. For example, a distributed SQL (Structured English Query Language) server used in production can be replaced with a local CSV file for testing. The mock file should have the same schema and distribution as the real data. These mock files can be periodically updated or linked to the real database via automated scripts. The framework uses an extensible DataTable component that can be accessed locally or remotely through a RESTful (Representational State Transfer) API (Application Programming Interface). This component can be customized by users to fit their specific requirements.
- API/Library mocking is useful when workflows call external APIs or libraries. Consider the example shown in Algorithm 6, which illustrates an example of such a call. Our framework supports this by offering an extendable table of classes and functions, like doServerRestAPI.getClientSalary, ToString, and getServerTimeOfDay. where an API is used to fetch the salary of a client, convert it to a string, and append a timestamp. This type of logic is common in real-world workflows but may introduce dependencies that are not desirable during testing. In this case, the user can register a mock implementation of the API or library functions, which will be called instead of the real ones. For example, the getClientSalary or getServerTimeOfDay functions can be overridden to return fixed values.
- Data types mocking is useful when working with complex or heavy types, such as System.Threading.Channels in C#. These types can be replaced with mock classes implemented in Python or C#, depending on the execution engine. This allows the user to simulate the behavior of certain components without introducing execution overhead or side effects. This mocking technique is aligned with modern software engineering practices such as dependency injection and inversion of control.

Algorithm 6 Example of multiple library calls within a node in workflow activity.

```
1: // Expression inside activity node
2: out = doServerRestAPI.getClientSalary(loggedID).ToString()
3:        + getServerTimeOfDay().ToString()
```

4.2.5. Generation of Test Cases

Based on evaluations and benchmarking with industry partners, we found that partitioning the generated test cases into smaller units at runtime represents a best practice for constructing extensive test corpus and efficiently leveraging the computational capacity of deployed RPA robots. The granularity of these partitions can be user-defined. This is crucial for continuous integration and regression testing, as the test framework and RPA robots may function in a producer–consumer model. In this setup, the framework acts as the producer by generating test case scenarios, while the RPA robots act as consumers, utilizing them simultaneously.

4.2.6. The Concolic Execution Engine

Our evaluations indicate that symbolic execution is more effective for small-scale RPA tasks. In contrast, concolic execution is better suited for achieving state-activity coverage in modern agile development environments, especially when graph G can be explored through loops or when data-dependent synchronization is required. Our framework incorporates the SAGE (System for Algebra and Geometry Experimentation) method's basic concepts.

To adapt concolic execution to the RPA domain, the proposed test is run through fuzzing with an RPA software robot, identifying branch points along the workflow path. At the end of the execution, the inputs are changed to a different set, resulting in different paths the next time. The software robots are operated remotely by another robot or a server procedure. Algorithm 7 shows the algorithm's main loop. The procedure uses a worklist (line 2) with inputs that span the workflow's variable space. This list prioritizes items based on their potential for increasing coverage, uncovering hidden states, and delving deeper into the state space tree. We allow users to specify the priority definition as an overridable function for maximum flexibility during testing. Common examples in the literature include increasing code coverage, identifying difficult scenarios, and addressing error-prone sections of source code. This inversion of the control technique, combined with annotation support, integrates human expertise into the testing process. The priority queue might begin with an initial set of input data, either randomly generated or from a test corpus. To optimize efficiency, it is crucial to add an initial set of input data to the worklist W during concurrent execution. This allows for early discovery of critical parts of the activity graph. The method would take longer to obtain the necessary coverage rate if it were not optimized. Our version of the architecture incorporates human knowledge into the production of initial input seeds, a feature known as "human-in-the-loop". To start testing, a user-supplied CSV file with inputs for the workflow's variables must be present. Rows specify test cases, and columns indicate the ordered set of variables in the workflow being tested. Missing data is acceptable, and only available data is used. Users can prioritize tests (last column in each row) and set them as key values in the worklist.

Algorithm 7 Main loop of a server or RPA robot for collecting concurrent execution tests.

```
1: // Initialize the starting set of input seeds
2: W ← {any initial method to fill starting set}
3: TestCorpus ← {W}
4: while W ≠ ∅ do
5:     R ← findRobotToExecuteTest()
6:     if R ≠ null then
7:         T_0 ← W.extract()
8:         R.newPath ← SMTPath(T_0)
9:         R.ExecutePath(R.newPath)
10:        NewInputs ← R.GenerateNewInputs(R.newPath)
11:        TestCorpus ← TestCorpus ∪ NewInputs
12:        W ← W ∪ NewInputs
13:    end if
14: end while
15: return TestCorpus
```

Line 5 of the main test coordination loop selects an RPA agent suitable for performing the test. In our usual implementation, this is the first available free agent. The function blocks until one of the robots becomes available. The robot executes the most promising input seed from the worklist (line 9). At this point, the path consists of a series of activities that take place when the input is executed. The path includes information on each branch point, including the SMT condition it operates under. In line 10, this information is used to modify the initial input text, allowing for different execution paths. The resulting TestCorpus includes both the original and newly discovered tests.

Algorithm 8 provides pseudocode for performing a path based on an input. In the workflow graph G, execution starts at the initial node. If the current node is not a branch decision (line 6), the execution will proceed to the next node. The executor evaluates the condition within the node (line 7) and returns True or False accordingly. If the branch node's condition does not rely on user-controlled variables, the execution will proceed with the condition evaluation result, as shown in line 11. If not, both the SMT assertion and its inverse are recorded in the route data structure for future processing (line 16). Finally, execution proceeds. The pseudocode summarizes all branch point assertions that potentially change pathways when the SMT solver changes the malformed input material associated with the branch conditions. It is worth noting that only the branch decision constraints can be changed to affect the execution path. The annotations set restrictions that are read-only (line 6). When a constraint is set due to a branching choice (Lines 8–14), the SMT solver (S) can potentially generate additional pathways. First, S's current stack context is stored (Line 9). Next, the inverse constraint is applied to S. If the modified constraints in S are now satisfiable (Line 12), a new input may lead to a different path of activities. Line 12 updates the set of inputs. To adjust a single branch condition with each new input, restore the previous SMT state and add the path evaluation decisions to the SMT solver. The execution behavior remains unchanged, save for the modification of the branch node for the new path.

Algorithm 8 Executing an input path in concolic mode and creating an `SMTPath` object.

```
1:  procedure EXECUTEPATH(path: SMTPath)
2:      currNode ← path.entryNode
3:      indexBranchAlongPath ← 0
4:      path.conditions_smt ← DataStoreTemplate.getVariablesAssertions()
5:      while currNode ≠ None do
6:          if currNode.type ≠ BRANCH then
7:              WorkflowExecutor.Execute(currNode)
8:              currNode ← currNode.next()
9:          else
10:             Result ← WorkflowExecutor.Execute(currNode)
11:             if ¬HasSymbolicVariables(currNode.condition) then
12:                 currNode.advance(Result)
13:                 continue
14:             end if
15:             if Result = True then
16:                 Taken_assert ← currNode.condition
17:                 Inverse_assert ← Not(currNode.condition)
18:             else
19:                 Taken_assert ← Not(currNode.condition)
20:                 Inverse_assert ← currNode.condition
21:             end if
22:             path.conditions_smt.add(Taken_assert)
23:             path.decisions_branch.Add(indexBranchAlongPath, Result, Inverse_assert)
24:             indexBranchAlongPath ← indexBranchAlongPath + 1
25:             currNode.advance(Result)
26:         end if
27:     end while
28:     StreamOut(currPath)
29: end procedure
```

4.3. All States Once Coverage Strategy

Our proposed technique is a unified white-box testing framework for RPA workflows, based on symbolic execution, concolic execution, and a custom coverage strategy ("All States Once"), described by Algorithm 9. The framework models an RPA workflow as a control-flow graph (CFG) where each node represents an activity, and branches correspond to conditional transitions. Symbolic execution systematically explores all feasible paths by collecting constraints at each branch and solving them using the Z3 satisfiability modulo theories (SMT) solver. Concolic execution complements this by executing workflows with real input data while collecting symbolic constraints along the path, which are then mutated to explore alternate paths. The "All States Once" strategy ensures that each node in the workflow is covered at least once, optimizing test coverage with reduced cost. This technique explores all execution paths within the workflow graph G, beginning from its set of entry nodes, referred to as Initial(G). The term "alternative term" in this context refers to the distinct sets of nodes located between different paths. Let Paths(G) denote the complete set of paths within G. Any two paths, P1 and P2, in Paths(G) are considered distinct if they differ by at least one node. Notably, paths that traverse the same loop a different number

of times are not treated as separate, as the technique does not differentiate based on the number of loop iterations. Conceptually, the aim of this strategy is to achieve coverage of all workflow nodes (activities) at least once. After collecting the set of all possible paths using a depth-first search (DFS) traversal, a SMT solver—specifically, Z3—is used to generate suitable test data for each path, based on the constraints accumulated along it.

The technique was developed based on prior symbolic/concolic methods used in software testing, but adapted to the unique structure of RPA workflows, which are visually defined, data-driven, and highly dependent on UI interactions. We worked closely with industry partners, particularly UiPath, to refine the approach to be practical and efficient on real workflows. The implementation is open-source, designed with extensibility in mind, and evaluated on six diverse RPA use cases, demonstrating its ability to generate relevant test cases, achieve full coverage, and reduce manual testing effort.

Algorithm 9 Symbolic execution using the "all states once" coverage strategy.

```
procedure SymbolicExecuteAllStatesOnce(G)
    setOfPaths ← getAllPaths(G, G.entry_node, [], {}, [])
    for all P ∈ setOfPaths do
        constraints ← getPathConstraints(P)
        S ← SMTSolver.solve(constraints)
        if S is valid then
            StreamOut(S)
        end if
    end for
end procedure
function getPathConstraints(P)
    conditions_smt ← [ ]
    for i ← 0 to |P| − 2 do
        node ← P[i]
        if node.type = BRANCH then
            nextNode ← P[i + 1]
            if node.nextNode(True) = nextNode then
                evalExpected ← True
            else
                evalExpected ← False
            end if
            if evalExpected then
                conditions_smt.add(node.condition)
            else
                conditions_smt.add(Not(node.condition))
            end if
        end if
    end for
    return conditions_smt
end function
function getAllPaths(G, currNode, currPath, visitedNodes, outAllPaths)
    currPath.add(currNode)
    visitedNodes.add(currNode)
    if currNode.next is None then
        outAllPaths.append(currPath)
        return
    end if
    anyValidSuccessor ← False
    for all S ∈ currNode.next do
        if S ∉ visitedNodes then
            anyValidSuccessor ← True
            getAllPaths(G, S, currPath + {S}, visitedNodes, outAllPaths)
        end if
    end for
    if not anyValidSuccessor then
        outAllPaths.append(currPath)
    end if
end function
```

5. Evaluation and Results

We evaluate our collaborative execution architecture on real-world applications that leverage RPA processes to automate repetitive, human-performed tasks. A brief overview of each application used in the evaluation is provided below.

5.1. Bank Loan

The first application, BankLoan, is designed to verify the type of a loan based on its amount and duration. Loans under 1000 are categorized as low volume, those between 1000 and 100,000 as medium volume, and loans exceeding 100,000 as high volume. For high-volume loans, an additional step is introduced to validate a PIN associated with the bank account. This step includes a retry mechanism—if the robot fails to provide the correct PIN (e.g., the expected PIN is "1234"), the workflow will fail after a predefined number of attempts. Regardless of the volume, each loan also has a duration attribute; loans with a term shorter than five years are considered short-term, while those with longer durations are classified as long-term (see Figure 6).

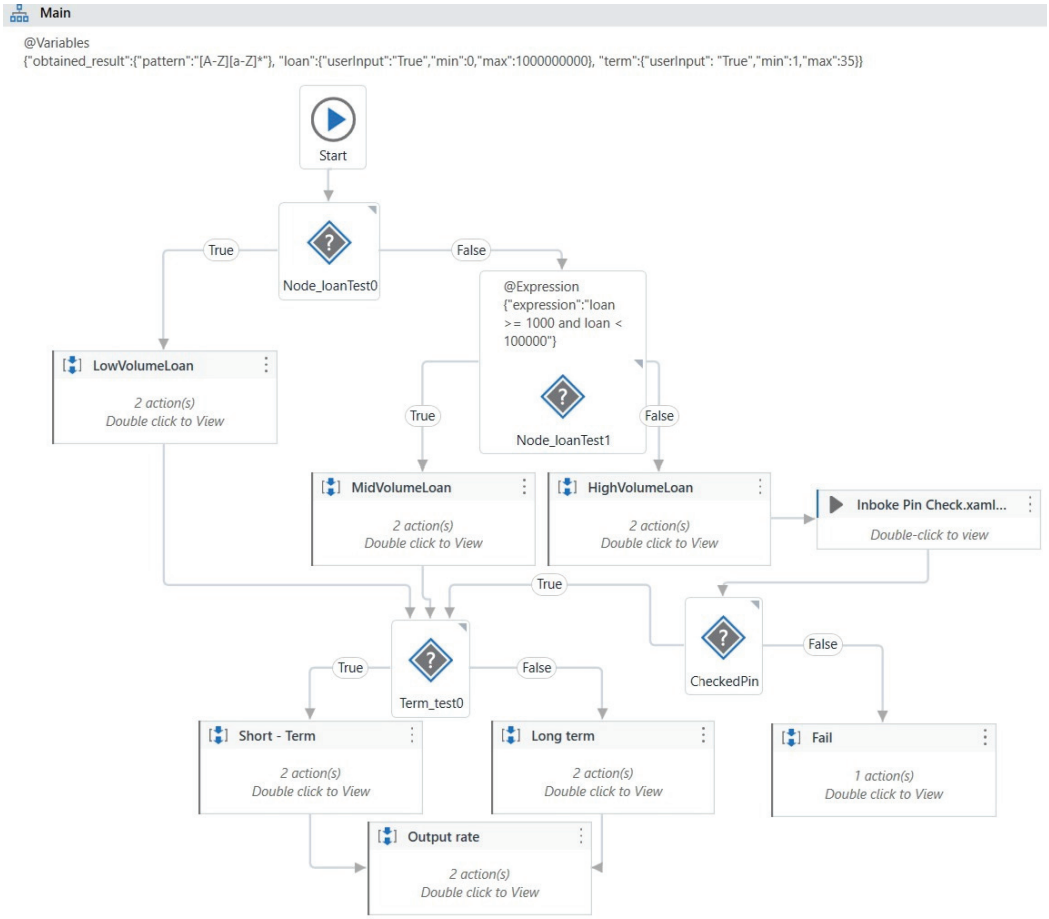


Figure 6. The main workflow of BankLoan application.

For this model, a dedicated test case was created. By executing it, testing data is generated, allowing full coverage of the model's activities. The workflow is represented by Figure 7.

In the given section, the test data for the workflow is generated by invoking the GenerateTestData.xaml workflow. This workflow begins by reading a configuration file, which contains the following parameters: the path to the Python script responsible for connecting all external components involved in the testing process, the working directory path, the Python home path. After reading the configuration, the workflow calls the Python

script, which orchestrates the external components and initiates the test data generation process. During the "when" phase, the generated test data is saved to CSV files, which will be consumed later during the execution of the workflows. In the "then" phase, the main workflow is invoked. In order to generate appropriate test data, the fuzzer component requires specific constraints for each variable. These constraints guide the fuzzer to provide input values that fall within valid intervals, match regular expressions, or represent specific types (e.g., Boolean values). To define such constraints, RPA developers can add annotations either at the workflow or activity level. In the BankLoan example, annotations are used in two main scenarios. In the first scenario, in the Main.xaml workflow, an annotation defines ranges for the main variables. For instance, the loan variable is constrained to values between 0 and 1,000,000,000, while the term variable must be between 1 and 35. These ranges allow the fuzzer to generate test data within the expected intervals. The second scenario is in the pin workflow where an annotation sets constraints for the number of retry attempts and the PIN values. The number of retries is limited to a range between 0 and 3, and the acceptable PIN values fall between 1111 and 9999. Once the annotations are processed, the generated data is saved in a CSV file named generatedTests.csv, which will be used during the testing sequence. The extracted data will be passed as parameters to the workflows during invocation.

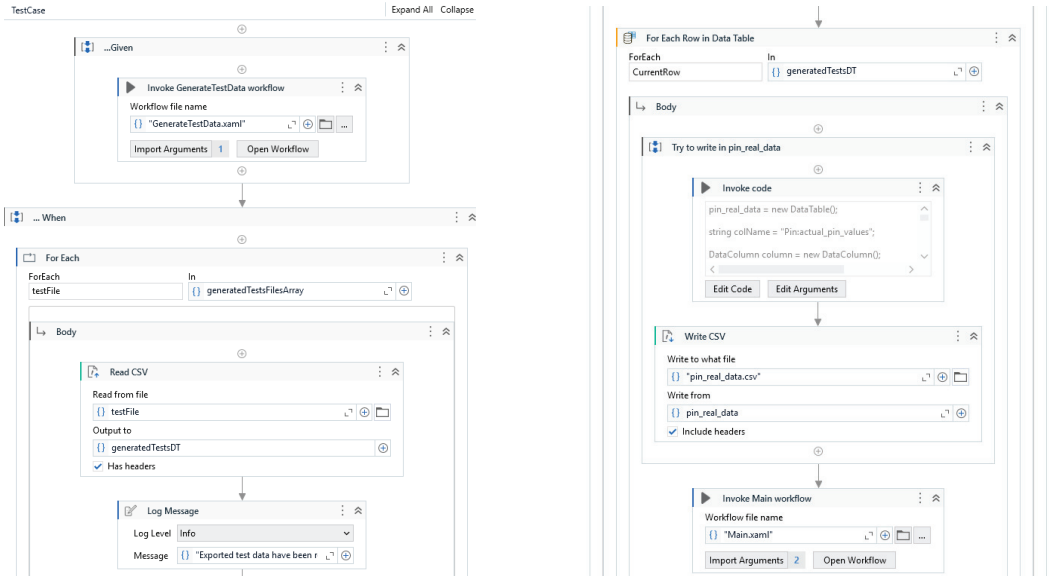


Figure 7. Test case for bank loan application.

5.2. Automating Loan Operations in Banking

Similar to the previous model, this one involves user interface interactions, according to Figure 8. It navigates to "https://uibank.uipath.com/" (accessed on 6 June 2025) to verify if a person qualifies for a loan with UiBank—a prototype website developed by UiPath Academy to demonstrate examples. Upon accessing the UiBank homepage, the robot clicks the "Products" button located in the upper-right corner, then selects "Loans" from the dropdown menu. On the loans page, the robot clicks the "Apply for a Loan" button. On the following page, several values are entered, including the applicant's email address, the loan amount requested, the loan term, the applicant's current year income, the applicant's age. If these values fall within certain predefined intervals, such as a minimum income required for the loan and an age of at least 18 years, the applicant qualifies to apply for the loan. If any of the conditions are not met, the individual is deemed ineligible to apply and must wait until all criteria are satisfied. The test data generated for this model includes the loan amount requested and whether or not the applicant is qualified for the loan.

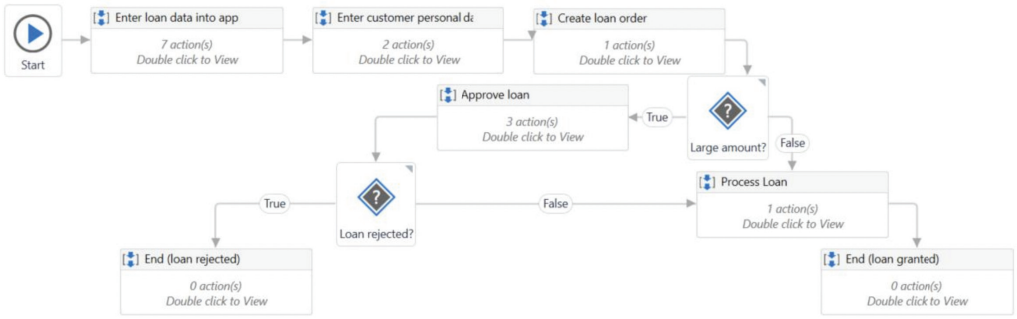


Figure 8. An example of RPA workflow that automates the loan operations for a banking process.

5.3. Application for Employee Management

In this application, the operations that can be performed are determined by the role of the user. The workflow interacts with a database containing employee details, including their name, salary, registration date, academic qualifications (specifically the presence of a master's degree), contract type, project assignment, leave requests, and notice periods. Administrators are permitted to modify the salaries of employees who have been with the organization for over one year and possess a master's degree. Conversely, users with human resources (HR) privileges can review and approve or deny employee leave requests. Leave requests are not granted if, for example, an employee with a short-term contract applies for unpaid leave or if the employee's notice period is less than five days. Furthermore, the workflow calculates the number of employees holding a master's degree and the distribution of employees across different projects. If the count of employees with a master's degree is less than or equal to two, the employer is encouraged to foster further educational development. Additionally, if fewer than two employees are assigned to a project, the workflow recommends the assignment of additional staff to the project. The workflow is represented by Figure 9.

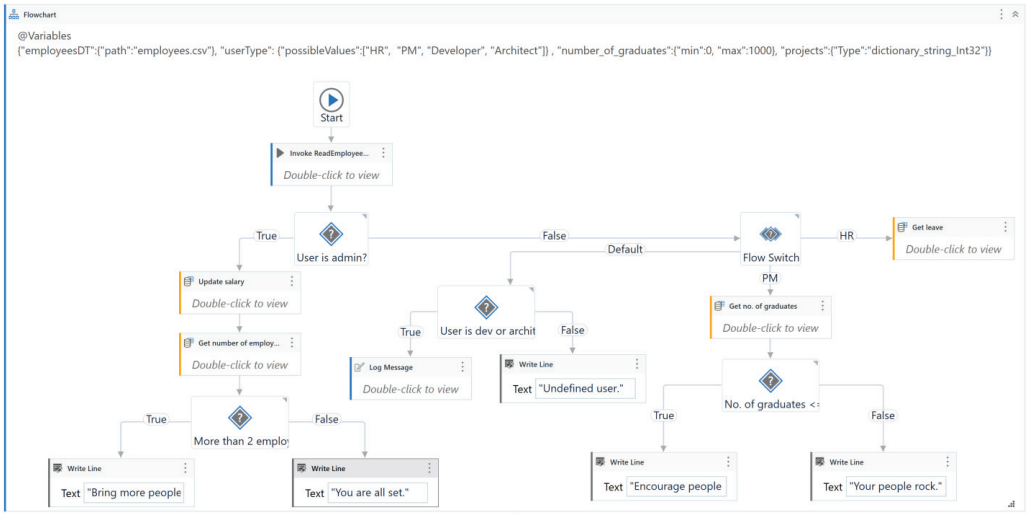


Figure 9. The main workflow of employee application.

A test case for this model is designed to generate test data that ensures comprehensive coverage of the workflow. The generated data include the type of user, the number of employees with a graduate degree, and the number of employees assigned to each project. This data allows for the evaluation of various conditions and actions within the workflow, ensuring that all potential scenarios are adequately tested.

5.4. A Hospital Services Application

The PrivateHospital model, illustrated in Figure 10, allows the user to select medical services and calculates the final price, incorporating discounts on specific services when applicable. Upon execution, the robot greets the user and requests their age. A list of available medical services is presented, and the user can select their preferred service. After each selection, the robot displays the price of the chosen service and prompts the user to decide whether they wish to add another service. If the user opts to add another service, the process repeats, and the price of the next service is added to the total cost. When the user decides to stop adding services, the final price is displayed, and the process concludes. At the outset, the user's age is requested to determine applicable discounts. Users under 18 years of age receive a 50% discount on selected services, while those over 65 years of age receive a 60% discount. Users between 18 and 65 years old are not eligible for any discounts. The generated data for a test case includes the user's age, the service selected from the input dialog (such as magnetic resonance imaging (MRI), neurosurgery consultation, or physical therapy), and the user's choice to add another service or not.

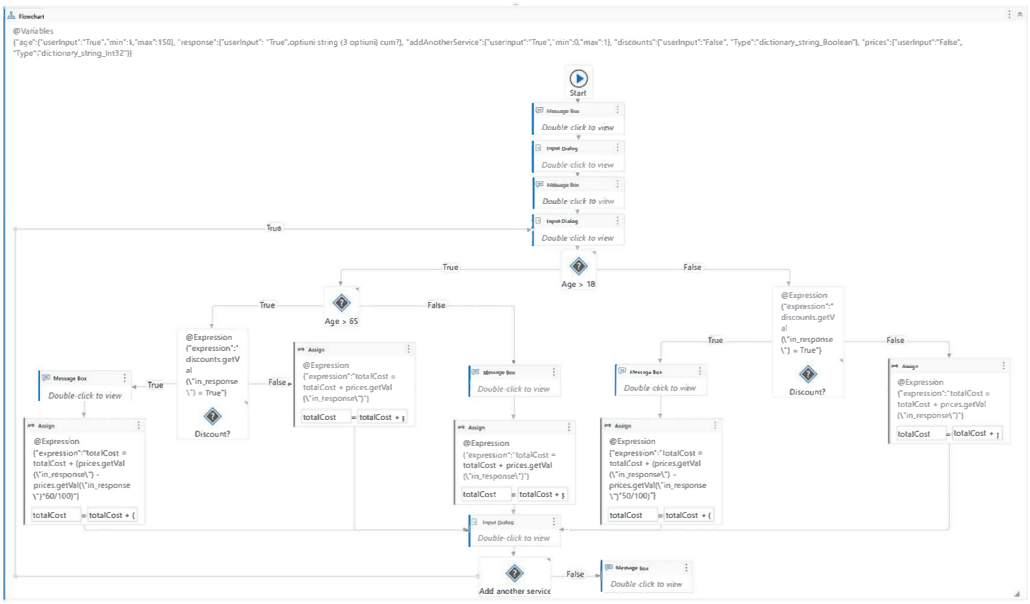


Figure 10. The main workflow of hospital application.

5.5. A Producer–Consumer Communication Pattern

The Producer and Consumer models are two distinct processes that interact through a queue in the Orchestrator, a web application designed to orchestrate the execution of repetitive business processes by facilitating the creation, monitoring, scheduling, and control of automated bots and workflows. The Producer model scans a local folder on the user's machine for PDF (Portable Document Format) files, which are invoices. It extracts key data such as client information, invoice number, due date, unit price, discount, total amount, and value added tax (VAT) number. The VAT number is determined by checking the country code from the invoice to apply the correct VAT value. If the payment due date exceeds 15 days, the invoice data is not added to the Orchestrator queue. Before inserting the data into the queue, the robot also checks a database to determine if the client has any unpaid invoices. If unpaid invoices exist, the robot includes the value of each outstanding invoice along with a penalty of 0.3% for each day the payment is overdue. The Consumer model reads the elements from the queue, processes them, and outputs the results into an Excel file. A test case for the Producer model can generate data for the payment due date, whether the vendor has unpaid invoices, the number of overdue days, and the country

code. A test case for the Consumer model can generate data for the "ShouldStop" flag provided by the Orchestrator, which is a Boolean value indicating whether the robot has been stopped by a user. When the flag is set to true, the Consumer should halt processing. Additionally, the test case can generate data for the number of transactions remaining in the queue and for the fields extracted from the invoices, such as vendor, invoice number, due date, unit price, discount, and total amount.

5.6. Spreadsheets _StateMachine—State Machine Data Processing Paradigm

The Spreadsheets_StateMachine is a model designed to process employee data from an Excel file. It reads the file, filters the data based on the salary column, and writes the filtered information into a new Excel file. The new file contains only the employees whose salary is less than 10,000. This model utilizes state machines to manage the processing steps, as indicated by its name. The state machine framework facilitates the structured flow of operations, ensuring that the data is efficiently filtered and stored according to the specified condition.

6. Metrics and Discussion

Table 1 presents various metrics reflecting the complexity of the applications under examination. These metrics include the total number of nodes (activities) and the number of branch points for each model. It is important to mention that connections within the applications that lead to cycles are considered to have an initial end node, which is counted as a branch point. This approach allows for a more comprehensive assessment of the complexity of each application's workflow, considering both linear and cyclic paths.

Table 1. The number of nodes (workflow activities) and branch points for each evaluated application.

Model Tested	Number of Nodes	Branch Nodes
BankLoanCheck	32	6
BankLoanRequest	20	11
EmployeesManagement	23	8
PrivateHospital	42	14
InvoicesProcessing	32	5
InvoicesManagementApp	39	8
Spreadsheets_StateMachine	8	3

The following two tables, Tables 2 and 3, present the metrics associated with the time required for each application to achieve 100% coverage, as well as the coverage obtained after running the fuzzer for the All states once strategy. The first table details the time performance for achieving complete coverage, while the second table outlines the extent of coverage attained by the fuzzer after processing all states in each application.

Table 2. The time needed for each application to obtain a 100% coverage.

Model Tested	Coverage Time in Minutes All States Once
BankLoanRequest	25.4 min
BankLoanCheck	18.33 min
EmployeesManagement	58.2 min
PrivateHospital	78.9 min
InvoicesProcessing	37.3 min
InvoicesManagementApp	70.2 min
SpreadsheetsStateMachine	20.4 min

Table 3. Coverage of the evaluated applications obtained after letting the fuzzer run for a fixed time of 60 min using the All states once execution strategy.

Model Tested for All States Once	Coverage for All States Once (%)
Bank loan	100%
BankLoanRequest	100%
EmployeesManagement	100%
PrivateHospital	100%
InvoicesProcessing	100%
InvoicesManagementApp	100%
SpreadsheetsStateMachine	100%

6.1. Metrics and Discussion for Concolic Execution

Table 4 presents the time required for the concolic technique to achieve full coverage of both activities (nodes) and branches. An important observation is that the time required to achieve full coverage is not directly proportional to the number of nodes, as demonstrated in the case of the InvoicesProcessing application.

To meet the demands of today's agile development processes, which include evaluating local changes before submission, continuous integration, and smoke testing, it is essential for the testing tool to provide adequate coverage within a short period of time. In order to assess this criterion, we employ the concolic execution technique to determine the depth of the testing process. The results obtained after executing the technique for one hour are presented in Table 5. This table provides an insight into how much coverage is achieved in terms of both activities and branches within the specified time frame.

Table 4. The time taken for each application to achieve 100% coverage.

Model Tested—Concolic	Coverage Time in Minutes Concolic
BankLoanCheck	33.91
BankLoanRequest	24.19
EmployeesManagement	57.30
PrivateHospital	81.43
InvoicesProcessing	45.72

Table 5. Coverage of the evaluated applications obtained after letting the fuzzer run for a fixed time of 60 min using the Concolic execution strategy.

Model Tested—Concolic	Coverage (%) Concolic
Bank loan	100%
BankLoanRequest	100%
EmployeesManagement	100%
PrivateHospital	92%
InvoicesProcessing	100%

6.2. Metrics and Discussion for Symbolic Execution

Table 6 displays the time required by the fuzzer to complete full coverage of the activities (nodes) and branches. We save the evaluation of the coverage of pairs of state values for future work in the next release of our fuzzing framework. One remark can be made at this point; the time necessary to obtain the full coverage stated is not directly related to the number of nodes. As an example, the time required to obtain full coverage for the InvoicesProcessing and InvoiceManagementApp models is substantially different, despite the underlying graphs being almost the same complexity. The fundamental reason

for this is that the later component requires particular input ranges and patterns supplied by a producer component to activate some branch points at the consumer component.

On the other hand, as we have seen in practice, obtaining strong metric coverage in a short period of time is critical for meeting today's requirements in agile development processes, such as continuous integration processes, smoke testing, evaluating local modifications before submission, and so on. To analyze the provided set of applications and our fuzzing framework in light of these requirements, we run the fuzzer on each application to determine how deep the testing process goes. Table 7 summarizes the findings.

Table 6. The time needed for each application to obtain a 100% coverage.

Model Tested—Symbolic	Coverage Time in Minutes Symbolic (%)
BankLoanRequest	30.23 min
BankLoanCheck	15.44 min
EmployeesManagement	59.4 min
PrivateHospital	73.10 min
InvoicesProcessing	29.14 min

Table 7. Coverage of the evaluated applications obtained after letting the fuzzer run for a fixed time of 60 min using the Symbolic execution strategy.

Model Tested—Symbolic	Coverage (%) Symbolic
Bank loan	100%
BankLoanRequest	100%
EmployeesManagement	69%
PrivateHospital	93%
InvoicesProcessing	100%

6.3. Complexity Comparison

Understanding the complexity of each symbolic execution strategy is important for choosing the right one for your goals. Complexity analysis helps estimate how long each algorithm might take and how much memory it might use. In Table 8, we compare the main algorithms in terms of their use of SMT solvers, execution time, and memory requirements. This comparison is based on the structure of the program being analyzed and the number of possible paths it can generate.

We use these terms:

- n: number of nodes in the program;
- b: average number of branches at a decision point;
- d: maximum depth of a path;
- S: cost of one SMT solver call;
- E: cost of executing a program node.

Table 8. Complexity analysis of symbolic execution algorithms based on time and memory.

Algorithm	Time	Memory
DFS Symbolic	$O(b^d \cdot d \cdot S)$	$O(d)$
Path Executor	$O(d \cdot (E + S))$	$O(d)$
Concolic	$O(d \cdot E)$	$O(d)$
All Paths Once	$O(b^d \cdot (d + S))$	$O(b^d \cdot d)$

6.4. Testing Recommendations Derived from Experimental Results

The results presented in the previous subsections provide valuable insights into the behavior of different testing strategies when applied to RPA workflows of varying complexity. Based on the analysis of node/branch structure, coverage time, and effectiveness under time constraints, we derive practical recommendations for RPA testing in Table 9. These are intended to support testers, tool developers, and process designers in optimizing test strategies for real-world automation workflows.

Table 9. Practical recommendations for RPA workflow testing derived from the experimental analysis.

Recommendation	Supporting Observation	When to Apply
Analyze branching structure early	Models with higher branch counts (e.g., PrivateHospital, BankLoanRequest) required significantly more time to reach full coverage.	During test planning or model inspection to allocate sufficient time and resources.
Use time-bounded fuzzing when model complexity is unknown	All models achieved 100% coverage with the "All States Once" strategy within 60 min.	In CI/CD pipelines or exploratory testing with limited time budgets.
Do not assume node count predicts test effort	InvoicesProcessing had fewer nodes yet longer coverage times due to input-dependent branches.	When estimating effort or prioritizing automation among workflows.
Symbolic and concolic techniques benefit from input-constraint modelling	Lower symbolic coverage for EmployeesManagement and PrivateHospital traced to unexercised input-dependent branches.	Whenever workflows rely on user data or dynamic behavior, integrate constraint-aware fuzzing.
Combine multiple testing strategies for comprehensive coverage	Concolic and symbolic testing explored deeper branches but with variable performance.	In complex workflows requiring high assurance, combine fuzzing with static analysis or symbolic engines.
Adopt short test iterations to support agile practices	All strategies provided strong coverage within a 60 min window, aligning with agile demands.	For fast feedback loops in development, smoke testing, and continuous integration.

These recommendations reinforce the importance of tailoring test strategies to the structural characteristics of the workflow models and the requirements of the testing environment. By leveraging both structural insights and empirical performance data, RPA testing can become more targeted, scalable, and aligned with modern development practices.

7. Conclusions and Future Work

This paper presented a unified framework for automated testing of RPA workflows using symbolic and concolic execution strategies. The approach models RPA processes as control-flow graphs and leverages SMT-based path exploration to generate high-coverage test suites. A novel "All States Once" coverage strategy was introduced to ensure practical completeness in node coverage with reduced computational overhead.

Through empirical evaluation on six real-world RPA applications—including banking workflows, invoice processing, and hospital services—we demonstrated that the proposed method achieves full node and branch coverage in all tested cases, with execution times ranging from 18.33 min to 78.9 min, depending on workflow complexity. Our results confirm that symbolic execution is better suited for smaller workflows requiring full path precision, while concolic execution scales better to broader coverage in time-constrained scenarios.

We also showed that annotations play a critical role in narrowing the input space and avoiding infeasible paths, improving solver efficiency and test relevance. The ability to mock APIs and inject domain-specific constraints made the framework adaptable to industrial use cases.

Building on the advancements presented in this study and informed by gaps identified in the current literature, several directions for future research emerge. The framework can be extended to accommodate more complex and dynamic RPA workflows. Our results raise questions around how to scale the execution strategies to very large workflows, and how to integrate learning-based techniques to improve path prediction and constraint solving in complex business environments. Machine learning can also be incorporated for automated test case generation, while prioritization holds significant potential to increase testing efficiency and further reduce manual efforts. One option is the automated selection of symbolic and concolic strategies, based on workflow characteristics such as depth, branching factor, and input space complexity. Another promising direction is to extend the framework to support data-driven RPA workflows, where control decisions depend on unstructured or external data sources. In addition, RPA-specific fault models could enhance the precision of test generation, especially in exception handling paths. These open research questions highlight important challenges that warrant further investigation to advance the state of automated testing in RPA.

We would like to express our gratitude to UiPath for their collaboration and support throughout the research and development of the proposed methods. This research was partially supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 334906.

Author Contributions: Conceptualization, C.P., M.C. and A.-N.S.; methodology, C.P., M.C. and A.-N.S.; software, C.P., M.C. and A.-N.S.; validation, C.P., M.C. and A.-N.S.; formal analysis, C.P., M.C. and A.-N.S.; investigation, C.P., M.C. and A.-N.S.; resources, C.P., M.C. and A.-N.S.; data curation, C.P., M.C. and A.-N.S.; writing—original draft preparation, C.P., M.C. and A.-N.S.; writing—review and editing, C.P., M.C. and A.-N.S.; visualization, C.P., M.C. and A.-N.S.; supervision, C.P., M.C. and A.-N.S.; project administration, C.P., M.C. and A.-N.S.; funding acquisition, C.P., M.C. and A.-N.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the project "Romanian Hub for Artificial Intelligence—HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021–2027, MySMIS no. 334906.

Data Availability Statement: The data and source code presented in this study are openly available at: https://github.com/unibuc-cs/rpa-testing (accessed on 6 June 2025).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Gartner. *Magic Quadrant for Robotic Process Automation Software*; Market Research Report, No. G00379618; Gartner Research: Stamford, CT, USA, 2019.
2. Gartner. *Magic Quadrant for Robotic Process Automation*; No. G00800237; Gartner Research: Stamford, CT, USA, 2024.

3. Cernat, M.; Staicu, A.; Stefanescu, A. Towards Automated Testing of RPA Implementations. In Proceedings of the 11th International Workshop on Automating Test Case Design, Selection, and Evaluation (A-TEST), Virtual, 8–9 November 2020; ACM: New York, NY, USA, 2020; pp. 21–24.
4. Chacón Montero, J.; Jimenez Ramirez, A.; Gonzalez Enríquez, J. Towards a Method for Automated Testing in Robotic Process Automation Projects. In Proceedings of the AST'19 Workshop, Hamburg, Germany, 19–20 February 2019; pp. 42–47.
5. Holmberg, M.; Dobslaw, F. An Industrial Case Study on GUI Testing with RPA. In Proceedings of the 2022 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), Valencia, Spain, 4–13 April 2022; pp. 199–206.
6. Yatskiv, S.; Voytyuk, I.; Yatskiv, N.; Kushnir, O.; Trufanova, Y.; Panasyuk, V. Improved Method of Software Automation Testing Based on the Robotic Process Automation Technology. In Proceedings of the 9th International Conference on Advanced Computer Information Technologies (ACIT), Ceske Budejovice, Czech Republic, 5–7 June 2019; pp. 293–296.
7. Cernat, M.; Staicu, A.; Stefanescu, A. Improving UI Test Automation Using Robotic Process Automation. In Proceedings of the 15th International Conference on Software Technologies (ICSOFT), Paris, France, 7–9 July 2020; ScitePress: Setúbal, Portugal, 2020; pp. 260–267.
8. Zhang, Z.; Chen, P.; Xie, T. WhiteFox: White-Box Compiler Fuzzing Empowered by Large Language Models. In Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis, Seattle, WA, USA, 17–21 July 2023; pp. 1–12.
9. Alshmrany, K.M.; Menezes, R.S.; Gadelha, M.R.; Cordeiro, L.C. FuSeBMC: A White-Box Fuzzer for Finding Security Vulnerabilities in C Programs. *arXiv* **2020**, arXiv:2012.11223.
10. Jaffar, J.; Maghareh, R.; Godboley, S.; Ha, X.L. TracerX: Dynamic Symbolic Execution with Interpolation. *arXiv* **2020**, arXiv:2012.00556.
11. Bardin, S.; Delahaye, M.; Kosmatov, N.; Laleau, R.; Le Gall, P. A Survey of New Trends in Symbolic Execution for Software Testing and Analysis. *Int. J. Softw. Tools Technol. Transf.* **2012**, *15*, 1–15.
12. Chen, B.; Havlicek, C.; Yang, Z.; Cong, K.; Kannavara, R.; Xie, F. CRETE: A Versatile Binary-Level Concolic Testing Framework. In Proceedings of the 21st International Conference on Fundamental Approaches to Software Engineering (FASE), Thessaloniki, Greece, 14–18 April 2018; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2018; Volume 10802, pp. 281–298.
13. Paduraru, C.; Staicu, A.N.; Stefanescu, A. Robotic Process Automation for the Gaming Industry. In Proceedings of the 18th International Conference on Software Technologies (ICSOFT), Rome, Italy, 10–12 July 2023.
14. Khankhoje, R. Robotic Process Automation (RPA) Towards Automation Testing. *Int. J. Adv. Res. Comput. Sci.* **2023**, *15*, 1930–1936. [CrossRef]
15. Yu, S.; Fang, C.; Tuo, Z.; Zhang, Q.; Chen, C.; Chen, Z.; Su, Z. Vision-Based Mobile App GUI Testing: A Survey. *arXiv* **2023**, arXiv:2310.13518.
16. Wickramasinghe, M.; Perera, I. Automating UI Issue Detection Using Convolutional Neural Networks and Machine Learning Models. In Proceedings of the 9th International Conference on Multidisciplinary Research, Colombo, Sri Lanka, 7–9 November 2024; pp. 1–6.
17. Wang, W.; Yang, W.; Xu, T.; Xie, T. VET: Identifying and Avoiding UI Exploration Tarpits. *arXiv* **2021**, arXiv:2102.06377.
18. Ravi, A. Exploring RPA (Robotic Process Automation) as a Means to Test and Develop User Interfaces. In *Purdue University Technical Report*; Submission for CNIT 58100-047—Fall; Purdue University: West Lafayette, IN, USA, 2021.
19. Baldoni, R.; Coppa, E.; D'Elia, D.C.; Demetrescu, C.; Finocchi, I. A Survey of Symbolic Execution Techniques. *ACM Comput. Surv.* **2018**, *51*, 50. [CrossRef]
20. De Moura, L.; Bjørner, N. Z3: An Efficient SMT Solver. In Proceedings of the 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS), Budapest, Hungary, 29 March–6 April 2008; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2008; Volume 4963, pp. 337–340.
21. Godefroid, P.; Levin, M.Y.; Molnar, D.A. SAGE: Whitebox Fuzzing for Security Testing. *Commun. ACM* **2012**, *55*, 40–44. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Roll/Tip-Over Risk Analysis of Agricultural Self-Propelled Machines Using Airborne LiDAR Data: GIS-Based Approach

Daniele Puri [1], Leonardo Vita [1,*], Davide Gattamelata [1] and Valerio Tulliani [2]

[1] National Institute for Insurance Against Accidents at Work (INAIL), Via Fontana Candida 1, Monte Porzio Catone, 00078 Rome, Italy; d.puri@inail.it (D.P.); d.gattamelata@inail.it (D.G.)
[2] Independent Researcher, 00100 Rome, Italy; vtulliani@gmail.com
* Correspondence: l.vita@inail.it

Abstract: Occupational Health and Safety (OHS) in agriculture is a critical concern worldwide, with self-propelled machinery accidents, particularly tip/roll-overs, being a leading cause of injuries and fatalities. In such a context, while great attention has been paid to machinery safety improvement, a major challenge is the lack of studies addressing the analysis of the work environment to provide farmers with precise information on field slope steepness. This information, merged with an awareness of machinery performance, such as tilt angles, can facilitate farmers in making decisions about machinery operations in hilly and mountainous areas. To address this gap, the Italian Compensation Authority (INAIL) launched a research programme to integrate georeferenced slope data with the tilt angle specifications of common self-propelled machinery, following EN ISO 16231-2:2015 standards. This study presents the first results of this research project, which was focused on vineyards in the alpine region of the Autonomous Province of Trento, where terrestrial LiDAR technology was used to analyze slope steepness. The findings aim to provide practical guidelines for safer machinery operation, benefiting farmers, risk assessors, and manufacturers. By enhancing awareness of tip/roll-over risks and promoting informed decision-making, this research aims to contribute to improving OHS in agriculture, particularly in challenging terrains.

Keywords: roll-over/tip-over risk; self-propelled machines; static stability; digital terrain model (DTM); geographic information system (GIS); farmer behaviour; risk assessment; machinery safety

1. Introduction

Occupational Health and Safety (OHS) in agriculture represents a significant global concern, despite increasingly stringent regulations in this sector. Notably, the majority of accidents are linked to the use of self-propelled machinery, such as tractors and harvesters. The improper operation of this equipment is responsible for numerous serious and fatal accidents in many countries [1–4]. This is especially relevant regarding the risk of roll-over incidents [5–7].

In Italy, although recent years have seen a decline in accidents, primarily due to reduced operations during the COVID-19 pandemic, the number of injuries and fatalities remains high compared to other sectors [8,9]. According to data collected by the Italian Compensation Authority (INAIL), in the five-year period of 2018–2022, deaths averaged 150 per year, with a peak of 171 in 2019 and a minimum of 137 in 2022, where the leading cause of death was the total or partial loss of control of the means of transport or moving equipment [10]. More specifically, the latter is the cause of more than 50% of fatal accidents

that occurred in 2022 as per the concluded investigations carried out by INAIL. Similar statistical incidences were observed in other countries [11–14], and in the European Union, agriculture represents one of the most impacted economic activities in terms of the number of fatal work accidents [15]. The causes of this phenomenon can be attributed to different roots [16,17]. First of all, the nature of agricultural activities, which are characterized by seasonality, outdoor work, overtime, etc., can be considered a risk factor that distinguishes them from other working contexts [18]. Further, the features of agricultural companies, which, in Italy, are mainly made of small-sized or family enterprises and part-time workers [19,20], should be considered. This aspect can be related to a lower level of safety education and training of farmers, leading to a low-risk perception and the tendency to neglect safety regulations and procedures [21,22]. When it comes to machinery safety, numerous studies have brought to light that all of the above factors contribute to the unsafe use of work equipment and, in particular, to that of self-propelled machinery, such as tractors and harvesters [23,24].

Another aspect that should be considered is related to the morphology of the territory: in Italy, only 23% of the territory is flat, while about 42% of the territory is made up of hills, and the remaining 35% is mountainous [25]. These features have a negative influence on agricultural activities as they make it difficult to use self-propelled equipment for fieldwork and the transport of products, leading to a higher probability of machinery roll-over events [26]. As noted by Facchinetti et al. [16], there is a strict correlation between the cases of tractor roll-over accidents and the territory where they happened: actually, the higher frequency of this type of accident occurred in mountainous and hilly areas, i.e., work environments characterized by slopes. Indeed, the presence of slopes causing an inadequate dynamic stability condition is considered the most significant factor contributing to overturning [27,28]. The majority of interventions aimed at dealing with this problem are focused on passive protection, i.e., the use of roll-over-protective structures (ROPSs), which, depending on the case, can be fixed or foldable/telescopic structures [29–31]. The use of ROPSs (as standalone structures or integrated into the machinery cabin) together with seatbelts can certainly help mitigate the consequences of accidents [32,33], but does not contribute to preventing roll-overs. To reduce the probability of occurrence, recent research has proposed proactive technical solutions, i.e., control systems that alert the driver when he/she is in a dangerous situation leading to the loss of stability or control loss of the vehicle [34]. For example, Casazza et al. [35] proposed a driver-assistance device that can signal critical stability conditions. Similarly, other researchers have developed alert systems in case of a loss of stability [36,37].

A third proposal to deal with this problem is related to farmers' information and training initiatives. This preventive approach was investigated, mainly focusing on two different issues. On the one hand, campaigns were promoted to spread the proper use of ROPSs (and seatbelts) and their use in old machinery (i.e., retrofitting initiatives) [38,39]. On the other hand, the proper use of machinery and the comprehension of safety instructions were also investigated to reduce farmers' risk-taking behaviour and attitudes [40]. As observed by O' Connor et al. [20], the lack of safety information is particularly significant for small-sized companies, which pose a regulatory challenge because assessing their occupational safety and health (OSH) standards on a regular basis demands significant resources. Moreover, it must be noted that although environmental factors affecting tractor safety (e.g., slopes, poor ground conditions, wet surfaces, and obstacles) increase the risk of roll-overs, most farmers tend to underestimate the risks posed by slopes, often believing their tractors can handle them [41]. This suggests a lack of awareness or a wrong perception that the risks are minimal, based on poor information and training [42]. To improve this situation, Qi et al. [43] underlined that it is crucial to develop and enhance

national and community information systems to inform farmers about natural hazards, with the goal of anticipating potential risks in farming activities. More specifically, safety information should be designed for specific areas to enhance awareness of potential risks in agriculture [44]. However, a few studies addressed this issue. Vigoroso et al. [45] examined the understanding of signals indicating hazardous slopes. Their study revealed that while farmers adopt protective behaviours when properly informed, they often struggle to process the information provided in machinery user manuals and road signs in practice.

In particular, to the authors' knowledge, there is a scarcity of research merging information on the slope steepness of fields with the characteristics of self-propelled machinery, i.e., the maximum slope angle at which the machinery can be used safely. Actually, different from the roads where signals can indicate the steepness, farmers have no precise information on the field slopes (e.g., a vineyard) where they have to carry out field works. Therefore, although they should be aware of the maximum operating angle of their machinery as specified by the manufacturer, they lack practical reference information when working in mountainous or hilly terrain. As stressed by Görücü et al. [46], farmers must accurately perceive tilt angles to prevent overturns, but research shows they struggle due to the complexity of factors involved and the lack of information provided. Recognizing potential instability is crucial for corrective action, yet sensing roll-over risk is challenging, especially with attached implements [47].

Following these research hints, the Italian Compensation Authority (INAIL) promoted a research programme aimed at providing specific information and thus training material on the characteristics of the slope steepness of agricultural fields in mountainous and hilly areas in Italy. These data were merged with information related to tilt angles provided by manufacturers of the most common self-propelled machinery, which is defined in accordance with the EN ISO 16231-2:2015 standard [48]. The final goal of this analysis is to provide practical guidelines for the safe use of self-propelled machinery for each specific area of a certain territory and made at the disposal of farmers operating in that area. This allows farmers to recognize the risks associated with the field's slope steepness, enabling them to make informed decisions about which equipment can be used safely and which should be avoided. It also helps them determine the appropriate maneuvers to perform and those that should be avoided to prevent roll-overs.

The current study presents the first phase of this project, which regarded the alpine region of the Autonomous Province of Trento (PAT) in the northern part of Italy, where vineyard cultivations were investigated and data provided by terrestrial LiDAR technology [49] allowed for 3D georeferencing and slope steepness definition for cultivated areas. Accordingly, this study addresses the absence of technical references to slope steepness for field works in mountainous and hilly areas, offering guidance not only to users and entrepreneurs conducting risk assessments of their equipment, but also to machinery manufacturers, which can provide more specific equipment. Occupational health and safety (OHS) considerations must be approached from both perspectives, focusing on developing effective strategies and tools to minimize accidents and injuries related to hazardous machinery. This aligns with Kogler et al. [6], who emphasized the importance of providing informational resources, such as guidelines, to enhance farmers' safety. Consequently, the research presented in this paper can contribute to expanding knowledge on the safe selection and use of self-propelled machinery in mountainous and hilly areas. The remainder of the paper is organized as follows: in the next sections, the research approach is described. Section 3 proposes the outcomes of the case study, while in Section 4, the achieved results are discussed. Section 5 concludes this paper by addressing further work.

2. Materials and Methods

The study aimed at combining data from the LiDAR system owned by the Autonomous Province of Trento (PAT) [49] and those related to tilt angles for self-propelled machinery provided by the EN ISO 16231-2 standard [48] in order to classify and categorize cultivated areas based on roll-over risk. Moreover, details of the scheme followed are illustrated in Figure 1.

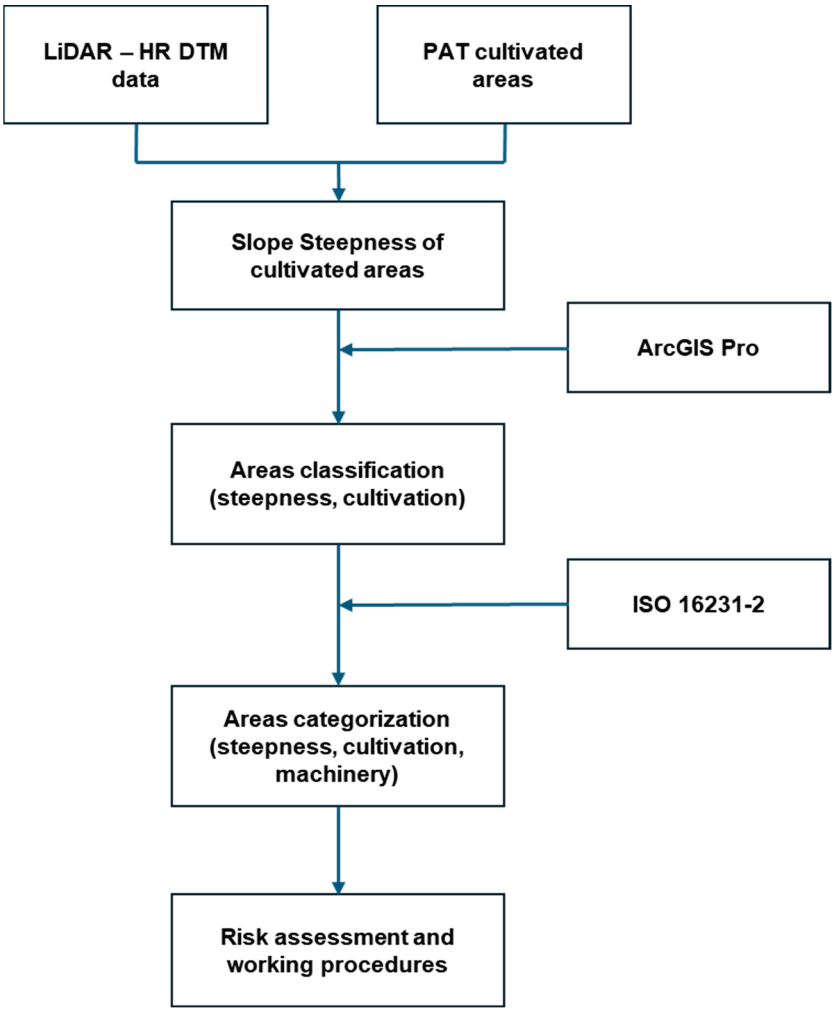


Figure 1. Scheme of the proposed approach.

The sample area is located in the northern part of Italy: as illustrated in Figure 2, the Autonomous Province of Trento (PAT) is an alpine region covering an area of 6210 km^2. The region's morphology features a diverse landscape, with approximately 70% of the total area situated above 1000 m above sea level (a.s.l.) and an average elevation of 1400 m a.s.l. The land use in the region reflects a typical mountain landscape, with forests—predominantly coniferous stands—covering approximately 70% of the total area. Rocky outcrops and bare ground account for 11.5% of the territory, while urbanized and agricultural areas are primarily concentrated in the valley floors, occupying around 19% of the province. Table 1 summarizes main agricultural activities in terms of: (1) utilized agricultural area (UAA), which takes into account the sum of the farm areas intended for agricultural production; (2) the number of enterprises involved in agricultural activities (including forestry), which are mainly small-sized and family-run companies [50].

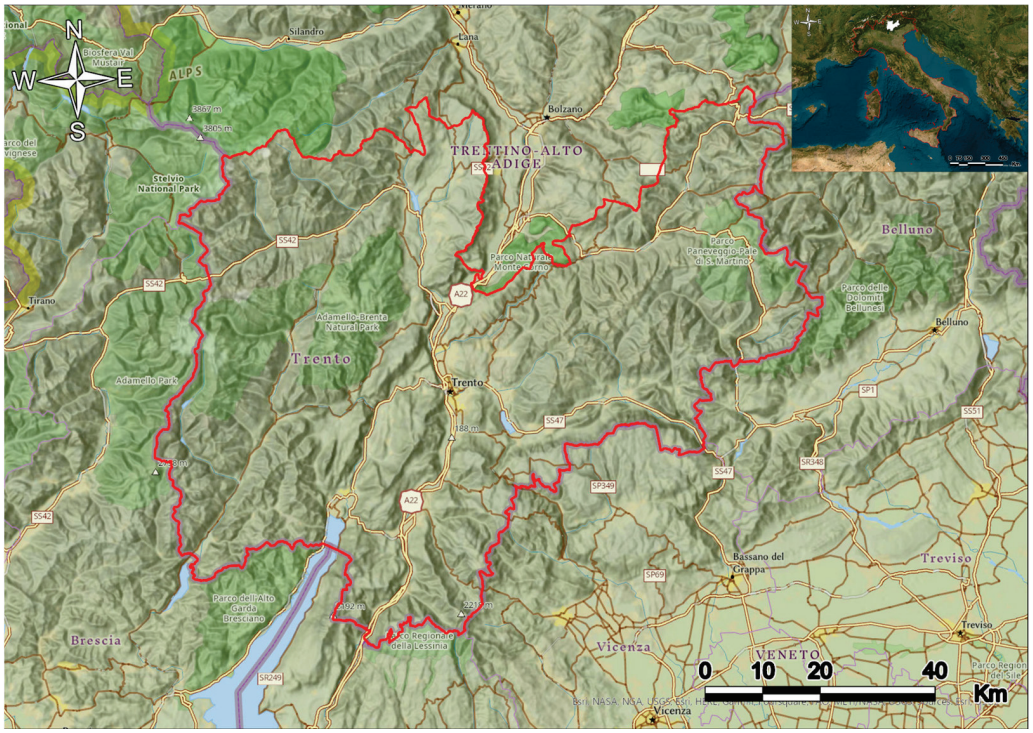


Figure 2. Study area—the Autonomous Province of Trento.

Table 1. Most relevant agricultural activities in the Autonomous Province of Trento (PAT).

Type of Activity	Utilized Agricultural Area (%)	Number of Companies (%)
Viticulture	5.8	36.3
Apple	8.1	28.0
Pasture	56.2	8.7

Due to the relevance of viticulture, this study focused on the characterization of vineyards.

2.1. Slope Steepness of Cultivated Areas

LiDAR data show great potential for agricultural applications, as demonstrated in the literature [51–53].

The Autonomous Province of Trento is covered by an HR-DTM (High-Resolution Digital Terrain Model) derived from airborne LiDAR (Light or Laser Detection and Ranging) data, which is freely available for download [49]. LIDAR is an active remote sensing technique used for topographic surveys by equipping an aerial vehicle with a laser scanner that captures data at extremely high speeds and resolutions. This system generates digital terrain models (DTMs) and digital surface models (DSMs) with a geometric resolution of 1×1 m. Specifically, the mapping process of the entire PAT was performed over different flights that occurred from 2013 to 2015, and subsequently in 2018, to remedy the lack of ground points in the most vegetated areas, and 46 flights were carried out. The 0.5 m $\times$ 0.5 m mesh grating was derived by the interpolation of the original data detected using the sensor (first pulse only), and is available in ASC and XYZ formats. The processed survey points have an altimetric accuracy corresponding to $+/-25$ cm with a confidence level of 95% (2 sigma) and a planimetric accuracy of $+/-50$ cm. The HR-DTM is composed of two distinct datasets, consisting of 25,201 and 400 blocks, respectively, each with a side length of 0.5 km and a pixel size of 0.5 m. The LIDAR data were merged with those

related to the land use provided by the reports of PAT to obtain information related to the morphological characteristics of the cultivated areas.

2.2. Areas Classification

Through the use of ESRI ArcGIS Pro 3.3.0 software [54], the next step consisted in the classification of the different areas depending on two parameters: the type of cultivation and the slope steepness. This allowed us to generate a slope dataset.

2.3. Areas Categorization

Based on the information previously collected, each area was categorized based on the type of machinery that can be used. Specifically, in this phase, a new classification of the slope dataset was performed with respect to the level of roll-over risk for each type of self-propelled machinery using the Reference Stability Static Angle (RSSA) Machine Operating Slope (MOS) values provided by the EN ISO 16231-2 standard. In Table 2, the values of MOS and RSSA for the most common self-propelled agricultural machinery are shown.

Table 2. MOS and RSSA for the most common self-propelled agricultural machinery (adapted from [EN ISO 16231-2]).

Machine Type	Lateral Roll/Tip-Over		Front/Rear Tip-Over	
	MOS (%)	RSSA (%)	MOS (%)	RSSA (%)
Combine harvester without slope compensation system	12	18	18	27
Combine harvester with slope compensation system	20	30	20	30
Combine harvester with body levelling system	30	45	30	45
Forage harvester	25	37.5	25	37.5
Field crop sprayer	15	22.5	25	37.5
Root crop harvester	10	15	15	22.5
Grape harvester without body levelling system	20	30	30	45
Grape harvester with body levelling system	30	45	30	45
Combine harvester without slope compensation system	12	18	18	27
Combine harvester with slope compensation system	20	30	20	30
Combine harvester with body levelling system	30	45	30	45
Forage harvester	25	37.5	25	37.5

It must be noted that RSSA is computed by multiplying the MOS values per safety factor, which allows engineers to consider a safety margin. In accordance with the EN ISO 16231-2 standard, this factor is equal to 1.5.

These values (expressed in percentages) are provided for both lateral and front roll-over/tip-over by the EN ISO 16231-2 standard. To be more precise, EN ISO 16231-2 defines MOS as the "value indicating, for each type of self-propelled machine and each direction, the maximum slope on which the machine is intended to work according to good agricultural practice", while the RSSA "provides the calculated slope on which the machine is required to be stable" [48].

2.4. Risk Assessment Data

The final step of the process consists of the definition of risk assessment data, which can be distinguished into two main outputs:

1. Zonal histograms were developed on the vector parcels to associate the amount of the area potentially exposed to roll-over/tip-over risk when using specific agricultural machines according to the cultivation type of each area.

2. Safety procedures were developed for the proper use of machinery in each area, including information concerning allowed maneuvers, allowed implements, etc., In other words, combining data provided by the EN ISO 16231-2 standard and the steepness of the field, it is possible to identify for each type of machinery the slope areas where the machine can work according to good agricultural practice in any direction, those where only certain maneuvers can be carried out safely, and those where the machine cannot be used.

3. Results

In this section, the experimental results are summarized as follows: in the first part, information related to the categorization procedure is provided, while in the second part of the section, the outcomes of a specific case study concerning an area dedicated to viticulture are reported.

3.1. Categorization

In the first phase of the analysis, data related to the morphology of the territory were acquired. In Figure 3, mountainous and hilly areas of the region were identified.

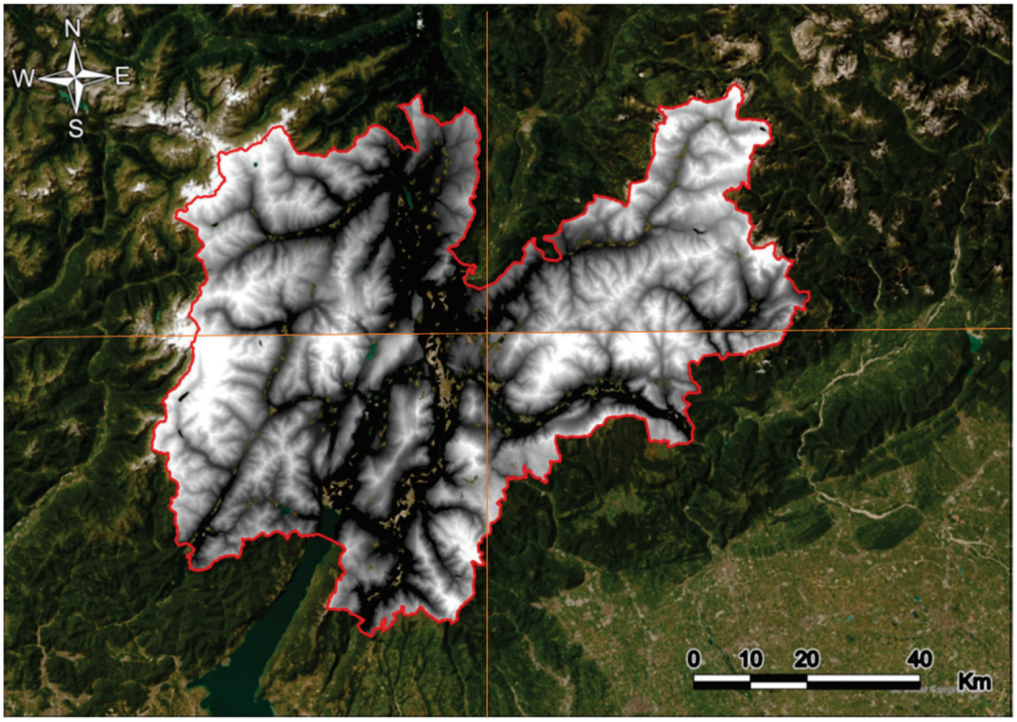


Figure 3. Autonomous Province of Trento digital elevation map to identify mountainous and hilly areas (dark grey) and flat ground (white/light grey).

Then, the slope classification was carried out: as shown in Figure 4, the lighter areas are those that have a lower slope (the white areas are those that have a slope less than or equal to 3%), while for greater slopes, the colour progressively darkens.

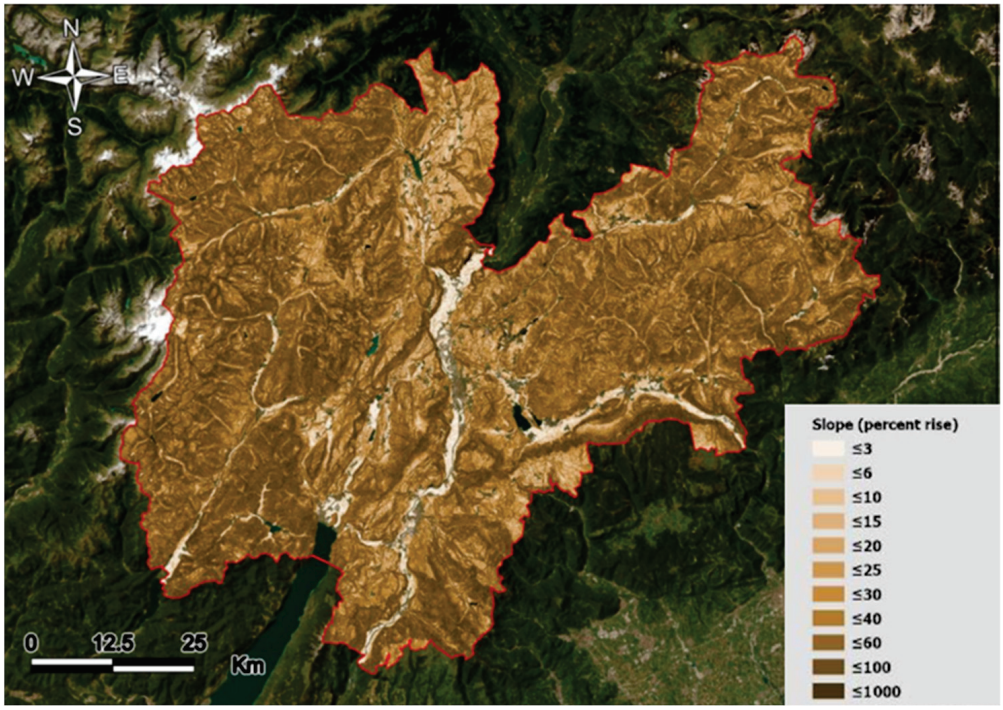


Figure 4. Definitive study area and relative slope (the slope steepness is expressed as a percentage).

Then, using the MOS values provided by EN ISO 16231-2, the different areas of the territory were categorized into three different categories:

- C1 (where there is no roll-over risk due to the slope steepness);
- C2 (where there is the risk of lateral roll-over);
- C3 (where there is a risk of both lateral and front/rear roll-over).

These data were merged with the different types of self-propelled machines reported in the EN ISO 16231-2: in Figure 5, a synthesis of this elaboration is reported considering the total surfaces expressed in hectares (e.g., for Type 1 machinery (Combine harvester without slope compensation system), the majority of the territory (525.735 hectares) belong to category C3, i.e., the most dangerous one).

CODE	MACHINE TYPE	AREA (ha) - ToT 597568					
		C1 -Area Ha Under Control	Area %	C2 - Area Ha - risk of lateral rollover	Area %	C3 - risk of both lateral and front/rear roll-over	Area %
Type 1	Combine harvester without slope compensation system	36093	6.04	35739	6.0	525735	88.0
Type 2	Combine harvester with slope compensation system	83354	13.9	n/a	n/a	514214	86.1
Type 3	Combine harvester with body levelling system	138328	23.1	n/a	n/a	459240	76.9
Type 4	Forage harvester (maize and grass)	111577	18.7	n/a	n/a	485991	81.3
Type 5	Field crop sprayer	61476	10.3	50101	8.4	485991	81.3
Type 6	Root crop harvester (beet, potato, carrot, onions)	36084	6.0	25393	4.2	536091	89.7
Type 7	Grape harvester without body levelling system	83354	13.9	54974	9.2	459240	76.9
Type 8	Grape harvester with body levelling system	138328	23.1	n/a	n/a	459240	76.9
Type 9	Cotton harvester	67042	11.2	n/a	n/a	530526	88.8
Type 10	Sugar cane harvester	36084	6.0	102244	17.1	459240	76.9
Type 11	Mower conditioner (same as for forage harvester)	111577	18.7	n/a	n/a	485991	81.3

Figure 5. Categorization of the territory in areas potentially exposed to different types of roll-over/tip-over risk (n/a = not applicable).

To put this information into practice, risk maps were derived as shown in Figures 6 and 7 which represent the false colour diagrams for the combine harvester without a slope compensation system (type 1 machine) and for the combine harvester with a body levelling system (type 2 machine), respectively.

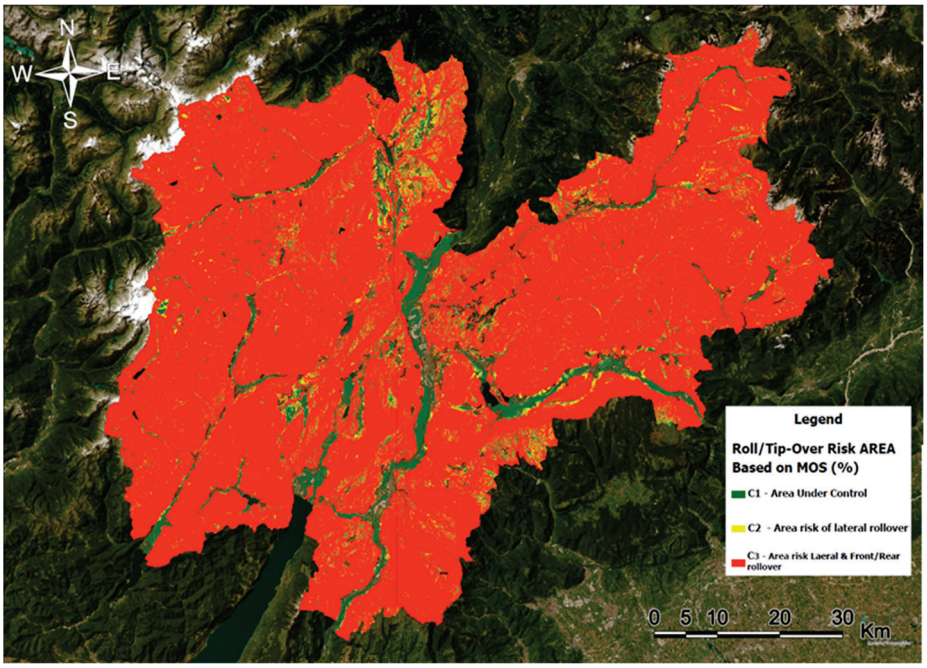


Figure 6. Map of roll-over risk level for combine harvester without slope compensation system (type 1 machine): green (category C1 area), yellow (category C2 area), and red (category C3 area).

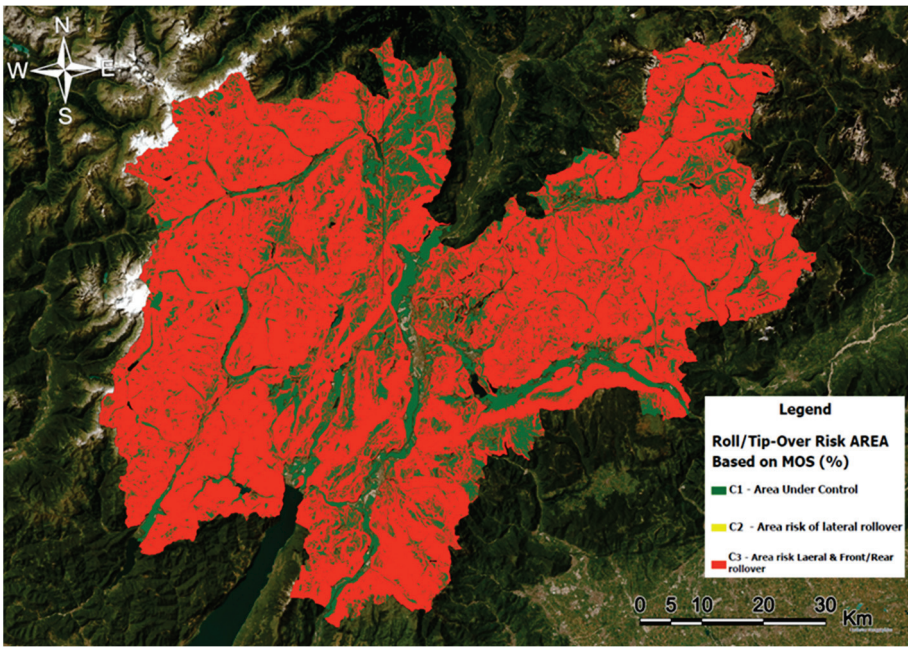


Figure 7. Map of roll-over risk level for combine harvester with body levelling system (type 3 machine): green (category C.1 area), yellow (category C.2 area), and red (category C.3 area).

Comparing the maps shown for type 1 and type 2 machines, it emerges that the roll-over risk is much higher for type 1: the red and yellow surfaces are significantly more widespread for type 1 machines than for type 2, while in the latter case the green areas are more widespread.

Thus, the final step of the categorization consisted in adding information on the type of cultivation by developing risk maps that consider three variables for each type of machinery: MOS, slope steepness, and cultivation type. In other words, using ArcGIS software, the information related to cultivation type and steepness is merged to create specific roll/tip-over risk maps.

3.2. Case Study

To delve deeper into this aspect of the study, a specific analysis was carried out that took into consideration the grape harvester and slope of vineyards, focusing on a restricted area delimited by the Municipality of Trento: in Figure 8, the positioning of this area (identified by the white boundary) into the PAT region (identified by the red boundary) is shown.

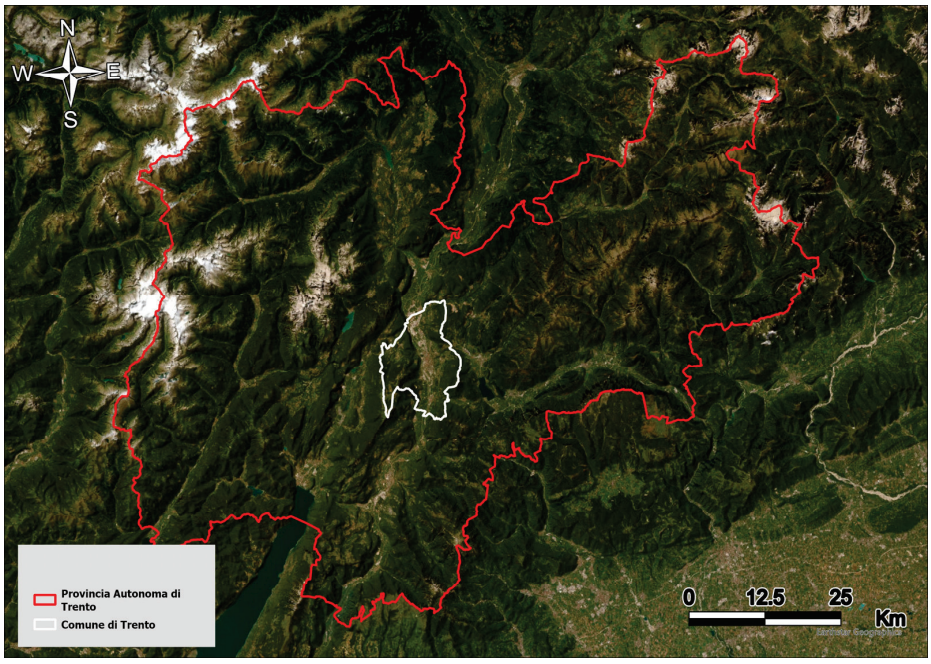


Figure 8. The municipality of Trento.

In this area, grape cultivation is very diffused, covering 1262.4916 hectares, i.e., almost 50% of the whole cultivated area of the municipality of Trento. In Figure 9, the map of the vineyards' location is represented: purple-coloured areas represent the vineyards, while the white border indicates the limits of the municipality.

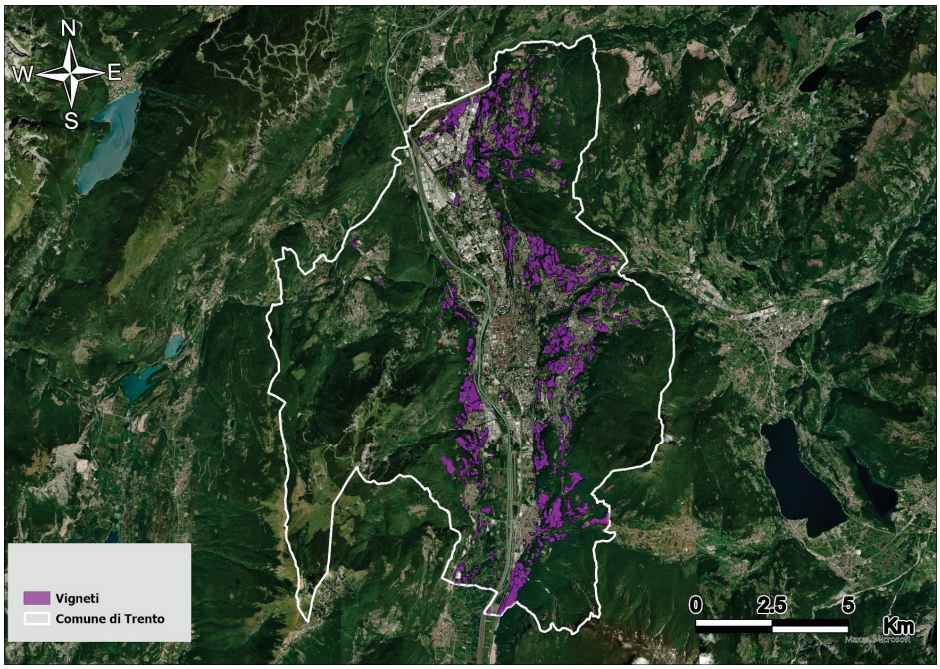


Figure 9. Vineyards in the municipality of Trento.

For this case study, the following machinery was chosen: (1) a grape harvester without a body levelling system (NO_BLS); and (2) a grape harvester with a body levelling system (W_BLS). The above-mentioned machines are most likely used in areas intended for grape farming. The total area dedicated to the cultivation of grapes was then analyzed to verify the slope of each field: these data were merged with the information related to MOS and RAAS of the machinery to determine the roll/tip-over risk. In Figure 10, data related to the risk level areas are reported for both kinds of machinery (NO_BLS and W_BLS).

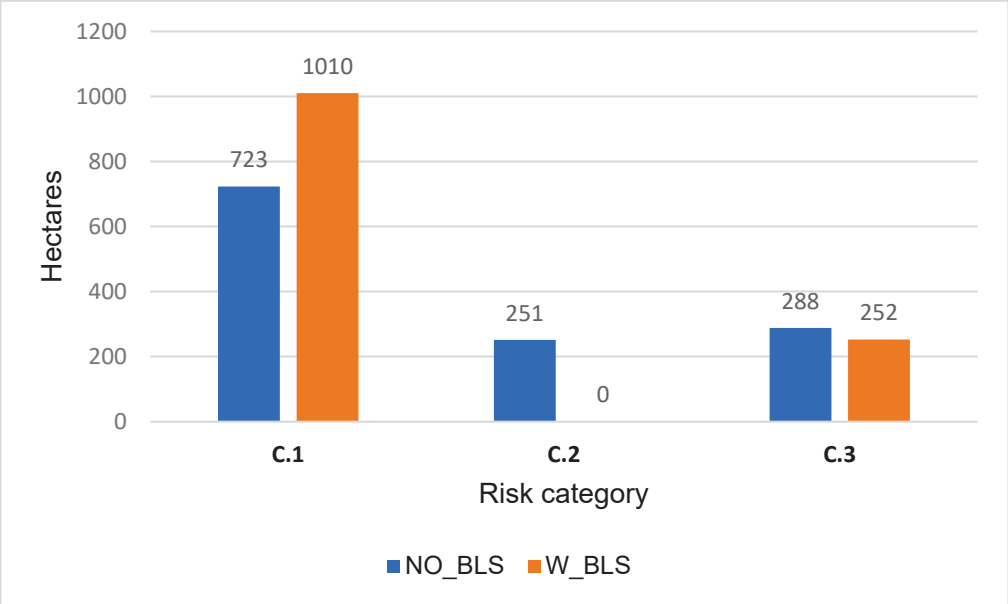


Figure 10. Area in hectares of the municipality of Trento where the use of grape harvesters with body levelling systems (orange) and without (blue) may expose farmers to roll-over risk according to risk category (C.1, C.2, and C.3).

Regarding grape harvesters equipped with an automatic levelling system (BLS), one can see that 1010 hectares represents the overall area where there is no roll-over risk due to the slope steepness. This also means that the harvester equipped with a body levelling system is not stable enough for safely working on 15% of the area, which is classified as C.3.

In more detail, the whole territory was categorized by taking into account the single plots of the municipality, as shown in Figure 11 (risk categories when using the harvester without a body levelling system (NO_BLS)) and Figure 12 (risk categories when using the harvester without a body levelling system (W_BLS)).

Looking at these figures, one can note how in mountainous and hilly areas, the slope rapidly changes, and when using machinery not equipped with the levelling system, the risk level changes in a few metres. (Figure 11). Furthermore, working in the same parcel with different versions of the same type of machinery implies a significant variation in the roll/tip-over risk.

Thus, the proper choice of work equipment is paramount, as well as working procedures: for instance, in yellow areas of Figure 11, which indicate risk category C.2, horizontal maneuvers (such as turning around) should be avoided. Alternatively, in the red zones, which indicate risk category C.3, harvesting should only be carried out manually.

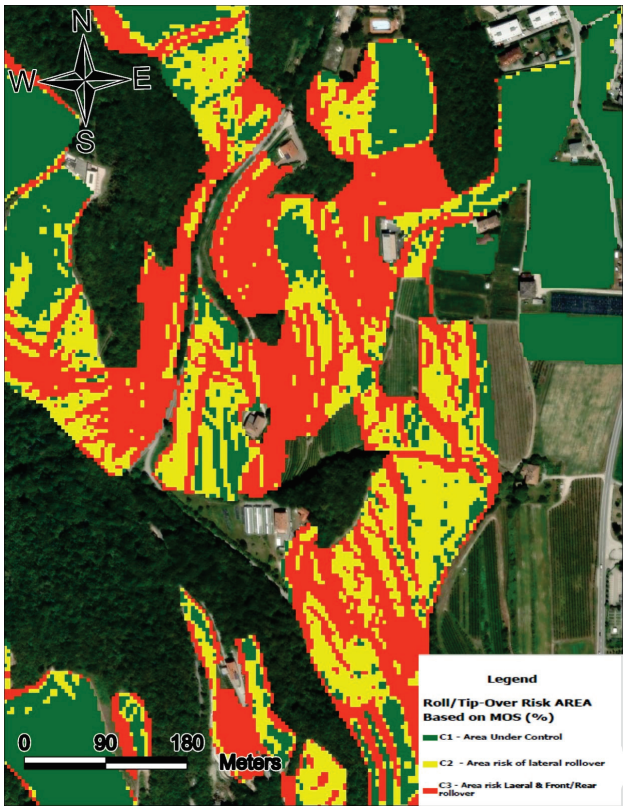


Figure 11. Details of categorization when using a grape harvester without a body levelling system (NO_BLS).



Figure 12. Details of categorization when using a grape harvester with a body levelling system (W_BLS).

4. Discussion

Recent advancements in machinery safety norms and standards have underscored an increasing awareness of occupational health and safety (OHS) from the perspectives of both equipment manufacturers and end users. These developments reflect a growing recognition of the hazardous nature of using agricultural self-propelled machinery when workers are frequently exposed to risks such as machinery roll/tip-overs [55]. Thus, significant challenges remain in the practical implementation of safety standards and working procedures [56]. This is particularly true concerning research integrating field slope steepness data with the operational characteristics of self-propelled agricultural machinery. Thus, a general contribution of this study relies on dealing with this gap in knowledge by presenting a methodology to provide farmers with specific information on the maximum safe working slope angles related to the parcels they farm depending on the machinery they use. Although this study presents the first results of a more thorough project, some practical findings can be outlined.

First, to effectively assess the roll-over risk of self-propelled agricultural machines, the proposed approach proves to be a valuable tool to enable the prediction of slope gradients in specific areas concerning plantation layouts and correlate them with the static stability angle of the agricultural machinery used on that terrain. By doing so, it becomes possible to generate a slope map that is tailored to both the specific topography and the types of machines operating in that environment.

In the literature, several studies have proposed advanced solutions integrating real-time detection mechanisms to enhance situational awareness and facilitate prompt communication of emergencies. For example, Bertacchini et al. [57] focused on advancements in driver assistance technologies: their research highlighted solutions that alert users in hazardous situations, such as instances where the vehicle experiences a loss of stability. Similarly, Liu and Koc [58] developed a detection and emergency notification system. These systems play a crucial role in accident prevention by providing timely notifications that enable drivers to take corrective actions before a critical situation escalates.

However, one of the key advantages of the proposed approach is its ability to provide a preliminary evaluation of whether a particular machine can safely operate on a given slope before reaching the area. This proactive assessment minimizes the risk of accidents by preventing machines from entering hazardous inclinations where stability could be compromised, thus avoiding risky situations.

Consequently, this tool offers a significant added value in enhancing safety measures, reducing the likelihood of roll/tip-overs, and preventing potential injuries to operators.

By integrating this predictive capability into agricultural planning and machine operation, farmers and agricultural professionals can make more informed decisions, ultimately improving both safety and efficiency in field operations. This output is in line with Cutini et al. [59], who pointed out that when using a self-propelled fruit harvester (SPFH) equipped with a limiting device on the pivoting axle, the vehicle maintains roll-over stability, even if one wheel becomes detached from the ground, thus fostering the proper choice of work equipment.

Moreover, when dealing with implements, substantial shifts in the centre of gravity occur, greatly increasing the risk of tipping, as stressed by several researchers [60–62] who observed that this risk is particularly pronounced on sloped terrain, where stability and traction may be further compromised. These findings highlight the importance of operational adjustments, especially in scenarios involving steep slopes and fluid-carrying equipment.

The importance of correct machine maneuvering is also a relevant issue that has emerged from the analysis, as in mountainous and hilly areas, the slope rapidly changes. In fields classified as C.2 (yellow areas), the risk of lateral overturns is high, and farmers must

follow paths parallel to the slope lines (red dotted line in Figure 13a) only. At the same time, this information can also be useful in the case of planning row alignment in novel plant cultivation (Figure 13b). Accordingly, proper training and information are needed to avoid risk-taking behaviours. This aligns with research findings of several studies underlining the fact that farmers are prone to incorrect maneuvering due to a mismatch between perceived and actual risk [63–67]. Moreover, it confirms the research findings by Cogato et al. [68], who emphasized that GIS mapping offers the ability to determine the potential mechanizability of a block in advance, thereby supporting operative and management decisions.

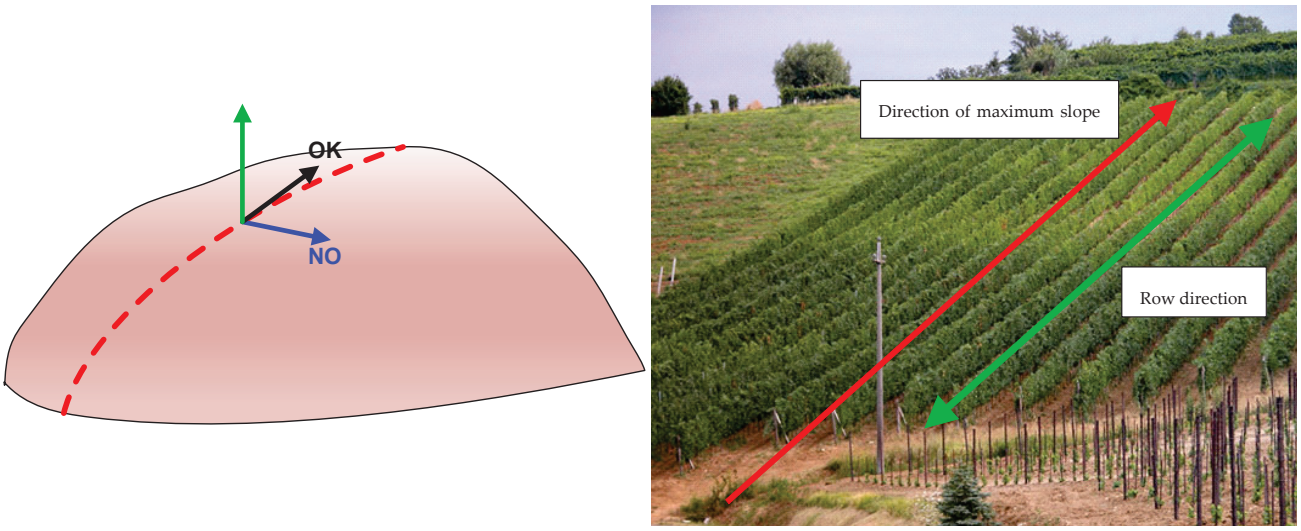


Figure 13. Slope lines for safe work activities in C.2 areas: (**a**) correct (OK) and incorrect (NO) working paths; (**b**) safe directions for planting new cultivations.

This information can be made available to farmers both through traditional means, such as farmers' associations and cooperatives, as well as by means of specific software tools that can be used individually before starting work activities. This type of information can be used to create practical, territory-specific guidelines to enhance safety awareness and accident prevention. These guidelines could help farmers determine which machinery can be used safely in specific areas, assess risks associated with slope steepness, and adopt appropriate operational solutions to minimize roll-over/tip-over risks. This output can reduce the lack of safety instructions when using self-propelled machinery, outlined by several studies [69–71]. This is consistent with Vigoroso et al. [45], who outlined the fact that when certain risks cannot be entirely eliminated, such as the risk of tractor roll/tip-overs due to steep slopes, additional safety measures must be implemented to alert operators, ensuring they are aware of high-risk areas. These warnings, combined with proper training and operational guidelines, help operators develop safer behaviours, such as adjusting their driving techniques, avoiding dangerous inclines, and using specialized equipment.

Furthermore, the development of specific applications to be used by farmers individually is in line with previous studies, where the availability of an app for preventive risk assessment before starting work can improve farmers' awareness of the correct use of machinery by means of more user-centred tools [72–75]. This issue represents a future advancement in current research, aligning with INAIL's institutional objectives.

Further improvements are also needed to better integrate the EN ISO 16231-2 data with those related to both the field area to be cultivated and the machinery features, since several parameters of the latter can influence its operative behaviour, such as the presence

of liquids, the condition of tyres, and the presence of ROPSs, altering the roll/tip-over risk level [76–78]. Overall, it must be noted that the proposed approach concerns the use of existing machinery, which, in most countries, such as those in the European Union, is designed in accordance with the slope limits indicated by the EN ISO 16231-2 standard. Nevertheless, the slope mapping obtained in this study could also be used as a reference to improve machinery design with the aim of augmenting its usability in mountainous areas. In addition, the proposed approach may also be used for assessing the stability of tractors combined with specific interchangeable equipment (e.g., flail mowers, sprayers, etc.). However, in this case, it is preliminarily necessary to define the reference stability angles for each combination of tractors with specific implements, considering that EN ISO 16231-2 does not apply.

5. Conclusions

Working in hilly and mountainous areas always presents a roll/tip-over risk for farmers when using self-propelled machinery. An accurate mapping of the slope steepness can allow them to be more aware of the level of risk, and consequently of the correct behaviour to follow. Today's information technology tools allow us to map cultivated areas by means of 3D georeferencing tools and classify them into different categories depending on slope steepness. The current study presents the first part of a national research project promoted by the Italian Compensation Authority (INAIL), and the findings showed that the proposed approach enables the prediction of terrain slopes within specific cultivated areas in relation to plantation layouts, and correlates them with the static stability angle of the machinery operating in those regions. In this way, it becomes possible to generate detailed slope maps tailored to particular areas and machine types, allowing operators to assess in advance whether a given machine can safely operate on a specific slope before physically reaching the location. Such preventive analysis cannot completely solve these problems, but can certainly serve to significantly enhance safety measures, reducing the likelihood of roll/tip-over accidents and minimizing the risk of injuries. However, this positive outcome is related to the case study context only, and further validation is necessary. In particular, additional effort is necessary to expand the database of mapped areas covered by high-resolution digital terrain models (HR-DTMs). Hence, further research is needed to improve the coverage degree of the terrain model, as well as to integrate an automated risk analysis procedure driven by artificial intelligence into user-centred tools. These advancements would streamline the mapping process, reducing the time required for preventive risk analysis and making the tool more practical for widespread adoption in agricultural safety management.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation; resources; data curation; writing—original draft preparation; writing—review and editing; visualization, supervision; project administration D.P., L.V., D.G. and V.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Additional data can be provided upon request to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ichihara, G.; Matsukawa, T.; Kitamura, F.; Yokoyama, K. Risk factors for occupational accidents in agricultural enterprises in Japan. *Ind. Health* **2019**, *57*, 627–636. [CrossRef]
2. Baraza, X.; Cugueró-Escofet, N. Severity of occupational agricultural accidents in Spain, 2013–2018. *Saf. Sci.* **2021**, *143*, 105422. [CrossRef]

3. Fargnoli, M.; Lombardi, M.; Haber, N.; Puri, D. The Impact of Human Error in the Use of Agricultural Tractors: A Case Study Research in Vineyard Cultivation in Italy. *Agriculture* **2018**, *8*, 82. [CrossRef]
4. Mishra, D.; Satapathy, S. An Integrated SWARA, QFD, and ISM Approach for Agricultural Injuries in India. *Int. J. Decis. Support Syst. Technol.* **2020**, *12*, 1–24. [CrossRef]
5. Ivascu, L.; Sarfraz, M.; Mohsin, M.; Naseem, S.; Ozturk, I. The Causes of Occupational Accidents and Injuries in Romanian Firms: An Application of the Johansen Cointegration and Granger Causality Test. *Int. J. Environ. Res. Public Health* **2021**, *18*, 7634. [CrossRef] [PubMed]
6. Kogler, R.; Quendler, E.; Boxberger, J. Analysis of occupational accidents with agricultural machinery in the period 2008–2010 in Austria. *Saf. Sci.* **2015**, *72*, 319–328. [CrossRef]
7. Biocca, M.; Aiello, L.; Baldoin, C. Rovitis 4.0: An autonomous robot for spraying in vineyards. In Proceedings of the Safety, Health and Welfare in Agriculture and Agro-Food Systems, SHWA 2020, Ragusa, Italy, 16–19 September 2020; Springer: Cham, Switzerland, 2020; pp. 176–185. [CrossRef]
8. Gattamelata, D.; Puri, D.; Vita, L.; Fargnoli, M. A Full Assistance System (FAS) for the Safe Use of the Tractor's Foldable Rollover Protective Structure (FROPS). *AgriEngineering* **2023**, *5*, 218–235. [CrossRef]
9. Fargnoli, M.; Lombardi, M. NOSACQ-50 for Safety Climate Assessment in Agricultural Activities: A Case Study in Central Italy. *Int. J. Environ. Res. Public Health* **2020**, *17*, 9177. [CrossRef]
10. INAIL, Occupational Accidents Statistics. Available online: https://www.inail.it/portale/it/inail-comunica/pubblicazioni/dati-inail/dati-inail-dettaglio.2024.03.dati-inail-anno-2024.html (accessed on 15 January 2025).
11. Kim, K.; Kim, H.; Lee, M.; Jung, W.; Choi, D. Incidence Rates of Agricultural Machine-Related Injuries in South Korea. *Int. J. Environ. Res. Public Health* **2022**, *19*, 15588. [CrossRef]
12. Mayrhofer, H.; Quendler, E.; Boxberger, J. Occupational incidents with self-propelled machinery in Austrian agriculture. *J. Agromed.* **2013**, *18*, 359–367. [CrossRef]
13. Antunes, S.M.; Cordeiro, C.; Teixeira, H.M. Analysis of fatal accidents with tractors in the Centre of Portugal: Ten years analysis. *Forensic Sci. Int.* **2018**, *287*, 74–80. [CrossRef] [PubMed]
14. Rautiainen, R.H.; Ledolter, J.; Donham, K.J.; Ohsfeldt, R.L.; Zwerling, C. Risk factors for serious injury in Finnish agriculture. *Am. J. Ind. Med.* **2009**, *52*, 419–428. [CrossRef]
15. Vigoroso, L.; Caffaro, F.; Tronci, M.; Fargnoli, M. Ergonomics and design for safety: A scoping review and bibliometric analysis in the industrial engineering literature. *Saf. Sci.* **2025**, *185*, 106799. [CrossRef]
16. Facchinetti, D.; Santoro, S.; Galli, L.E.; Pessina, D. Agricultural Tractor Roll-Over Related Fatalities in Italy: Results from a 12 Years Analysis. *Sustainability* **2021**, *13*, 4536. [CrossRef]
17. Chowdhury, M.; Ali, M.; Habineza, E.; Reza, M.N.; Kabir, M.S.N.; Lim, S.-J.; Choi, I.-S.; Chung, S.-O. Analysis of Rollover Characteristics of a 12 kW Automatic Onion Transplanter to Reduce Stability Hazards. *Agriculture* **2023**, *13*, 652. [CrossRef]
18. Vigoroso, L.; Caffaro, F.; Micheletti Cremasco, M.; Cavallo, E. Innovating Occupational Safety Training: A Scoping Review on Digital Games and Possible Applications in Agriculture. *Int. J. Environ. Res. Public Health* **2021**, *18*, 1868. [CrossRef]
19. Fargnoli, M.; Parrella, E.; Costantino, F.; Tronci, M. Hybrid solutions for agricultural vehicles: A comparative life cycle analysis from the users' standpoint. *J. Clean. Prod.* **2024**, *485*, 144406. [CrossRef]
20. O'Connor, T.; Kinsella, J.; O'Hora, D.; McNamara, J.; Meredith, D. Safer tomorrow: Irish dairy farmers' self-perception of their farm safety practices. *J. Saf. Res.* **2022**, *82*, 450–458. [CrossRef] [PubMed]
21. Fargnoli, M.; Vita, L.; Gattamelata, D.; Laurendi, V.; Tronci, M. A reverse engineering approach to enhance machinery design for safety. In *DS 70, Proceedings of the DESIGN 2012, the 12th International Design Conference*; Dubrovnik, Croatia, 21–24 May 2012, Marjanovic, D., Storga, M., Pavkovic, N., Bojcetic, N., Eds.; International Design Conference: Dubrovnik, Croatia, 2012; pp. 627–636, ISBN 978-953-7738-17-4.
22. Caffaro, F.; Micheletti Cremasco, M.; Roccato, M.; Cavallo, E. It does not occur by chance: A mediation model of the influence of workers' characteristics, work environment factors, and near misses on agricultural machinery-related accidents. *Int. J. Occup. Environ. Health* **2017**, *23*, 52–59. [CrossRef]
23. Franceschetti, B.; Rondelli, V.; Capacci, E. Lateral Stability Performance of Articulated Narrow-Track Tractors. *Agronomy* **2021**, *11*, 2512. [CrossRef]
24. Capacci, E.; Franceschetti, B.; Guzzomi, A.; Rondelli, V. Energy Absorption in Actual Tractor Rollovers with Different Tire Configurations. *Int. J. Environ. Res. Public Health* **2021**, *18*, 6517. [CrossRef] [PubMed]
25. Fredi, P.; Lupia Palmieri, E. Morphological Regions of Italy. In *Landscapes and Landforms of Italy. World Geomorphological Landscapes*; Soldati, M., Marchetti, M., Eds.; Springer: Cham, Switzerland, 2017. [CrossRef]
26. Cecchini, M.; Zambon, I.; Monarca, D.; Piccioni, F.; Marucci, A.; Colantoni, A. Spatial Analysis for Detecting Recent Work Accidents in Agriculture in Italy. In *Innovative Biosystems Engineering for Sustainable Agriculture, Forestry and Food Production*; Coppola, A., Di Renzo, G., Altieri, G., D'Antonio, P., Eds.; MID-TERM AIIA 2019; Lecture Notes in Civil Engineering; Springer: Cham, Switzerland, 2020; Volume 67. [CrossRef]

27. Vita, L.; Gattamelata, D.; Pessina, D. Retrofitting Agricultural Self-Propelled Machines with Roll-Over and Tip-Over Protective Structures. *Safety* **2021**, *7*, 46. [CrossRef]
28. Kogler, R.; Quendler, E.; Boxberger, J. Occupational accidents with agricultural machinery in Austria. *J. Agromed.* **2016**, *21*, 61–70. [CrossRef] [PubMed]
29. Gattamelata, D.; Vita, L.; Fargnoli, M. Machinery Safety and Ergonomics: A Case Study Research to Augment Agricultural Tracklaying Tractors' Safety and Usability. *Int. J. Environ. Res. Public Health* **2021**, *18*, 8643. [CrossRef]
30. Wang, L.; Liu, F.; Song, Z.; Ni, Y.; He, Z.; Zhai, Z.; Zhu, Z.; Zhou, Z.; Song, Z.; Li, Z. Advances in tractor rollover and stability control: Implications for off-road driving safety. *Comp. Electron. Agric.* **2024**, *226*, 109483. [CrossRef]
31. Kim, W.S.; Siddique, M.A.A.; Kim, Y.-J.; Jung, Y.-J.; Baek, S.-M.; Baek, S.-Y.; Kim, Y.-S.; Lim, R.-G. Simulation of the Rollover Angle of a Self-Propelled Radish Harvester for Different Load Conditions. *Appl. Sci.* **2022**, *12*, 10733. [CrossRef]
32. Fargnoli, M.; Lombardi, M. Safety Vision of Agricultural Tractors: An Engineering Perspective Based on Recent Studies (2009–2019). *Safety* **2020**, *6*, 1. [CrossRef]
33. Alfaro-Lopez, J.R.; Perez-Ezcurdia, A.; Latorre-Biel, J.-I.; Arana-Navarro, I.; Benito-Amurrio, M.; Villanueva-Roldán, P. The Integration of Mechanical Energy Absorbers into Rollover Protective Structures to Improve the Safety of Agricultural Tractors in the Event of Rollover. *Agriculture* **2024**, *14*, 1050. [CrossRef]
34. Gao, J.; Yin, C.; Yuan, G. Development of an Anti-rollover Warning and Active Intervention Controller for Wheeled Tractor. *Recent Pat. Eng.* **2024**, *18*, 55–66. [CrossRef]
35. Casazza, C.; Martelli, R.; Rondelli, V. Evaluation of a commercial tractor safety monitoring system using a reverse engineering procedure. *J. Agric. Saf. Health* **2016**, *22*, 215–225. [CrossRef]
36. Sun, D.; Chen, D.; Wang, S.; Wang, X. A dynamic instability detection and prediction system for high clearance tractor. *IFAC-Pap.* **2016**, *49*, 50–54. [CrossRef]
37. Chowdhury, M.; Islam, M.N.; Iqbal, M.Z.; Islam, S.; Lee, D.-H.; Kim, D.-G.; Jun, H.-J.; Chung, S.-O. Analysis of Overturning and Vibration during Field Operation of a Tractor-Mounted 4-Row Radish Collector toward Ensuring User Safety. *Machines* **2020**, *8*, 77. [CrossRef]
38. Cavallo, E.; Langle, T.; Bueno, D.; Tsukamoto, S.; Görücü, S.; Murphy, D. Rollover Protective Structure (ROPS) Retrofitting on Agricultural Tractors: Goals and Approaches in Different Countries. *J. Agromed.* **2014**, *19*, 208–209. [CrossRef]
39. Tinc, P.J.; Ayers, P.; May, J.J.; Purschwitz, M.A.; Park, S.; Bayes, B.; Sorensen, J. Implementing a National Tractor Safety Program: Using "Whole System in a Room" to Mobilize Partners and Implement Solutions. *J. Agromed.* **2016**, *21*, 127–131. [CrossRef] [PubMed]
40. Glasscock, D.J.; Rasmussen, K.; Carstensen, O.; Hansen, O.N. Psychosocial factors and safety behaviour as predictors of accidental work injuries in farming. *Work Stress* **2006**, *20*, 173–189. [CrossRef]
41. Irwin, A.; Poots, J. Investigation of UK farmer go/no-go decisions in response to tractor-based risk scenarios. *J. Agromed.* **2018**, *23*, 154–165. [CrossRef]
42. Caffaro, F.; Cremasco, M.M.; Roccato, M.; Cavallo, E. Drivers of farmers' intention to adopt technological innovations in Italy: The role of information sources, perceived usefulness, and perceived ease of use. *J. Rural Stud.* **2020**, *76*, 264–271. [CrossRef]
43. Qi, X.; Yao, X.; Cong, X.; Li, S.; Han, M.; Tao, Z.; Yang, X.; Qi, X.; Shi, F.; Wang, S. Profile and risk factors in farmer injuries: A review based on Haddon matrix and 5 E's risk reduction strategy. *Front. Public Health* **2024**, *12*, 1322884. [CrossRef]
44. Cividino, S.R.S.; Pergher, G.; Gubiani, R.; Moreschi, C.; Da Broi, U.; Vello, M.; Rinaldi, F. Definition of a Methodology for Gradual and Sustainable Safety Improvements on Farms and Its Preliminary Applications. *Agriculture* **2018**, *8*, 7. [CrossRef]
45. Vigoroso, L.; Caffaro, F.; Cavallo, E. Warning against Critical Slopes in Agriculture: Comprehension of Targeted Safety Signs in a Group of Machinery Operators in Italy. *Int. J. Environ. Res. Public Health* **2019**, *16*, 611. [CrossRef]
46. Görücü, S.; Cavallo, E.; Murphy, D. Perceptions of Tilt Angles of an Agricultural Tractor. *J. Agromed.* **2014**, *19*, 5–14. [CrossRef] [PubMed]
47. Tillapaugh, J.A.; Murphy, D.J.; Sommer, I.J.; Garvey, P.M. Operator assessment of tractor roll angle with and without a tractor stability visual feedback device. *J. Agric. Saf. Health* **2010**, *16*, 249–264. [CrossRef] [PubMed]
48. *ISO 16231-2:2015*; Self-Propelled Agricultural Machinery. Assessment of Stability—Part 2: Determination of Static Stability and Test Procedures. International Organization for Standardization: Genevan, Switzerland, 2015. Available online: https://www.iso.org/standard/61583.html (accessed on 22 January 2025).
49. Provincia Autonoma di Trento. Stem-WebGIS. Available online: https://siat.provincia.tn.it/stem/ (accessed on 22 January 2025).
50. Provincia Autonoma di Trento. PAT Statistics. Available online: http://www.statistica.provincia.tn.it/ (accessed on 22 January 2025).
51. Jin, S.; Sun, X.; Wu, F.; Su, Y.; Li, Y.; Song, S.; Guo, Q. Lidar sheds new light on plant phenomics for plant breeding and management: Recent advances and future prospects. *ISPRS J. Photogramm. Remote Sens.* **2021**, *171*, 202–223. [CrossRef]

52. Iqbal, J.; Xu, R.; Sun, S.; Li, C. Simulation of an Autonomous Mobile Robot for LiDAR-Based In-Field Phenotyping and Navigation. *Robotics* **2020**, *9*, 46. [CrossRef]
53. Aguiar, A.S.; dos Santos, F.N.; Sobreira, H.; Boaventura-Cunha, J.; Sousa, A.J. Localization and Mapping on Agriculture Based on Point-Feature Extraction and Semiplanes Segmentation From 3D LiDAR Data. *Front. Robot. AI* **2022**, *9*, 14. [CrossRef]
54. ArcGIS Pro. Available online: https://pro.arcgis.com/en/pro-app/latest/get-started/get-started.htm (accessed on 22 January 2025).
55. Cecchini, M.; Cossio, F.; Marucci, A.; Monarca, D.; Colantoni, A.; Petrelli, M.; Allegrini, E. Survey on the status of enforcement of European directives on health and safety at work in some Italian farms. *J. Food Agric. Environ.* **2013**, *11*, 595–600.
56. Zambon, I.; Piergentili, A.; Salvati, L.; Monarca, D.; Matyjas-Łysakowska, P.; Colantoni, A. Applied Research for a Safer Future: Exploring Recent Job Accidents in Agriculture, Italy (2012–2017). *Processes* **2018**, *6*, 87. [CrossRef]
57. Bertacchini, A.; Napoletano, G.; Dondi, D.; Larcher, L.; Pavan, P. Optimized energy-aware wireless system for identification of the relative positioning of articulated systems in the free space. *IEEE Sens. J.* **2014**, *14*, 1682–1692. [CrossRef]
58. Liu, B.; Koc, A.B. Field tests of a tractor rollover detection and emergency notification system. *J. Agric. Saf. Health* **2015**, *21*, 113–127. [CrossRef]
59. Cutini, M.; Brambilla, M.; Bisaglia, C.; Melzi, S.; Sabbioni, E.; Vignati, M.; Cavallo, E.; Laurendi, V. A Study of the Lateral Stability of Self-Propelled Fruit Harvesters. *Agriculture* **2017**, *7*, 92. [CrossRef]
60. Bietresato, M.; Pavan, S.; Cozzi, G.; Sartori, L. A numerical approach for evaluating and properly setting self-propelled forage harvesters. *Trans. ASABE* **2013**, *56*, 5–14. [CrossRef]
61. Jang, M.-K.; Hwang, S.-J.; Nam, J.-S. Simulation Study for Overturning and Rollover Characteristics of a Tractor with an Implement on a Hard Surface. *Agronomy* **2022**, *12*, 3093. [CrossRef]
62. Kang, S.; Kim, Y.; Park, H.; Kim, Y.; Woo, S.; Uyeh, D.D.; Ha, Y. Rollover Safety and Workable Boundary Suggestion of an Agricultural Platform with Different Attachments. *Agriculture* **2022**, *12*, 1148. [CrossRef]
63. Fargnoli, M.; Lombardi, M.; Puri, D. Applying Hierarchical Task Analysis to Depict Human Safety Errors during Pesticide Use in Vineyard Cultivation. *Agriculture* **2019**, *9*, 158. [CrossRef]
64. Cil, M.; Gedik, T. Research on factors affecting the risk-taking behavior of personnel working in the forest products sector. *Int. J. Occup. Saf. Ergon.* **2022**, *28*, 2315–2323. [CrossRef]
65. Špička, J. Socio-Demographic Drivers of the Risk-Taking Propensity of Micro Farmers: Evidence from the Czech Republic. *J. Entrepren. Emerg. Econ.* **2020**, *12*, 569–590. [CrossRef]
66. Fargnoli, M.; Lombardi, M. Safety Climate and the Impact of the COVID-19 Pandemic: An Investigation on Safety Perceptions among Farmers in Italy. *Safety* **2021**, *7*, 52. [CrossRef]
67. Cecchini, M.; Bedini, R.; Mosetti, D.; Marino, S.; Stasi, S. Safety Knowledge and Changing Behavior in Agricultural Workers: An Assessment Model Applied in Central Italy. *Saf. Health Work* **2018**, *9*, 164–171. [CrossRef]
68. Cogato, A.; Pezzuolo, A.; Sørensen, C.G.; De Bei, R.; Sozzi, M.; Marinello, F. A GIS-Based Multicriteria Index to Evaluate the Mechanisability Potential of Italian Vineyard Area. *Land* **2020**, *9*, 469. [CrossRef]
69. Fargnoli, M.; Lombardi, M.; Haber, N. A fuzzy-QFD approach for the enhancement of work equipment safety: A case study in the agriculture sector. *Int. J. Reliab. Saf.* **2018**, *12*, 306–326. [CrossRef]
70. Cutini, M.; Forte, G.; Maietta, M.; Mazzenga, M.; Mastrangelo, S.; Bisaglia, C. Safety-Critical Manuals for Agricultural Tractor Drivers: A Method to Improve Their Usability. *Agriculture* **2017**, *7*, 67. [CrossRef]
71. Nag, P.K.; Gite, L.P. Farm Accidents and Injuries. In *Human-Centered Agriculture. Design Science and Innovation*; Nag, P.K., Gite, L.P., Eds.; Springer: Singapore, 2020; pp. 175–204. [CrossRef]
72. Caffaro, F.; Bagagiolo, G.; Micheletti Cremasco, M.; Vigoroso, L.; Cavallo, E. Tailoring Safety Training Material to Migrant Farmworkers: An Ergonomic User-Centred Approach. *Int. J. Environ. Res. Public Health* **2020**, *17*, 2104. [CrossRef] [PubMed]
73. Rubino, F.M.; Puri, D.; Fargnoli, M.; Lombardi, M.; Mandić-Rajčević, S.; Colosio, C. Sicurpest: A Prototype of a User-Friendly Tool for Preventive Risk Assessment of Pesticide Use in Agriculture. *Toxics* **2025**, *13*, 89. [CrossRef] [PubMed]
74. Marucci, A.; Colantoni, A.; Zambon, I.; Egidi, G. Precision Farming in Hilly Areas: The Use of Network RTK in GNSS Technology. *Agriculture* **2017**, *7*, 60. [CrossRef]
75. Radočaj, D.; Plaščak, I.; Heffer, G.; Jurišić, M. A Low-Cost Global Navigation Satellite System Positioning Accuracy Assessment Method for Agricultural Machinery. *Appl. Sci.* **2022**, *12*, 693. [CrossRef]
76. Majdan, R.; Abrahám, R.; Tkáč, Z.; Drlička, R.; Matejková, E.; Kollárová, K.; Mareček, J. Static Lateral Stability of Tractor with Rear Wheel Ballast Weights: Comparison of ISO 16231-2 (2015) with Experimental Data Regarding Tyre Deformation. *Appl. Sci.* **2021**, *11*, 381. [CrossRef]

77. Melzi, S.; Sabbioni, E.; Vignati, M.; Cutini, M.; Brambilla, M.; Bisaglia, C.; Cavallo, E. Multibody model of fruit harvesting trucks: Comparison with experimental data and rollover analysis. *J. Agric. Eng.* **2018**, *49*, 92–99. [CrossRef]
78. Rondelli, V.; Capacci, E.; Franceschetti, B. Evaluation of the Stability Behavior of an Agricultural Unmanned Ground Vehicle. *Sustainability* **2022**, *14*, 15561. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

On the Specimen Design, Physical Properties and Geometry Effect on Heat Generation and Thermal Gradient in Ultrasonic Fatigue

Felipe Klein Fiorentin [1,*], Rita Dantas [2,3,4], Jorge Wolfs Gil [2,3], Andrea Piga Carboni [1], Thiago Antonio Fiorentin [1] and Abílio Manuel Pinho de Jesus [2,3]

[1] Department of Mobility Engineering, UFSC (Federal University of Santa Catarina), Joinville 89219-600, Brazil; andrea.piga@ufsc.br (A.P.C.)
[2] Institute of Science and Innovation in Mechanical and Industrial Engineering, Campus da FEUP, Dr. Roberto Frias Street, 4200-465 Porto, Portugal; rdantas@inegi.up.pt (R.D.)
[3] Faculty of Engineering, University of Porto, Dr. Roberto Frias Street, 4200-465 Porto, Portugal
[4] Institute for Sustainable Construct, 4200-465 Porto, Portugal
* Correspondence: felipe.fiorentin@grad.ufsc.br

Abstract: Performing fatigue characterisation is often an expensive task, being both time consuming and expensive. Taking that into account, ultrasonic fatigue testing is an interesting solution, since it can be thousands of times faster than traditional experiments. In ultrasonic fatigue testing, excitation frequencies are in the order of magnitude of 20 kHz, while common fatigue testing frequencies are typically approximately a few hundreds of Hz. Although promising, ultrasonic fatigue testing has some challenges, like high strain rates, heat generation and complex specimen design. Regarding the latter, since the working principle of ultrasonic fatigue tests relies on exciting the specimen in one of its natural frequencies, finding a specimen geometry to resonate at this given frequency might be challenging. Additionally, some materials often present challenges associated with high temperature during tests. The goal of this paper is to provide guidelines for specimen design, encompassing the effects of critical factors and their impact on important test parameters, like temperature and dimensions. The proposed methodology developed a parameter able to quantify the heat generation severity during ultrasonic fatigue testing for several materials based on their physical properties. Moreover, the effects of the geometry and material properties in the temperature during loading cycles, with special focus on thermal gradients were enumerated.

Keywords: ultrasonic fatigue; specimen design; gigacycle; temperature effect

1. Introduction

In order to understand material behaviour after being subjected to a very high number of cycles, from 10^7 to 10^{10} (or beyond), different machines and testing techniques have to be developed, leading to so-called ultrasonic fatigue testing (alternative names are used like very-high-cycle fatigue testing, VHCF, or gigacycle fatigue testing). In order to achieve such large numbers of cycles, testing frequencies have to be greatly increased—otherwise, the testing times would be prohibitive. For example, a standard fatigue machine operating at 20 Hz would take more than 1.6 years to perform 10^9 cycles in only one specimen. Machines operating using servo hydraulic actuators or electrical drives (e.g., rotating bending machines) are typically not able to reach frequencies above 1 kHz. Consequently, ultrasonic fatigue testing systems typically use another excitation device, a piezo-electric

actuator. Such systems are able to achieve frequencies above tens of kHz, with 20 kHz being the most common excitation frequency [1]. However, in order to maximise stress amplitudes, these systems excite the specimen in one of its resonance frequencies. This intrinsic characteristic of ultrasonic fatigue testing leads to some challenges regarding specimen design, due to the resonance constraint and stress analysis, since there is no external loading on the specimen, and the stresses are a consequence of the specimen's inertial forces. Due to this working principle, total fracture of the specimen typically does not occur, since the specimen's natural frequency is reduced by the crack length (which reduces the cross-section stiffness) [2].

Due to the high excitation frequencies, with some machines operating at dozens of kHz, specimens may experience significant heat generation and excessive temperature during fatigue tests. Some authors reported noticeable heat generation for certain materials, like cast iron, and temperatures during tests above 150 °C [3]. The impact of temperature in fatigue properties is well known, requiring specimen cooling during testing in such conditions.

Additionally, since the specimen must be cooled, temperature gradients are present at the testing sections of the specimens (regions where the stress amplitudes are at a maximum), which can be very abrupt for some materials. High temperatures will also change the mechanical properties of the materials, like the elasticity modulus, and density, which will change the specimen's natural frequency [4]. This is especially critical in ultrasonic fatigue testing, since the machine can only operate within a certain narrow frequency range, and the test will stop if the specimen's natural frequency is outside this range.

The heat generation is directly associated with the elastic energy available. A fraction of this energy will be dissipated and converted into heat [5]. This dissipative term is proportional to a material damping property, like the damping ratio or loss factor. For some materials, this damping property is nearly invariant within the elastic energy [6]. However, other materials present significant variations with increments in stress amplitudes, where this dissipative term can increase by one or two orders of magnitude. This is critical for ultrasonic fatigue testing, since higher stress amplitudes will lead to severe heat generation during tests. A comprehensive study was developed by Visnapuu et al. [7], where several steel and cast-iron alloys were analysed. Some materials, like the AISI 1045, presented a nearly constant damping ratio, virtually independent of stress levels. Other materials, like AISI 4140 and a grey cast iron, presented abrupt variations in damping parameters (more than 10 times) with an increase in stress levels. The authors found a strong relation between localised plasticity (more intense in higher stresses) and the increase in the damping ratio. This damping variation due to stress amplitudes was also observed during ultrasonic fatigue testing for a cooper specimen [8]. The authors have measured the damping ratio using the exponential decay during free vibration. They have observed that if a shorter period of time was used for measuring the damping ratio (higher average stress amplitudes), the measured damping ratio would be higher than when measurements were performed for longer periods (smaller average amplitudes). Variations larger than 25% in the damping ratio were observed depending on the total measurement time. Similar results were observed for a nickel superalloy and 316L stainless steel specimens [9], where higher damping ratios were measured for higher stress amplitudes during ultrasonic fatigue testing.

In order to reduce the effect of high temperatures and its impact on the fatigue results, different approaches might be applied. One of the most popular is intermittent testing, where the specimen is loaded for a certain period, experiences heat generation and a temperature increase, and a stop for a determined time in order to cool down (idle) [10]. This waiting time (machine off) may vary from a few hundred milliseconds to

several seconds, the latter for materials experiencing severe heat increasing during tests. Some authors have observed that specimens subjected to longer waiting times, and lower temperatures as a consequence, presented superior fatigue performance than specimens with shorter idling times [11]. Other interesting work observed severe temperatures during tests of the 316L austenitic steel [12], and very long waiting times (5 s) were applied in order to avoid excessive heating during tests and its impact on the fatigue results.

Alternative methods are related to specimen cooling, where heat is extracted by convection. The most widely used method is cooling the specimen testing section (the region where the temperatures are higher) with compressed air. However, for severe cases, this cooling method is not enough, and some works applied different solutions, like liquid nitrogen cooling or water, although it is important to understand and avoid the impact of the cooling system in the fatigue properties. For example, water cooling is not suitable for materials with low corrosion resistance, since the specimen is subjected to this corrosive environment for extended periods of time.

The increments in temperature typically have a negative impact on the fatigue behaviour of a material (except for cryogenic conditions) [13]. Therefore, it is essential to have information regarding the temperature during fatigue tests. The temperature distribution is also an impacting factor in fatigue [14]. Some authors have found that the temperature gradient has a strong effect on the initiation and propagation of fatigue cracks [15], and this damage mechanism is often referred to as temperature gradient mechanical fatigue (TGMF) [16]. In an interesting study conducted by Sun et al. [17], the authors compared the fatigue behaviour of a nickel superalloy for two specimen geometries, one with a very uniform temperature gradient, and the other presenting a much greater temperature gradient (approximately 20 °C/mm). The specimens were subjected to the same loading levels and maximum temperatures, and they found that the specimens with the higher temperature gradients presented a much shorter fatigue life (by an entire order of magnitude). High temperature gradients can occur during ultrasonic fatigue testing, especially for materials with lower thermal conductivities, and are exacerbated during the convection cooling (typically used during ultrasonic tests), making the surface temperature much cooler than the inner region.

Since excessive heat generation is a frequent problem during ultrasonic fatigue characterisation, mainly associated with the high excitation frequencies, the main goal of this research is to provide a guideline regarding specimen design, with special emphasis on the heat generation. In addition, it is intended to provide basic information regarding materials most susceptible to the excessive heat generation. Moreover, based on analytical formulation, a discussion regarding the effect of each material parameter effect on the resonance frequencies and specimen geometry will be drawn. While the main focus of the present work is related to metallic materials, some examples regarding polymers will be presented. Additionally, the present discussion may be expanded for non-metallic materials, like ceramics and composites.

2. Analytical Formulation and Discussion

The present section will provide a discussion regarding the material property effect on the specimen dimensions (especially constrained by the natural frequency requirement for ultrasonic fatigue), and on the energy dissipated during fatigue cycles. Additionally, a discussion encompassing the effect of specimen dimensions on the temperature field during fatigue cycling will also be presented. For this analysis, a FEM model will be used to present the effect of several parameters on the temperature distribution.

2.1. The Effect of Specimen Geometry on the Specimen's Natural Frequency

Due to its working principle, the ultrasonic fatigue specimen must be designed to have one of its natural frequency coincident to the excitation frequency of the machine. Typically, the 1st longitudinal frequency of a specimen is the one of interest, and during tests, this vibration mode will be excited. Figure 1a presents the stress and displacement field for a typical hourglass specimen (1st longitudinal vibration mode), based on [1], while its main dimensions are present in Figure 1b. Even though the present discussion will focus on the hourglass specimens, an analogous analysis could be performed for dog-bone, flat or other specimen geometries.

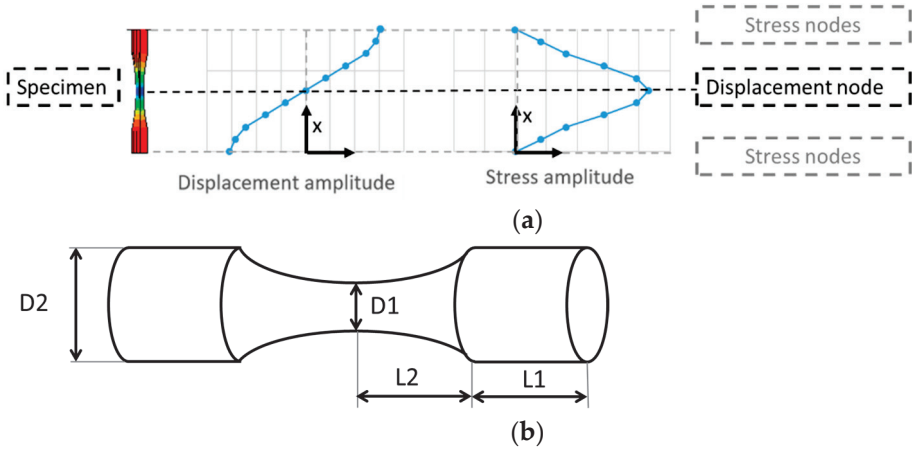


Figure 1. Hourglass ultrasonic fatigue specimen: (**a**) stress and displacement distribution along specimen length; (**b**) main dimensions.

Analytical formulations for hourglass specimens regarding their dimensions, stress, strain and displacement fields for a desired resonance frequency are available (for example, the study in [1] presents formulations for finding *L1* to reach a desired resonance frequency, f_n). The contribution of each dimension can be analysed by simply using the natural angular frequency equation for a 1 degree-of-freedom (D.O.F.) system

$$\omega_n = \sqrt{\frac{K}{M}}, \tag{1}$$

where K is the stiffness, and M is the mass. Looking at the stress and displacement fields in Figure 1a, it can be observed that the central portion of the specimen is associated with higher stress amplitudes, and lower displacement amplitudes. Therefore, this region will have a higher impact on the stiffness component, K, of natural frequency. On the other hand, the outer portions of the specimens are the regions with higher displacement amplitudes, while the stress ones are the smallest. Therefore, the outer region of the specimen will have a greater impact on the inertial terms (M).

Moreover, the stiffness of a bar (axial tension/compression loading only) can be written as

$$K = \frac{EA}{L}, \tag{2}$$

where E is the elasticity modulus, A represents the cross-section area and L is the total length. Each specimen dimension impact can be summarised as:

- *D1*: Increasing *D1* will greatly improve specimen stiffness, resulting in a higher natural frequency. Moreover, increasing *D1* will also increase the specimen cross section, reducing its stress amplitudes.
- *D2*: When *D2* is incremented, it will increase the specimen inertial term (associated with *M*), therefore decreasing its natural frequency. Moreover, a higher *D2* will increase

the stress levels of the specimen, since it will be subjected to a higher force (higher inertial component).

- *L1*: A higher *L1* will result in a less stiff specimen (the stiffness of a bar is inversely proportional to its length, as Equation (2) shows); therefore, it will reduce the specimen's natural frequency. Moreover, it will also increase the specimen mass, reducing its resonance frequency even further. This additional mass will also lead to an increase in stress levels at the central section of the specimen.
- *L2*: By being at the central section of the specimen, *L2* will have a greater impact on the specimen stiffness component. A larger *L2* will lead to a lower natural frequency, due to its inverse impact on the specimen stiffness. It is important to avoid very small values of *L2*, since this can lead to high stress concentration levels due to the abrupt changes in the cross section (from *D2* to *D1* in a short length).

It is important to emphasise that when changing one dimensions, other dimensions (at least one) will have to be adjusted, in order to maintain the desired resonance frequency.

2.2. The Effect of Material Properties on the Specimen's Natural Frequency

Only two material properties are influent in the specimen's natural (undamped) frequency [1]: the elasticity modulus, E, and the density, ρ. The stiffness of a specimen is directly proportional to its elasticity modulus. Therefore, a higher elasticity modulus will lead to stiffer specimens, which will lead to higher natural frequencies, as shown in Equation (1). On the other hand, the material density will have an opposite effect; a higher density will lead to lower natural frequencies, since it increases the inertial component (M). Moreover, higher-density materials will also increase the stress levels at the specimen testing section for a given geometry, since the inertial component is increased.

When incorporating the damping component, the damped natural frequency can be written as

$$\omega_d = \omega_n \sqrt{1 - \xi^2}\,, \tag{3}$$

where ξ is the damping ratio. Since the damping ratio for metals is typically a value much smaller than 1, and this component contribution is squared, ω_d is very close to ω_n, and for most applications they are considered equal.

Material Properties and Specimen Geometry

The present section aims to provide some discussion regarding the effect of material properties in the ultrasonic fatigue specimen dimensions. For the present comparison, a 20 kHz longitudinal frequency will be used for design. Moreover, the same testing section diameter was used for all specimens, 6 mm.

Four materials were analysed, with two of them metals, one representing steel and other representing aluminium alloys, and the remaining two polymers, polypropylene and vinylester. These materials with very discrepant properties were chosen in order to make more noticeable their impact on the specimen final geometry. Table 1 presents the material properties used for the specimen design, following the analytical approaches proposed in [1], the elasticity modulus (or the flexural modulus for polymers) and density.

Table 1. Material properties for ultrasonic fatigue specimen design.

Material	Flexural or Elasticity Modulus (GPa)	Density (kg/m^3)
Steel [9]	185.0	8000
Aluminium [18]	65.0	2700
Polypropylene [19]	1.2	910
Vinylester [20]	3.0	1120

Figure 2 presents the obtained geometries for all four materials, where y is the profile (radius), and x is the specimen coordinate (starting at centre). For simplicity, only half of the length of the specimen was drawn (symmetric). The most notable difference can be seen between metallic and polymeric materials. Due to their relatively low elastic modulus, for a certain geometry, these materials would result in components with lower stiffness (by one or two orders of magnitude), as shown in Equation (2). Therefore, the geometry must be shortened in order to increase its overall stiffness. The lower density (when comparing to metals) has an opposite effect—it will increase the specimen's natural frequency (for a given geometry), as shown in Equation (1). However, since the differences in the elastic modulus are much greater than in the density, the overall consequence will be a shorter specimen. For example, polypropylene has a density approximately 10 times lower than steel, but its elasticity modulus is more than 100 times lower than the steel.

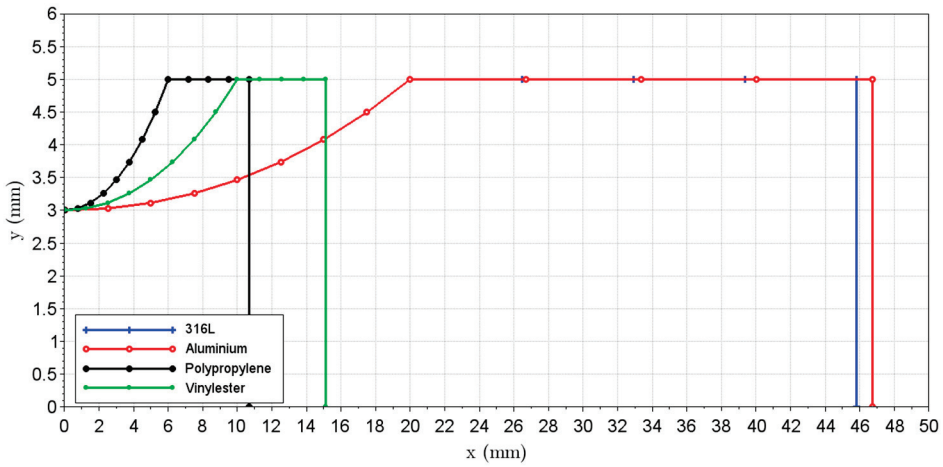


Figure 2. Specimen geometry (profile and length) for different materials.

When comparing aluminium and steel, a fascinating behaviour can be seen, as the specimens geometries are very similar to each other. This can be explained by looking at the material properties, Table 1, and the natural frequency equation, Equation (1). Stiffness (K) is proportional to the elasticity modulus, and mass (M) varies linearly with the specimen density. Therefore, the natural frequency will vary according to the square root of the elasticity modulus and density ratio. Since the elasticity modulus of steel is approximately 3 times that of aluminium, with the same pattern the density, this relation will remain similar to both materials. Consequently, for a given geometry, steel and aluminium specimens will present a very similar natural frequency.

Table 2 presents the total specimen length for all four materials (not encompassing the additional thread length, present in some ultrasonic fatigue systems). Moreover, the maximum possible stress amplitude was also computed. For that case, a maximum displacement amplitude of 50 µm was used (common value for some commercial ultrasonic machines). When comparing steel and aluminium, even though the specimen dimensions are very similar, the maximum stress amplitude for the steel specimen is nearly 3 times higher than that of aluminium. When looking at the material properties, Table 1, it can be seen that the steel density is approximately three times higher than that of aluminium. Consequently, steel will present higher inertial forces (for the same geometry), increasing the maximum stress amplitude at the testing section. For example, if higher stress amplitudes were desired for testing an aluminium alloy, changes in the specimen geometries must be performed, like reducing the testing section diameter.

Table 2. Specimen total length and maximum stress amplitudes.

Material	Specimen Total Length (mm)	Maximum Stress (MPa), @ Max. Disp. of 50 μm
Steel	91.6	350.0
Aluminium	93.4	121.6
Polypropylene	21.3	10.2
Vinylester	30.3	14.1

When looking at polymeric materials, since the obtained geometries were smaller, a lower inertial contribution will be presented. The inertial term is decreased even further by the lower density (when comparing with the analysed metals, for example). The obtained maximum stress amplitudes for the polymeric materials (for the present geometry and material properties) are an entire order of magnitude lower those of the metal. It is worth mentioning thin the present analysis only encompasses the material property effects on the specimen geometries, and changes in specimen geometries might be necessary to achieve the necessary stress levels for performing VHCF tests in these materials.

2.3. Important Parameters Affecting Heat Generation and Specimen Temperatures

Energy dissipation is an intrinsic characteristic of real-life systems. Fatigue systems are no different. Every loading cycle encompasses stored energy, a certain quantity of which is dissipated per cycle. Materials subjected to loads under the elastic regime presents an elastic potential energy (per unit of volume) described as

$$\mathrm{E}_e = \frac{\sigma\varepsilon}{2}, \tag{4}$$

where σ and ε are the stress and strain amplitudes, respectively. According to Hooke's law, at the linear elastic regime, the stress can be written as a function of the elasticity modulus and strain, as

$$\sigma = E\,\varepsilon. \tag{5}$$

Therefore, the elastic energy can be written only as a function of stress and the elasticity modulus,

$$\mathrm{E}_e = \frac{\sigma^2}{2\,E}. \tag{6}$$

The dissipated energy per cycle can be written as function of the available elastic energy, E_e, and a material dissipation property, loss factor (η), as

$$\mathrm{E}_d = 2\pi\eta\mathrm{E}_e. \tag{7}$$

For low values of damping, the loss factor can be directly associated with the damping ratio [21],

$$\eta = 2\xi. \tag{8}$$

In most mechanical systems, due to relatively low damping (small values of zeta), the dissipated energy is often ignored, since it does not lead to significant dissipative effects (like an increase in temperature or noise). However, in ultrasonic fatigue tests, these dissipative effects are often very significant. Figure 3 presents an example of the elastic energy (blue) and the dissipated energy for a given value of the damping ratio (green). The ordinate axis represents the sum of elastic (F_e) and dissipative forces (F_d), while abscissas represents the displacement during a cycle. Although the dissipated energy per cycle is very low (compared to the total elastic energy available), typical excitation frequencies exceed thousands of cycles per second. Most ultrasonic fatigue machines operate at 20 kHz; consequently, the low, dissipated energy per cycle became very relevant

due to the high number of cycles in short periods of time. Looking at Equations (4)–(8), the dissipated energy per unity of volume and per cycle can be written only as a function of a material damping property (like the loss factor, for example), and the stress level. Moreover, the dissipated energy per unit of time (power) is also known if material volume and frequency are both known during ultrasonic fatigue tests.

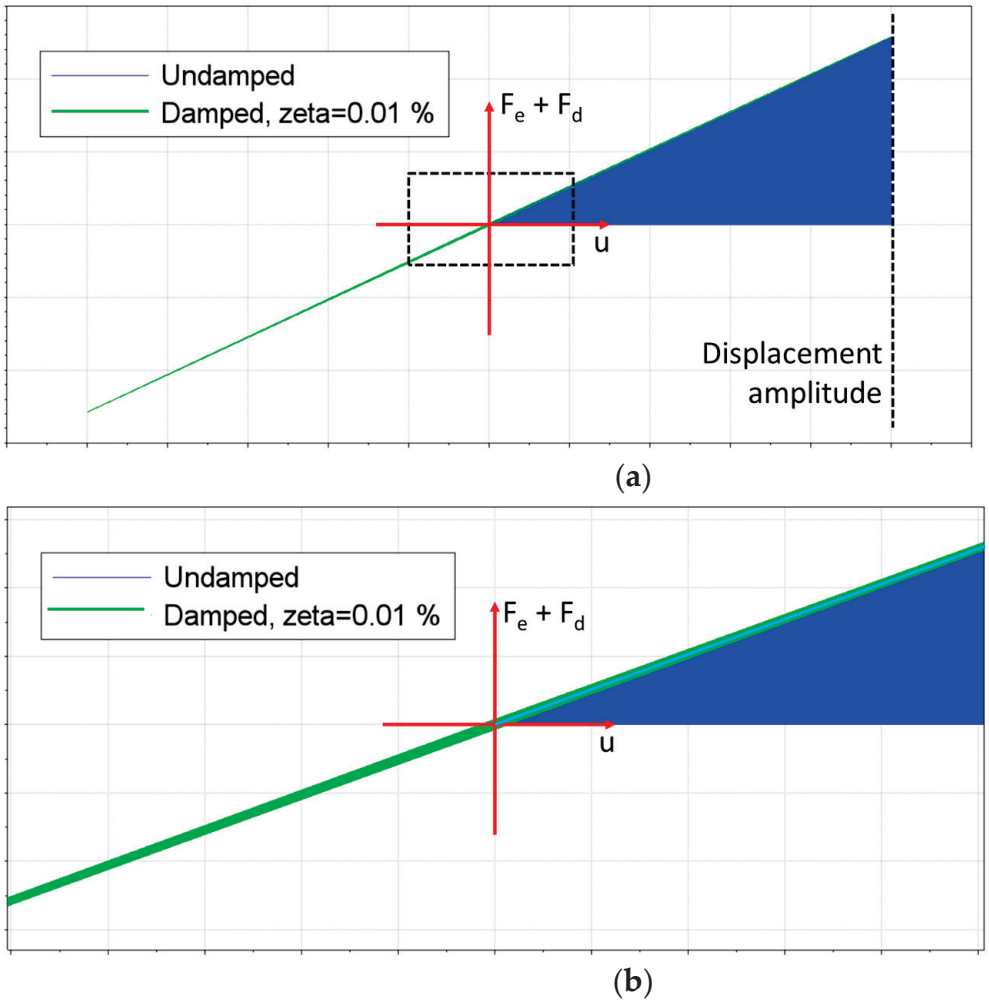


Figure 3. Hysteresis loop, energy dissipation: (**a**) Elastic and dissipated energy, overview; (**b**) Zoomed view.

Figure 4 presents a schematic view of the heat generation and transfer of an ultrasonic fatigue specimen. A symmetry condition at the central portion of the specimen was used for simplicity. The dissipated elastic energy is mainly converted into heat. Since the central portion of the specimen presents higher stress levels, this will also lead to higher energy generated at these regions. The generated energy will be transferred to other parts of the specimen by conduction, where the material thermal conductivity (k) plays a major role. Moreover, at the specimen surface, energy will be dissipated via convection, where the convection coefficient (h), surface and ambient temperature (T_i and T_{amb}, respectively) and surface area (A_l) will dictate how much heat is transferred. If high surface temperatures are present, it is also important to take into account the radiation, which will not be covered in the present analysis. Often, forced convection is used in order to cool down the specimen during ultrasonic fatigue tests. Compressed air is the most common solution regarding cooling. However, in severe cases, liquid nitrogen can be used.

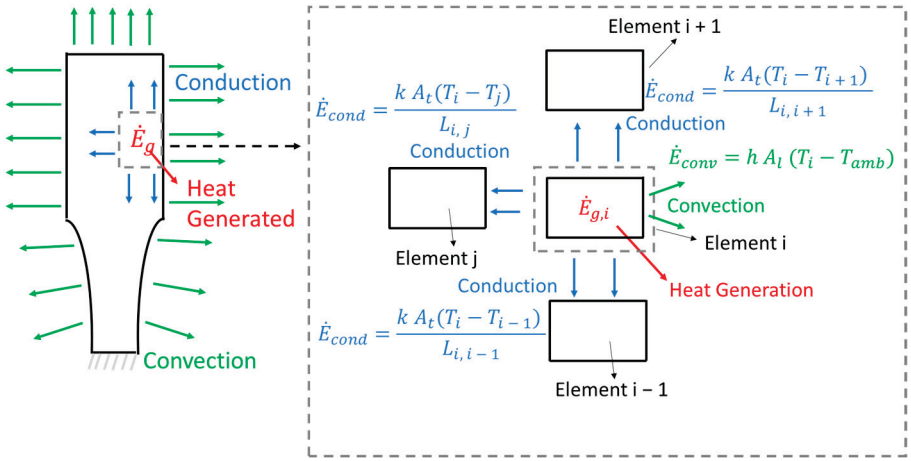


Figure 4. Heat generation and transfer.

Additional material properties are relevant for transient heat transfer analysis, like the density (ρ) and specific heat (cp). A higher specific heat will require more energy to increase the temperature of a certain portion (mass) of a material. When comparing two materials with the same specific heat, a higher density, as the property implies, will increase the amount of material mass for a given volume, which will also increase the amount of energy required to increase the temperature.

Heat Generation Severity for Different Materials

As previously discussed, the heat generation during ultrasonic fatigue tests and temperatures are a function of several material properties, and the specimen geometry too. The main goal of the present section is to provide a guideline regarding how critical a given material is regarding the heat generation during ultrasonic fatigue testing. Moreover, as previously discussed, material properties (namely the elasticity modulus and density) will affect the heat generation and temperature distribution; therefore, this effect should not be neglected. However, the present section will not encompass the geometry effect on heat generation and the temperature field.

In order to quantify this heat severity, a material parameter was developed, being a function of material properties, which will be referred here as *Hs*. For simplicity, three material properties were used, namely:

- Loss factor, η: the amount of dissipated mechanical energy (and converted into heat) is proportional to the product of the elastic energy and the loss factor. When a material presented a high loss factor, a larger amount of the available mechanical energy will be converted into heat. Alternatively, the damping ratio could be used, since there is a linear relation between it and the loss factor.
- Fatigue limit, S_f: As was observed in Equations (4)–(6), the elastic energy is proportional to the square of the stress levels. Therefore, a higher stress level will result in higher elastic energy, and due to the dissipative terms, a higher amount of energy will be converted into heat. Materials with a higher fatigue limit will have to be subjected to high stress amplitudes during tests in order to present fracture (or a crack nucleation and growth up to a macroscopic size) during ultrasonic fatigue tests.
- Thermal conductivity, k: This parameter is directly related to how thermal energy can be transferred within a given material. A high thermal conductivity is beneficial for ultrasonic fatigue tests, since the heat generated at the testing section (higher stress levels) is easily transferred from this region.

Since higher values of fatigue limit and loss factor contribute to heat generation, these material parameters were added to the upper portion (dividend) of the heat severity parameter, H_s, as

$$H_s = \frac{\eta \, S_f}{k}. \tag{9}$$

On the other hand, the thermal conductivity was positioned as a divisor, since a higher value will be beneficial for the heat transfer, reducing the temperatures during tests. Other material properties, like the specific heat, could be included in a more complete analysis. For materials presenting a strong variation in damping ratios due to stress levels, it is important to use proper values of ξ, according to the test amplitudes during ultrasonic fatigue testing.

In order to understand how some materials might present high heat generation and temperatures during tests, a few metallic alloys were analysed and compared with each other using the heat severity parameters, *Hs*. A total of eight materials were selected, and are presented in Table 3. Three steels were used, encompassing a high strength one, 51CrV4, a structural steel, S690, and an austenitic stainless steel, 316L. Additionally, a nickel superalloy, a cast iron, an aluminium, a titanium and a magnesium alloy were analysed. All material properties were obtained from an engineering database [22], except when referred otherwise.

Table 3. Material properties for heat severity analysis.

Material	Thermal Conductivity ($\mathbf{W{\cdot}m^{-1}{\cdot}^{\circ}C^{-1}}$)	Loss Factor	Fatigue Limit (MPa)
Stainless Steel, 316L	14	4.0×10^{-3}	170 [23]
High Carbon Steel, 51CrV4	45	2.4×10^{-4}	660 [24]
Low alloy Steel, S690QL	19	8.0×10^{-4}	520 [25]
Nickel Superalloy, Inconel 625	10	8.0×10^{-4} [9]	450 [26]
Cast Iron, EN-GJS-700	31	6.0×10^{-3}	290 [27]
Aluminium, Al 6061	167	5.6×10^{-3} [28]	97
Titanium, Ti4Al6V	7	4.0×10^{-4} [29]	510
Magnesium, AZ61	70	8.0×10^{-3}	160 [30]

Figure 5 presents the obtained results for heat generation severity for the eight analysed materials, obtained using Equation (9). The cast iron EN-GJS-700 was computed as the most severe regarding "testability" in ultrasonic fatigue, mainly due to its very high damping properties (such property makes cast iron strong candidates for application in vibrating systems [18], for example) and moderate fatigue limit (leading to high stress amplitudes during fatigue tests). Next, 316L appeared as the second most severe case, mainly associated with its low conductivity (common for austenitic stainless steels) and high loss factor. This observation is in good agreement with the literature [12], where some researchers related high waiting times between loading cycles (idle) in order to avoid excessive heating.

Inconel 625 also presented one of the highest heat severity parameters during tests, due to a combination of low thermal conductivity and high mechanical properties. Once more, the literature data reporting this characteristic are also available, on researchers related issues regarding the heat generation for this alloy [26], where a specific water-cooling system was developed in order to perform ultrasonic fatigue tests in this super alloy. On the other side of the spectrum, aluminium alloy Al 6061 and high-strength steel 51CrV4 presented the lowest heat severity parameter for ultrasonic fatigue testing. The first is mainly associated with the high thermal conductivity of the aluminium, and a relative low fatigue limit (leading to moderate stress amplitudes during tests). The second, although

presenting high mechanical properties, presented a very low loss factor, leading to only a small amount of the elastic energy being converted into heat.

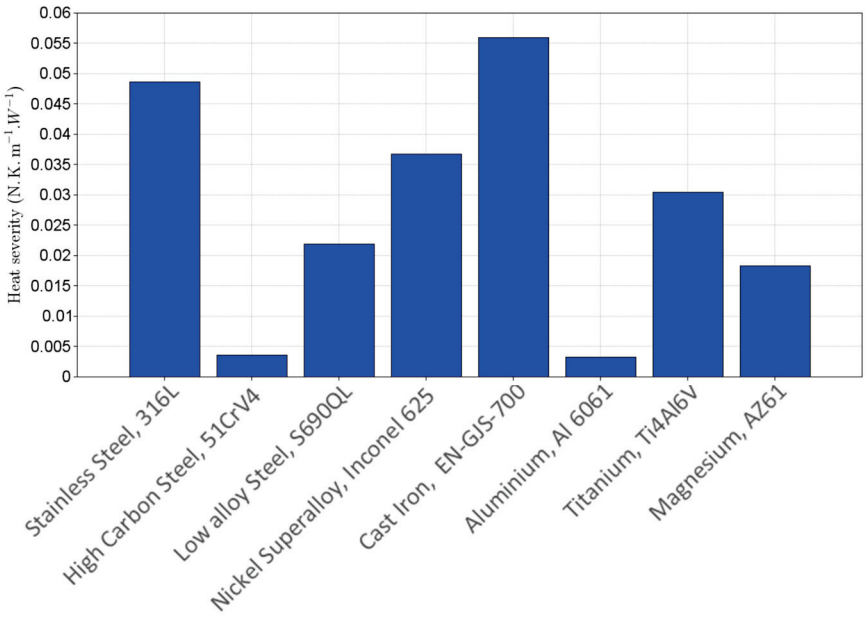
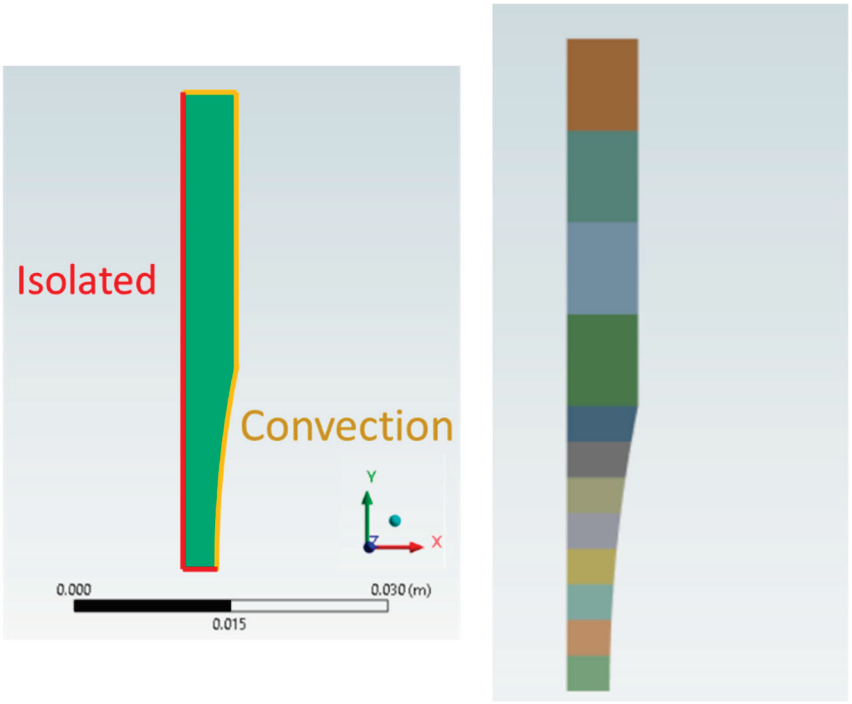


Figure 5. Heat generation severity during ultrasonic fatigue testing.

It is important to emphasise that the goal of the present list is to provide a guideline, and some properties, especially the fatigue limit and damping ratio, might change significantly for a given alloy due to differences in its microstructure due to different heat treatments (presence of martensitic or ferritic structures for example). Residual stresses will also result in differences in these properties.

3. FEM Analysis

In order to quantify and better represent how certain material properties, specimen geometry aspects and surrounding ambient conditions affect the temperature and its distribution along the specimen, FEM analysis will be used. Since the stress/strain field and loading frequency (which is equal to the natural specimen frequency in the ultrasonic fatigue test, 20 kHz in the present case) are known, the heat generated at each specimen section can be computed and will be used as an input for the heat transfer analysis. A steady-state analysis will be performed, using axisymmetric condition. Second-order quadrilateral elements (Quad8) were used, with eight nodes. Regarding the mesh, an average size of 0.1 mm was used. Figure 6 presents a schematic view of the boundary conditions, combining convection boundaries at the outer surfaces (yellow) and isolated edges due to symmetry (red). Moreover, heat generation due to the previously discussed dissipative term was used, as a function of the stress amplitudes (variable along the specimen length) and excitation frequency. A total of 12 subregions were used regarding the heat generation (each of them having an average heat generation for that area), with 8 of them located at the variable cross section (higher stress levels and higher heat generated, consequently) and 4 at the outer region (smaller stress amplitudes). Regarding the latter, 20 kHz will be used, since it is the most used frequency in ultrasonic fatigue specimens.



(**a**) (**b**)

Figure 6. Heat transfer model and boundary conditions: (**a**) boundary conditions; (**b**) heat generation regions.

Regarding material properties, convection parameters and testing conditions, Table 4 presents the baseline that will be used. A stainless steel, 316L, was used as the baseline for material properties. In order to understand the material property effect, like the thermal conductivity of a high-carbon steel, for example, or the effect on the thermal distribution, only that property will be changed. Moreover, all simulations will be performed using the same specimen geometry, except when the change in geometry is the main focus of this study. The same is valid for testing condition (excitation frequency and stress amplitude).

Table 4. Material properties, test condition and convection information.

Property	Value
Elasticity modulus (GPa)	185
Density (kg·m^{-3})	7960
Thermal conductivity (W·m^{-1}·°C^{-1})	10
Damping ratio (%)	0.2
Ambient temperature (°C)	21
Convection coefficient (W·m^{-2}·°C^{-1})	10,000
Excitation frequency (kHz)	20
Maximum stress amplitude (MPa)	350

In order to analyse the temperature variation at the central section of the specimen (where the maximum stresses and temperatures are present), a ratio between the surface temperature (T_s) and the inner section (core, T_c) of the specimen will be analysed, written as

$$R_t = \frac{T_s}{T_c} \tag{10}$$

where a R_t value closer to 1 represents a relatively homogeneous temperature distribution at the centre (region of interest, due to the maximum stress amplitudes). For the present work, the ratios were computed using temperatures in °C.

3.1. The Effect of Specimen Diameter

Two specimen geometries were used for comparison. The first have a testing section diameter of 6 mm, which will often be referred simply as "D6". This geometry will also be used for all other comparison cases. Regarding the second geometry, a testing diameter of 3 mm was used, referred to as "D3". As stated before, due to the resonance frequency condition, it is not possible to only change one dimension, since it would lead to a different natural frequency. Therefore, other dimensions were also changed. The specimens were designed to resonate at 20 kHz (1st longitudinal vibration mode) using the formulation provided by Bathias and Paris [1]. Table 5 presents the specimen dimensions used, following the notation presented in Figure 1b.

Table 5. Specimens main dimensions.

Dimensions	Geometry 1, D6	Geometry 2, D3
D1 (mm)	6.0	3.0
D2 (mm)	10.0	10.0
L1 (mm)	25.8	11.0
L2 (mm)	20.0	17.0

Regarding the analytical stress and displacement fields, Figure 7 presents their amplitudes along the specimen length (being x = 0 the central portion of the specimen). As stated before, the maximum stress amplitude was defined as 350 MPa. It is important to emphasise that, due to differences in specimen dimensions, the specimen will naturally present different displacement amplitudes for the same stress levels. This is an important parameter in ultrasonic fatigue testing, often referred to as specimen gain (frequently written in MPa/μm), providing the relation (for a given geometry and material properties) of the applied displacement and the resulting stress.

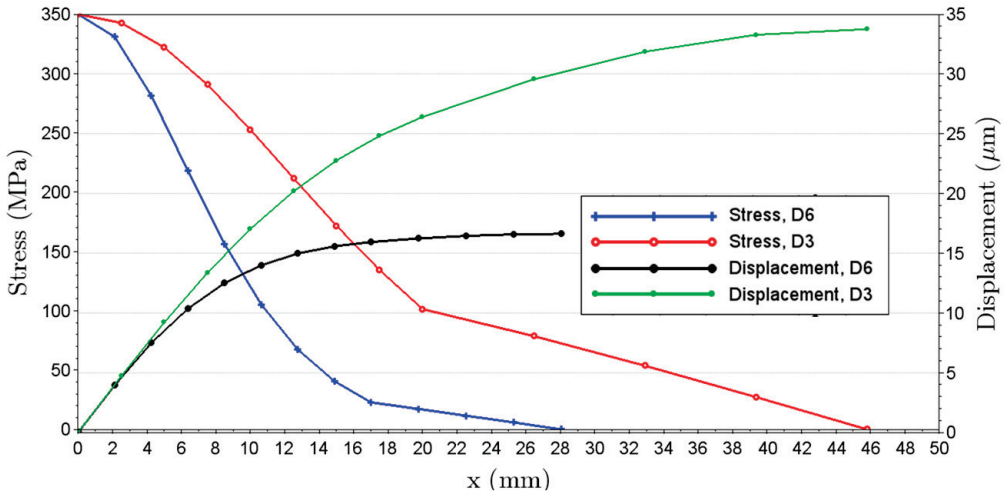


Figure 7. Stress and displacement amplitudes along specimen length, for specimens D6 and D3.

3.2. The Effect of Convection

The main objective of the present section is to quantify the effect of the cooling conditions in the specimen temperature. Two convention coefficients were used, 10 $W \cdot m^{-2} \cdot {}^{\circ}C^{-1}$, representing the natural coefficient and 10,000 $W \cdot m^{-2} \cdot {}^{\circ}C^{-1}$, representing force convection (typical value found for compressed air at the testing section of the specimen). Being convection the main contributor for heat exchange during fatigue tests, the present analysis aims to observe the differences between surface and inner temperatures of the specimens. Moreover, the effects of the convection coefficient will also be analysed

in combination with other parameters, such as the material thermal conductivity, which will be detailed in the following section. Both D3 and D6 specimens will be analysed and compared.

3.3. The Effect of Thermal Conductivity

The effect of thermal conductivity in the temperature distribution of a specimen will be analysed. The first case will be carried out using a thermal conductivity of 10 $W{\cdot}m^{-1}{\cdot}^{\circ}C^{-1}$, a typical value associated with stainless steels with low conductivity. The second case will be analysed for a thermal conductivity of 70 $W{\cdot}m^{-1}{\cdot}^{\circ}C^{-1}$, common values for high-carbon steels, but considerably lower than good conductors, like cooper (values of approximately 400 $W{\cdot}m^{-1}{\cdot}^{\circ}C^{-1}$). The specimens with testing section diameters of 6 and 3 mm will be investigated.

3.4. The Effect of Stress Levels and the Damping Ratio

As previously discussed, based on the analytical formulation of dissipated energy, from Equations (4)–(8), the elastic energy is a function of stress levels, while the dissipated energy depends on this elastic energy and a dissipative material property, like the damping ratio or the loss factor. Moreover, it was observed that for some materials, the damping ratio greatly increases according to the increment in stress amplitudes.

Figure 8 was developed based on the measured damping ratio (using the exponential decay) for a 316L ultrasonic fatigue specimen, as observed by Fiorentin et al. [9]. The main goal of the present analysis is to provide a comparison between three cases. The first (case 1) was a lower stress amplitude, 175 MPa (half the maximum stress cases), and a damping ratio of 0.7×10^{-3}, as shown in Figure 8. The other two cases will present the same maximum stress amplitude, 350 MPa, but will present different damping ratios. We will consider a damping ratio independent of the stress amplitude; therefore, the same 0.7×10^{-3} will be used (case 2), while the remaining case will encompass the effects of a variable damping ratio, with a value of 2.7×10^{-3} for the maximum stress amplitude (case 3). All cases will be performed using the larger VHCF specimen (D6), with forced convection (h = 10,000 $W{\cdot}m^{-2}{\cdot}^{\circ}C^{-1}$).

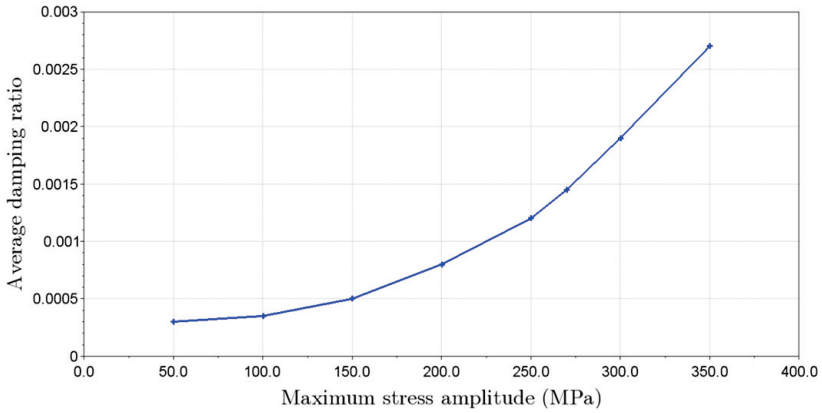


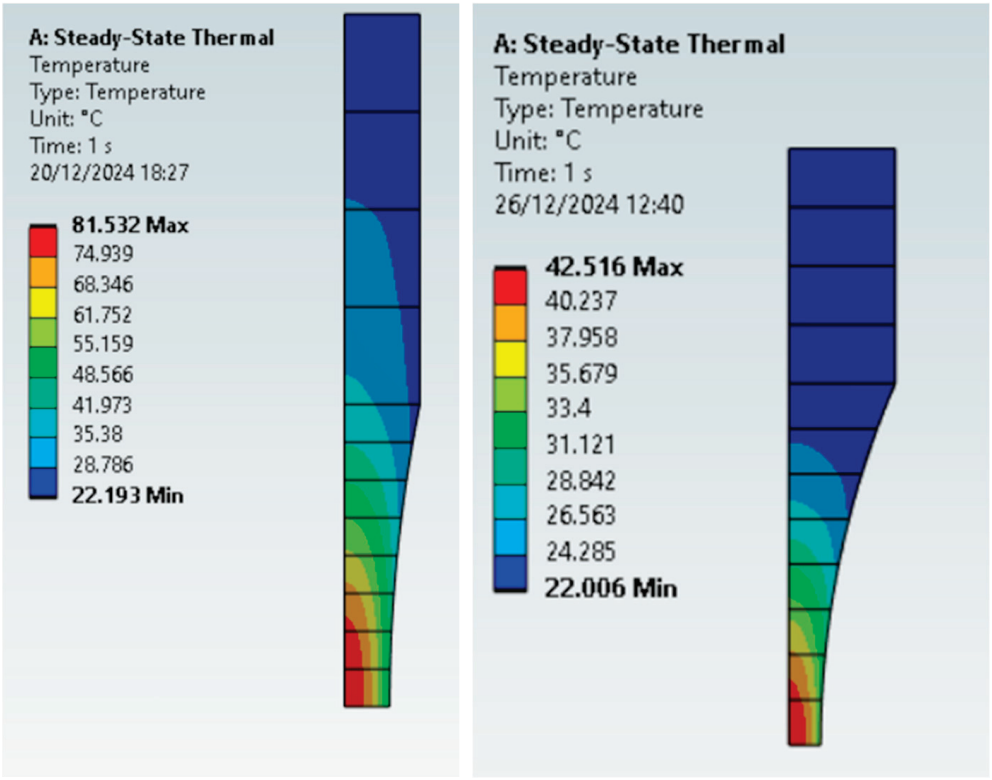
Figure 8. The damping ratio according to stress amplitudes for a 316L stainless steel.

4. Numerical Results and Discussion

The present section will provide the obtained numerical results, encompassing the impact of material properties, specimen dimensions, testing parameters and cooling conditions in the specimen heat generation and temperatures.

4.1. The Effect of Testing Section Diameter in the Specimen Temperature Distribution

Figure 9a,b present the thermal field for the specimens with testing section diameters of 6 and 3 mm, respectively. Although the maximum stress amplitudes were the same (and consequently the same dissipated energy per volume), with that of the specimen being D3 much smaller, this will lead to a lower total heat generated. Moreover, the heat cannot be transferred efficiently from the inner section of the larger specimen. The maximum temperature for the large specimen was 81 °C (at its inner section of the central portion of the specimen), while the smaller specimen presented a maximum temperature of 42 °C. When analysing the ratio between surface and core temperatures (R_t), the smaller specimen (D3) presented a more homogeneous temperature distribution at the maximum stress amplitude region, R_t = 0.79, while the larger specimen presented large differences, R_t = 0.58. Depending on how severe these temperature differences are, crack nucleation might occur at different locations than what it was expected for some materials, like in the specimen centre (regions with the highest temperatures).



(**a**) (**b**)

Figure 9. The effect of specimen dimensions in the temperature field during tests for: (**a**) a 6 mm diameter, D6; (**b**) a 3 mm diameter, D3.

It is important to emphasise that, even though a smaller specimen can be beneficial when looking at the heat generation and temperature aspects, reducing the specimen size will have stronger implications at the fatigue results. There is a known size effect on fatigue, where typically smaller specimens will present longer fatigue lives, since there is a smaller portion of the specimen subjected to higher stress amplitudes, and this smaller region will present lower probability of occurrence of larger defects (detrimental to fatigue life). Therefore, when designing a specimen, its dimensions must be representative (large enough), but the thermal factor must be taken into account for some materials. Moreover, drastically reducing the specimen cross section will also decrease its stiffness in the other direction (for example, the specimen will be prone to bending or lateral vibration), which can cause vibration in unwanted directions/modes.

4.2. The Effect of Convection in the Specimen Temperature Distribution

The present section will provide a brief discussion regarding the effect of natural and force convection in the specimen temperature during tests. It is important to note that ultrasonic fatigue tests using natural convection are often not practical, and the goal of the present analysis is only to demonstrate the effects of cooling and convection at the temperatures during tests, and its distribution, and how important an effective specimen cooling is.

Figure 10a,b present the temperature fields for a forced and natural convection, respectively, for the D6 specimen. It can be seen that, for such levels of heat generation, relying on natural convection for specimen cooling is not feasible. Due to the relatively low heat extraction due to natural convection, a homogeneous temperature at the central section can be found, with a relation between surface and inner temperature very close to 1 (R_t = 0.99).

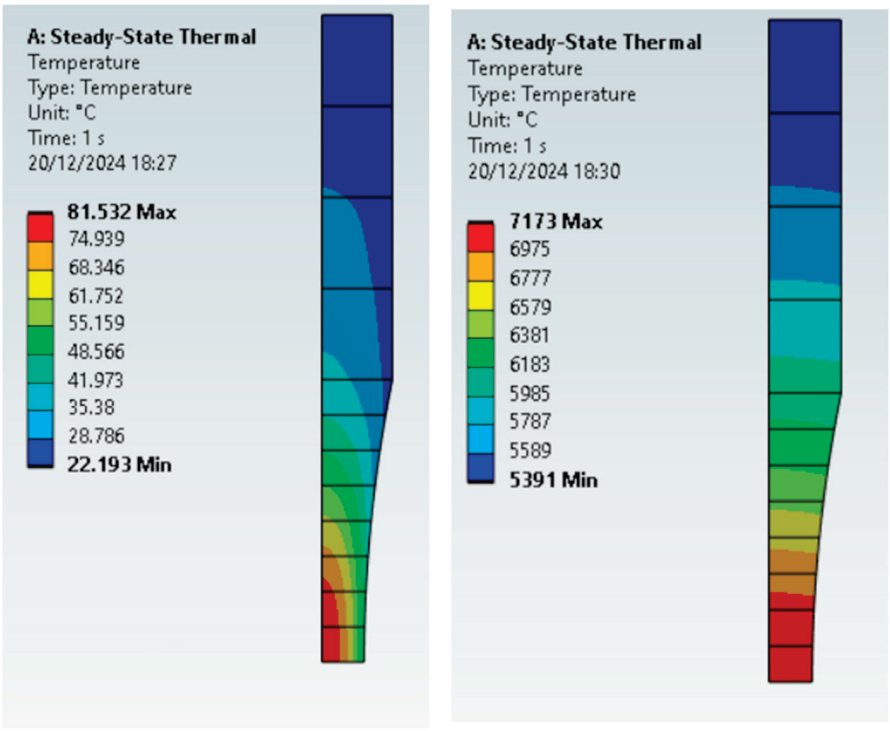
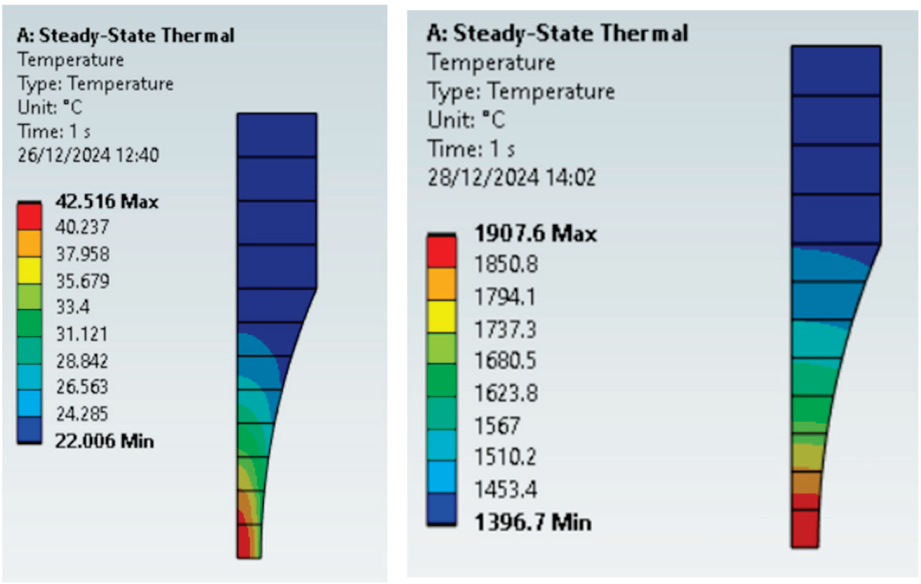


Figure 10. The effect of forced and natural convection in the temperature field during tests, specimen D6: (**a**) forced convection; (**b**) natural convection.

An analogous investigation was performed for the smaller specimen, D3, and the comparison between force and natural convection is presented in Figures 11a and 11b, respectively. Once more, when comparing the larger and smaller specimens, the latter presented lower temperatures in all the analysed cases. In most cases, temperature increases in the ultrasonic fatigue testing are seen as a disadvantage when the goal is to characterise the material in ambient temperature. However, this intrinsic characteristic of ultrasonic fatigue testing might be beneficial in some cases. Some materials will generate so much heat that it is possible to perform tests in high temperatures using only the specimen dissipated energy, without the need for an external heating device, if an efficient temperature measurement and control system is present.

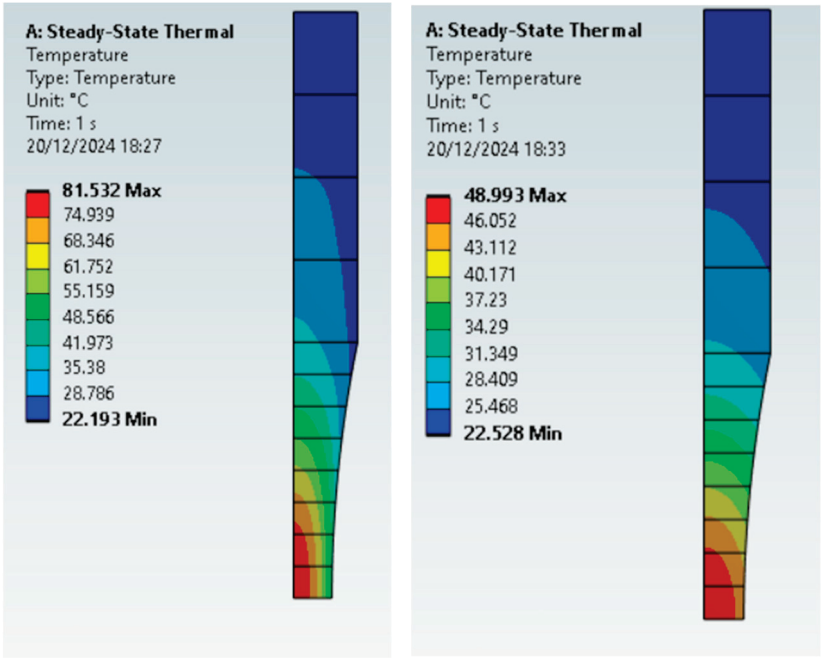


(**a**) (**b**)

Figure 11. The effect of forced and natural convection in the temperature field during tests, specimen D3: (**a**) forced convection; (**b**) natural convection.

4.3. The Effect of Thermal Conductivity on the Specimen Temperature Distribution

Next, an analysis encompassing the effects of thermal conductivity in the specimen distribution will be performed. All analyses were performed using the forced convection coefficient. Figure 12a,b present the temperature field for low and high thermal conductivity, respectively, for the larger specimen. When looking at the high conductivity case, lower temperatures were found. Moreover, at the testing section, the surface and inner temperature are much closer to each other, resulting in an R_t = 0.90, while the lower conductivity resulted in a R_t = 0.58.

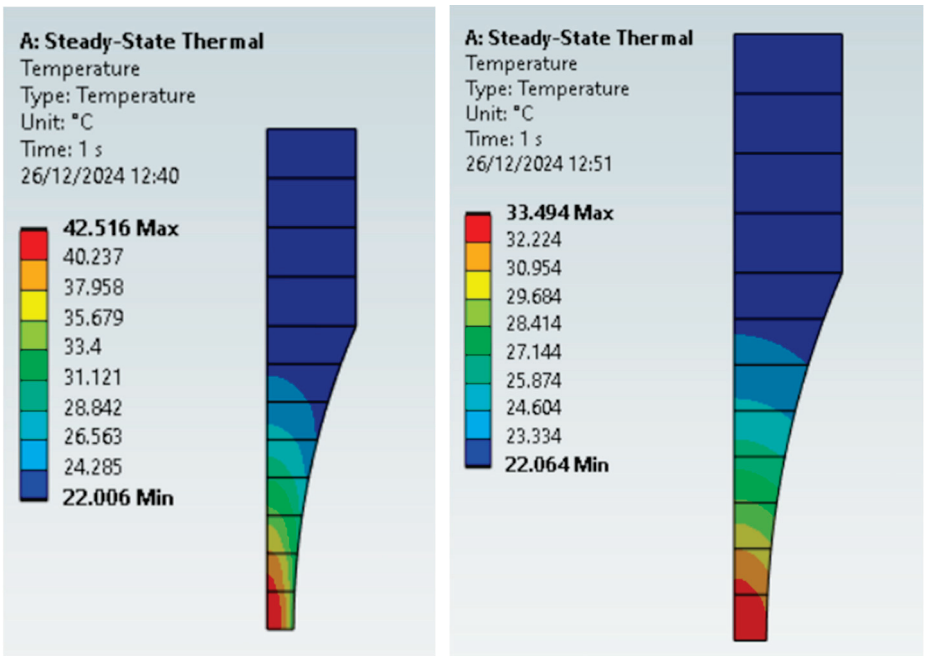


(**a**) (**b**)

Figure 12. The effect of thermal conductivity in the temperature field during tests, specimen D6: (**a**) k = 10 W.m^{-1}.°C^{-1}; (**b**) k = 70 W.m^{-1}.°C^{-1}.

When monitoring specimen temperature during tests, these measurements are performed using the surface of the specimen, by using pyrometers or thermocouples, for example). For materials with low thermal conductivity, the measurements (surface) might be very discrepant for some specimen temperatures, especially its core. Therefore, it is important to be aware of these discrepancies, which became greater as the thermal conductivity of the specimen material is reduced.

Regarding the smaller specimen, Figure 13a,b present the temperature field for lower and higher thermal conductivity, respectively. A very homogeneous temperature field was observed for the larger value of k, resulting in a ratio between the surface and inner temperature of 0.97.

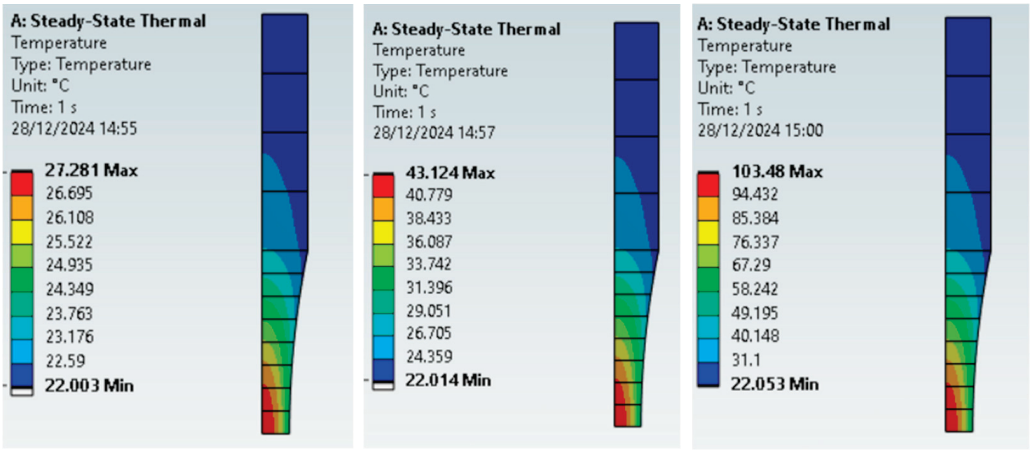


(**a**) (**b**)

Figure 13. The effect of thermal conductivity in the temperature field during tests, specimen D3: (**a**) k = 10 $W \cdot m^{-1} \cdot {}^{\circ}C^{-1}$; (**b**) k = 70 $W \cdot m^{-1} \cdot {}^{\circ}C^{-1}$.

4.4. The Effect of Stress Amplitude and Damping Properties in the Specimen Temperature

The effect of stress amplitude and the damping ratio will be discussed in the present section. Results from the simulations are presented in Figure 14. When comparing cases 1 and 2, where the only changes happen at the stress levels (the same damping ratio), Figures 14a and 14b, respectively, show where the maximum temperature increased by approximately 50% (27.3 °C for the lower stress level and 43.1 °C for the higher amplitude). The elastic energy increased by a factor of 4, since the stress levels were duplicated, and Equation (6) shows that the elastic energy varies with the squared stress amplitude. However, more heat is also extracted via convection, since the temperatures are higher. When the comparison is performed between a constant damping ratio approach and a variable one (dependent on stress levels), Figures 14b and 14c, respectively, it is clear that large errors in temperature estimates will be present if the wrong assumption regarding the damping ratio is used (more than 100% for the present case, when looking at maximum temperatures).



(**a**) (**b**) (**c**)

Figure 14. The effect of stress amplitudes and the damping ratio in the temperature field during tests, specimen D6: (**a**) stress amplitude of 175 MPa; (**b**) stress amplitude of 350 MPa, constant damping ratio; (**c**) stress amplitude of 350 MPa, variable damping ratio.

5. Conclusions and Future Work

The present work quantified the influence of certain material properties on heat generation and temperature increments during ultrasonic fatigue tests. Material property, like the damping ratio and thermal conductivity, effects on temperature increases and distribution during tests were discussed and quantified. Moreover, the effect of specimen dimensions and cooling characteristics on the temperature field distribution was investigated. By combining analytical approaches and numerical simulations, the present work was able to quantify the contribution of each material parameter and specimen main dimensions in the dissipated energy and temperature generated during tests.

By analysing the contribution of each material parameter on the heat generation during tests, it was possible to develop a guideline for specimen design and testing. The main contributions of the present work can be summarised as:

- Heat generation severity: a material parameter (based on relevant material properties) was developed to quantify whether a given material is prone to present high temperatures during ultrasonic fatigue tests. Such information might provide valuable data for engineers during the design of specimens, avoiding unpredicted challenges, which would only be detected after specimen manufacturing and during fatigue tests.
- Thermal gradient: The observation of high-temperature variations within the testing section (differences from the inner section of the specimen, higher temperatures, and the surface, lower temperatures) for low-conductivity materials may raise awareness regarding diverging fatigue results due to the temperature gradient effect. Moreover, higher cooling rates due to convection (forced convection and/or colder fluids, like liquid nitrogen) will also lead to a heterogeneous temperature field.
- The specimen geometry effect on temperature: The maximum temperatures and temperature gradients are strongly dependent on the specimen geometries. Larger specimens are more prone to higher temperature gradients, with the inner section of the specimens reaching the highest temperatures. This is especially critical because, temperatures are measured at the surface of the specimens, leading to underestimations of the temperatures during tests.

For future work, it would be interesting to expand the material analysis beyond metals, providing further guidelines and insights for polymeric and composites materials, for example. Moreover, the damping ratio dependence of stress levels is not fully well understood (possibly due to localised/microsplasticity), but its impact on the temperatures

during tests cannot be neglected; therefore, further investigation for several materials is needed. Additionally, encompassing more material parameters that affect the severity of heat generation during the ultrasonic fatigue testing would be interesting, where different weights could be used for parameters presenting higher influences on the heat generation and thermal distribution.

Furthermore, analysing the effects of severe temperature gradients found in some material (for example, a very hot inner portion of the specimen for a material with low thermal conductivity) could provide valuable information regarding the crack nucleation phenomenon for different temperatures. Last, the excessive heat generation observed in some metals could be used to perform high-temperature tests without needing an external heat source device.

Author Contributions: Conceptualisation, F.K.F. and A.M.P.d.J.; methodology, R.D. and A.P.C.; investigation, F.K.F., R.D. and A.P.C.; writing—original draft, F.K.F. and R.D.; software, J.W.G.; formal analysis, J.W.G.; writing—review and editing, J.W.G., A.P.C. and T.A.F.; funding acquisition, A.M.P.d.J.; resources, T.A.F. and A.M.P.d.J.; supervision, A.P.C., T.A.F. and A.M.P.d.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by FEDER and National Funds (FCT), reference PTDC/EME-EME/31307/2017, through the Add.Strength project entitled "Enhanced Mechanical Properties in Additive Manufactured Components". In addition, this research was also funded by FEDER and National Funds (FCT), reference PTDC/EME-EME/7678/2020, through the GCYCLEFAT project entitled "Giga-Cycle Fatigue Behaviour of Engineering Metallic Alloy". Moreover, this research was also funded by the Hi-rEV project (reference C644864375-00000002)—entitled "Recuperação do Setor de Componentes Automóveis" cofinanced by Plano de Recuperação e Resiliência (PRR), República Portuguesa through NextGeneration EU.

Institutional Review Board Statement: This study does not contain any studies with human or animal subjects performed by any of the authors.

Data Availability Statement: Data available on request due to restrictions, e.g., privacy or ethical.

Conflicts of Interest: All authors declare that they have no conflicts of interest.

References

1. Bathias, C.; Paris, P.C. *Gigacycle Fatigue in Mechanical Practice*; CRC Press: Boca Raton, FL, USA, 2004.
2. Fiorentin, F.K.; Dantas, R.; Darabi, R.; Lesiuk, G.; Figueiredo, M.A.V.; de Castro, P.T.; Jesus, A.M.P. Relation Between Striations Spacing and Fatigue Crack Growth Rates for Additive Manufactured Inconel 625. *Fatigue Fract. Eng. Mater. Struct.* **2024**, *48*, 454–466. [CrossRef]
3. Xue, H.; Wagner, D.; Ranc, N.; Bayraktar, E. Thermographic analysis in ultrasonic fatigue tests. *Fatigue Fract. Eng. Mater. Struct.* **2006**, *29*, 573–580. [CrossRef]
4. Grigorescu, A.; Hilgendorff, P.-M.; Zimmermann, M.; Fritzen, C.-P.; Christ, H.-J. Cyclic deformation behavior of austenitic Cr–Ni-steels in the VHCF regime: Part I–Experimental study. *Int. J. Fatigue* **2016**, *93*, 250–260. [CrossRef]
5. Ono, K. Dynamic Viscosity and Transverse Ultrasonic Attenuation of Engineering Materials. *Appl. Sci.* **2020**, *10*, 5265. [CrossRef]
6. Galuppo, W.C.; Santana, P.; Alves, F.; Nóbrega, J.M. Extending the Finite Area Method for enhanced simulation of deformable membranes and its application to extrusion blow moulding. *Thin-Walled Struct.* **2024**, *203*, 112184. [CrossRef]
7. Visnapuu, A.; Nash, R.W.; Turner, P.C. *Damping Properties of Selected Steels and Cast Irons*; US Department of the Interior, Bureau of Mines: Pittsburgh, PA, USA, 1987; Volume 9068.
8. Lage, Y.; Cachão, H.; Reis, L.; Fonte, M.; Freitas, M.D.; Ribeiro, A.A. A damage parameter for HCF and VHCF based on hysteretic damping. *Int. J. Fatigue* **2014**, *62*, 2–9. [CrossRef]
9. Fiorentin, F.K.; Reis, L.; Lesiuk, G.; Reis, A.; Jesus, A.M.P. A Predictive Methodology for Temperature, Heat Generation and Transfer in Gigacycle Fatigue Testing. *Metals* **2023**, *13*, 492. [CrossRef]
10. Mayer, H. Recent developments in ultrasonic fatigue. *Fatigue Fract. Eng. Mater. Struct.* **2016**, *39*, 3–29.
11. Yu, Y.; Gu, J.; Xu, L.; Shou, F.L.; Bai, B.Z.; Liu, Y.B. Very high cycle fatigue behaviors of Mn–Si–Cr series Bainite/Martensite dual phase steels. *Mater. Des.* **2010**, *31*, 3067–3072. [CrossRef]

12. Xiong, Z.; Naoe, T.; Futakawa, M. Effect of artificial defects on the very high cycle fatigue behavior of 316L stainless steel. *Metals* **2019**, *9*, 412. [CrossRef]
13. Juijerm, P.; Altenberger, I. Effect of temperature on cyclic deformation behavior and residual stress relaxation of deep rolled under-aged aluminium alloy AA6110. *Mater. Sci. Eng. A* **2007**, *452–453*, 452–475. [CrossRef]
14. Xu, J.; Lu, J.; Yuan, H. Effects of temperature gradients on thermomechanical fatigue of nickel-based supearlloy. *Int. J. Fatigue* **2024**, *184*, 108297. [CrossRef]
15. Jones, J. Thermo-Mechanical Fatigue—In the Gas Turbine Engine. *Encycl. Mater. Met. Alloys* **2022**, *1*, 476–484.
16. Sahith, M.S.; Giridhara, G.; Kumar, R.S. Development and analysis of thermal barrier coatings on gas turbine blades—A Review. *Mater. Today Proc.* **2018**, *5*, 2746–2751. [CrossRef]
17. Sun, J.; Yuan, H.; Vormwald, M. Thermal gradient mechanical fatigue assessment of a nickel-based superalloy. *Int. J. Fatigue* **2020**, *135*, 105486. [CrossRef]
18. Ashby, M.F. *Materials Selection in Mechanical Design*; Butterworth-Heinemann: Oxford, UK, 1999; p. 433.
19. Ahmed, T.; Mobarak, Y. Predictable Models and Experimental Measurements for Electric Properties of Polypropylene Nanocomposite Films. *Int. J. Electr. Comput. Eng.* **2016**, *6*, 120–129.
20. Salamt-Talab, M.; Delzendehrroy, F.; Akhavan_Safar, A.; Safari, M.; Bahrami-Manesh, H.; da Silva, L.F.M. Environmental effects on mode II fracture toughness of unidirectional E-glass/vinyl ester laminated composites. *Sci. Eng. Compos. Mater.* **2021**, *28*, 382–393. [CrossRef]
21. Petrone, G.; D'Alessandro, V.; Franco, F.; de Rosa, S. Damping evaluation on eco-friendly sandwich panels through reverberation time (RT60) measurements. *J. Vib. Control.* **2015**, *21*, 3328–3338. [CrossRef]
22. Matweb. Matweb. MatWeb, LLC, 2024. Available online: https://www.matweb.com/ (accessed on 30 December 2024).
23. Chen, W.; Spätig, P.; Seifert, H.P. Fatigue behavior of 316L austenitic stainless steel in air and LWR environment with and without mean stress. In Proceedings of the MATEC Web of Conferences, Poitiers, France, 27 May–1 June 2018.
24. Gomes, V.M.G.; Fiorentin, F.K.; Dantas, R.; Silva, F.G.A.; Correia, J.A.F.O.; Jesus, A.M.P. Probabilistic Modelling of Fatigue Behaviour of 51CrV4 Steel for Railway Parabolic Leaf Springs. *Metals* **2025**, *15*, 152. [CrossRef]
25. Dantas, R.; Gouveia, M.; Silva, F.G.A.; Fiorentin, F.; Correia, J.A.C.; Lesiuk, G.; Jesus, A.M.P. Notch effect in very high-cycle fatigue behaviour of a structural steel. *Int. J. Fatigue* **2023**, *177*, 107925. [CrossRef]
26. Fiorentin, F.K. High-Cycle and Ultra-High Cycle Fatigue Behaviour of Directed Energy Deposited Inconel 625. Ph.D. Thesis, Universidade do Porto, Porto, Portugal, 2023.
27. Bleicher, C.; Wagener, R.; Kaufmann, H.; Melz, T. Fatigue strength of nodular cast iron with regard to heavy-wall applications. *Mater. Test.* **2015**, *57*, 723–731. [CrossRef]
28. Colakoglu, M. Factors effecting internal damping in aluminium. *J. Theor. Appl. Mech.* **2004**, *42*, 95–105.
29. Lee, Y.Y.; Welsch, G. Young's modulus and damping of Ti6Al4V alloy as a function of heat treatment and oxygen concentration. *Mater. Sci. Eng. A* **1990**, *128*, 77–89. [CrossRef]
30. Chamos, A.N.; Pantelakis, S.G.; Haidemenopoulos, G.N.; Kamoutsi, E. Tensile and fatigue behaviour of wrought magnesium alloys AZ31 and AZ61. *Fatigue Fract. Eng. Mater. Struct.* **2008**, *31*, 812–821. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Accuracy Analysis of Complex Transmission System with Distributed Tooth Profile Errors

Min Zhang [1], Zhijing Zhang [1,*], Jian Xiong [1,2] and Xiao Chen [1]

[1] School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China; m.zhang@bit.edu.cn (M.Z.); cielo@bit.edu.cn (J.X.); chenxiao@bit.edu.cn (X.C.)
[2] EPSRC Future Metrology Hub, University of Huddersfield, Huddersfield, Queensgate HD1 3DH, UK
* Correspondence: zhzhj@bit.edu.cn

Abstract: Tooth profile errors are the internal excitations that cause gear meshing errors, which are critical error factors affecting gear transmission accuracy. In existing studies, it is usually regarded as a constant or random distribution function. However, the actual machined tooth profile error is not a constant, so this estimation is inconsistent with the actual situation, resulting in an inaccurate evaluation of transmission accuracy. This paper proposes a method for representing tooth profile errors using distribution errors (including systematic and random errors), and a mathematical model of distributed tooth profile errors is presented. The contact stresses of the complex transmission system were compared with those obtained by formulas, proving that tooth profile errors increase contact stress. A method for calculating gear meshing error is proposed to evaluate the actual output accuracy of the complex transmission system. Compared with the test, the output accuracy is reduced by 13.8% under the temperature environment and distributed tooth profile errors. The proposed methods can accurately predict the transmission accuracy of precision transmission systems at the design stage and provide theoretical support for reducing systematic and random errors at the gear machining stage.

Keywords: complex transmission system; distributed tooth profile error; gear meshing error; transmission accuracy; contact stresses

1. Introduction

Transmission accuracy is one of the important indexes for evaluating the performance of gear transmission, usually expressed by transmission error, which can reflect the accuracy and stability of gears in transmitting motion and power. As pointed out in [1], the difference between the transmission accuracy of the gear system and the ideal accuracy is unavoidable due to multiple sources of errors in the machining process, such as manufacturing errors, workpiece mounting errors, and the transmission chain errors of the machine tool. Therefore, the existing research on transmission error mainly focuses on how to accurately determine the transmission error under the influence of load, error, friction, thermal deformation, and other parameters [2,3], as well as how to reduce the transmission error through design and machining [4–8]. Among them, they aim to improve the gear profile by modification techniques to reduce the transmission error. There has been considerable research on tooth profile modification techniques. Accordingly, in the author's previous work [9], a study on the quantitative influence of tooth pitch error, gear runout error, and assembly error on transmission error was carried out. However, as a mechanism relying on the sliding and rolling of teeth of non-trivial geometries [10], the contact state of the tooth surfaces is greatly affected by the tooth profile error. Therefore, analyzing gear transmission accuracy by considering tooth profile errors is necessary, especially regarding the effect of distribution errors of tooth profiles on transmission error. It is also a problem that the authors have been trying to study.

Gear dynamics has received increasing attention since Tuplin first proposed the spring–mass vibration model [11]. Many investigations have focused on tooth profile errors from the perspective of dynamic equations [12–16], in which tooth profile errors are regularly incorporated into dynamic equations as displacement-type functional excitations occurring at interfaces to form the unloaded static transmission error (STE). For instance, in [12], the author regarded the tooth profile error as a random tooth profile error and incorporated it into a nonlinear random model. On the other hand, Liu [17] simulated the tooth profile error excitation with pseudo-random numbers. Concerning the tooth flank error, Yuan's work [18,19] involved the distributed tooth flank error, but only the meshing clearance was considered as the distributed tooth flank error. The effect of tooth profile error on transmission error was analyzed using the above method.

Some other existing literature has addressed tooth profile errors in terms of gear tolerances, thin slice gears, and undesired tooth profile errors [20–24], e.g., of which [20], manufacturing errors were treated as random errors. Thus, a random tooth profile error distribution was given for each tooth. In [21,22], the tooth profile deviation distribution was obtained by discretizing the teeth. In [23,24], the authors focused on the effect of tooth profile error on meshing stiffness. The former defined the tooth profile error as the amount of tooth profile modification, while the latter represented the tooth profile error as the sum of the sinusoidal curves. In the calculation, the amplitude was given by the determined value.

The abovementioned research on tooth profile errors needs to be revised, and the current understanding of gear errors still needs to be completed. Recently, in another research field, the study of evaluating the transmission errors of harmonic reducers for robotics based on actual machined tooth profiles has attracted the authors' attention. For this paper, the tooth profiles of the harmonic drive were measured using a commercial non-contact CMM, the pitch errors were extracted using various mathematical methods, and the angular transmission error of the harmonic drive was predicted considering three-dimensional tooth engagement [25]. The above study predicts that the prediction of transmission errors based on actual machined tooth profiles becomes the future trend for research and improvement of transmission errors, regardless of spur or harmonic gearing.

Therefore, this paper aims to simulate the effect of actual machining errors on tooth profile errors, including position errors and form errors according to the characteristics of precision machinery manufacturing, and then assess the impact of such distributed tooth profile errors on the transmission error of the complex transmission system. First, the mathematical modeling method is used to simulate the tooth profile slope deviation (positional error) caused by systematic errors and the tooth profile form error caused by random errors in gear machining errors, making the obtained involute tooth surface as close as possible to being in line with the actual. Secondly, the effects of distributed tooth profile errors on contact stresses are evaluated based on the established three-dimensional solid model containing distributed tooth profile errors and finite element dynamics calculations. Furthermore, the evaluation method of gear meshing error is proposed, which provides an effective method for evaluating the transmission accuracy of the gear transmission system. Finally, based on the purpose of obtaining the three-dimensional point cloud data of the actual tooth surface, a general-purpose contact CMM was used to measure the experimental gears and finite element calculations were carried out on the experimental gears, which verified the significance of the involute tooth profile error modeling in this paper.

This paper is organized into six sections. Section 2 contains the involute profile error modeling, including systematic and random errors. Section 3 uses the finite element dynamics method to calculate the complex transmission system with distributed tooth profile error. The effects of distributed tooth profile errors on the contact stress and gear meshing errors are determined. The experiment in Section 4 verifies the presented approach's validity. Section 5 discusses the benefits of the proposed methods in this paper. Section 6 concludes this paper.

2. Geometric Model of Complex Transmission System Containing Distributed Tooth Profile Errors

This paper takes the complex transmission system of the joint for a large space manipulator as the research object. The reason is that as an essential transmission of the joint, its planetary drive has many stages and large gear ratios. In addition, the space environment and stress state differ from the ground environment. Gear tooth profiles for the complex transmission system are all involutes. Their manufacturing methods include molding, profiling, and generating. Currently, the most commonly used method is generating, such as gear shaping, hobbing, and shaving, of which the first two are the most widely used gear processing methods. The factors that affect gear machining accuracy include errors in machine tools, cutting tools, gear blanks, and mounting. These errors affect the accuracy of the processed gear according to a specific law. For example, machining errors in hobbing can lead to six profile errors: flanks out of the edge, asymmetry, profile angle error, periodic error, root cutting, and root fillet error. Figure 1a shows the asymmetry of tooth profile (also called skewed teeth) generated during gear hobbing, which is formed due to the following factors. Note the poor radial performance of the hob blade after grinding, that is, the rake angle of the hob blade is not equal to zero, which results in non-radial error γ of the front blade, as shown in Figure 1b.

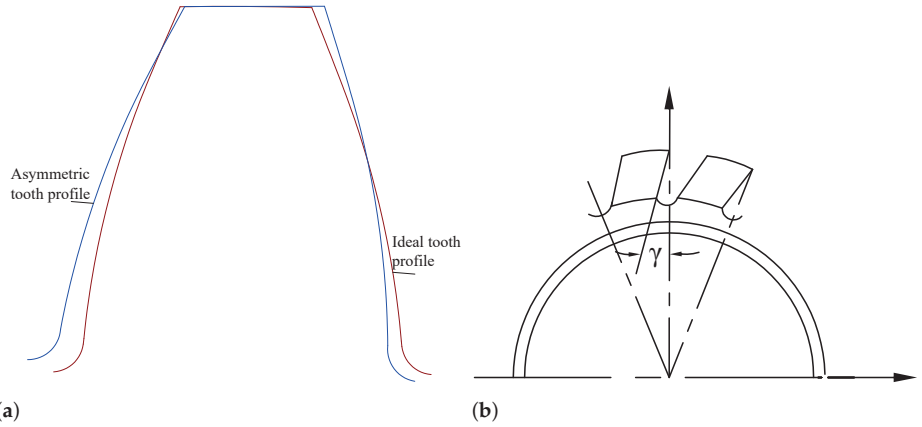


Figure 1. Formation of skewed teeth. (**a**) Tooth profile asymmetry. (**b**) Positive rake angle.

Therefore, the actual tooth profile deviates from the ideal form, and the actual gear is not an ideal involute gear. However, the gear system mainly relies on the contact meshing of the tooth surface to transmit motion and power, and the contact state between two tooth surfaces is greatly affected by the tooth profile error. For a high-precision complex transmission system, the meshing excitation due to tooth profile errors will be a crucial factor affecting the transmission accuracy. Accordingly, this section establishes a three-dimensional solid model of the complex transmission system with distributed tooth profile errors through the modeling and analysis of tooth profile errors, providing a three-dimensional geometric model for the subsequent finite element analysis in this paper.

2.1. Modeling of Involute Tooth Profile Error

Given the above analysis, the machining error that forms tooth profile error can be composed of systematic and random errors. The geometric error of the tooth surface caused by systematic error is deterministic, which can describe the distribution law of the geometric error produced by the same machining process. In contrast, random errors lead to differences between machined gears. Therefore, to establish an involute cylindrical gear that matches the machining error, we divided the involute tooth profile error into systematic and random errors in the subsequent modeling and provided their modeling formulas, respectively.

In the design of involute gears, tooth profile accuracy is one of the items used to evaluate the accuracy of the tooth flanks on the same side of the gear teeth. It is generally expressed by the total deviation ΔF_α of the tooth profile. As presented in Figure 2, the total deviation of tooth profile refers to the distance between the two design tooth profile traces encompassing the actual tooth profile trace within the range of the profile count [26]. To satisfy the requirements of transmission stability, the total deviation is specified to follow the condition [27]:

$$\Delta F_\alpha \leq F_\alpha \tag{1}$$

where F_α is the total tolerance value of the tooth profile, which is determined according to the design requirements.

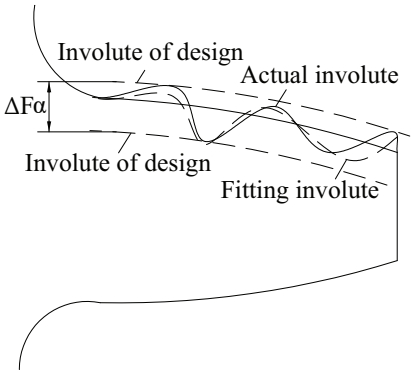


Figure 2. Schematic diagram of involute tooth profile.

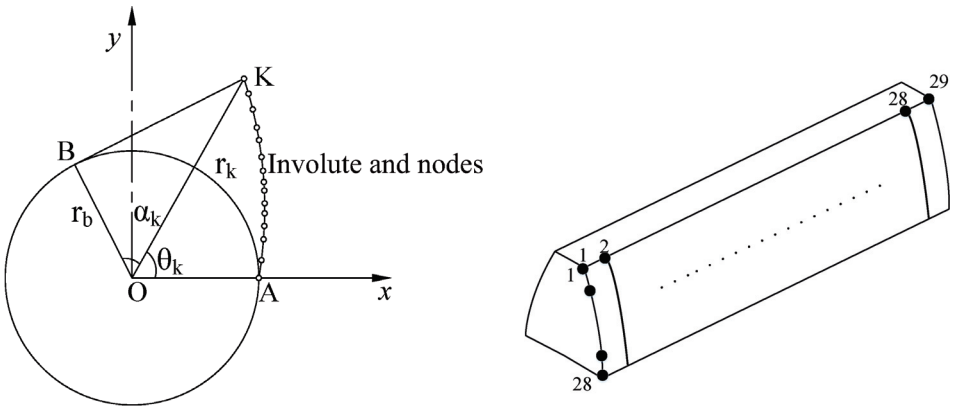
The total deviation of the tooth profile can be decomposed into two parts, expressed by the tooth profile form error representing the form accuracy and the tooth profile slope deviation representing the direction accuracy, respectively. Tooth profile form error is mainly formed by high-frequency vibration of the machining system and the form error of the cutting tool of the forming gear. Tooth profile slope deviation, also known as pressure angle deviation, is related to the hob cutter's pressure angle, the base circle's diameter deviation, and the gear workpiece's geometric eccentricity during the cutting process. Consequently, in gear machining, systematic errors mainly cause tooth profile slope deviation, while random errors mostly form tooth profile form errors. The mathematical models of the systematic and random errors that form the involute tooth profile errors will be described in detail as follows.

Figure 3 refers to the formation of the actual tooth flank, where Figure 3a is the formation of the involute curve, and its polar coordinate equation takes the form of:

$$\begin{cases} r_k = \frac{r_b}{\cos \alpha_k} \\ \theta_k = \tan \alpha_k - \alpha_k \end{cases} \tag{2}$$

where r_b is the radius of the base circle and r_k refers to the radius vector of point k on the involute. The θ_k denotes the development angle, which is a function of the pressure angle α_k of point k. Since the pressure angles at each point on the involute are unequal, a change in α_k will lead to a variation in θ_k. Therefore, the involute position error e_p caused by systematic errors in the machining process can be expressed as a function of the change in pressure angle $\Delta\alpha_k$.

$$e_p = \Delta r_k(\Delta\alpha_k) \tag{3}$$



(**a**) (**b**)

Figure 3. Formation of actual involute tooth flank. (**a**) Involute tooth profile curve. (**b**) Involute tooth flank.

That is, the actual radial vector $r_k{}'$ is as follows:

$$r_k{}' = r_k + \Delta r_k(\Delta\alpha_k) \tag{4}$$

In practice, we usually calculate the variation in the pressure angle α on the pitch circle, which is also the tooth profile angle error $\Delta\alpha$. It can be obtained by combining Equation (5) with Equation (6).

$$L_\alpha = 0.92 \times \left(\sqrt{r_a^2 - r_b^2} - r_b \tan\alpha + \frac{h_a^* m}{\sin\alpha} \right) \tag{5}$$

where L_α is the range of values for evaluating tooth profile deviation, r_a refers to the radius of the addendum circle, h_a^* denotes the addendum coefficient, and m is the modulus of the gear.

$$\Delta\alpha = -\frac{\Delta f_\alpha}{L_\alpha \tan\alpha} \tag{6}$$

where Δf_α is the tooth profile slope deviation.

Then, the point cloud data (x',y',z) on the involute containing the systematic errors can be obtained by converting the polar coordinates into Cartesian coordinates.

On the other hand, the theoretical parameter equation of an involute is given as follows:

$$\begin{cases} x = r_b \cos\theta + r_b\theta \sin\theta \\ y = r_b \sin\theta - r_b\theta \cos\theta \end{cases} \tag{7}$$

where θ refers to the rolling angle at any point of the involute, which can be computed by $\theta{=}\theta_k + \alpha_k$; thus, the tooth profile form error caused by random errors can be expressed as:

$$\begin{cases} \Delta x = \Delta r_b \cos(\Delta\theta) + \Delta r_b \Delta\theta \sin(\Delta\theta) \\ \Delta y = \Delta r_b \sin(\Delta\theta) - \Delta r_b \Delta\theta \cos(\Delta\theta) \end{cases} \tag{8}$$

When calculating, it is assumed that the variation Δr_b in the radius of the base circle and the variation $\Delta\theta$ in the rolling angle follow a normal distribution.

$$\begin{cases} \Delta r_b \in \mathrm{N}\left(\mu_{\Delta r_b}, \sigma_{\Delta r_b}\right) \\ \Delta\theta \in \mathrm{N}(\mu_{\Delta\theta}, \sigma_{\Delta\theta}) \end{cases} \tag{9}$$

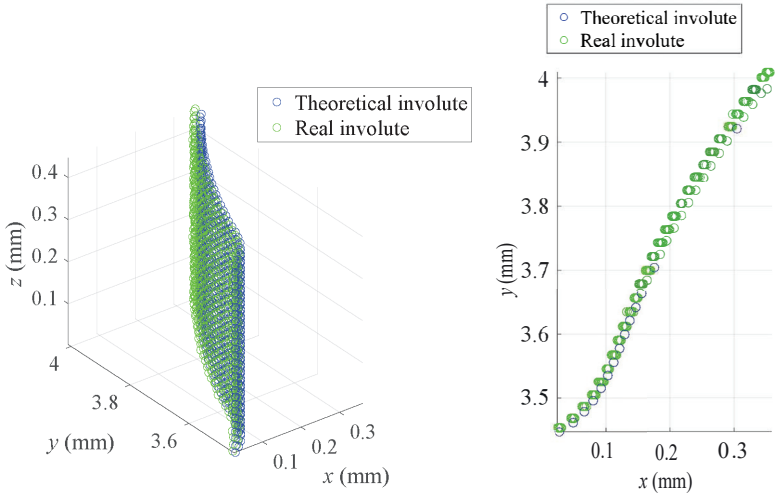
where $\mu_{\Delta r_b}$ and $\mu_{\Delta\theta}$ are the mean values of the variation in the base circle radius and the variation in rolling angle, respectively. $\sigma_{\Delta r_b}$ and $\sigma_{\Delta\theta}$ are their standard deviations, respectively.

Accordingly, the points (X',Y',Z) on the actual involute are determined by the following formula under the combined action of systematic and random errors.

$$\begin{cases} X' = x' + \Delta x \\ Y' = y' + \Delta y \\ Z = z \end{cases} \tag{10}$$

2.2. Generation of Error Tooth Surface

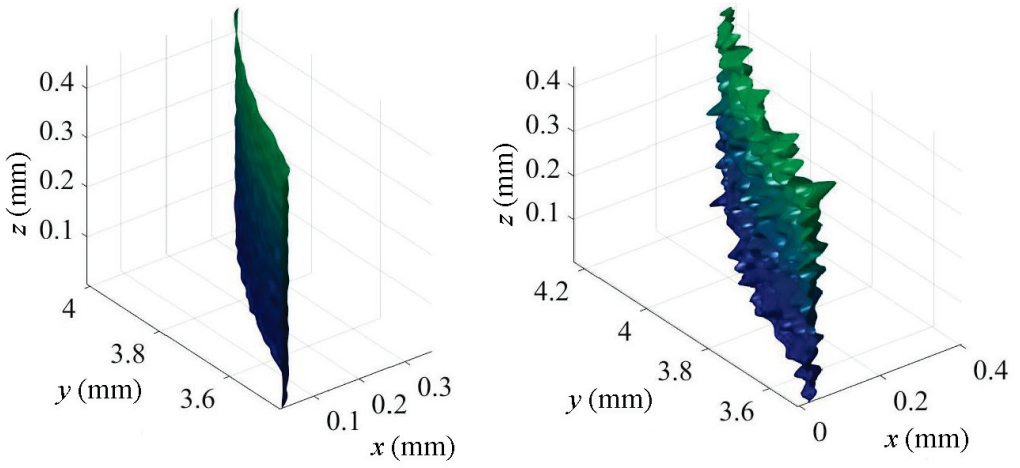
Figure 3b displays a tooth flank with 28 points along the involute direction and 29 involutes along the face width direction. When using the above equations to generate error tooth flanks that include tooth profile errors, the first problem to be solved is to determine the point cloud data for the $m \times n$ matrix. The specific steps are detailed below. In the first step, the first theoretical involute curve is acquired using Equation (7), and the (x_0, y_0, z_0) coordinates of the m points are obtained. It is worth noting that the z coordinate of the first involute is 0, and then n such involutes are determined along the z direction of the face width. Subsequently, combined with Equation (2) to Equation (6), the data matrix of $m \times n$ considering systematic errors is determined. Furthermore, random errors can be obtained by combining Equation (8) with Equation (9), thus determining the point cloud data on the actual tooth flanks based on Equation (10). The results are plotted in Figure 4a. To better compare the ideal and error points, their xy views are depicted in Figure 4b.



(**a**) (**b**)

Figure 4. Comparison between ideal and error points. (**a**) Ideal and error points. (**b**) xy view of ideal and error points.

Zhang [28] used NURBS surface reconstruction to represent geometric error. However, in the case of gears, the establishment of error surfaces based on NURBS numerical modeling has yet to be thoroughly studied. Hence, for the generated point cloud data, this paper creatively uses the NURBS numerical model for the first time to establish an error surface for gears with distributed tooth profile errors, as depicted in Figure 5. Among them, Figure 5b is the tooth flank magnified ten times along x and y, respectively, to display the form error of the tooth profile.



(**a**) (**b**)

Figure 5. Actual involute tooth flank and its error amplified. (**a**) Interpolated 3D surface. (**b**) Error amplified.

2.3. Verification of Surface Modeling Accuracy

The distance d between the set of points on the actual involute surface and the set of points on the theoretical involute surface is used to verify that the geometric error of the tooth profile is within the total profile tolerance.

$$d = \sqrt{(x_i - x_{0i})^2 + (y_i - y_{0i})^2} \tag{11}$$

where (x_i, y_i) are points on the actual involute surface and (x_{0i}, y_{0i}) represent points on the theoretical involute surface.

Based on Equation (11), the maximum value d_{Max} and the minimum value d_{Min} are determined, respectively, which makes the actual tooth profile errors F'_α satisfy the following condition:

$$F'_\alpha = |d_{\mathrm{Max}}| + |d_{\mathrm{Min}}| \leq F_\alpha \tag{12}$$

2.4. Geometric Model Of Complex Transmission System

Based on Section 2.2, a NURBS mathematical model is employed to represent the actual tooth flanks with distributed tooth profile errors, providing a basis for integrating the error tooth flank with the CAD design model. Furthermore, the gear teeth with distributed tooth profile errors were obtained, as shown in Figure 6. The 3D geometric model of the gear containing distributed tooth profile errors was obtained using CAD modeling software (PTC Creo Parametric 3.0 M060). Finally, we obtained the geometric model of the complex transmission system, which includes a fixed-axis gear train and a planetary gear train, as shown in Figure 7. Figure 7a displays the fixed-axis gear train, which includes an input gear 1, two pinions 2, a fixed carrier, and a double gear 3. Figure 7b is the planetary gear train, which contains a gear 3′, a gear 4, a gear 5, an output gear 6, and a fixed ring 7. Figure 7a,b are connected via double gear 3. For convenience of analysis, the fixed carrier is omitted in Figure 7a. At the same time, the internal gear of gear 3 is also omitted in Figure 7b, and only the external gear 3′ is retained.

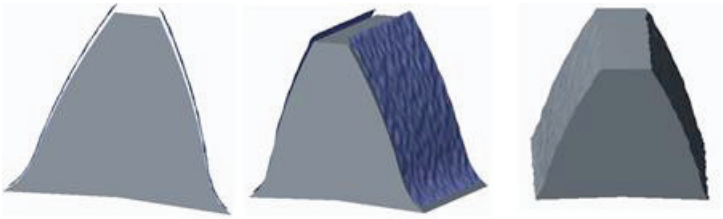
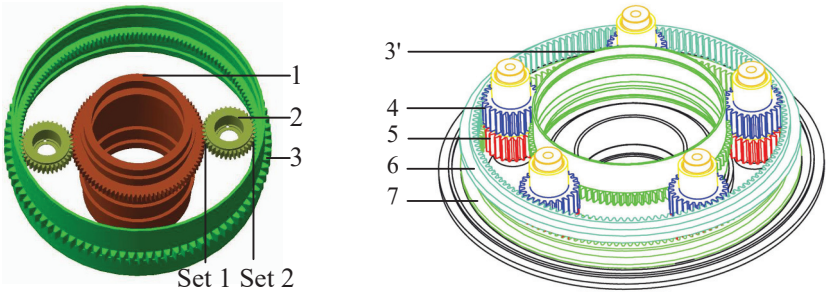


Figure 6. The gear teeth with distributed tooth profile error.

Moreover, given the sensitivity analysis in the author's previous study [9], we know that the gear errors of the planetary gear system, especially gears 4 and 5, have a significant impact on the whole transmission system. Hence, the planetary gear system is the focus of this paper. To depict the planetary gear system, the planetary carrier and output shaft in Figure 7b are omitted, but they all exist in the actual analysis. The basic parameters and profile accuracy of each gear are presented in Tables 1 and 2, respectively.



(**a**) (**b**)

Figure 7. Complex transmission system. (**a**) Fixed-axis gear train. (**b**) Planetary gear train.

Table 1. Parameters of each gear.

Gears	Module (mm)	Number of Teeth	Width (mm)	Pressure Angle (°)
Gear 1	1	91	8	20
Gear 2	1	35	9	20
Gear 3	1	161	8	20
Gear 3′	2	90	11	20
Gear 4	2	25	24	20
Gear 5	2	24	24	20
Gear 6	2	140	22	20
Gear 7	2	139	22	20

Table 2. Profile accuracies of each gear.

Gears	Tooth Profile Tolerance (μm)	Tooth Profile Form Error (μm)	Tooth Profile Slope Deviation (μm)	Tooth Line Tolerance (μm)
Gear 1	8.5	6.5	5.5	9
Gear 2	7.5	5.5	4.6	9
Gear 3′	10	7.5	6	9
Gear 3	14	11	9	11
Gear 4	7.5	5.5	4.6	9
Gear 5	7.5	5.5	4.6	9
Gear 6	14	11	9	11
Gear 7	14	11	9	11

3. Finite Element Dynamic Model

The primary purpose of this section is to analyze the contact dynamics of the planetary gear system under the combined effect of distributed tooth profile error, load, and temperature field. Accordingly, the contact analysis should be performed in two steps. First, a finite element dynamic model is developed using Dynamic/Explicit. Second, the contact stress and gear meshing error are determined based on the simulation results. As for dynamic analysis, the masses and inertias of the wheels are shown in Table 3. The preprocessing settings for the finite element model are reported in Table 4.

Table 3. Masses and inertias of each gear.

Gears	Masses (Kg)	Inertias (Kg· mm^2)
Gear 1	0.448	819
Gear 2	0.041	11.1
Gear 3	1.236	7070
Gear 3′	1.236	7070
Gear 4	0.7	356
Gear 5	0.645	356
Gear 6	1.797	31,100
Gear 7	1.309	25,300

Table 4. Finite element preprocess settings.

Modeling Steps	Parameters	Numerical Value
Solver settings	Software	ABAQUS (Abaqus/CAE 2021)
	Step	Dynamic
	Solver	Explicit
Mesh	Element	C3D8R
Property	Ultra-high-strength stainless steel (gear)	Young's modulus 190 GPa
		Poisson's ratio 0.305
		Mass density 7.9 g/cm^3
		Expansion coeff 1.1 × 10^{-5}
	Titanium alloy (carrier and shaft)	Young's modulus 105 GPa
		Poisson's ratio 0.34
		Mass density 4.51 g/cm^3
		Expansion coeff 1.09 × 10^{-5}
Interaction	Interaction	Surface-to-surface contact (Explicit)
	Interaction property	Coulomb friction coeff 0.1 and hard contact
	Constraints	Coupling and tie
Boundary condition	Temperature	From −20°C to 50°C
	Fixed boundary	Bottom of support
	Rotate or translate	Gears and shafts

3.1. Influence of Distributed Tooth Profile Error on Contact Stress

First, we illustrate the influence of distributed tooth profile error on the gear contact stress by using the example of the fixed-axis gear train of the complex transmission system. The idea is to use the gear contact stress calculation method given in Section 4.1.2 of the national standard GB/T 3480-1997 [29] to obtain the calculated contact stress. However,

this must not take into account the distributed tooth profile error. Then, the above results are compared and analyzed with those obtained by the finite element calculation.

Whether Hertz's contact theory [30] or the national standard is applied to calculating the contact stress of involute cylindrical gears, the effect of temperature on the contact stress is not considered. Therefore, the temperature field was not considered in the finite element dynamic model of the error model of the fixed-axis gear train. For comparison, Table 5 shows the calculated contact stresses obtained based on the formulas and the maximum contact stresses obtained by finite element analysis.

Gear meshing is a line contact that starts from the tooth root of the driving gear pushing the tooth tip of the driven gear. Therefore, the driving gear's root and the driven gear's tip are prone to stress concentrations. In addition, when the gears are just coming into a mesh, or the single and double teeth alternately mesh, the maximum value of the stress variation is easily reached.

Table 5. Comparison of calculated contact stresses with maximum contact stresses of FEM.

Parameters		Calculated Contact Stress (MPa)	FEM (MPa)	Relative Error (%)
Set 1	Gear 1	557.8427	619.4	11.03
	Gear 2	572.0677	559.2	2.25
Set 2	Gear 2	458.9514	426.8	7.53
	Gear 3	458.9514	556.6	18.99

Table 5 clearly shows that the numerical results for gears 1 and 3 are larger than those calculated by the formulas, with relative errors of 11.03% and 18.99%, respectively. The reason is that the stress concentration occurs on the root of gear 1 and the tip of gear 3. On the other hand, the finite element dynamic model considers the distributed tooth profile error, so the local contact stress is larger than those calculated by formulas. More importantly, the calculated contact stress can only evaluate the stress-bearing limit of the gear as a whole, whereby it cannot obtain the stress concentration point and the maximum contact stress when the stress concentration occurs. In contrast, the FEM is closer to the actual situation of the gear system. Consequently, when designing gears, we should consider the distribution of tooth profile error on the whole tooth surface rather than simply determining the gear's bearing limit based on an ideal geometric model or a fixed formula.

3.2. Influence of Distributed Tooth Profile Error on Transmission Accuracy

In the space environment, many factors affect the joint's output accuracy and operation stability, such as the temperature field, load, and parts errors. After we have analyzed and proved that the distributed profile errors significantly impact the contact stress of the gear train, we need to analyze the transmission accuracy of the complex transmission system under the effects of distributed profile error, temperature, and load. In addition, we propose a method to obtain the meshing errors by combining the finite element analysis results. Then, at the end of this section, based upon the author's previous work on this large space manipulator [9], the output accuracy and stability of this large space manipulator in the actual service process are comprehensively evaluated.

As depicted in Figure 8, it is a meshing pair of involute tooth profiles. When the pair of tooth profiles engage at any position, the common normal line passing through the contact points is the same line N_1N_2, which indicates that all the engagement points of the pair of involute tooth profiles from engagement to disengagement are on this line. Therefore, line N_1N_2 is the trajectory of the contact points of the tooth profile in a fixed plane, also known as the line of action.

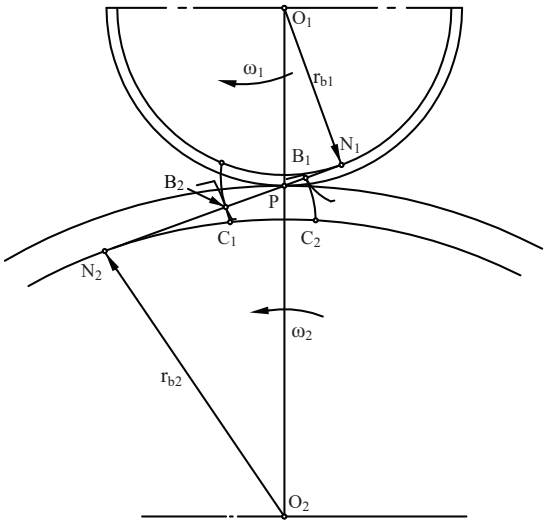


Figure 8. Meshing diagram of involute tooth profile: O_1O_2 is the two gears' centerline; P is the intersection of O_1O_2 and N_1N_2; B_1B_2 is the actual line of action; r_{b1} and r_{b2} are the gears' base circle radius, and ω_1 and ω_2 are their angular velocities; C_1 and C_2 are the two gears' starting points of the involute curves.

Subsequently, the coordinates of all nodes on the involute where the maximum contact stress occurs are extracted according to the principle of invariant positive pressure direction of the involute tooth profile. Based on the original node coordinates and the deformed node coordinates, the radius of the circle where each node on the involute is located is obtained, which corresponds to the radial r_k forming the involute, as presented in Figure 3a. As a result, the radius difference δr_j before and after the gear deformation can be determined. For illustration, Table 6 and Table 7 display the involute node radius changes for gears 3′ and 4, respectively.

Table 6. Change in involute node radius of gear 3′.

No Temperature			With Temperature		
Before Deformation (mm)	**After Deformation (mm)**	**Radius Difference (mm)**	**Before Deformation (mm)**	**After Deformation (mm)**	**Radius Difference (mm)**
87.50001	87.50137	0.001361	87.49996	87.50254	0.002589
87.96042	87.9628	0.002373	87.96038	87.96388	0.003504
88.41834	88.42083	0.002494	88.41836	88.42268	0.004324
88.87398	88.87638	0.002402	88.87395	88.87846	0.004508
89.32725	89.32971	0.002463	89.32725	89.33252	0.005268
89.77824	89.77896	0.000716	89.77831	89.7825	0.004185
90.22696	90.22597	−0.00099	90.227	90.23014	0.003144
90.67344	90.67352	7.72×10^{-5}	90.67351	90.67691	0.003402
91.11783	91.11916	0.001337	91.11783	91.12363	0.005799
91.56003	91.56075	0.000724	91.56	91.56513	0.005124
92.00001	92.0012	0.001194	92.00003	92.00616	0.006127

Since the gear meshing error occurs along the line of action, i.e., the normal direction of the involute tooth profile, the following formula can determine the gear meshing error. The pressure angle α_k and the development angle θ_k can be determined by Equation (2).

$$\delta = \delta r_j \cdot \theta_k \tag{13}$$

Table 7. Change in involute node radius of gear 4 when gear 4 is meshed with gear 3′.

No Temperature			With Temperature		
Before Deformation (mm)	**After Deformation (mm)**	**Radius Difference (mm)**	**Before Deformation (mm)**	**After Deformation (mm)**	**Radius Difference (mm)**
23.8238	23.90013	0.076329	24.2386	24.33186	0.093254
24.23868	24.30746	0.06878	24.64855	24.72476	0.076204
24.64866	24.69927	0.050611	25.0533	25.11462	0.061315
25.05332	25.08933	0.036009	25.45233	25.49516	0.042833
25.45229	25.46973	0.017433	25.8455	25.86815	0.022646
25.84553	25.84219	−0.00334	26.23244	26.23121	−0.00123
26.23242	26.20512	−0.02729	26.61312	26.5875	−0.02562
26.61316	26.56096	−0.0522	26.98747	26.93468	−0.05279

Combining the radius difference δr_j of the involute node with Equations (2) and (13), it is possible to determine the meshing errors for each pair of gears under the excitation of distributed profile error, rotational speed, and external load in the absence and presence of temperature variations, as illustrated in Figures 9–11.

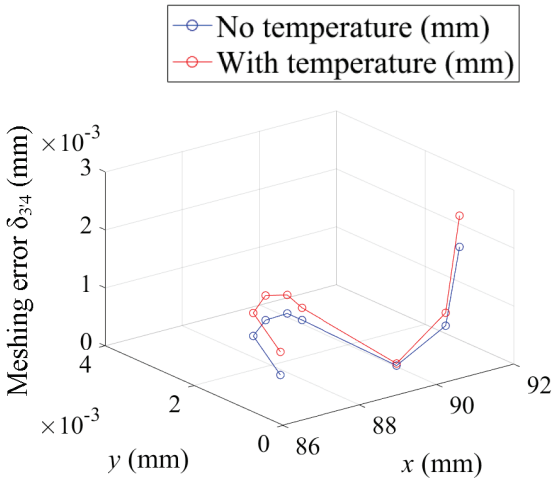


Figure 9. Meshing error of gears 3′ and 4.

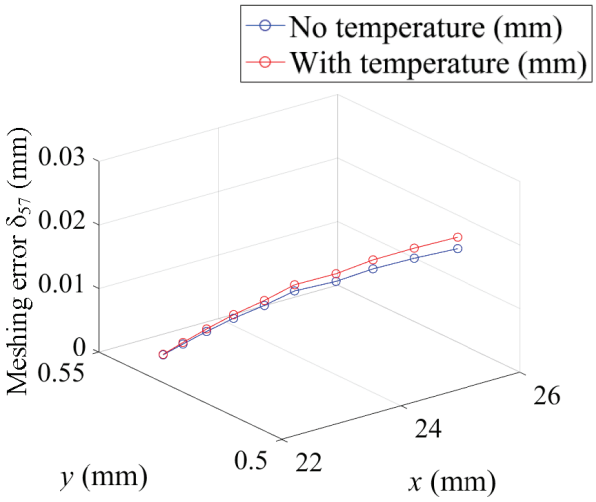


Figure 10. Meshing error of gears 5 and 7.

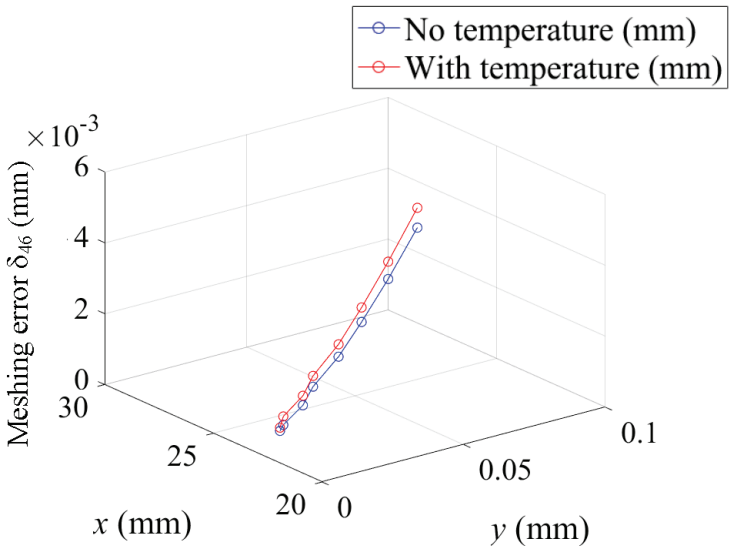


Figure 11. Meshing error of gears 4 and 6.

Because two gears are involved at each meshing point location, the position coordinates xy of the meshing points are used to represent the node positions. The z coordinates of Figures 9–11 represent the meshing errors. The relationship between the position of the meshing points and the meshing errors is clearly illustrated.

Referring to Figures 9–11, it can be observed that regardless of temperature, the tooth profile errors will affect the variation in the transient ratio in the gear mesh transmission, i.e., it causes the variation in the transient transmission error, which is the reason why the tooth profile error is an internal excitation of the gear transmission. To minimize this internal excitation, it is necessary to make the tooth surfaces in uniform contact as much as possible, instead of non-uniform contact. Meanwhile, the meshing errors of each pair of gears in the planetary gear system are higher than those without temperature variation, which indicates that the temperature has a significant effect on the transmission accuracy of the system.

Accordingly, in the cases of coupling tooth profile error, load, and temperature field, the transmission error of the system can be calculated through the linear displacement Equation (14) of gear meshing error.

$$\Delta\delta \cdot r_b = x \tag{14}$$

where $\Delta\delta$ refers to the angular displacement of the gears' torsional motion and r_b is the radius of the base circle.

In this work, the transmission error with the coupling factors was predicted to be 0.192′. Combined with the transmission error TE′ based on machining and assembly errors in the author's previous investigation [9], the total transmission error for the complex transmission system was 1.5021′. Compared with the experimental results on transmission accuracy in the same reference [9], it demonstrates that the joint's transmission error is 13.8% higher than that of the test due to the action of geometric errors, assembly errors, tooth profile errors, load, and temperature field.

The results prove that, under the comprehensive influence of the above error factors, the actual output accuracy of the complex transmission system of the joint for a large space manipulator is 13.8% lower than that of the test accuracy during the actual space operation. Consequently, the operation accuracy and stability of the large space manipulator in service will be seriously affected when completing the scheduled task.

4. Experiment

Since the complex transmission system studied in this paper is expensive and involves many gears, no spare objects are available for experimentation. Therefore, to explore the effect of machining error on the tooth profile error, this paper took the first-stage gear pair of a three-stage gearbox as the experimental object, as shown in Figure 12. Their basic parameters and profile accuracies are denoted in Table **??** and Table 9, respectively.

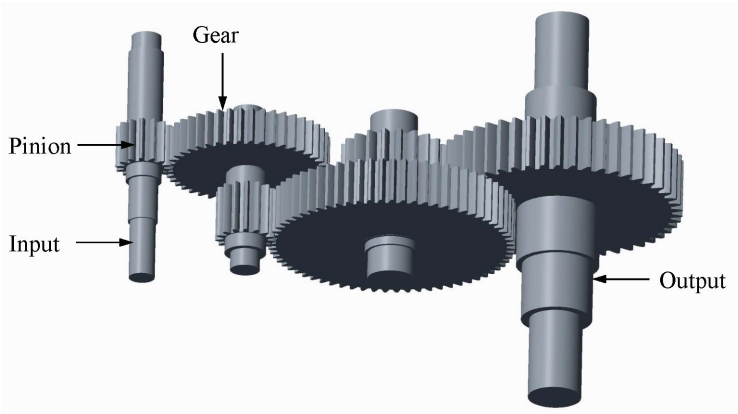


Figure 12. Three-stage gear system.

Table 8. Parameters of measured gears.

Gears	Module (mm)	Number of Teeth	Width (mm)	Pressure Angle (°)
Pinion	3	14	36	20
Gear	3	44	30	20

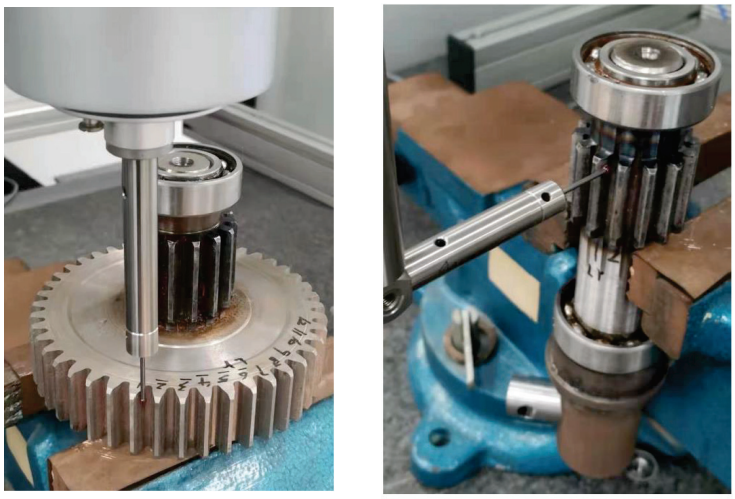
Table 9. Profile accuracies of measured gears.

Gears	Tooth Profile Tolerance (μm)	Tooth Profile form Error (μm)	Tooth Profile Slope Deviation (μm)	Tooth Line Tolerance (μm)
Pinion	20	19	15	18
Gear	25	19	15	18

4.1. Measurement of Gear Tooth Surfaces

As we all know, a special gear measuring machine is usually used for measurement after gear machining. Then, its geometric accuracy is evaluated to determine whether the machining error meets the requirements for gear accuracy. But it cannot obtain the three-dimensional point cloud data of the tooth surface. The author's team has been researching error and precision assembly technology and theory for a long time. It has accumulated many bases in obtaining the surface topography of parts, including selecting measurement methods and developing measurement strategies [28,31–34]. To compare the effects of different measurement methods on the measurement data of the assembly surface of precision machined parts, Zhang [33] designed an experimental sample and analyzed the flatness measured on the sample surface using a triggered CMM, a continuous scanning profilometer, an auto-zoom three-dimensional surface measuring instrument, and a metrological X-CT. The following conclusions were drawn: CMMs not only have the advantages of a simple calibration process, good measurement repeatability, minor systematic errors, and controllable sampling interval but also have the advantages of flexible probes, adaptable to a wide range of features, such as cylindrical and semi-circular surfaces, as well as low requirements for measurement environments. It has also been verified that the CMM is more suitable for acquiring geometric form error information. In addition, the evaluation of the advantages and disadvantages of a measurement method

mainly depends on whether it can adapt to the surface features, whether it can carry out efficient and high-precision measurements, whether it can improve measurement efficiency, and whether it can reduce the amount of data collected under the premise of guaranteeing the measurement accuracy [35]. Therefore, the measurement experiments in this section use a CMM (Brown & Sharpe Mistral 7107, with a measurement uncertainty of (1.5 + L/300) μm, the measurement system of PC-DMIS, and a 40 mm stylus with ball tip 2 mm in diameter), rather than a specialized gear measuring machine, to measure the surface topography of the experimental gears, as described in Figure 13.

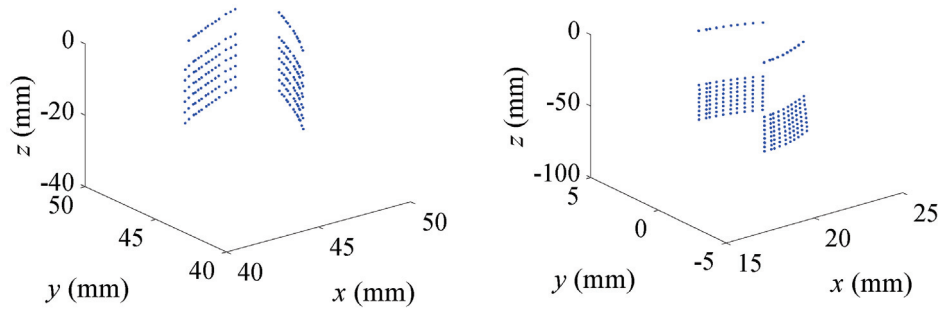


(**a**) (**b**)

Figure 13. Measurement experiment of tooth flanks. (**a**) Gear. (**b**) Pinion.

First, in developing the measurement scheme for the gear surface, the layout scheme of the measurement points is shown in Figure 3b. Furthermore, a fixed coordinate system needs to be established for the measurement, and each tooth is measured in this coordinate system. In terms of the clamping method, since several tooth surfaces of the gear need to be measured simultaneously, to facilitate one-time measurement and reduce the measurement error generated by part clamping, we need a fixture to fix the gear to ensure that it is not moved during the measurement process. Considering that the experimental gear is a gear shaft, a fixed shaft part can be used to fix the gear, so a vise was chosen for this experiment. It is undeniable that mounting a vise will introduce mounting error, which will be a part of the measurement error of the tooth surface measurement data. After obtaining the measurement data, we will process it, including gross error removal, systematic error removal, random error filtering, etc., and interpolate the measurement data points to obtain matrix data that can be modeled using NURBS surfaces.

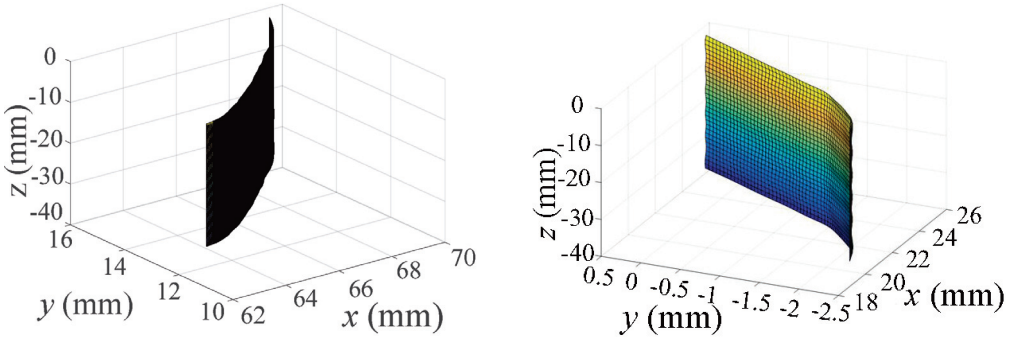
The original measurement data of the gear pair are reported in Figure 14. Generally, the accuracy of a gear is evaluated when measuring its tooth flanks. Although more parameters can be used to evaluate the geometric accuracy of gears, such as tooth pitch deviation, tooth profile deviation, helix deviation, tangential integrated deviation, one-tooth tangential integrated deviation, etc., this paper focuses on studying the impact of geometric distribution errors on the gear transmission accuracy. Therefore, after measuring the point cloud data of the tooth surface using the CMM, we are more concerned about the error distribution state of the whole tooth surface. Other geometric errors are not the focus of this paper, so they are not evaluated in this paper.



(**a**) (**b**)

Figure 14. Original measurement data of tooth flanks. (**a**) Gear. (**b**) Pinion.

In terms of the measured data, based on Section 2.2, the NURBS interpolated 3D surfaces of the actual measured gears were established, as presented in Figure 15. It can be noticed that the errors of the tooth flanks are mainly concentrated in the y direction, that is, the normal direction of the involute. Finally, the actual gear models with distributed tooth profile errors were determined by integrating error tooth surfaces with their CAD design models.



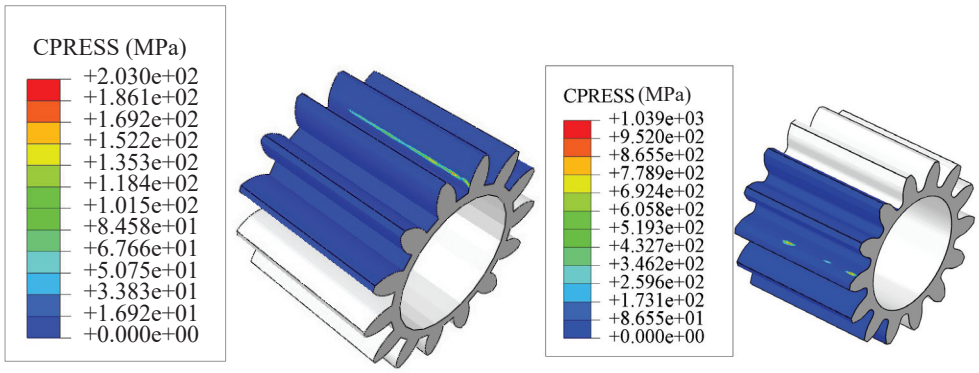
(**a**) (**b**)

Figure 15. NURBS interpolated 3D surfaces of gear and pinion (The colors only represent changes in height in the z-direction and have no other special meanings). (**a**) Gear. (**b**) Pinion.

4.2. Finite Element Analysis of Measured Gears

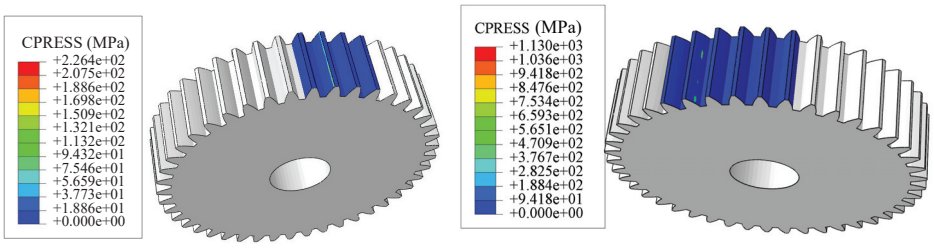
This section is to evaluate the contact stresses on the measured gear pairs. Although the measurement experiments were performed for only one gear pair, the finite element calculation was performed for the three-stage gear system. The applied torque load at the output end is 200 Nm, the friction coefficient is 0.2, and the rest of the finite element preprocessing method is similar to that of Section 3. For comparative analysis, Figures 16 and 17 report the maximum contact stress contours for the ideal and error gears. Because finite element software like ABAQUS has no unit information when inputting material parameters, load parameters, etc., the users only need to input a numerical value. However, the users need to ensure consistency in the dimensions of the input data for each physical quantity and know the units of the calculation results. To better understand Figures 16 and 17, the authors manually added the units MPa.

Attending to Figures 16 and 17, the contact stresses are uniform along the face width direction for the ideal pinion and gear, while for the error pinion and gear, the contact stresses are discontinuous. The results of the stress comparison between the two figures are presented in Table 10.



(**a**) (**b**)

Figure 16. Contact stress contours of pinion along face width direction. (**a**) Ideal pinion. (**b**) Error pinion.



(**a**) (**b**)

Figure 17. Contact stress contours of gears along face width direction. (**a**) Ideal gear. (**b**) Error gear.

Table 10. Comparisons of contact stresses between ideal gears and error gears.

Parameters	Ideal Pinion	Error Pinion	Multiple	Ideal Gear	Error Gear	Multiple
CPRESS/MPa	203	1039	5.12	226.4	1130	4.99

To further illustrate the relationship between tooth profile error and contact stress distribution, Figure 18 is introduced. It depicts the distribution of tooth profile error along the face width z direction for the gear's measurement data (one of the rows), where the red line indicates no error and the blue line refers to the tooth profile error. Combining Table 10 and the stress distribution state of the gears, it can be concluded that the distribution law of the contact stress is related to the distribution law of the tooth profile error. The maximum contact stress will appear where the tooth flanks are convex or concave, that is, where the tooth profile error is greatest. The existence of local maximum contact stress will increase the gear wear. Therefore, studying systematic and random errors can enable us to reduce these two types of errors at the machining stage by means of tooth profile modification and other methods.

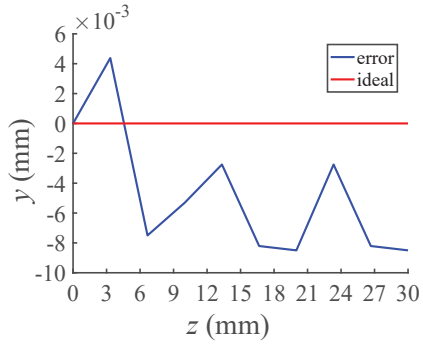


Figure 18. Distribution of tooth profile error along face width direction based on measurement data.

5. Discussion

The purpose of our study is to propose a modeling method for involute tooth profile error, which can simulate the actual tooth flanks formed by gear machining errors. The systematic errors formed by different gear machining processes are different, and the tooth profile form errors formed by the same machining method are also different due to different random errors. The actual involute tooth profile formed by the above reasons can only be determined through modeling within the tooth profile tolerance. In summary, when designing a gear transmission system, if the machining process of this gear and the variations in its base circle radius and rolling angle are known, the accurate tooth profile of the gear can be simulated using the modeling method proposed in this paper, and a three-dimensional solid error model of the gear can be established. Through methods such as finite element calculations, we can obtain more realistic transmission accuracy. Especially for fine-pitch precision gears, the impact of systematic and random errors during the machining process is more significant due to the smaller allowable range of tooth profile tolerance, tooth profile slope deviation, and tooth profile form error. At this point, if we need to predict the transmission performance of the fine-pitch gear systems at the design stage, the superiority of the method proposed in this paper can also be demonstrated.

On the other hand, tooth profile errors will aggravate tooth wear and affect the transient transmission error of gears. To minimize these effects, some measures should be taken. For mass-produced gears, it is impossible to measure each gear, but we can measure a few gears to find the systematic error that conforms to the machining law. Then, the tooth profile modification can be utilized to remove that systematic error. For precision gears processed in small quantities, due to the small quantity, we can measure each gear to find systematic and random errors. At this point, the systematic and random errors can be removed by profile modification. Therefore, the benefit of the involute profile error modeling proposed in this paper is that we must first understand the systematic and random errors formed during machining and then eliminate them as much as possible. Moreover, for precision complex transmission systems that involve a large number of gears, according to the meshing error calculation method proposed in this paper, the quantitative impact of each pair of gears on the transmission error of the whole transmission system can be determined, which is conducive to the optimization design of essential gears.

6. Conclusions

This paper investigates the effects of distributed tooth profile errors (including systematic and random errors) on the transmission accuracy and contact stress of a complex transmission system. The systematic error is mainly caused by the pressure angle deviation. The random error is primarily caused by the variation in base circle radius and the variation in rolling angle, and their mathematical models are established, respectively. Then, the finite element contact dynamic model is established. In view of the simulation results, the contact stresses of the fixed-axis gear train are compared with the calculated contact stresses to illustrate the impact of the distributed tooth profile errors on the contact stresses. Furthermore, based on the numerical simulation model, we propose a calculation method for the gear meshing error. According to the calculation results, we have found that non-uniform tooth profile errors will increase the variation in transient transmission errors, thereby affecting gear transmission errors.

The main conclusions of this study are summarized as follows.

(1) Since the pressure angle at each point on the involute is not equal, the positional error of the involute due to systematic errors in gear machining can be expressed as a function of the variation in the pressure angle. The change in the tooth profile form of the involute due to random errors in gear machining can be expressed as a function of the change in base circle radius and the change in roll angle. These are presented in the modeling of involute profile errors proposed in this paper.

(2) Compared with the calculated contact stresses obtained by the formulas, the contact stresses with distributed tooth profile error are relatively higher. Through experimental

verification, we have determined that the contact stresses are non-uniformly distributed along the face width direction and the maximum contact stresses appear in the convex and concave positions of the tooth surface. Moreover, the transmission error with distributed tooth profile error is larger than the transmission error without considering the tooth profile error previously studied by the authors. Furthermore, comparing the transmission errors under the influence of temperature with the transmission errors considering only tooth profile errors, it is found that temperature exacerbates the meshing error fluctuations.

(3) The gear meshing error extraction method proposed in this paper can further obtain the transmission error of the gear system, thereby determining the meshing gear pairs that have the most significant impact on the transmission accuracy of the system. For instance, according to the results of the analysis of transmission accuracy in this paper, the meshing error formed by gears 5 and 7 is the governing factor affecting the transmission error. This result is consistent with the sensitivity analysis in the authors' previous study.

Author Contributions: Conceptualization and methodology, M. Z., Z.Z. and J.X.; writing—original draft, M.Z.; writing—review and editing, M.Z. and J.X.; formal and resources, X.C.; software and validation, M.Z.; supervision, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant numbers U22B2088, 52205512, and 52305530).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, J.; Wang, S.; Ma, C.; Xia, C. Crucial errors identification of worm grinding process of face gears based on tooth flank posture error model and their influences on tooth flank errors. *Int. J. Adv. Manuf. Technol.* **2022**, *119*, 6177–6192. [CrossRef]
2. Karma, V.K.; Maheshwari, G. *Determination of Transmission Error and Mesh Stiffness in Spur Gear Pair Using CAD-FEM Integration Approach*; IOP Publishing Ltd.: Tamil Nadu, India, 2023. [CrossRef]
3. Pedrero, J.I.; Sánchez, M.B.; Pleguezuelos, M. Analytical model of meshing stiffness, load sharing, and transmission error for internal spur gears with profile modification. *Mech. Mach. Theory* **2024**, *197*, 105650. [CrossRef]
4. Velex, P.; Chapron, M.; Fakhfakh, H.; Bruyère, J.; Becquerelle, S. On transmission errors and profile modifications minimising dynamic tooth loads in multi-mesh gears. *J. Sound Vib.* **2016**, *379*, 28–52. [CrossRef]
5. Chen, H.; Zhu, Y.; Mao, J.; Wang, X. Influence of tooth profile modification on design transmission error of involute spur gear. *J. Huazhong Univ. Sci. Technol. (Natural Sci. Ed.)* **2018**, *46*, 91–97. [CrossRef]
6. Sanchez, M.B.; Pleguezuelos, M.; Pedrero, J.I. Influence of profile modification on the transmission error of spur gears under surface wear. *Mech. Mach. Theory* **2024**, *191*, 105473. [CrossRef]
7. Chen, Z.; Ji, P. Research on the variation of mesh stiffness and transmission error for spur gear with tooth profile modification and wear fault. *Eng. Fail. Anal.* **2020**, *122*, 105184. [CrossRef]
8. Zhou, D.; Guo, Y.; Yang, J.; Zhang, Y. Study on the Parameter Influences of Gear Tooth Profile Modification and Transmission Error Analysis. *Machines* **2024**, *12*, 316. [CrossRef]
9. Zhang, M.; Zhang, Z.; Shi, L. A new assembly error modeling and calculating method of complex multi-stage gear transmission system for a large space manipulator. *Mech. Mach. Theory* **2020**, *153*, 103982. [CrossRef]
10. Ottewill, J.R.; Neild, S.A.; Wilson, R.E. An investigation into the effect of tooth profile errors on gear rattle. *J. Sound Vib.* **2010**, *329*, 3495–3506. [CrossRef]
11. Tuplin, W.A. *Gear Design*, 1st ed.; Machinery Publishing Co. Ltd.: London, UK, 1962.
12. Xun, C.; Long, X.; Hua, H. Effects of random tooth profile errors on the dynamic behaviors of planetary gears. *J. Sound Vib.* **2018**, *415*, 91–110. [CrossRef]
13. Gregory, R.W.; Harris, S.L.; Munro, R.G. Dynamic Behaviour of Spur Gears. *Proc. Inst. Mech. Eng.* **1963**, *178*, 207–226. [CrossRef]
14. Theodossiades, S.; Natsiavas, S. Periodic and chaotic dynamics of motor-driven gear-pair systems with backlash. *Chaos Solitons Fractals* **2001**, *12*, 2427–2440. [CrossRef]
15. Özgüven, H.; Houser, D.R. Dynamic analysis of high speed gears by using loaded static transmission error. *J. Sound Vib.* **1988**, *125*, 71–83. [CrossRef]
16. Mark, W.D. Use of the generalized transmission error in the equations of motion of gear systems. *J. Mech. Des. Trans. ASME* **1987**, *109*, 283–291. [CrossRef]
17. Liu, C.; Qin, D.; Liao, Y. Dynamic model of variable speed process for herringbone gears including friction calculated by variable friction coefficient. *J. Mech. Des. Trans. ASME* **2014**, *136*, 041006. [CrossRef]

18. Yuan, B.; Chang, S.; Liu, G. Quasi-static analysis based on generalized loaded static transmission error and dynamic investigation of wide-faced cylindrical geared rotor systems. *Mech. Mach. Theory* **2019**, *134*, 74–94. [CrossRef]
19. Yuan, B.; Chang, L.; Liu, G. An efficient three-dimensional dynamic contact model for cylindrical gear pairs with distributed tooth flank errors. *Mech. Mach. Theory* **2020**, *152*, 103930. [CrossRef]
20. Bonori, G.; Pellicano, F. Non-smooth dynamics of spur gears with manufacturing errors. *J. Sound Vib.* **2007**, *306*, 271–283. [CrossRef]
21. Wang, Q.; Zhang, Y. A model for analyzing stiffness and stress in a helical gear pair with tooth profile errors. *J. Vib. Control.* **2016**, *23*, 272–289 . [CrossRef]
22. Wang, Q. An Investigation for Analyzing the Dynamic Characteristics and Fatigue Reliability of a Gear System. Ph.D. Thesis, Northeast University, Shengyang, Liaoning Province, China, 2015.
23. Chen, Z.; Zhou, Z.; Zhai, W. Improved analytical calculation model of spur gear mesh excitations with tooth profile deviations. *Mech. Mach. Theory* **2020**, *149*, 103838. [CrossRef]
24. Chen, Z.; Shao, Y. Mesh stiffness calculation of a spur gear pair with tooth profile modification and tooth root crack. *Mech. Mach. Theory* **2013**, *62*, 63–74. [CrossRef]
25. Kim, B.S.; Jeong, S.T.; Ahn, H.J. The Prediction of the angular transmission error of a harmonic drive by measuring noncontact tooth profile and considering three-dimensional tooth engagement. *Int. J. Precis. Eng. Manuf.* **2023**, *24*, 371–378. [CrossRef]
26. *GB/T 10095.1-2022*; Shi, Z.; Wang, Z.; Yu, B. Cylindrical gears-ISO System of Flank Tolerance Classification—Part 1: Definitions and Allowable Values of Deviations Relevant to Flanks of Gear Teeth. State Administration for Market Regulation and Standardization Administration of China, CNKI, Beijing, China, 2022.
27. Liu, X. (Ed.) *Involute Cylindrical Gear*; China Planning Press: Beijing, China, 2004.
28. Zhang, Z.; Zhang, Z.; Jin, X. A novel modelling method of geometric errors for precision assembly. *Int. J. Adv. Manuf. Technol.* **2017**, *94*, 1139–1160. [CrossRef]
29. *GB/T 3480-1997*; Calculation Methods of Load Capacity for Involute Cylindrical Gears. State Technology Supervision Administration, CNKI: Beijing, China, , 1997.
30. Hertz, H. Uber die beruhrung fester elastlseher korper. *Reine Angew. Math.* **1882**, *92*, 156–171. [CrossRef]
31. Zhang, Q.; Zhang, Z.; Jin, X.; Zeng, W. Entropy-based method for evaluating spatial distribution of form errors for precision assembly. *Precis. Eng.* **2019**, *60*, 374–382. [CrossRef]
32. Zhang, Q.; Xin, J.; Zhihua, L.; Zhang, Z. A new approach of surfaces registration considering form errors for precise assembly. *Assem. Autom.* **2020**, *40*, 789–800. [CrossRef]
33. Zhang, Q. Research on Assembly Accuracy Prediction for Precision Instruments. Ph.D. Thesis, School of Mechanical Engineering, Beijing Institute of Technology, Beijing, China, 2020.
34. Guo, H.; Zhang, Z.; Xiao, M.; Liu, H. Tolerance optimization method based on flatness error distribution. *Int. J. Adv. Manuf. Technol.* **2021**, *113*, 279–293. [CrossRef]
35. Ding, H.; Zhu, L.; Xiong, Z. Survey on coordinate measurement, geometric modeling and RP or NC generation from measured data points. *Chin. J. Mech. Eng.* **2003**, *39*, 10. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Design and Experimental Research of a Non-Destructive Detection Device for High-Precision Cylindrical Roller Dynamic Unbalance

Zhuangya Zhang, Baorun Yang, Mingde Duan, Ruijie Gu *, Shijie Liang and Yang Chen

School of Mechatronics Engineering, Henan University of Science and Technology, Luoyang 471003, China; zhangzhuangya@126.com (Z.Z.); 15036621751@163.com (B.Y.); duan_mingde@163.com (M.D.); 18437905090@163.com (S.L.); chen_yang12138@163.com (Y.C.)

* Correspondence: iapte@haust.edu.cn

Abstract: Due to their small size and light mass, small precision cylindrical rollers present challenges in dynamic unbalance detection, including difficulties in measurement and the risk of surface damage. This paper proposes a non-destructive detection device for assessing the dynamic unbalance of small precision cylindrical rollers. The device utilizes an air flotation support method combined with resonance amplification to indirectly measure the dynamic unbalance. A dynamic model of the air flotation tooling-cylindrical roller vibration system was developed to explore the relationship between the vibration parameters of the air flotation tooling and the dynamic unbalance of the cylindrical roller. Modal analysis and harmonic response analysis were performed, revealing that the amplitude of the vibration system at resonance could be detected using the sensor. Additionally, modal testing was conducted to determine the natural frequency of the system. A non-destructive detection platform was constructed for testing the dynamic unbalance of cylindrical rollers. Microscopic observation of the roller surface before and after testing confirmed that the device successfully performs non-destructive detection of dynamic unbalance.

Keywords: small precision cylindrical rollers; dynamic unbalance; vibration system; non-destructive detection

1. Introduction

Recent advancements in aerospace, high-speed rail, and other industries have heightened the focus on bearings, which are crucial components of mechanical systems [1–3]. The performance of rotating machinery is often contingent upon the quality of bearings, with cylindrical roller bearings being favored in industrial applications for their superior load-bearing capacity and longevity. As technology progresses, the demands on cylindrical roller bearings have escalated, particularly in high-precision equipment, where bearings must withstand high speeds and impacts while maintaining excellent motion characteristics and an extended service life [4,5]. Manufacturing and assembly processes can introduce variations such as material inconsistencies, shape asymmetries, and internal defects, leading to deviations between the geometric center of rotation and the center of gravity of cylindrical rollers, resulting in dynamic unbalance [6]. In high-speed, high-precision applications, this dynamic unbalance generates periodic forces on the inner and outer raceways and the cage of the bearing, which can lead to operational issues, noise, and other detrimental effects. In severe cases, it may cause damage to the bearing surfaces or lead to catastrophic failure of the cage, significantly impairing bearing performance and reducing its lifespan [7–9].

Currently, common dynamic balancing machines, both domestically and internationally, are primarily designed for large-volume and high-mass rotary bodies such as automobile wheel hubs, machine tool spindles, and motor rotors. In contrast, cylindrical

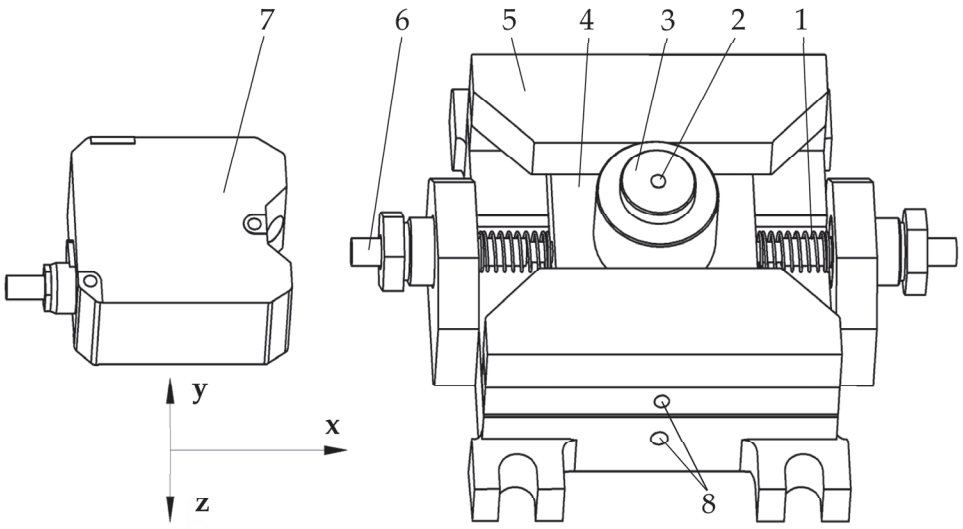
rollers, which have relatively smaller volume and mass, present challenges in dynamic unbalance detection. Issues such as detection difficulties and the risk of damaging the surface during detection are more prevalent with these smaller components. Thearle et al. [10] and Baker et al. [11] conducted studies on the detection of dynamic unbalance of rotors and proposed the influence coefficient method for the cases of double-sided, dual-sensors, single-speed, and known points. Assuming that the dynamic unbalance of the rotor is proportional to the vibration feedback it excites, the influence coefficient can be derived by measuring the vibration value at each test point and measuring the dynamic unbalance of the cylindrical roller [12,13]. Louis et al. [14] further modified the influence coefficient method, applied the four-run method to double-sided dynamic unbalance detection, and proposed a new method for detecting double-sided dynamic unbalance of rotor systems without phase response. Bishop and Gladwell [15] proposed the modal balance method, based on the classical vibration theory, which decomposes the dynamic unbalance of the rotor into many rinks of dynamic unbalance components according to the main vibration pattern [16,17]. Kellenberger [18] proposed the N + 2 plane method of modal balancing, in which N balancing planes are used to balance the deflection of the rotor, and two additional correction planes are required if the dynamic reaction force of the rotor is also required to be zero. Abbasi et al. [19] presented a hybrid algorithm integrating the salp swarm algorithm and Nelder–Mead algorithms for detecting unbalance magnitude and phase as the unbalance parameters. Gu et al. [20] of Chongqing University investigated in depth the detection device of the dynamic unbalance influence coefficient and the hardware structure and software construction of the detection system in the computer-aided dynamic unbalance detection machine through the built-in vibration sensor and photoelectric sensor of the dynamic unbalance detection machine. Cui et al. [21,22] and Sui et al. [23] from Henan University of Science and Technology designed a MEMS-based detection system for dynamic unbalance in small-mass cylindrical rollers. This system uses V-block support and a pulley drive to address issues associated with traditional detection systems, such as large swing frame mass, high cost, and low sensitivity.

Conventional dynamic unbalance detection methods predominantly rely on contact-based techniques, which often lead to issues such as interference from additional vibration signals and potential damage to the cylindrical roller's surface. This paper addresses these challenges by focusing on precision small-mass cylindrical rollers and proposes a non-destructive detection method specifically designed for small dynamic unbalances. A dynamic model of the air flotation tooling-cylindrical roller vibration system (referred to as the vibration system) is developed to explore the relationship between the air flotation tooling and the dynamic unbalance of the cylindrical roller. The high stiffness of the air film supports the cylindrical rollers without causing damage, allowing the centrifugal forces arising from dynamic unbalance to be transferred to the vibration system with minimal interference. By collecting and processing the displacement signals of the vibration system, accurate data on the dynamic unbalance of cylindrical rollers can be obtained. This approach effectively addresses the difficulties associated with detecting dynamic unbalance in small cylindrical rollers and offers a novel solution for the dynamic unbalance detection of other micro and small rotating components.

2. Dynamic Unbalance Non-Destructive Detection Device

2.1. Overall Structure of the Detection Device

The dynamic unbalance detection device is shown in Figure 1. The detection device consists of the displacement sensor, cylindrical roller, and air flotation tooling (including springs, air flotation sleeve, air flotation slide, air flotation guide, detection rod, air inlet interface, and other components).

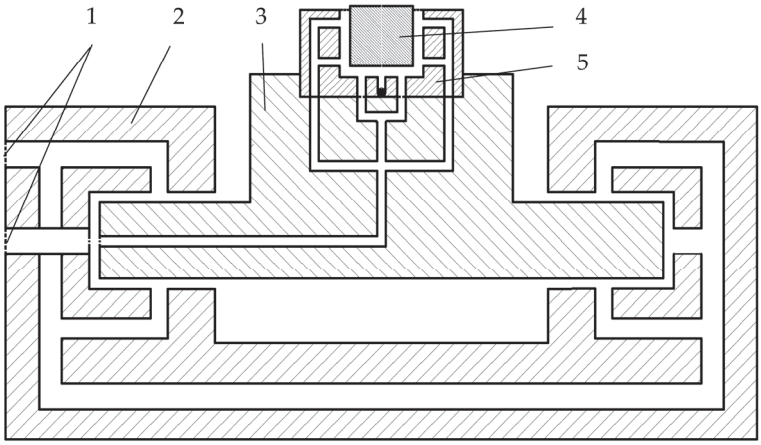


1-Spring; 2-Cylindrical roller; 3-Air flotation sleeve; 4-Air flotation slide; 5-Air flotation guide; 6-Detection rod; 7-Displacement sensor; 8-Air inlet interface.

Figure 1. Schematic diagram of the structure of the detection device.

The dynamic unbalance detection process is as follows: compressed air is passed into the air flotation tooling from the air inlet interface, so that the air flotation slide is suspended in the air flotation guide, and the cylindrical roller is suspended in the air flotation sleeve, as shown in Figure 2. The red represents the air supply system for the air flotation slide, while the green represents the air supply system for the air flotation sleeve. The cylindrical roller to be detected with high-speed rotation is put into the air flotation sleeve. Due to the high rigidity and low damping bearing characteristics of the air film, the cylindrical roller can keep the high-speed rotation for a long time [24]. The cylindrical roller with dynamic unbalance will generate a periodic centrifugal force when rotating at high speed, and this centrifugal force will drive the air flotation slide through the film [25] to vibrate under the support and guidance of the air flotation guide, and the amplitude will reach the maximum when the speed of the cylindrical roller (i.e., the centrifugal excitation frequency) is the same as the natural frequency of the air flotation slide. The displacement sensor collects the vibration signal of the end face of the detection rod, imports it into the information analysis system, carries out the subsequent data processing, and then obtains the dynamic unbalance of the cylindrical roller after the system is solved to complete the detection of the dynamic unbalance of the cylindrical roller.

The dynamic unbalance detection device proposed in this paper uses the amplification effect of resonance on amplitude to convert the small unbalance into an easily detectable mechanical vibration displacement. During the detection process, the air flotation sleeve provides non-contact air flotation support for the detected cylindrical rollers, which excludes the influence of the outside world on the detection of cylindrical roller dynamic unbalance, protects the surface of cylindrical rollers, and improves the detection accuracy [26]. By replacing different air flotation sleeves, the detection of different types of cylindrical rollers can be satisfied.



1-Air inlet interface; 2-Air flotation guide; 3-Air flotation slide;
4-Cylindrical roller; 5-Air flotation sleeve.

Figure 2. Schematic diagram of the principle of air flotation support.

2.2. Detection Principle

The vibration system comprises the detection rod, air flotation slide, air flotation sleeve, cylindrical roller, and spring. The air flotation guide constrains five degrees of freedom of the air flotation slide, thereby allowing it to move freely along the x-axis. As the cylindrical roller rotates within the air flotation sleeve, it induces motion in the vibration system along the x-direction. The mass of the vibrating system is collectively constituted by the detection rod, air flotation slide, air flotation sleeve, and cylindrical roller, with the spring contributing a small, negligible mass. The vibration system is abstracted and simplified to obtain a vibration system dynamics model that is convenient for calculation and analysis. The model can reflect the dynamics of the detection device and the dynamic response relationship between the detection device and the dynamic unbalance of the cylindrical roller to provide a theoretical basis for analyzing the motion characteristics of the vibration system.

Let the eccentric mass of the cylindrical roller be m. The mass eccentricity distance is e and the dynamic unbalance is defined as U. When the cylindrical roller rotates with angular velocity ω, the centrifugal force F generated by the eccentric mass is as follows:

$$F = U\omega^2 = me\omega^2 \tag{1}$$

The cyclic centrifugal force on the air flotation sleeve in the direction of vibration of the air flotation slide is as follows:

$$F(t) = F\cos(\omega t + \varphi) = me\omega^2\cos(\omega t + \varphi) \tag{2}$$

where φ is the initial phase of the force on the air flotation sleeve. The reference position for the phase is in the x-positive direction.

Figure 3 illustrates a simplified dynamic model of the vibrating system, with k_0 and c_0 denoting the stiffness and damping coefficients of the air film in the air flotation sleeve, and k and c representing the stiffness and damping coefficients of the vibrating system in the x-direction. Given that the air film stiffness significantly surpasses the spring stiffness, an approximation can be made that, under the influence of the excitation force, the air film behaves as purely rigid. The excitation force exclusively induces deformation in the spring, while the damping coefficient c_0 solely impacts the rotation of the cylindrical roller. Consequently, the vibrating system exhibits a stiffness coefficient along the x-direction, denoted as k, and a damping coefficient, denoted as c. Notably, the stiffness coefficient k can

be reasonably approximated as the superposition of the stiffness of two springs depicted in Figure 1.

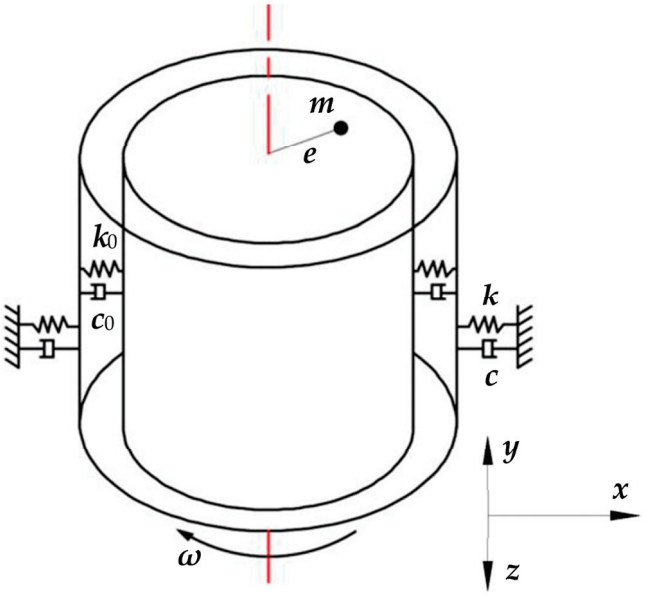


Figure 3. Simplified dynamic model of a vibration system.

Establishing a Cartesian coordinate system with the guide length direction designated as the x-direction, the air flotation guide constrains the air flotation slide in five degrees of freedom. When the cylindrical roller undergoes angular velocity ω, under the approximation that the vibration system can be treated as a single-degree-of-freedom system with damped forced vibration, the dynamic unbalance-induced excitation force compels the slide to undergo periodic vibrations along the x-direction. The resulting differential equation of motion is as follows:

$$\frac{d^2x}{dt^2} + 2\delta\frac{dx}{dt} + 4\pi^2 f_0^2 x = \frac{H}{M}\sin(2\pi f t) \tag{3}$$

$$\delta = \frac{c}{2M} \tag{4}$$

where c is the vibration system damping coefficient; f_0 is the vibration system's natural frequency; H is the excitation force amplitude;M is the vibration system mass; f is the excitation force frequency, and δ is the damping coefficient, the general solution of the differential equation is:

$$x = Ae^{-\delta t}\sin(\sqrt{(2\pi f_0)^2 - \delta^2 t} + \theta) + b\sin(2\pi f t - \varepsilon) \tag{5}$$

where A and θ are the integration constants; ε is the phase angle of the forced vibration behind the excitation force, determined by the initial conditions of the vibration.

The damped forced vibration consists of two parts: the first part is the free vibration whose frequency is the natural frequency, and the second part is the forced vibration whose frequency is the frequency of the excitation force. Since there is always damping in the actual vibration system, the free vibration part will decay down quickly, so we focus on the second part of the steady state forced vibration. The amplitude b is as follows:

$$b = \frac{H}{4\pi M\sqrt{\pi^2\left(f_0^2 - f^2\right)^2 + \delta^2 f^2}} \tag{6}$$

Because of the small cylindrical roller dynamic unbalance, the excitation force generated is also very weak, and the vibration system displacement information is difficult to accurately detect. As can be seen from Equation (6), when the frequency of the excitation force tends to the vibration system's natural frequency, the amplitude increases significantly; that is, $f \to f_0$, and the vibration system resonance is as follows:

$$f = \frac{1}{2\pi}\sqrt{4\pi^2 f_0^2 - 2\delta^2} \tag{7}$$

When the frequency of the excitation force satisfies Equation (7), the maximum amplitude $b_{\max}$ is obtained as follows:

$$b_{\max} = \frac{U\left(4\pi^2 f_0^2 - \frac{c^2}{2M^2}\right)}{c\sqrt{4\pi^2 f_0^2 - \frac{c^2}{M^2}}} \tag{8}$$

That is, the air flotation slide has the maximum amplitude at resonance, the use of resonance to amplify the vibration response can better extract the vibration signal of the cylindrical roller dynamic unbalance, and Equation (8) is converted to get the dynamic unbalance calculation formula as follows:

$$U = \frac{b_{\max} c\sqrt{4\pi^2 f_0^2 - \frac{c^2}{M^2}}}{4\pi^2 f_0^2 - \frac{c^2}{2M^2}} \tag{9}$$

Through the inverse of Equation (9), to find the dynamic unbalance you need to know the natural frequency of the vibration system. The vibration system is simplified to a simple spring-mass system, establishing a single-degree-of-freedom system vibration differential equation and solving the vibration system's natural frequency as follows:

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{k}{M}} \tag{10}$$

where k is the vibration system stiffness coefficient, which is approximated as the superposition of the two spring stiffness coefficients in Figure 1.

3. Vibration System Simulation Analysis

In order to obtain the natural frequency of the vibration system and the motion characteristics of the vibration system under the action of external excitation force, the harmonic response simulation analysis of the vibration system is carried out by the modal superposition method. The basic idea is to carry out the modal analysis first to obtain the natural frequency of the vibration system and then carry out the harmonic response analysis on the results of the analysis.

The 3D modeling software SolidWorks 2023 was used to build the 3D model of the air flotation tooling, and the 3D model was imported into the model module of Ansys Workbench to set the material parameters of the vibration system. The material is aluminum alloy, the modulus of elasticity is 70 GPa, Poisson's ratio is 0.33, and the density is 2770 kg/m^3. The 3D model is shown in Figure 4, after completion.

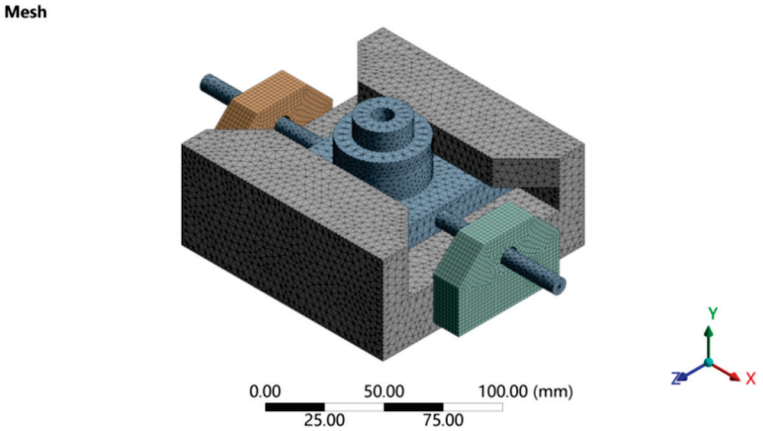


Figure 4. Mesh division model.

The boundary conditions are set for the finite element model after the mesh is divided. The frictionless contact between the air flotation slide and the air flotation guide is set (the air film is considered a frictionless rigid support for the air flotation slide), and the x-direction is the vibration direction of the air flotation slide, so the x-direction degrees of freedom are retained, and displacement constraints are applied to the y-direction and z-direction. The spring stiffness is set to 5 N/mm, and the harmonic response analysis uses a damping ratio model with a damping ratio of 0.02. After setting up, the solution is calculated to extract the first six mode shapes system. The first six orders are shown in Figure 5.

As shown in Figure 5, the first six orders of modes of the vibrating system are 19.039 Hz, 1091.4 Hz, 1096.7 Hz, 1138.3 Hz, 1142.0 Hz, and 6567.8 Hz. The analysis results show that the first-order mode of the vibration system vibrates along the x-direction with a frequency of about 19.039 Hz.

After the end of the modal analysis, a harmonic response analysis is performed, and the force load is added in the x-direction of the vibration system as the excitation force of the vibration system. The frequency of the excitation force is the rotational speed of the cylindrical roller, the magnitude of the excitation force simulates the centrifugal force generated by the cylindrical roller at different rotational speeds, and the numerical magnitude can be obtained from Equation (1). Set the frequency range of excitation force 0–50 Hz and execute the solution. After the solution is completed, the displacement values of the vibration system in the x-direction are extracted. The results are shown in Figure 6.

The harmonic response analysis employs a damping ratio model with a damping ratio of 0.02. This analysis aims to assess the feasibility of the proposed device rather than providing a precise quantitative evaluation, and thus, the results should be interpreted with this in mind. The excitation force induced by the rotation of the cylindrical roller causes the vibration system to oscillate. The frequency of this excitation force corresponds to the rotational speed of the cylindrical roller, and the magnitude of the excitation force is equivalent to the centrifugal force exerted by the cylindrical roller. The centrifugal force can be calculated using Equation (1), where $\omega = 2\pi f$, f represents the frequency of the excitation force, and me denotes the dynamic unbalance of the cylindrical roller. By substituting the rotational speeds of the cylindrical roller ranging from 0 to 50 Hz into Equation (1), the excitation force at each frequency of the vibration system can be determined. The harmonic response analysis was performed for each Hz increment, totaling 50 calculations.

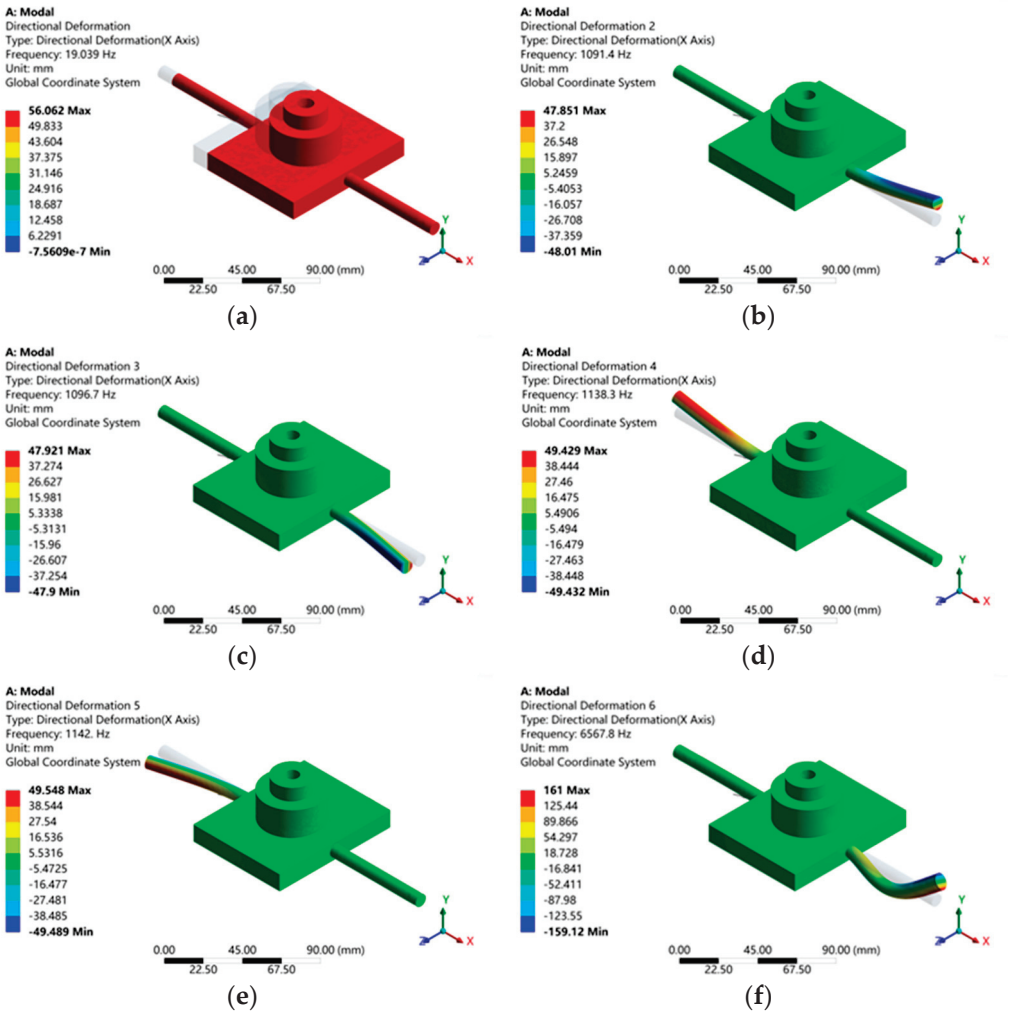


Figure 5. The first six vibration modes of the air flotation slide. (**a**) First−order vibration diagram. (**b**) Second−order vibration diagram. (**c**) Third−order vibration diagram. (**d**) Fourth−order vibration diagram. (**e**) Fifth−order vibration diagram. (**f**) Sixth−order vibration diagram.

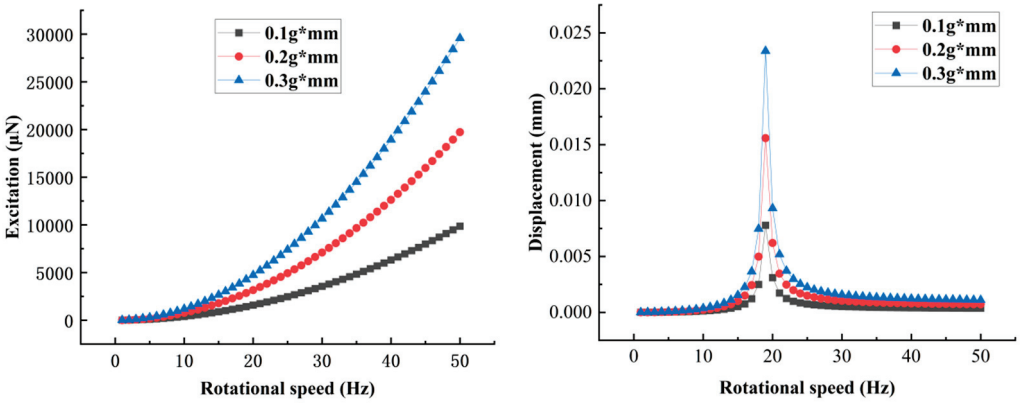


Figure 6. Harmonic response analysis.

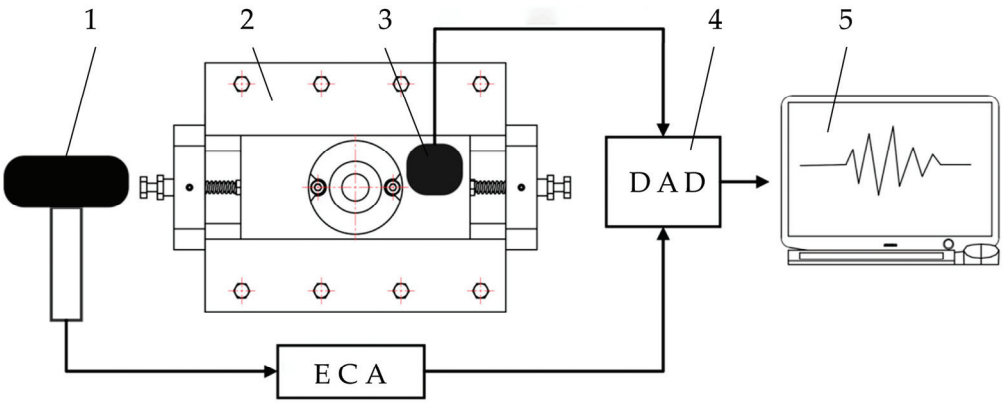
From Figure 6, it can be seen that the amplitude gradually increases in the range of 0–19 Hz, and when it reaches near 19 Hz (near the resonance point), the amplitude reaches the maximum, and the amplitude near the resonance point has a significant increase compared with the amplitude when it is not resonant. Thus, it can be theoretically concluded

that it is highly feasible to use the resonance amplification effect to detect the dynamic unbalance of cylindrical rollers. Without using the resonance amplification effect, the excitation force caused by the cylindrical roller dynamic unbalance is very weak, and it is difficult to make the air flotation slide produce significant displacement in the x-direction. Although this micro-displacement information can be collected by using ultra-high-precision displacement sensors, it is not feasible to improve the sensor accuracy in one go, considering the cost and external input interference.

4. Test

4.1. Vibration System Modal Test

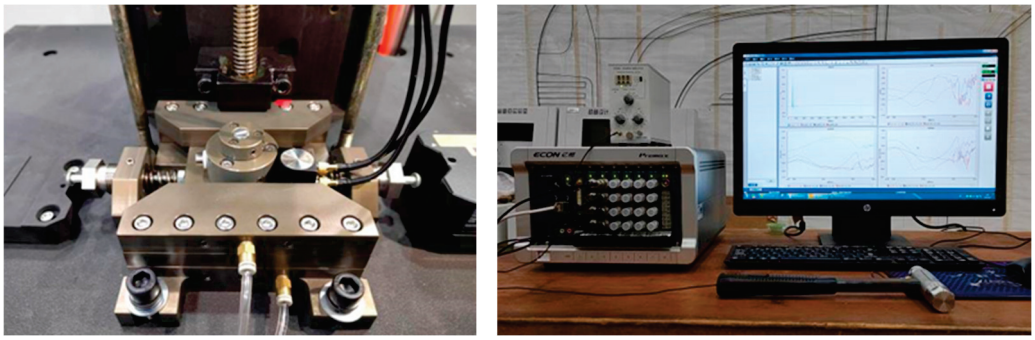
To obtain the actual natural frequency of the vibration system and verify the accuracy of the theoretical calculation and simulation analysis results, the modal test of the vibration system is required. The modal test system mainly consists of the force hammer, air flotation tooling, three-way acceleration sensor, data acquisition device, data analyzer, and electric charge amplifier (air supply conditions are consistent with Section 4.2). The block diagram of the testing system is shown in Figure 7.



1-Force hammer; 2-Air flotation tooling; 3-Three-way acceleration sensor; 4-Data acquisition device; 5-Air flotation sleeve.

Figure 7. Modal test system.

Test method: Since the structure of the vibration system is simple and the modal test only focuses on the vibration frequency of the air flotation slide along the x-direction, a single-point excitation and single-point vibration pick-up test can meet the needs. The three-way acceleration sensor (EA-YD-193) was used to collect the required test data, and the excitation method was chosen from the force hammer (LC-40A). The excitation point of the force hammer is the end face of the detection rod, and the direction is along the x-direction. The built vibration system dynamics test platform is shown in Figure 8.



(**a**) Vibration system prototype. (**b**) Test instruments.

Figure 8. Vibration system dynamics test platform.

The individual spring specifications of the vibration system under test were as follows: the wire diameter–outside diameter–length was 1 mm–12 mm–20 mm, the spring stiffness coefficient was 2.2525 N/mm, and the vibration system mass was 356 g. The modal data were collected several times using Text EngineX software, and the collected data were imported into Modal Genius software for analysis. The three-way acceleration sensor had a mass of about 18 g, which is much less than the mass of the vibration system (356 g). Its location is shown in Figure 7. The attachment of the three-way accelerometer to the air flotation slide is bound to introduce additional disturbances to the vibration system, but due to its small mass, the disturbances it introduces to the modal test can be acceptable. According to the analysis of the modal parameters by the analysis software, the natural frequency of the vibration system in the x-direction is 18.32 Hz.

4.2. Cylindrical Roller Dynamic Unbalance Detection

To verify the practicality of the cylindrical roller dynamic unbalance detection device, we built a non-destructive testing platform for cylindrical roller dynamic unbalance detection, which mainly consists of the cylindrical roller, driving device, air flotation tooling, laser displacement sensor (LK-H008, 0.0001 mm), and laser tachometer (UT372, 0.1 rpm) [27,28]. The laser displacement sensor mainly completes the vibration signal collection, and the laser tachometer is used to monitor the cylindrical roller speed in real time.

Before performing dynamic unbalance detection on cylindrical rollers, it is essential to calibrate the rollers for dynamic unbalance. Six cylindrical rollers, which are precision-machined with the same weight and dimensional specifications, were selected and labeled from 0 to 5. The end faces of rollers 1 through 5 were laser-etched to reduce their weight, and a specific gradient of dynamic unbalance was manually introduced. The roller labeled 0 was retained as a reference for comparison. Multiple measurements of the reduced weight and eccentricity were conducted to determine the calibration dynamic unbalance of the cylindrical rollers.

The de-weighting method is as follows: operate the CNC system of the laser etching machine for centering, find the center of the cylindrical roller as the coordinate origin, and then set the etching point position. The distance between the etching point position and the center of the circle is the eccentricity distance. The rollers after etching are shown in Figure 9.

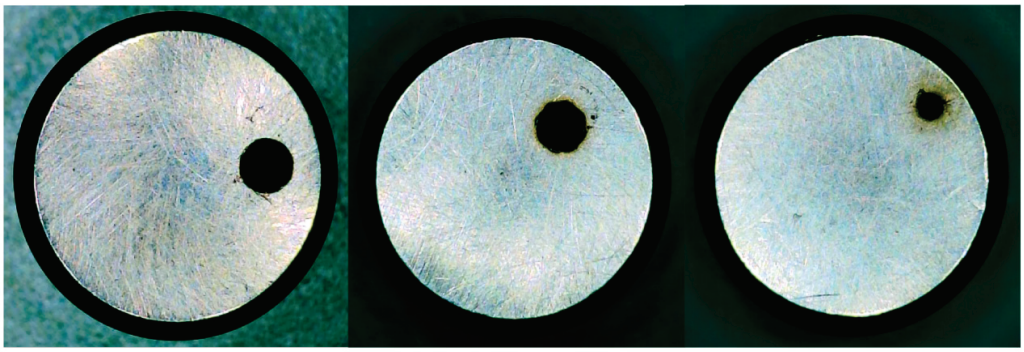


Figure 9. De-weighting diagram.

As shown in Table 1, the cylindrical roller dynamic unbalance mark was completed. In the cylindrical roller dynamic unbalance detection test, adjust the air flotation sleeve air supply pressure of 0.15 MPa, air flotation guide air supply pressure of 0.3 MPa, spring specifications, modal test with the same spring, and the laser displacement sensor installation to ensure that the sensor surface and the distance between the detection rod is 8 ± 0.5 mm. The dynamic unbalance of the cylindrical roller non-destructive detection test platform is shown in Figure 10.

Table 1. Cylindrical roller calibration dynamic unbalance.

Markings	De-Weight Mass/g	Eccentricity/mm	Dynamic Unbalance /g·mm
No. 0	0.0000	0.0000	0.00000
No. 1	0.0148	2.6813	0.03968
No. 2	0.0191	2.6920	0.05142
No. 3	0.0385	2.4647	0.09489
No. 4	0.0477	2.8076	0.13392
No. 5	0.0638	2.9043	0.18530

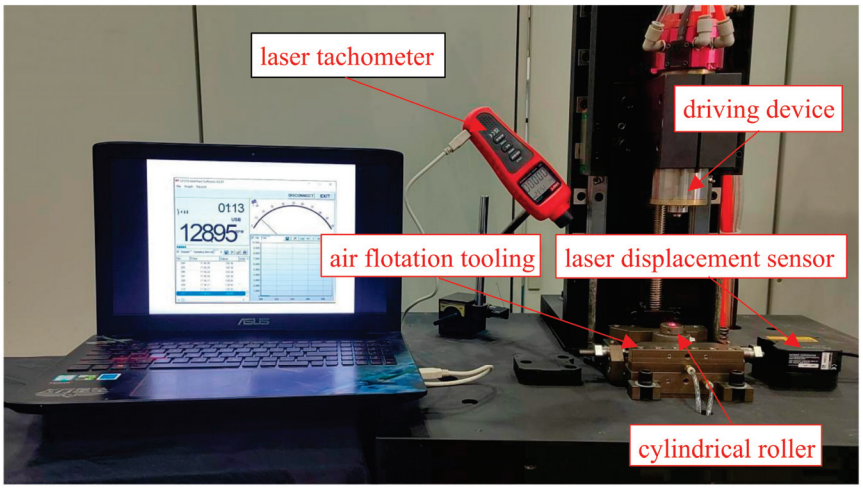


Figure 10. Non-destructive testing platform for cylindrical roller dynamic unbalance detection.

At the beginning of the test, the cylindrical roller to be detected is put into the air flotation sleeve, the cylindrical roller is kept in suspension in the air flotation sleeve, the driving device descends and drives the cylindrical roller to rotate at high speed when the rotational speed reaches 6000 rpm, the driving device is detached from the cylindrical roller, and the cylindrical roller relies on its inertia to rotate freely at high speed in the air flotation sleeve. The laser displacement sensor starts to collect vibration signals, the collected vibration signals are transmitted to the computer through the data acquisition device, and the collected signals are processed by the analysis software to get the cylindrical roller dynamic unbalance detection data. The displacement detection data for cylindrical rollers No. 2 and No. 0 are shown in Figure 11.

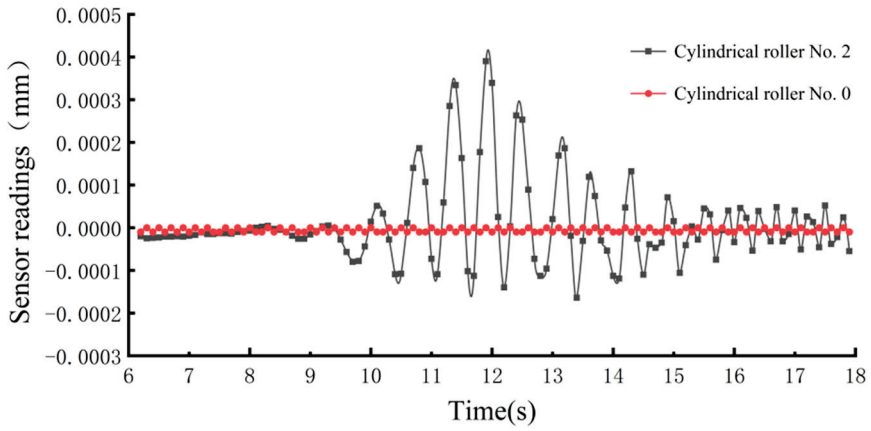


Figure 11. Displacement detection results.

The cylindrical roller in the air flotation sleeve speed is gradually reduced, and the excitation force frequency is also gradually reduced (to a rotational speed that is the frequency of the cylindrical roller excitation force) when the rotational speed is close to the vibration of the system's natural frequency of 18.32 Hz resonance. At this time, the rotational speed is 1150 rpm (corresponding to the frequency of the excitation force of 18.64 Hz). Repeat the test 10 times for cylindrical rollers Nos. 1–5 under the same test conditions respectively, extract the amplitude data, observe the amplitude deviation of cylindrical rollers under the same dynamic unbalance, and judge the stability of the test results. The curve is shown in Figure 12.

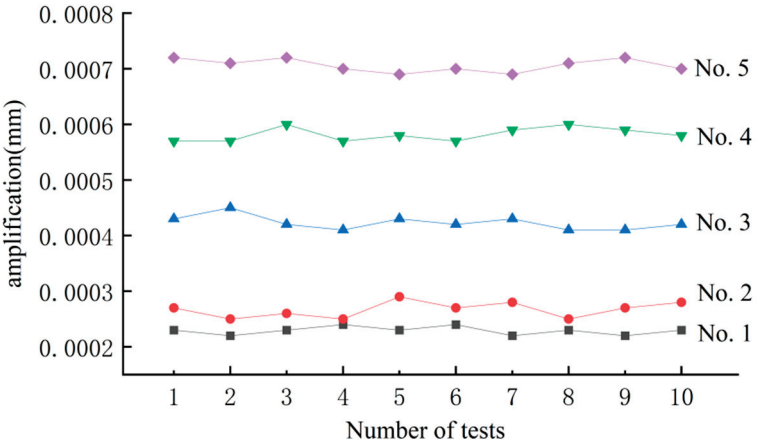


Figure 12. Repeatability test results.

As shown in Figure 12, the amplitude of cylindrical rollers with the same dynamic unbalance was basically the same under the test platform, with good repeatability, and with the increase of the dynamic unbalance, the amplitude detected by the sensor increases and also maintains a certain degree of repeatability, which shows that the test data have a certain degree of reliability.

Ten sets of amplitude values detected for each cylindrical roller were averaged to obtain the average value of the amplitude of the vibration system under different dynamic unbalance, as shown in Table 2.

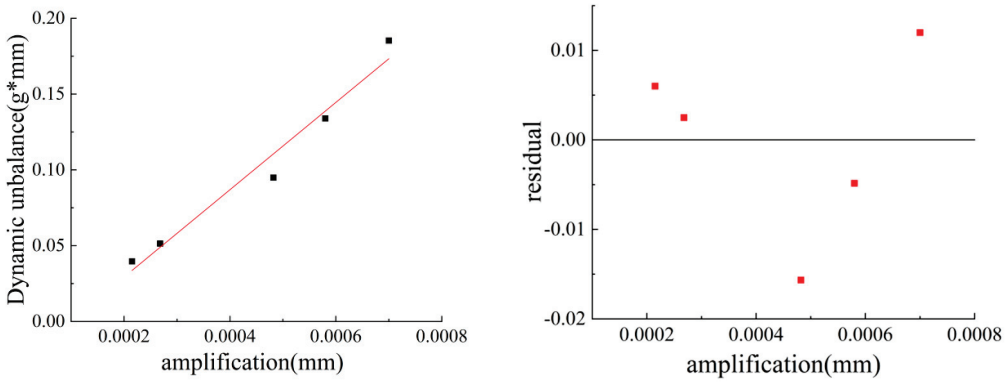
Table 2. The average values of the amplitude of the vibration system.

Cylindrical Roller Markings	No. 1	No. 2	No. 3	No. 4	No. 5
Average Amplitude/mm	0.000215	0.000268	0.000482	0.00058	0.0007

Substituting the cylindrical roller dynamic unbalance and the corresponding amplitude of Nos. 1–5 and the known parameters of the vibration system into Equation (9), the corresponding equations of cylindrical roller dynamic unbalance and amplitude are obtained approximately:

$$U = \frac{c\sqrt{212.819 - 0.127 \times c^2}}{1679.232 - 0.5c^2} b_{\max} \tag{11}$$

The damping coefficient c can be considered constant since the damping of the air film on the air flotation slide is extremely small and the spring undergoes only a uniform micro-deformation. From Equation (11), the function is a linear function, which is linearly fitted using Origin software 2023, and the fitting results are shown in Figure 13.



(**a**) Fitting curve. (**b**) residual.

Figure 13. Fitting results of the formula for the dynamic unbalance.

From Figure 13, the slope of the fitted function is 287.91323 ± 29.955, the intercept is 0.02823 ± 0.01453, and the variance of the fitted function is 0.96855, which is greater than 0.9, so the fit is better. The residual plot was obtained by making a scatter plot with the average amplitude of the vibration system as the independent variable and the residual as the dependent variable. The more random the distribution of the residuals, the more reliable the regression results obtained, and the fitted function is more reliable, as can be seen in Figure 13. Therefore, the fitted function can be considered reliable and the relationship between the amplitude and the dynamic unbalance under this vibration system is given in the following equation.

$$U = 287.91323 b_{\max} - 0.02823 \tag{12}$$

The average value of the measured amplitude is substituted into Equation (12) to calculate the dynamic unbalance, which is compared with the calibrated dynamic unbalance, and the result is shown in Figure 14.

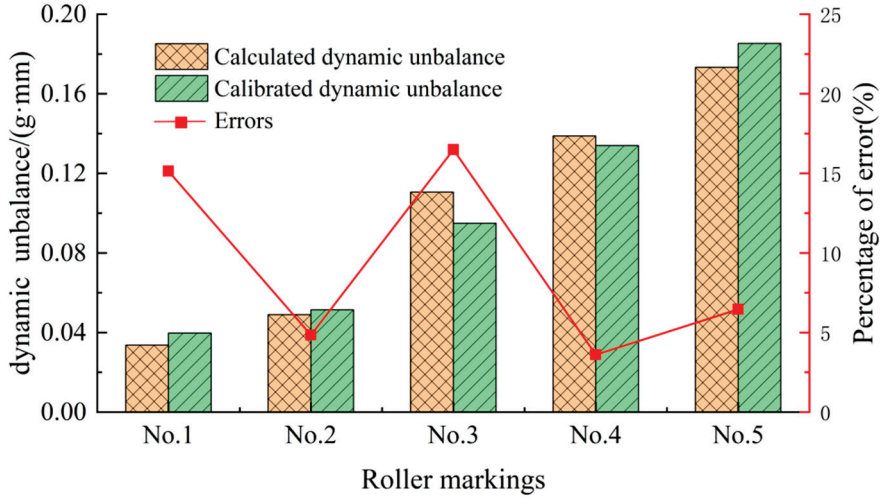


Figure 14. Error histogram.

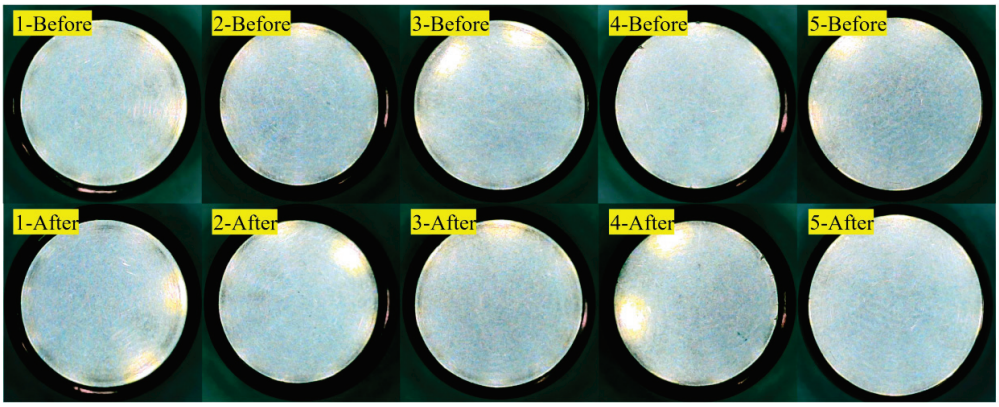
As can be seen from Figure 14, the calculated value of the dynamic unbalance is basically close to the calibrated value of the dynamic unbalance, with a maximum error of 16.3%, and the fitted function represents the relationship between the amplitude and the dynamic unbalance better. The error was analyzed for the following reasons: the calculation formula is a simplified ideal model, and the actual vibration system is different from the ideal model because of the structure; the external environment will have a certain impact on the measurement of the laser sensor during the measurement, and the resulting

amplitude is mixed with vibrations generated by other factors; the number of samples of the fitting function is not enough, and the fitted results will have a certain degree of error; in addition, the damping characteristics of the vibrating system are not very stable, which contributes to the fluctuations in the curves in Figure 12.

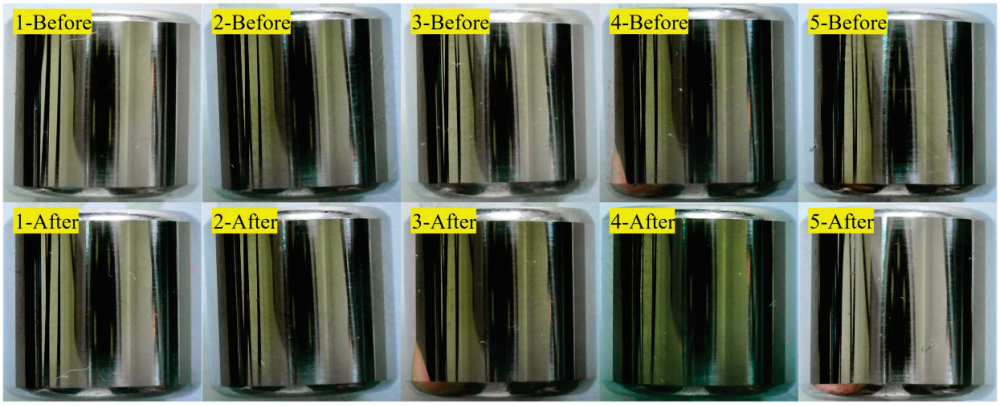
4.3. Cylindrical Roller Non-Destructive Detection Test Verification

Cylindrical rollers need to be put into use in the bearings after the dynamic unbalance detection, so there cannot be any damage to the surface of the cylindrical roller during the detection process, which needs to be calibrated to see if the detection device realizes non-destructive detection of the surface of the cylindrical roller.

All the cylindrical roller samples before and after the test of dynamic unbalance detection were put under the microscope to observe and record the surface quality of the cylindrical roller, and the comparison of the surface quality of the cylindrical roller before and after the detection is shown in Figure 15.



(**a**) The cylindrical roller end surface.



(**b**) The cylindrical roller cylindrical surface.

Figure 15. Comparison of surface quality of cylindrical roller before and after detection.

As can be seen from Figure 15, the end face and column surface of the cylindrical roller after the detection are not different from those before the detection, the cylindrical roller moves smoothly and noiselessly in the air flotation sleeve during the whole test process, and it is considered that the detection device can realize the non-destructive detection of the dynamic unbalance of the cylindrical roller.

5. Conclusions

The precision cylindrical roller's existing micro-dynamic unbalance problem is one of the key factors affecting the performance and service life of rolling bearings, and at present, the industry commonly uses means of dynamic unbalance detection, so there is a

large detection difficulty and detection of the surface damage and other problems. These problems seriously constrain the development of the non-destructive detection technology of dynamic unbalance. In order to solve these problems, this paper proposes proposed a new detection method, established a dynamic model of the vibration system, and carried out a modal analysis and harmonic response analysis of the vibration system. The analytical results show that the detection device can detect the micro-dynamic unbalance of the cylindrical roller.

To verify the feasibility of the proposed method, a non-destructive testing platform for cylindrical roller dynamic unbalance was constructed. The experimental results show that the cylindrical roller with dynamic unbalance has a significant amplitude value when its rotational speed is close to the natural frequency of the vibration system, which proves that the detection device can realize the detection of the dynamic unbalance of the cylindrical roller. Microscopic observation tests were carried out on the surface of the cylindrical roller before and after the detection, and the results showed that there was no obvious damage to the surface of the cylindrical roller, which proved that the detection device could realize non-destructive detection of the cylindrical roller's dynamic unbalance.

The non-destructive detection device of cylindrical roller dynamic unbalance proposed in this paper adopts a non-destructive air flotation bearing method, which reduces the damage to the surface of the cylindrical roller compared with the traditional destructive contact bearing method. In addition, due to the detection principle based on the resonance amplification effect, the detection accuracy of the device is greatly improved compared to the traditional "soft support" detection method. In the future, we will strive to optimize and improve the structure of the detection device, in order to significantly improve the detection stability of the device and reduce the manufacturing cost, as well as to meet the more demanding needs of the non-destructive detection of dynamic unbalance.

6. Patents

1. Duan, M.; Liang, S. A Kind of Cylindrical Roller Bearing Roller High-Speed Rotation Driving Device and Driving Method. CN115979513A, 18 April 2023.
2. Duan, M.; Liang, S. A Device and Method for Driving a Cylindrical Roller to Rotate at High Speed in an Air-Floating State. CN115950584A, 11 April 2023.
3. Duan, M.; Liang, S. A High-Speed Bearing Small Mass Cylindrical Roller Rotation Speed Obtaining Device. CN217059169U, 26 July 2022.

Author Contributions: Writing—original draft preparation, Z.Z. and B.Y.; writing—review and editing, B.Y. and Y.C.; data curation, S.L. and B.Y.; validation, M.D. and Z.Z.; conceptualization, R.G. and S.L.; methodology, Y.C.; software, S.L. and Y.C.; supervision, R.G. and M.D.; funding acquisition, R.G. project administration, Z.Z. and M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 52005158, and the Henan Province science and technology research project (Research on key technology of high-speed micro cylindrical roller dynamic unbalance non-destructive detection system), grant number 242102220078. The APC was funded by Henan University of Science and Technology.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. He, C.; Zhang, J.; Geng, K.; Wang, S.; Luo, M.; Zhang, X.; Ren, C. Advances in ultra-precision machining of bearing rolling elements. *Int. J. Adv. Manuf. Technol.* **2022**, *122*, 3493–3524. [CrossRef]

2. Xu, F.; Ding, N.; Li, N.; Liu, L.; Hou, N.; Xu, N.; Chen, X. A review of bearing failure Modes, mechanisms and causes. *Eng. Fail. Anal.* **2023**, *152*, 107518. [CrossRef]
3. Li, X.; Horie, M.; Kagawa, T. Study on the basic characteristics of a vortex bearing element. *Int. J. Adv. Manuf. Technol.* **2013**, *64*, 1–12. [CrossRef]
4. Wang, D.; Yuan, J.; Hu, L.; Lyu, B. Multidimensional study on the wear of high-speed, high-temperature, heavy-load bearings. *Materials* **2023**, *16*, 2714. [CrossRef]
5. Hou, X.; Diao, Q.; Liu, Y.; Liu, C.; Zhang, Z.; Tao, C. Failure Analysis of a Cylindrical Roller Bearing Caused by Excessive Tightening Axial Force. *Machines* **2022**, *10*, 322. [CrossRef]
6. Cui, Y.; Deng, S.; Niu, R.; Chen, G. Vibration effect analysis of roller dynamic unbalance on the cage of high-speed cylindrical roller bearing. *J. Sound Vib.* **2018**, *434*, 314–335. [CrossRef]
7. Tiwari, R.; Chakravarthy, V. Simultaneous estimation of the residual unbalance and bearing dynamic parameters from the experimental data in a rotor-bearing system. *Mech. Mach. Theory* **2009**, *44*, 792–812. [CrossRef]
8. Harsha, S.P. Nonlinear dynamic analysis of an unbalanced rotor supported by roller bearing. *Chaos Solitons Fractals* **2005**, *26*, 47–66. [CrossRef]
9. Wang, A.; Yao, W. Theoretical and numerical studies on simultaneous identification of rotor unbalance and sixteen dynamic coefficients of two bearings considering unbalance responses. *Int. J. Control Autom. Syst.* **2022**, *20*, 1971–2007. [CrossRef]
10. Therale, E.L. Dynamic balancing of rotating machinery in the field. *ASME J. Appl. Mech.* **1934**, *56*, 745–753. [CrossRef]
11. Baker, J.G. Methods of rotor-unbalance determination. *ASME J. Appl. Mech.* **1939**, *61*, A1–A6. [CrossRef]
12. Zhang, X.; Jiao, H.; Hu, D. Research progress on field dynamic balancing methods for rotating machinery. *J. Mech. Electr. Eng.* **2021**, *38*, 1367–1377. (In Chinese)
13. Zhang, L.; Duan, Z.; Li, D. Research progress of on-site dynamic balancing technology. *Chem. Eng. Mach.* **2012**, *39*, 690–694. (In Chinese)
14. Louis, J. Two-plane balancing of a rotor system without phase response measurements. *Trans. ASME J. Vib. Acoust. Stress Reliab. Des.* **1987**, *109*, 162–167.
15. Bishop, R.E.D.; Gladwell, G.M.L. The Vibration and Balancing of an Unbalanced Flexible Rotor. *J. Mech. Eng. Sci.* **1959**, *1*, 66–70. [CrossRef]
16. Zou, D.; Zhao, H.; Liu, G. Application of augmented Kalman filter to identify unbalance load of rotor-bearing system: Theory and experiment. *J. Sound Vib.* **2019**, *463*, 68–69. [CrossRef]
17. Parkinson, A.G. Balancing of rotating machinery. *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.* **1991**, *205*, 61–63. [CrossRef]
18. Kellenberger, W. Should a flexible rotor be balanced in N or (N+2) planes. *J. Eng. Ind.* **1972**, *94*, 548–560. [CrossRef]
19. Abbasi, A.; Firouzi, B.; Sendur, P. Identification of unbalance characteristics of rotating machinery using a novel optimization-based methodology. *Soft Comput.* **2022**, *26*, 4831–4862. [CrossRef]
20. Gu, Q.; Tang, Y. Research on computer-aided dynamic balancing machine measurement and analysis system. *Mech. Sci. Technol.* **1992**, *4*, 24–27.
21. Cui, Y.; Deng, S.; Yang, H.; Zhang, W.; Niu, R. Effect of cage dynamic unbalance on the cage's dynamic characteristics in high-speed cylindrical roller bearings. *Ind. Lubr. Tribol.* **2019**, *71*, 1125–1135. [CrossRef]
22. Cui, Y.; Deng, S.; Ni, Y.; Chen, G. Effect of roller dynamic unbalance on cage stress of high-speed cylindrical roller bearing. *Ind. Lubr. Tribol.* **2018**, *70*, 1580–1589. [CrossRef]
23. Sui, X.; Liu, C.; Li, J.; Xue, Y.; Yu, Y.; Cui, Y. Laser-based measurement for micro-unbalance of cylindrical rollers of the high-speed precision rolling bearings. *Clust. Comput.* **2019**, *22*, S9159–S9167. [CrossRef]
24. Liu, D.; Liu, Y.; Chen, S. *Hydrostatic Gas Lubrication*, 1st ed.; Harbin Institute of Technology: Harbin, China, 1990; pp. 51–121.
25. Chen, G.; Ge, Y.; Lu, Q.; Zhang, W.; Wang, S. Air film pressure field characteristics of aerostatic thrust bearing with orifice blockage. *Int. J. Adv. Manuf. Technol.* **2023**, *124*, 4317–4328. [CrossRef]
26. Liang, S.; Duan, M.; Zhang, Z. Development and Investigation of Non-Destructive Detection Drive Mechanism for Precision Type Cylindrical Roller Dynamic Unbalance. *Appl. Sci.* **2023**, *13*, 13266. [CrossRef]
27. Venturini, S.; Cavallaro, S.P.; Vigliani, A. Experimental Techniques for Flywheel Energy Storage System Self-discharge Characterisation. In Proceedings of the International Conference of IFToMM ITALY, Turin, Italy, 11–13 September 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 183–191.
28. Jin, X.; Liu, Y. Numerical and Experimental Analysis for the Dynamics of Flawed–Machining Rod–Disk Rotor with Inner Misalignment. *Machines* **2022**, *10*, 355. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Hybrid Brushless Wound-Rotor Synchronous Machine with Dual-Mode Operation for Washing Machine Applications

Sheeraz Ahmed [1], Qasim Ali [1], Ghulam Jawad Sirewal [2], Kapeel Kumar [1] and Gilsu Choi [2,*]

[1] Department of Electrical Engineering, Sukkur IBA University, Sukkur 65200, Pakistan; sheerazahmed.mef18@iba-suk.edu.pk (S.A.); qasim-ali@iba-suk.edu.pk (Q.A.); kapeelkumar.meres23@iba-suk.edu.pk (K.K.)

[2] Department of Electrical and Computer Engineering, Inha University, Incheon 22212, Republic of Korea; gjsirewal@inha.ac.kr

* Correspondence: gchoi@inha.ac.kr

Abstract: This paper proposes a hybrid brushless wound-rotor synchronous machine (HB-WRSM) with an outer rotor topology that can operate as a permanent magnet synchronous machine (PMSM), as well as an HB-WRSM. In the first part, the existing brushless wound-rotor synchronous machine (BL-WRSM) is modified into a hybrid model by introducing permanent magnets (PMs) in the rotor pole faces to improve the magnetic field strength and other performance variables of the machine. In the second part, a centrifugal switch is introduced, which can change the machine operation from HB-WRSM to PMSM. The proposed machine uses an inner stator, outer rotor model with 36 stator slots and 48 poles, making the stator winding a concentrated winding. The HB-WRSM is utilized for dual-speed applications such as washing machines that run at low speed (46 rpm) and high speed (1370 rpm). For high speed, to have a better efficiency and less torque ripple, the machine is switched to PMSM mode using a centrifugal switch. The results are compared with the existing BL-WRSM. A 2D model is simulated using ANSYS Electromagnetics Suite to validate the machine model and performance.

Keywords: brushless operation; finite element analysis (FEA); hybrid excitation harmonic winding; outer rotor wound-rotor synchronous machine; washing machine

1. Introduction

Today's electrical machines are extremely diverse in type and application, ranging from direct-current machines of a few horsepower used in portable devices and household appliances to MW-scale AC synchronous generators used in power plants that convert natural potential energy into electrical energy. Permanent magnet (PM) machines have been widely used in a variety of applications requiring a high power output performance due to their high efficiency, excellent power factor, and high power density [1]. However, recent challenges such as a high cost, scarcity, and risk of demagnetization have led researchers to explore alternatives that use fewer or no PMs. The magnets used in the PM machines contain rare earth materials (REMs) such as neodymium (Nd) and dysprosium (Dy), which are increasingly vulnerable to supply risks and pose environmental concerns in their production [2,3]. PM synchronous machines (PMSMs) face not only demagnetization issues but also fixed magnetic fields.

An alternative to PM machines is a wound-rotor synchronous machine (WRSM), which is a more cost-effective solution with a controllable rotor flux capability. WRSMs feature rotors that contain copper windings as field windings, which require a separate

power supply. Traditional WRSMs typically use slip rings and brushes to energize the rotor main field windings. However, this approach results in power losses in the form of heat and wasted energy, which can reduce the overall efficiency of the machine. In addition, maintenance and sparking issues make the use of slip rings and brushes undesirable. Despite these drawbacks, WRSMs offer several advantages, including the elimination of expensive PMs, simple variable speed control, and the ability to adjust the field current. These benefits allow for a higher efficiency under light-load conditions for variable-speed operations. Unlike WRSMs that use brushes to transfer power to the rotor, brushless WRSMs (BL-WRSMs) eliminate the need for brushes, reducing maintenance needs and the risk of mechanical failure. The transition to brushless designs addresses the long-standing need to eliminate brushes and slip rings from the system. In addition, the use of concentrated windings in BL-WRSMs improves harmonic component control, resulting in better performance and energy efficiency.

Previous studies have explored various machine topologies to enable a brushless operation through alternative stator winding configurations, inverter arrangements, and power control strategies. Each approach offers distinct advantages and trade-offs in terms of system performance, complexity, and cost. The impact of different stator winding designs on machine behavior has been extensively analyzed [4–8], with particular focus on specialized configurations such as open-end and semi-open stator windings [9,10]. These topologies, especially when driven by coils operating at different frequencies, can generate magnetic motive force (MMF) harmonics that support a brushless operation. However, they often require additional stator coils and inverters, leading to increased system complexity and costs.

Alternatively, the use of single-, dual-, or multi-inverter systems has also been investigated to improve the performance and flexibility of brushless machines [10–14]. Some researchers have proposed modified inverters that operate at different frequencies to optimize system performance [15,16]. These inverters power harmonic windings and can be operated independently or in combination. This approach improves operational flexibility but increases complexity, weight, and cost due to the need for additional inverters. Other topologies use thyristor switches, single-phase inverters, and rectifiers to regulate the power injection for brushless operation [12,17,18]. Thyristor switches are used to control the injected supply power for brushless operation, and rectifiers can also be used for the same purpose. However, these components introduce increased power losses, system complexity, and costs, making them difficult to implement in cost-sensitive applications.

This work examines the design, analysis, and performance of a hybrid brushless WRSM (HB-WRSM). This study seeks to offer a comprehensive understanding of the present advancements in hybrid brushless machines and their potential for future growth and utilization in industries such as renewable energy generation, electric vehicles, and industrial automation [19,20]. The HB-WRSM provides several advantages over conventional electric machines, such as a higher torque density, improved efficiency, and reduced maintenance requirements. These advantages make it an ideal choice for a variety of applications, including electric vehicles and industrial automation systems. The proposed machine incorporates PMs on the rotor pole faces of a conventional BL-WRSM. By embedding magnets in the outer rotor, the motor can operate at higher speeds, which broadens its potential range of applications and solves the efficiency issues commonly observed in BL-WRSMs [19]. This HB-WRSM configuration is particularly suitable for applications that require different performance characteristics depending on the mode of operation, such as washing machines (e.g., wash and spin modes). The washing machine controller is assumed to limit the operating current and protect against short circuits that could completely demagnetize the magnet [21]. The performance of the proposed HB-WRSM is evaluated by comparing it with a BL-WRSM and conventional WRSM (C-WRSM) [22].

2. The Latest Brushless Technology and Its Applications

WRSMs are a type of synchronous machine in which the rotor is equipped with field windings supplied through brushes and slip rings. WRSMs have several advantages over other types of electric machines. Notably, they do not require expensive permanent magnets, which reduces the overall system cost and makes them suitable for a wide range of applications. In addition, the controllability of the rotor excitation current allows for flexible speed operation and improved efficiency under various load conditions from medium to high speeds. Despite these advantages, WRSMs face several limitations. The biggest of these is their reliance on brushes and slip rings, which introduces mechanical complexity, increases maintenance requirements, and leads to additional power losses and a lower overall efficiency.

2.1. The Latest Brushless Technology

In this section, we first provide a detailed analysis of the characteristics of HB-WRSMs based on a comprehensive review of the literature on various brushless techniques explored in previous research and commercial applications since 2015.

In [19], the authors proposed a brushless solution based on a modified inverter design that utilizes time harmonics in contrast to the conventional approach. This approach allows the armature winding to supply both fundamental and third harmonic current components. However, it requires a complex inverter design and control technique. In [15], the author proposed a brushless technique that incorporates additional thyristor switches connected in parallel with the three-phase armature winding. These switches are activated at the zero-crossing points of each phase current, generating a third harmonic component.

In [23], the three-phase supply currents to the stator windings are phase-shifted by 180 degrees. By continuously switching these currents through a thyristor during each electrical period, a third harmonic component is generated. This harmonic component is subsequently rectified using a diode, as illustrated in Figure 1a. In [24], a novel method to switch between C-WRSM and BL-WRSM is implemented using a thyristor drive circuit; however, this increases the cost and volume of the drive system, as shown in Figure 1b.

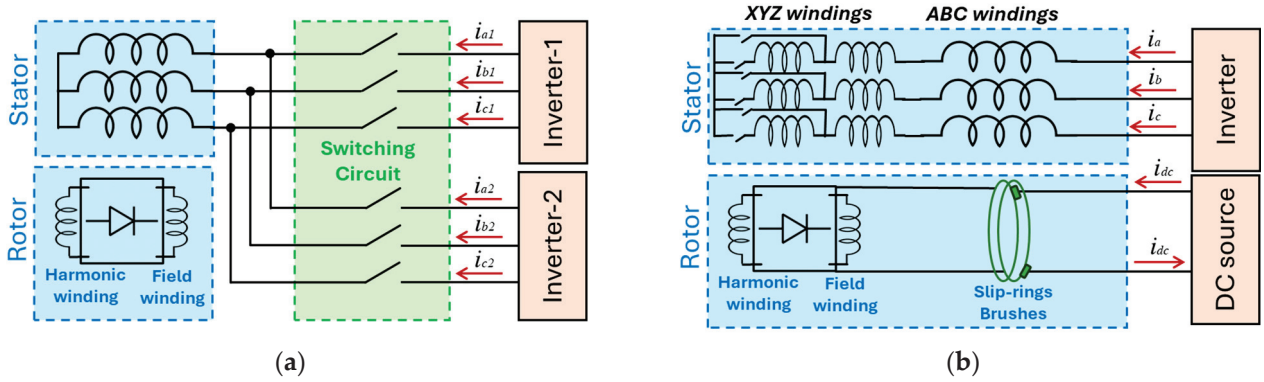


Figure 1. Dual inverter topology: (**a**) schematic diagram; (**b**) dual-mode operation of BL-WRSM.

In [25], a BL-WRSM topology was proposed utilizing a three-phase rectifier. The stator winding receives a standard three-phase current from the inverter, and the rectifier links the output to the neutral of the Y-connected armature winding. This design generates an additional third harmonic MMF component in the airgap, which is then used to induce an harmonic voltage and current in the harmonic winding. This harmonic current, in turn, drives the rotor field winding, allowing brushless operation, as shown in Figure 2.

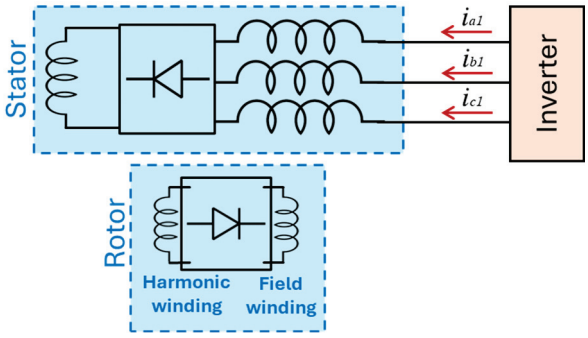


Figure 2. Uncontrolled three-phase rectifier-based BL-WRSM.

The literature review suggests that the WRSM is a strong candidate to replace PMSMs because it provides comparable results on key performance metrics in comparisons between the two machines. In [13], two inverters were used to supply three-phase currents to two separate sets of stator windings. Based on this topology, a hybrid variant that incorporates PMs to improve high-performance performance was introduced in [16]. This configuration received some attention, leading to a new stator topology in [26], where two sets of stator windings with different numbers of series turns were employed. This topology, however, relied on a single inverter to supply a three-phase current to the stator windings, which increased the control complexity and copper losses and consequently reduced the overall efficiency compared to a C-WRSM. Shortly after, another hybrid variant with consequent magnets on the rotor poles was introduced in [10], improving machine performance. Further optimization of this machine topology was carried out using a genetic algorithm in [27].

2.2. Dual Speed Applications

The ever-increasing need for electrical energy, along with environmental concerns, has raised the demand for variable-speed drives in low-cost, high-volume applications such as electric fans, washing machines, and home appliances. Currently, many of these applications use simple constant-speed drives, resulting in a degraded efficiency. They can benefit from variable-speed operation, which offers significant energy savings, increased dependability, and provides better process control. The main reason variable-speed drives have not penetrated such cost-sensitive applications is that they are more expensive than fixed-speed drives [28].

Washing operations can be separated into two major cycles: the wash cycle and the spin-dry cycle. During the washing cycle, the drum rotates at a relatively low speed, between 50 and 60 rpm, allowing clothes to tumble for effective cleaning. In contrast, the spin-dry cycle involves removing residual water from the laundry by utilizing centrifugal forces, achieved by drying the drum at higher speeds, typically from 800 to 1600 rpm [29]. The torque requirements differ significantly between the two cycles: the wash cycle demands a torque of approx. 15–20 Nm, while the spin-dry cycle requires a lower torque of around 3–5 Nm, as illustrated in Figure 3 [30].

Conventional washing machines were driven using belts, which generated significant losses and noise. The first direct-driven motor for washing machines was developed by Fisher and Paykel Appliances Ltd., Auckland, New Zealand [31]. The commercial model released in 2004 was a 36-slot, 48-pole PM machine operating at 1100 rpm [31]. Since then, the external rotor 36-slot, 48-pole configuration has been used by various manufacturers [32]. Figure 4a–d show different brands employing the 36/48 topology with segmented rotor magnets.

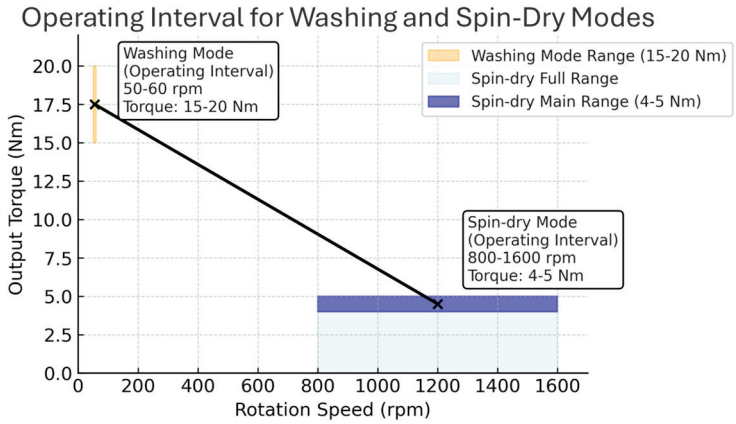


Figure 3. Typical washing machine requirements.

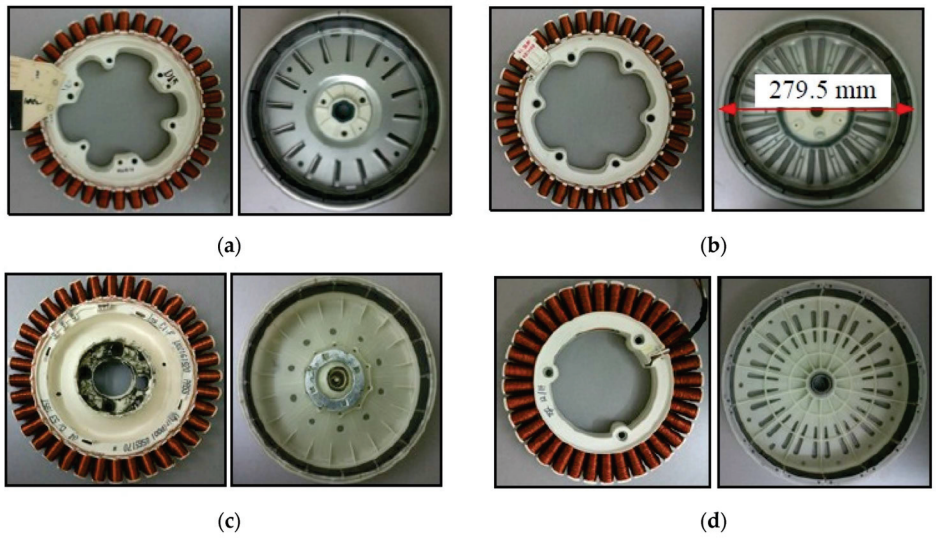


Figure 4. Stator and rotor structure of 4 commercial washing machines: (**a**) Samsung 36-slot/48-pole external rotor model; (**b**) LG 36-slot/48-pole external rotor model; (**c**) Whirlpool 36-slot/48-pole external rotor model; (**d**) Whirlpool Maytag 36-slot/48-pole external rotor model.

3. Simulation Results

The proposed machine uses an outer rotor, inner stator design with a novel winding topology. The 2D diagrams of the baseline machines are shown in Figure 5. In the proposed HB-WRSM design, the stator is equipped with three-phase windings placed in 36 slots, and it features 48 poles with PMs on the rotor pole faces, as shown in Figure 5c.

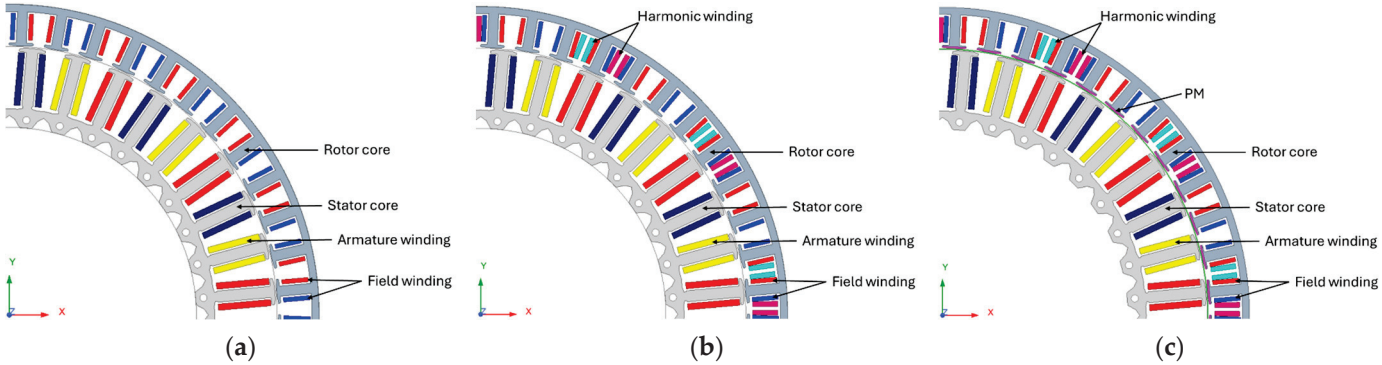


Figure 5. Structure of the 3 baseline machines: (**a**) C-WRSM; (**b**) BL-WRSM; (**c**) HB-WRSM.

For brushless operation, the stator must be equipped with specialized windings that produce MMF to excite the rotor's main field winding. Thus, the rotor is designed with two sets of windings: one referred to as the harmonic or excitation winding, and the other as the main field winding. Although various techniques for achieving brushless operation have been proposed, as in the literature review, the brushless operation technique used in this study is shown in Figure 6 to provide better clarity for the reader.

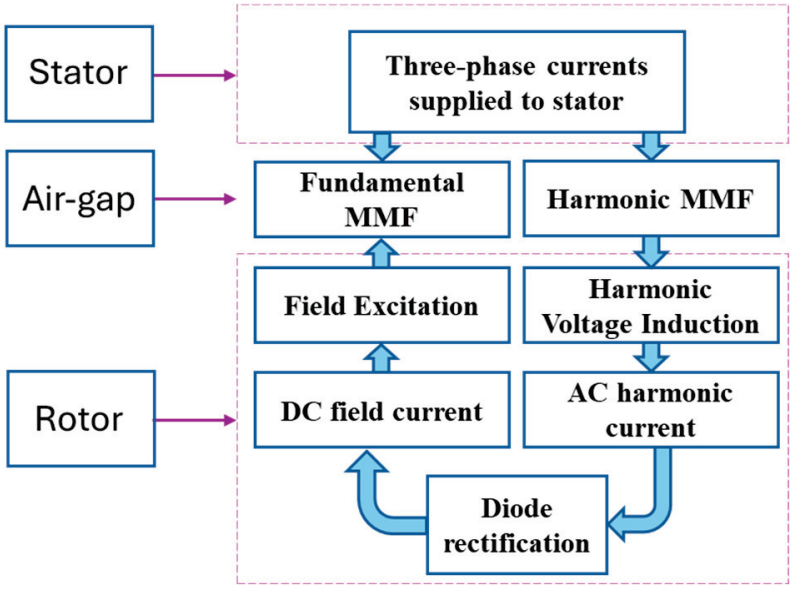


Figure 6. Working principle of BL-WRSM.

To achieve brushless operation in the proposed topology, the stator three-phase winding is fashioned to be a double-layer concentrated winding, as shown in the winding configuration diagram in Figure 7. The number of poles in the field winding should match that of the stator's fundamental MMF component, which is 48 poles. With a 48-pole fundamental component, the excitation winding needs to have 24 poles, as illustrated in Figure 5. The rotor excitation winding pole pitch is two times that of the field winding as it intercepts the 0.5th harmonic airgap magnetic field.

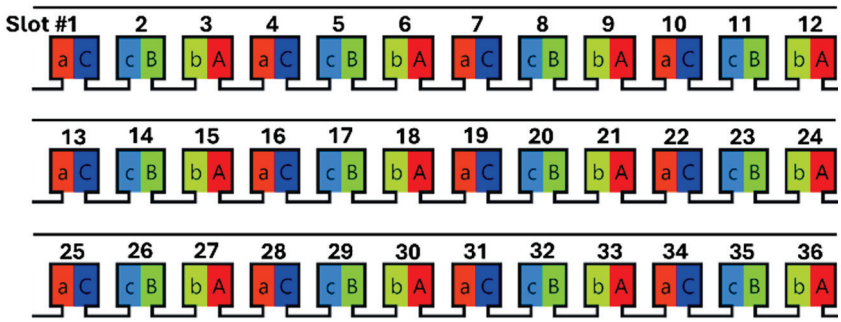


Figure 7. Stator winding configuration.

For brushless operation, the stator MMF must generate two sets of dominant frequency components: a fundamental component with an harmonic order of 1 and a sub-harmonic component with an harmonic order of 0.5. As shown in Figure 6, the stator windings of the proposed machine are capable of generating this sub-harmonic component without the need for additional stator windings and/or additional inverters. As evidenced in the literature review, some papers have used additional inverters and switching devices to

generate the desired harmonic components to achieve brushless operation, whereas this paper uses only one inverter and one set of concentrated three-phase stator windings.

The airgap harmonic field spectrum of a machine can be derived from its winding function, which represents the spatial distribution of conductors relative to the pole pitch. This function captures the collective contribution of all conductors across the machine's airgap. By applying a Discrete Fourier Transform (DFT), typically implemented via the Fast Fourier Transform (FFT) algorithm, the winding distribution can be decomposed into its harmonic components, as illustrated in Figure 8 [33]. Figure 8 shows that the 0.5th-order sub-harmonic component of the proposed machine is higher than the fundamental (1st-order) component. The rotational speed of the sub-harmonic component in the air gap is given as below:

$$n_{s(h)} = \frac{n_s}{h} = \frac{120f}{h * p} \tag{1}$$

where $n_{s(h)}$ represents the rotating speed of the harmonic component, n_s is the fundamental synchronous speed, h is the harmonic number, f is the supplied frequency, and p is the number of poles. Using (1), the synchronous speeds of both the fundamental and sub-harmonic components are shown in Table 1. The fundamental synchronous speed will be 150 rpm when using the 48 poles in the proposed model at a fundamental frequency of 60 Hz. Observing the 0.5th-order sub-harmonic component in Figure 8, the rotating speed of the sub-harmonic component is 300 rpm, which is twice the speed of the fundamental component.

Table 1. MMF components and their rotating speeds.

Fundamental Frequency	Harmonic Order	Number of Poles	Synchronous Speed
60 Hz	0.5	24	150 r/min
60 Hz	1	48	300 r/min

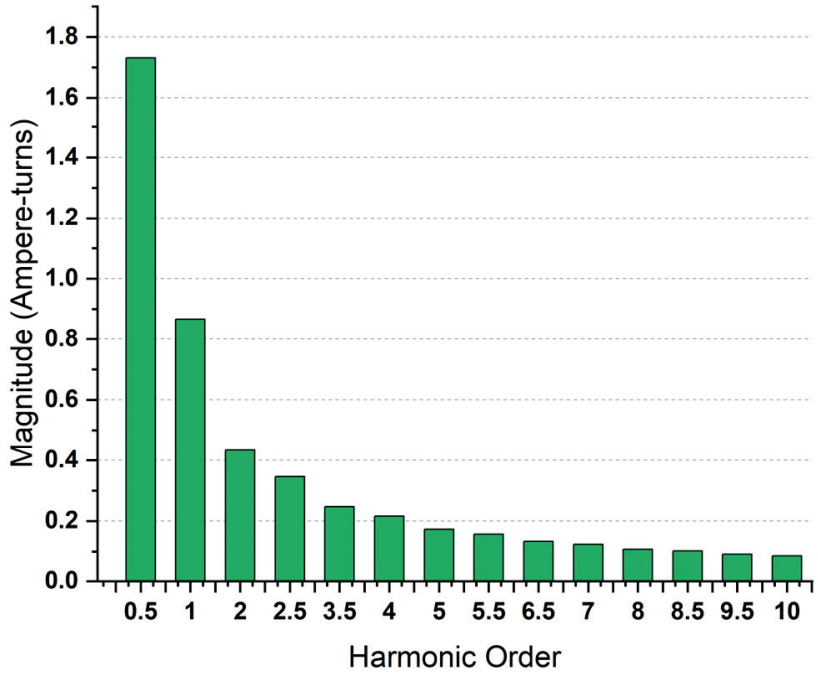


Figure 8. Three-phase MMF harmonic spectrum.

The fundamental MMF component seen by the rotor manifests as a DC component, while the sub-harmonic MMF appears as an AC component. To utilize this time-varying

magnetic flux and generate an induced voltage, the rotor is equipped with an additional winding for field excitation. The voltage induced in the excitation winding is rectified via a rotating rectifier mounted on the rotor, producing a DC current that energizes the rotor's field windings. A diagram of the rotor circuit, showing both the main field winding and the harmonic/excitation winding, is presented in Figure 9.

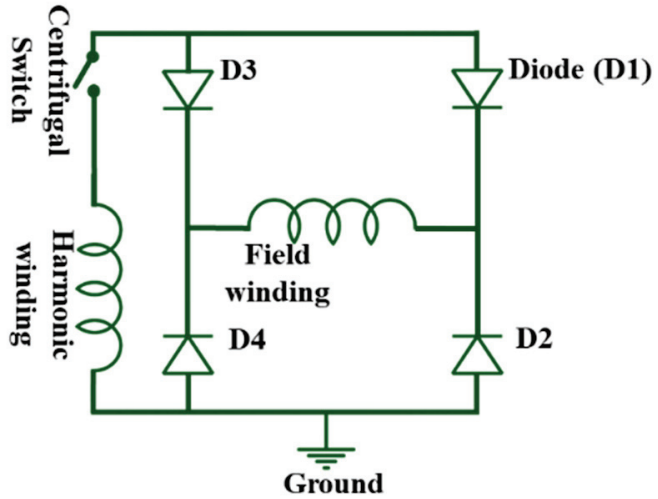


Figure 9. HB-WRSM rotor circuit.

The design parameters of the proposed HB-WRSM topology and the C-WRSM and BL-WRSM are summarized in Table 2. As mentioned earlier, the proposed machine is used in a dual-speed washing machine, which initially operates in the washing mode at 46 rpm (18.4 Hz) and then switches to the drying mode at 1370 rpm (548 Hz). Therefore, the torque output and field current generation are different in both modes, which will be discussed in later sections.

Table 2. Key parameters of the baseline machines.

Parameter	Unit	C-WRSM	BL-WRSM	HB-WRSM
		Value		
Number of slots	-	36	36	36
Field winding poles	-	48	48	48
Harmonic winding poles	-	-	24	24
Stator OD/ID	mm	265/186	265/186	265/186
Rotor OD/ID	mm	308/267	308/267	308/267
Air-gap length	mm	1	1	1
Axial length	mm	24	24	24
Washing mode speed	r/min	46	46	46
Drying mode speed	r/min	1370	1370	1370
Frequency in washing mode	Hz	18.4	18.4	18.4
Frequency in drying mode	Hz	548	548	548
Number of magnets	-	-	-	48
Magnet material	-	-	-	NdFe30

4. Machine Performance Analysis

To evaluate the performance of the proposed HB-WRSM, 2D FEA simulations were conducted using ANSYS Maxwell software (Ver. 2022 R2). The simulations were performed with a time step of 1.2 ms, and an automated mesh with a maximum element length of 6 mm was applied. The output torque and power losses were obtained directly from the simulations, while the input power and efficiency were calculated analytically via post-processing. The main performance metrics are the output torque and current generated in the rotor field winding. The performance of two existing topologies (i.e., C-WRSM and BL-WRSM) has been simulated and presented in [34], and the results are compared with the simulation results of the HB-WRSM proposed in this study. For fair comparison, the input power and all other machine parameters have been kept the same for both the C-WRSM and BL-WRSM.

4.1. No-Load Performance

Figure 10a,b show the no-load voltage waveforms simulated at 46 rpm for the C-WRSM and the HB-WRSM. As for field winding excitation, a field current of 5.83 A is required to produce a no-load voltage of approximately 10.2 Vrms. In the case of the HB-WRSM with PM excitation only, the same voltage level is achieved, as shown in Figure 10b.

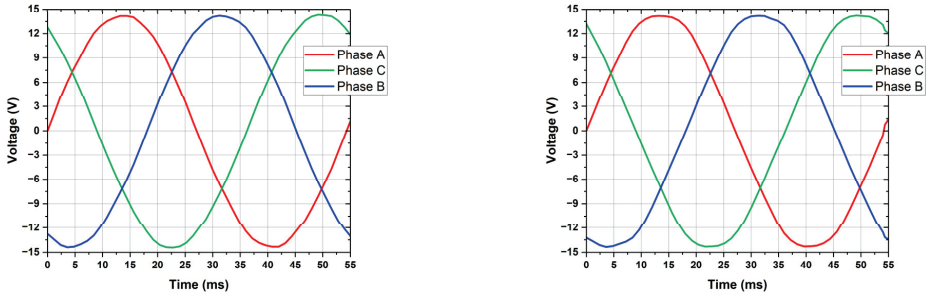


Figure 10. No-load back-emf voltage waveforms: (**a**) C-WRSM; (**b**) HB-WRSM with PM excitation only.

4.2. Washing Mode Performance

Figure 11a–c present the simulated torque waveforms for the C-WRSM, BL-WRSM, and HB-WRSM, respectively. In the C-WRSM, the torque waveform exhibits a reduced torque ripple due to the minimal impact of the tooth harmonic fields. In contrast, the torque waveform of the BL-WRSM shows a significant ripple, largely attributed to the current ripple in the rectified field current supplied by the harmonic winding. The field winding currents generated in the HB-WRSM, as shown in Figure 11d, contribute to additional losses in both the windings and the machine's core.

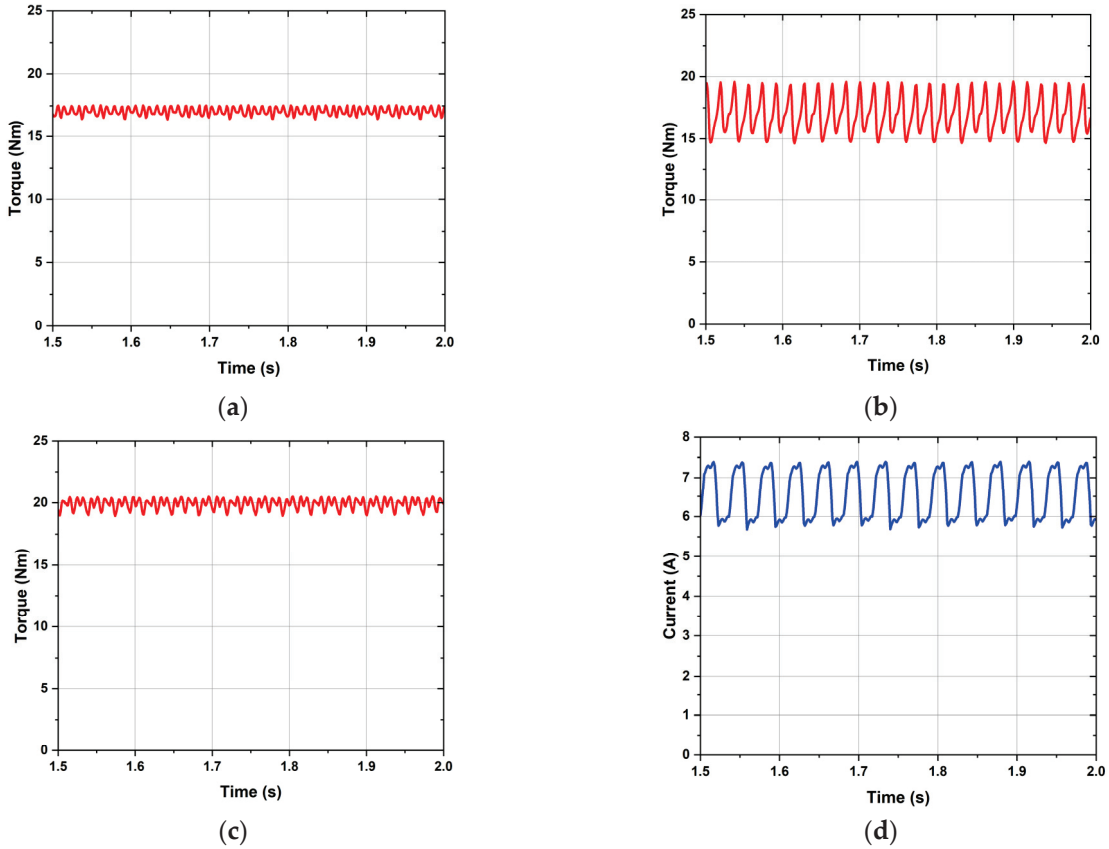


Figure 11. Torque waveforms: (**a**) C-WRSM in washing mode; (**b**) BL-WRSM in washing mode; (**c**) HB-WRSM in washing mode; (**d**) field current waveform of HB-WRSM.

To mitigate torque ripple and losses in the BL-WRSM, magnets are added to the rotor pole faces. This modification reduces the dependence on harmonic excitation, resulting in a lower torque ripple and reduced copper losses during the washing mode. The efficiency of the HB-WRSM during the washing mode increases from 35.35% to 50.02%, and the average torque reaches the highest value of the three designs.

Figure 12a shows that the flux density of the HB-WRSM in wash mode remains below 1.5 T, which is an acceptable level for most magnetic core regions. Figure 12b shows the mesh plot of the HB-WRSM. The mesh settings are the same for the other models.

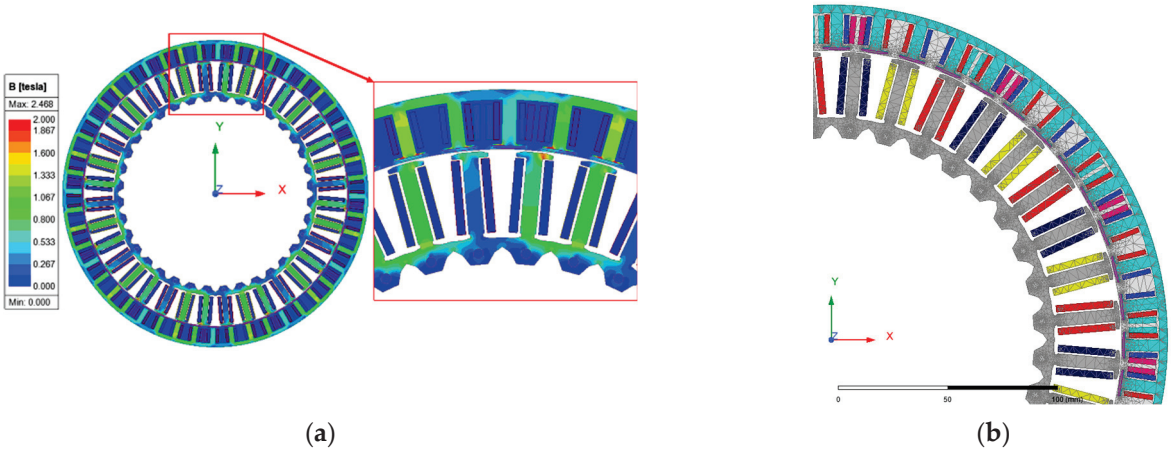


Figure 12. HB-WRSM (**a**) flux density contour in washing mode, (**b**) mesh plot.

4.3. Dry-Mode Performance

Figure 13a–c show the simulation torque waveforms in drying mode at 1370 rpm for the C-WRSM, BL-WRSM, and HB-WRSM, respectively.

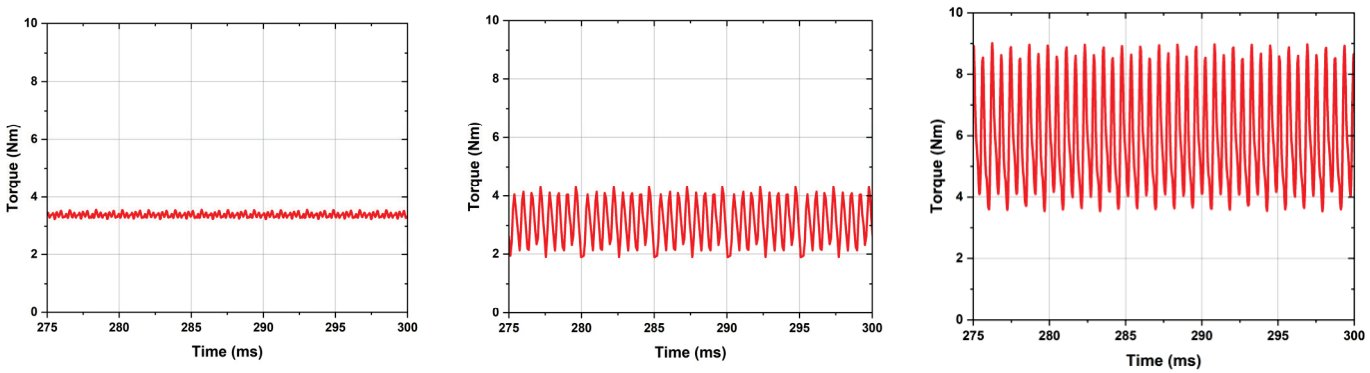


Figure 13. Torque waveforms in dry mode: (**a**) C-WRSM; (**b**) BL-WRSM; (**c**) HB-WRSM.

As can be seen in the figures, the torque waveform of the C-WRSM in drying mode meets the average torque requirements, with excellent torque ripple performance. However, the torque waveform of the BL-WRSM in Figure 13b shows significant torque ripple due to the field current, similar to that in Figure 11d. As shown in Table 3, the torque ripple further increases at higher speeds due to the interaction between the fundamental magnetic field and the field current.

Table 3. Comparison of average torque and torque ripple.

Fundamental Frequency	Washing Mode	Dry Mode	Washing Mode	Dry Mode
	Torque [Nm]		Torque Ripple [%]	
C-WRSM	17.14	3.37	5.22	8.81
BL-WRSM	16.39	3.04	25.12	39.77
HB-WRSM	19.41	5.46	7.86	93.0

When the HB-WRSM operates in dry mode at 1370 rpm, the electromagnetic induction level increases, resulting in a higher rotor current and a corresponding increase in copper losses. The additional copper loss, amounting to 254.5 W, significantly reduces the machine's efficiency to 67.7%, as shown in Table 3.

Figure 14 shows the flux contour plot of the HB-WRSM in dry mode, revealing elevated saturation levels in the rotor tooth and pole shoe areas. This saturation is attributed to the combined fluxes generated by both the rotor magnets and harmonic excitation. Although the average torque of the HB-WRSM is approx. 60 to 80% higher than the other two designs, the required output torque in dry mode is below 5 Nm, a level that can be achieved by magnet torque only. As a result, the HB-WRSM operates in dry mode as a PMSM without harmonic excitation by utilizing the centrifugal switch in Figure 13 that disconnects the harmonic winding from the field winding.

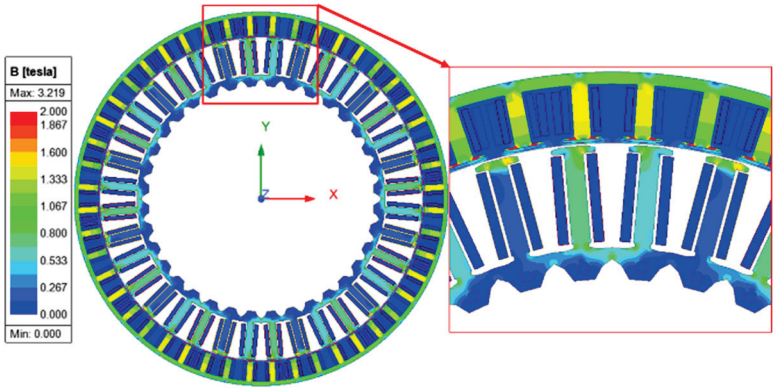


Figure 14. HB-WRSM flux density contour in dry mode.

4.4. Performance Comparison

The average output torque and torque ripple for each topology are compared in Table 3. The HB-WRSM topology demonstrates improved torque in both the washing and drying modes. However, a notable increase in torque ripple is also observed, particularly during the drying mode. To address this issue, a fourth operational mode, referred to as the PMSM mode, will be introduced in the machine.

5. PMSM Operation Mode

To mitigate the high torque ripple observed in the HB-WRSM in dry mode, a centrifugal switch is introduced into the machine's main electrical circuit, as shown in Figure 9. This switch allows the harmonic windings to be isolated from the main field windings, effectively de-energizing the rotor windings. This switch is activated when the washing machine is switched from a low-speed washing mode to a high-speed dry mode at 1370 rpm. When the machine switches to dry mode, the switch disconnects the windings, and the PMs embedded in the rotor poles solely generate rotor flux. In this configuration, the machine operates as a PMSM.

In this section, we present the results obtained for the HB-WRSM topology when operating in PMSM mode only at 1370 rpm. Figure 15 shows the flux contour plot of the HB-WRSM operating as a PMSM in dry mode, with lower saturation levels in the rotor tooth and pole shoe regions compared to Figure 14. The observed torque in dry mode is shown in Figure 16. The torque ripple is significantly reduced to 20.24%, compared to 91.24% for the HB-WRSM configuration. Therefore, the machine achieves optimal performance in PMSM mode during the drying phase.

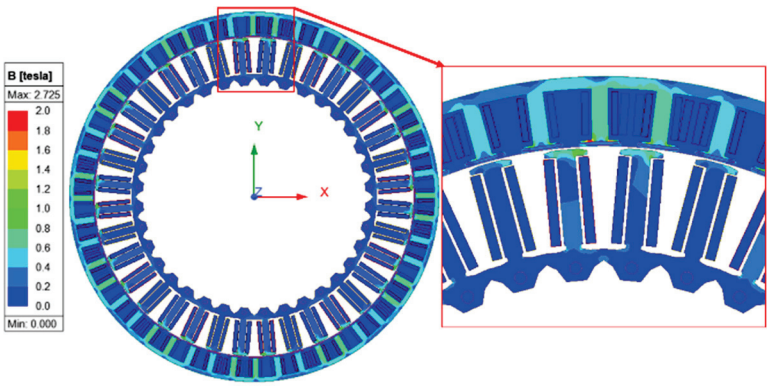


Figure 15. PMSM flux density contour in dry mode.

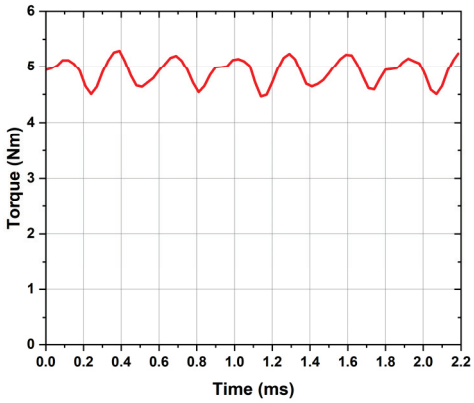


Figure 16. Torque waveform in dry mode during PMSM mode.

The average torque of the proposed HB-WRSM in PMSM mode is reduced by 10%, yet it remains significantly higher than the two conventional designs, exceeding the C-WRSM and BL-WRSM by 46% and 62%, respectively. In addition, the efficiency of the proposed machine configuration is greatly enhanced, increasing from 67.68% to 92.48%, largely due to the reduced losses in the rotor windings and the machine's core.

Table 4 provides a detailed comparison of performance metrics across all machine topologies. The overall performance of the proposed HB-WRSM with PMSM mode enabled outperforms the conventional WRSMs in all key metrics, including average torque, torque ripple, and efficiency.

Table 4. Performance comparison of the three machine topologies.

Parameter	C-WRSM		BL-WRSM		HB-WRSM		PMSM Mode
	Wash	Dry	Wash	Dry	Wash	Dry	Dry
Average torque [Nm]	17.14	3.37	16.34	3.03	19.41	5.47	4.91
Torque ripple [%]	5.22	8.81	25.12	39.77	7.86	93.0	20.0
Harmonic current [A]	0	0	5.70	3.63	1.78	10.86	0
Field current [A]	5.83	4.16	6.26	4.87	1.85	13.64	0
Output power [W]	82.54	484.25	78.90	435.80	93.46	783.87	704.13
Core loss [W]	1.62	18.72	1.45	11.59	1.17	86.32	23.76
Rotor copper loss [W]	41.12	20.93	55.48	31.99	4.87	254.46	0
Stator copper loss [W]	87.33	33.48	87.33	33.48	87.33	33.48	33.48
Efficiency [%]	38.82	86.87	35.35	84.97	50.02	67.68	92.48

6. Conclusions

This research introduced a novel HB-WRSM topology designed to deliver optimal performance at two distinct operating speeds for washing machines: a low-speed washing mode and high-speed drying mode. This study evaluated whether the HB-WRSM outperforms the C-WRSM and BL-WRSM at both operating points. To minimize torque ripple and copper losses, the HB-WRSM can function as a PMSM mode during drying operation. To achieve this, the HB-WRSM was redesigned to operate as a hybrid brushless system when a centrifugal switch is closed and transition to a PM synchronous machine when the switch is open.

One of the key advantages of the proposed HB-WRSM is its use of an inherent sub-harmonic magnetic field as an excitation source, which allows for reduced permanent magnet content, while increasing excitation levels at low speeds to achieve the necessary high torque. In drying mode, the rotor windings are disconnected via the centrifugal

switch, eliminating copper losses, with permanent magnets providing sufficient excitation to achieve the required low torque.

Future research will focus on reducing torque ripple and losses through optimized or novel design strategies employing advanced techniques. Cost optimization may also be explored by incorporating a consequent-pole rotor configuration. Although transient phenomena—such as torque and current fluctuations during speed transitions—are critical to the overall performance, they have not yet been thoroughly investigated. Future work will therefore include detailed transient analyses and a comprehensive optimization across all the considered cases to enhance efficiency and dynamic performance.

Author Contributions: Conceptualization, S.A., Q.A. and G.J.S.; methodology, S.A.; software, S.A. and K.K.; validation, S.A., Q.A. and K.K.; formal analysis, S.A.; investigation, S.A. and K.K.; writing—original draft preparation, S.A. and Q.A.; writing—review and editing, G.J.S. and G.C.; supervision, Q.A., G.J.S. and G.C.; project administration, G.C.; funding acquisition, G.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by an Inha University Research Grant.

Data Availability Statement: The original contributions presented in this study are included in the article; further inquiries can be directed to the corresponding author.

Acknowledgments: During the preparation of this manuscript/study, the authors used ChatGPT (GPT-4-turbo version) for the purposes of grammatical proofreading. The authors have reviewed and edited the output and take full responsibility for the content of this publication. The AI-powered proofreading tool was used to identify typos and grammatical errors, and explanations were requested for the corrections provided.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Du, Z.S.; Lipo, T.A. Cost-Effective High Torque Density Bi-Magnet Machines Utilizing Rare Earth and Ferrite Permanent Magnets. *IEEE Trans. Energy Convers.* **2020**, *35*, 1577–1584. [CrossRef]
2. Cheramin, M.; Saha, A.K.; Cheng, J.; Paul, S.K.; Jin, H. Resilient NdFeB magnet recycling under the impacts of the COVID-19 pandemic: Stochastic programming and Benders decomposition. *Transp. Res. Part E Logist. Transp. Rev.* **2021**, *155*, 102505. [CrossRef]
3. Kaya, E.E.; Kaya, O.; Stopic, S.; Gürmen, S.; Friedrich, B. NdFeB Magnets Recycling Process: An Alternative Method to Produce Mixed Rare Earth Oxide from Scrap NdFeB Magnets. *Metals* **2021**, *11*, 716. [CrossRef]
4. Chakraborty, C.; Rao, Y.T.; Bhattacherjee, H. Brushless induction excited synchronous motor (BinSyM): A new motor for high power applications. In Proceedings of the 2016 IEEE 7th Power India International Conference (PIICON), Bikaner, India, 25–27 November 2016; pp. 1–6. [CrossRef]
5. Yao, F.; An, Q.; Sun, L.; Lipo, T.A. Performance Investigation of a Brushless Synchronous Machine with Additional Harmonic Field Windings. *IEEE Trans. Ind. Electron.* **2016**, *63*, 6756–6766. [CrossRef]
6. Chen, H.; Tang, J.; Liu, Y.; Jiang, B.; Boscaglia, L. Electromagnetic Performance Investigation of a Brushless Electrically Excited Synchronous Machine for Long-Distance Heavy-Duty Electric Vehicles. *IEEE Trans. Transp. Electrific.* **2025**, *11*, 225–235. [CrossRef]
7. Hussain, A.; Atiq, S.; Kwon, B. Consequent-Pole Hybrid Brushless Wound-Rotor Synchronous Machine. *IEEE Trans. Magn.* **2018**, *54*, 8206205. [CrossRef]
8. Mondal, A.K.; Basak, S.; Chakraborty, C. An Optimal Excitation Control Technique for Brushless Induction Excited Synchronous Motor (BINSYM). *IEEE Trans. Ind. Electron.* **2025**, *72*, 111–121. [CrossRef]
9. Sirewal, G.J.; Ayub, M.; Atiq, S.; Kwon, B.-I. Analysis of a Brushless Wound Rotor Synchronous Machine Employing a Stator Harmonic Winding. *IEEE Access* **2020**, *8*, 151392–151402. [CrossRef]
10. Yao, F.; An, Q.; Gao, X.; Sun, L.; Lipo, T.A. Principle of Operation and Performance of a Synchronous Machine Employing a New Harmonic Excitation Scheme. *IEEE Trans. Ind. Appl.* **2015**, *51*, 3890–3898. [CrossRef]
11. Jiang, B.; Tang, J.; Liu, Y. Extended-Kalman-Filter-Based Field Current Estimation for Brushless Electrically Excited Synchronous Machines Using Stator Current Measurements. *IEEE Trans. Transp. Electrific.* **2025**, *11*, 5042–5054. [CrossRef]

12. Pallantla, M.; Kumar, P. A Line-Commutated Thyristor Bridge Emulated Rotating Power Electronic Converter (LCTEPC) with Mode-Switching Algorithm for Brushless Exciter Applications. In Proceedings of the 2025 IEEE Texas Power and Energy Conference (TPEC), College Station, TX, USA, 10–11 February 2025; pp. 1–6. [CrossRef]
13. Ali, Q.; Atiq, S.; Lipo, T.A.; Kwon, B. PM Assisted, Brushless Wound Rotor Synchronous Machine. *J. Magn.* **2016**, *21*, 399–404. [CrossRef]
14. Zhang, Z.; Shi, L.; Yang, L.; Li, W.; Ma, J.; Li, R.; Yin, Z.; Du, G. Identification Methods for Rotating Diode Rectifier Open Circuit Failures Without Rotor Position Information in Brushless Exciter Systems of Wound Field Synchronous Machines. *IEEE Trans. Ind. Electron.* 2025; *in press*. [CrossRef]
15. Sirewal, G.J.; Lipo, T.A.; Kwon, B. Torque ripple reduction in brushless wound rotor synchronous machine by two-phase excitation winding. *Int. J. Appl. Electromagn. Mech.* **2019**, *59*, 765–773. [CrossRef]
16. Jawad, G.; Ali, Q.; Lipo, T.A.; Kwon, B.-I. Novel Brushless Wound Rotor Synchronous Machine With Zero-Sequence Third-Harmonic Field Excitation. *IEEE Trans. Magn.* **2016**, *52*, 8106104. [CrossRef]
17. Lee, J.; Lee, D.; Nam, K. Inductance Measurement and Parameter Identification for a Permanent Magnet Synchronous Motor Drive. *IEEE Trans. Ind. Electron.* **2010**, *57*, 2520–2528. [CrossRef]
18. Nonaka, S.; Kawaguchi, T. Excitation scheme of brushless self-excited-type three-phase synchronous machine. *IEEE Trans. Ind. Appl.* **1992**, *28*, 1322–1329. [CrossRef]
19. Kano, Y. Design optimization of brushless synchronous machines with wound-field excitation for hybrid electric vehicles. In Proceedings of the 2015 IEEE Energy Conversion Congress and Exposition (ECCE), Montreal, QC, Canada, 20–24 September 2015; pp. 2769–2775. [CrossRef]
20. Park, H.-J.; Lim, M.-S. Design of High-Power Density and High Efficiency Wound-Field Synchronous Motor for Electric Vehicle Traction. *IEEE Access* **2019**, *7*, 46677–46685. [CrossRef]
21. Li, J.; Li, Z.; Zhang, J.; Zhao, S.; Cheng, F.; Qian, C.; Hu, X.; Zhou, G. Automated Monitoring of the Uniform Demagnetization Faults in Permanent-Magnet Synchronous Motors: Practical Methods and Challenges. *Sustainability* **2023**, *15*, 16326. [CrossRef]
22. Dorrell, D.G. Are wound-rotor synchronous motors suitable for use in high efficiency torque-dense automotive drives? In Proceedings of the IECON 2012—38th Annual Conference on IEEE Industrial Electronics Society, Montreal, QC, Canada, 25–28 October 2012; pp. 4880–4885. [CrossRef]
23. Yao, F.; Sun, D.; Sun, L.; Lipo, T.A. Dual Third Harmonic-Current Excitation Principle of a Brushless Synchronous Machine Based on Double Three-Phase Armature Windings. In Proceedings of the 2019 22nd International Conference on Electrical Machines and Systems (ICEMS), Harbin, China, 11–14 August 2019; pp. 1–4. [CrossRef]
24. Salo, M.; Tiitinen, J. Vector Controlled PWM Current-Source-Inverter-Fed Permanent Magnet Synchronous Motor Drive with a Low-Pass Filter. *IEEE Trans. Ind. Appl.* **2002**, *38*, 1392–1400. [CrossRef]
25. Bukhari, S.S.H.; Ahmad, H.; Sirewal, G.J.; Ro, J.-S. Simplified Brushless Wound Field Synchronous Machine Topology Based on a Three-Phase Rectifier. *IEEE Access* **2021**, *9*, 8637–8648. [CrossRef]
26. Hussain, A.; Kwon, B. Novel single inverter fed brushless wound rotor synchronous machine. In Proceedings of the 2017 IEEE International Magnetics Conference (INTERMAG), Dublin, Ireland, 24–28 April 2017. [CrossRef]
27. Gieras, J.F.; Wang, C.; Kamper, M.J. *Axial Flux Permanent Magnet Brushless Machines*; Springer: Berlin/Heidelberg, Germany, 2008. [CrossRef]
28. Salminen, P.; Pyrhönen, J.; Nerg, J. Design Aspects of High-Speed Synchronous Machines. *IEEE Trans. Ind. Electron.* **2006**, *53*, 1419–1426. [CrossRef]
29. Tang, L.; Rahman, M.F.; Zhong, L.; Lim, K.W. A Novel Direct Torque Control for Interior Permanent-Magnet Synchronous Machine Drive with Low Ripple in Torque and Flux—A Speed-Sensorless Approach. *IEEE Trans. Ind. Electron.* **2004**, *51*, 759–767. [CrossRef]
30. Xu, L.; Ruxi, S.; Zhao, L. A New Rotor Position Detection Method for Permanent Magnet Synchronous Machine Drive at Zero and Very Low Speed. *IEEE Trans. Ind. Electron.* **2009**, *56*, 2017–2023. [CrossRef]
31. Direct Drive Motor Specification. Fisher & Paykel Technologies. Available online: https://www.fisherpaykeltechnologies.com/technology/direct-drive-motors (accessed on 9 March 2025).
32. In Geun Ahn. Stator of Motor and Washing Apparatus Having the Same. Available online: https://patents.google.com/patent/US20110016929 (accessed on 9 March 2025).

33. Scuiller, F. General, Compact and Easy-to-Compute Winding Factor Formulation. *IET Electr. Power Appl.* **2020**, *14*, 1430–1437. [CrossRef]
34. Ahmed, S.; Siddiqi, M.R.; Ali, Q.; Yazdan, T.; Hussain, A.; Hur, J. Brushless Wound Rotor Synchronous Machine Topology Using Concentrated Winding for Dual Speed Applications. *IEEE Access* **2023**, *11*, 119560–119567. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Development of a Fault-Tolerant Permanent Magnet Synchronous Motor Using a Machine-Learning Algorithm for a Predictive Maintenance Elevator

Vasileios I. Vlachou [1] and Theoklitos S. Karakatsanis [2,*]

[1] School of Electrical and Computer Engineering, National Technical University of Athens, 15780 Athens, Greece; vasilvlachou@mail.ntua.gr
[2] Department of Production and Management Engineering, Democritus University of Thrace, 67100 Xanthi, Greece
* Correspondence: thkarak@pme.duth.gr; Tel.: +30-25410-79363

Abstract: Elevators serve as essential vertical transportation systems for both passengers and heavy loads in modern buildings. Electromechanical lifts have become the dominant choice due to their performance advantages over hydraulic systems. A critical component of their drive mechanism is the Permanent Magnet Synchronous Motor (PMSM), which is subject to mechanical and electrical stress during continuous operation. This necessitates advanced monitoring techniques to ensure safety, system reliability, and reduced maintenance costs. In this study, a fault-tolerant PMSM is designed and evaluated through 2D Finite Element Analysis (FEA), optimizing key electromagnetic parameters. The design is validated through experimental testing on a real elevator setup, capturing operational data under various loading conditions. These signals are preprocessed and analyzed using advanced machine-learning techniques, specifically a Random Forest classifier, to distinguish between Normal, Marginal, and Critical states of motor health. The model achieved a classification accuracy of 94%, demonstrating high precision in predictive maintenance capabilities. The results confirm that integrating a fault-tolerant PMSM design with real-time data analytics offers a reliable solution for early fault detection, minimizing downtime and enhancing elevator safety.

Keywords: permanent magnet machine; fault diagnosis; condition monitoring; short circuit; fault tolerance; predictive maintenance; machine learning

1. Introduction

In recent years, the need for energy efficiency, environmental protection, and pollutant reduction has been one of the main driving forces behind the development of electromobility. Many efforts have focused on the electrification of ships, aerospace, electric vehicles, and elevators [1]. Today, the need for vertical passenger transportation and the handling of large loads has made elevators a necessity. Electromechanical elevators have become the dominant choice due to their high energy efficiency potential [2]. The energy saving rate depends on the weight of the car, the counterweight, the load it is required to transport during operation, and the number of status calls it receives [3].

The main difference between electromechanical and hydraulic elevators lies in the drive system, which in electromechanical systems is based on electric motors [4]. Passenger demands for safe and direct movement, high-speed travel, transition precision, and low energy consumption have led industries to prefer this choice [5]. In this context, the

dominant market technology for elevator systems is the PMSM. Gearless PMSMs offer significant advantages over geared Induction Motors (IMs), such as high efficiency, suitable drive systems with current and speed control, noise reduction, and reduced mechanical stress [6].

Many PMSM manufacturers place great emphasis on motor design, particularly the stator and rotor, to meet specific requirements based on the rated operating load of the chamber and the maximum drive speed [7]. The continuous development of magnetic materials has led research to focus on incorporating various materials into magnets, such as NdFeB and ferrite [8]. The use of multi-objective optimization [9] and prototype design criteria [10] leads to appropriate parameterization, exploration of comparative topologies, and a better understanding of machine behavior at different operational points. As a result, numerical FEA allows for an accurate examination of the magnetic field [11].

The pressure exerted on the electric motor, which can lead to significant electrical and mechanical failures, creates potential safety issues for passengers [12]. To ensure proper motor operation, various techniques have been proposed to monitor the machine's condition in real time using smart sensors [13]. The collected data can then be processed using either traditional techniques, such as Fast Fourier Transform (FFT), Short-Time Fourier Transform (STFT), Continuous Wavelet Transform (CWT) [14], or more advanced methods based on machine-learning algorithms and multimodal data [15].

Since many failures are caused by sensors, algorithms have been developed using advanced filters such as the Extended Kalman Filter (EKF) to estimate the currents in the d-q components and compare healthy and faulty current signals, achieving effective categorization [16]. Open circuit faults result in changes in key parameters such as torque, along with intense variations and noise, compromising system safety. To maintain system operation even with reduced performance, a fault-tolerance mechanism using two key fault detection indicators has been proposed [17].

Recent studies have shown that, at low voltages (48 V), the majority of faults (66%) are electrical, while at high voltages (380 V) bearing failures account for the highest percentage (75%) [18]. For fault tolerance in electrical faults, multi-phase motors have been proposed, capable of operating even with reduced phases. Addressing winding faults like Inter Turn Short Circuit (ITSC) remain challenging, and mathematical models based on stator resistance, inductance, mutual inductance, and magnetic flux have been developed using the Particle Swarm Optimization (PSO) algorithm [19].

Changes in motor speed result in shifts in characteristic frequencies, causing frequency ambiguity. In such cases, vibration signals from various bearing faults are used to ensure accurate fault diagnosis. This is done by combining Local Mean Decomposition (LMD) and Linear Predictive Filtering (LPF) to reduce noise [20]. Another major mechanical issue is rotor eccentricity, which may be static, dynamic, or mixed. Some diagnostic techniques include load torque observers, motor oscillation analysis, and current spectral analysis [21]. More advanced frequency analyses, such as current spectrum sideband analysis and instantaneous active power calculations are used to detect rotor bar breakage due to material fatigue or excessive heat [22].

Early detection of such faults, which could lead to complete motor failure, requires continuous monitoring of the motor's condition. This is achieved by integrating smart sensors, Internet of Things (IoT) technology, and data transmission to cloud platforms where large data volumes can be stored and processed [23]. The key advantages of condition monitoring include extended PMSM lifetime, reduced maintenance costs through the prevention of unexpected shutdowns, and optimized energy efficiency via continuous monitoring [24]. A reliable methodology for machine condition monitoring is FEA, which,

by analyzing the PMSM's electromagnetic responses, identifies the location and nature of faults, such as demagnetization [25].

Although significant research has been conducted into the application of modern techniques for fault diagnosis, the use of machine-learning methods remains in its early stages. Machine-learning algorithms for analyzing large volumes of motor data have proven highly effective for early ITSC fault detection in windings. Methods with high accuracy have been proposed using both small datasets, like Support Vector Machines (SVMs) and larger datasets requiring multiple Convolutional Neural Network (CNN) measurements [26]. Various approaches have been developed to reduce computational costs and limit complex manual feature extraction. With the advantages of deep learning, automatic feature extraction from data has been achieved, improving classification through the use of Stacked Autoencoder (SAE) [27].

A hybrid model combining supervised and unsupervised learning has been considered particularly innovative, as it optimizes CNN parameters while avoiding the omission of critical data that could compromise prediction accuracy [28]. Similarly, Deep Belief Network (DBN) models using Restricted Boltzmann Machines (RBMs) have been used to diagnose bearing wear and coil short-circuits, while Recurrent Neural Networks (RNNs) are considered suitable for time-series analysis of current and vibration signals for diagnosing non-stationary PMSM faults [29]. Another important factor is the effect of noise on bearing fault diagnosis. Preprocessing signals at various noise levels has shown that Continuous Wavelet Transform (CWT) provides the best feature extraction and machine fault classification accuracy, reaching up to 99.53% [30].

Aiming to study a comprehensive elevator monitoring system, this paper investigates a low-speed, high torque PMSM. In particular, a high-performance PMSM design is proposed for elevator applications. In addition, the functional and magnetic characteristics of the motor are analyzed using FEA and compared with experimental results. The implementation of smart sensors allows for real-time monitoring and data collection. The data are processed using a prototype deep-learning algorithm for predictive maintenance, issuing alerts based on the probability of fault occurrence. This ensures uninterrupted system operation, as the predictive analysis is based on actual fault data.

The main innovation of this study lies in combining electromagnetic analysis through finite element methods to make the motor tolerant to potential short circuit faults and in employing a specialized deep-learning algorithm for predictive maintenance and fault detection, both electrical and mechanical. The proposed methodology uses this framework to effectively differentiate between normal and faulty operating conditions, ensuring passenger safety. The remainder of this paper is organized as follows. Section 2 presents the PMSM design methodology, the development of a mathematical model, and its experimental validation through a real elevator system. It also includes harmonic distortion analysis under various fault conditions and introduces diagnostic techniques for fault identification. Section 3 focuses on the FEM analysis and evaluates the performance of the motor under fault-tolerant configurations. Furthermore, it outlines the fault diagnosis and predictive maintenance methodology based on machine learning. Section 4 discusses the results and proposes future research directions to improve diagnostic accuracy and system reliability. Finally, Section 5 concludes the study by summarizing the main findings and highlighting the effectiveness of the proposed integrated approach in ensuring reliable elevator operation.

2. Design Methodology for the PMSM

Depending on the technical specifications of the building and the load that the elevator is required to support during operation, the design parameters of the electrical motor are

determined. To evaluate the benefits of using a gearless PMSM in a typical elevator installation, the following design process is presented.

2.1. Permanent Magnet Motor Topology

The motor analyzed is a three-phase PMSM rated at 5.1 kW, with a rotational speed of 160 rpm. It is characterized by its low electrical frequency of 16 Hz and high electromagnetic torque of 305 Nm, making it well-suited for low-speed, high-torque applications such as elevator systems. The general motor topology includes 12 poles and 36 slots, which provides good torque smoothness and contributes to reducing cogging effects. Figure 1 presents the cross-section view (1/4 section) of the PMSM, highlighting the arrangement of stator teeth, windings, and permanent magnets. It is a surface-mounted permanent magnet motor, utilizing NdFe-B magnets with a remanent flux density of 1.23 T to ensure strong magnetic performance and reliability over long periods of operation. To minimize iron losses, M235-25A laminations (Qingdao Lian Ge Special Steel Co., Ltd., Qingdao, China) were used in the stator core, offering a good balance between magnetic permeability and low core losses at the given operating frequency. Additionally, the motor was designed with considerations for thermal dissipation and mechanical robustness to withstand the demands of frequent elevator cycles. The nominal parameters of the PMSM used in this application are presented in Table 1, while the geometric features tailored to the elevator system's specifications are provided in Table 2.

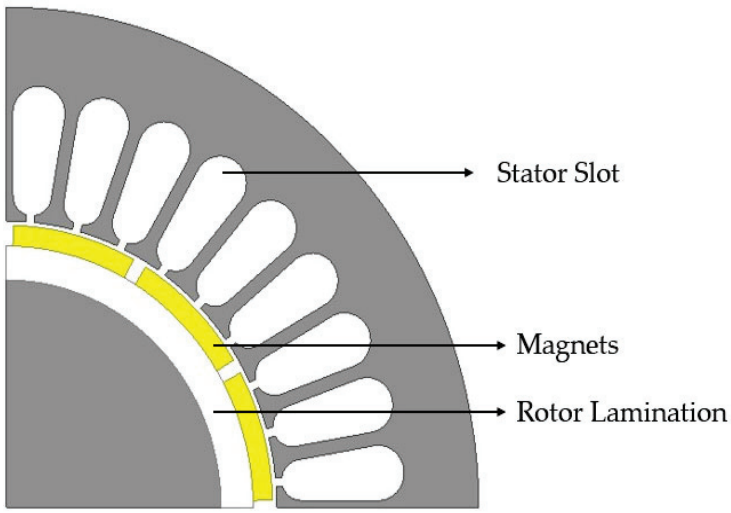


Figure 1. Cross section view (1/4 section) of a PMSM.

Table 1. Nominal electrical and mechanical characteristics of the PMSM.

Parameters	Symbol	Value
Output power	P_{out}	5.1 kW
Input power	P_{in}	6 kW
Frequency	f	16 Hz
Efficiency	E	85%
Rated current	I_r	11 A
Rated speed	n	160 rpm
Rated torque	T_r	305 Nm
Windings resistance	R_s	1 Ω
Rated voltage	V_r	380 V
Power factor	cosφ	0.80
Moment of inertia	J	0.35 Kgm^2

Table 2. Design variables and geometric parameters of the PMSM topology.

Parameters	Symbol	Value
Stator outer diameter	D_o	220 mm
Stator inner diameter	D_s	126 mm
Rotor outer diameter	D_r	124 mm
Rotor inner diameter	D_{ir}	100 mm
Axial length	L	350 mm
Shaft diameter	D_{shaft}	60 mm
Number of pole pairs	p	6
Number of slots	Q_s	36
Airgap length	L_g	1 mm
Slot opening width	b_{so}	3 mm
Slot width at the base	b_{s1}	8.6 mm
Slot width at the top	b_{s2}	12.5 mm
Stator tooth shoe height	H_{so}	1.5 mm
Stator curvature height	H_{s1}	2.5 mm
Slot total height	H_{s2}	18 mm
Magnet thickness	l_m	4.5 mm

2.2. Experimental Validation

A comparative investigation of the PMSM's performance was conducted using an operational elevator installation. To validate the characteristics of the designed PMSM, a prototype motor was constructed and tested, as shown in Figure 2, incorporating the following components: 1. VVVF Inverter, 2. Uninterruptible Power Supply (UPS), 3. Vibration sensor, 4. Interior Permanent Magnet Motor, 5. Control Board, 6. Energy Analyzer, 7. Data Collector, 8. Magnetic Field Current Transformers, 9. Raspberry pi Module 4G, 10. PC for Data Preprocessing, 11. Power Quality Analyzer.

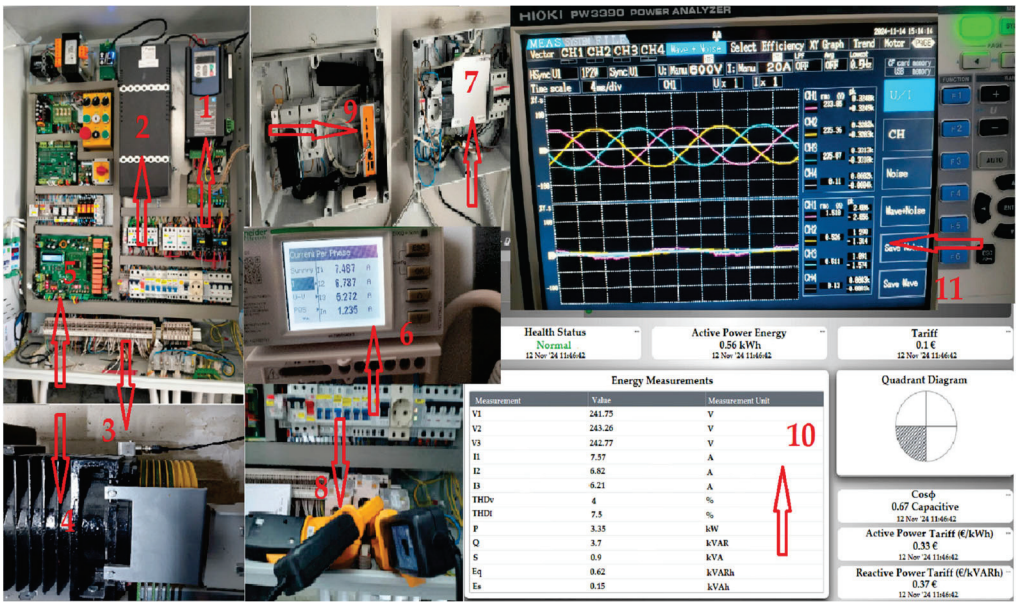


Figure 2. Experimental setup of the elevator testing platform.

Numerous Hansford HS422ST0105406 vibration sensors (Hansford Sensors Ltd., High Wycombe, UK) were strategically installed—including on axial points and the support frame—to detect vibration modes contributing to system instability, in accordance with ISO 20816-3 [31]. A Schneider Electric PM3250 (Schneider Electric SE, Rueil-Malmaison, France) energy analyzer was installed in the control panel for current and power consumption

measurements. Collected data were securely transmitted to the cloud for time- d frequency-domain analysis.

Motor characteristics were recorded under various load conditions at the rated speed. By operating the PMSM as a generator, the electromotive force (Back-EMF) was measured using a Hioki PW3390 power quality analyzer (Hioki E.E. Corporation, Ueda, Japan). Torque, current, and magnetic field distribution were analyzed using 2D FEA and compared with the experimental results.

At full load and nominal current (10 A), the output torque was 298 Nm, with a torque ripple of 6.13%, as illustrated in Figure 3. Simulated and experimental torque values aligned well under nominal conditions. Due to rotor inertia, the direct measurement of torque ripple is challenging, but smooth operation indicates enhanced system performance.

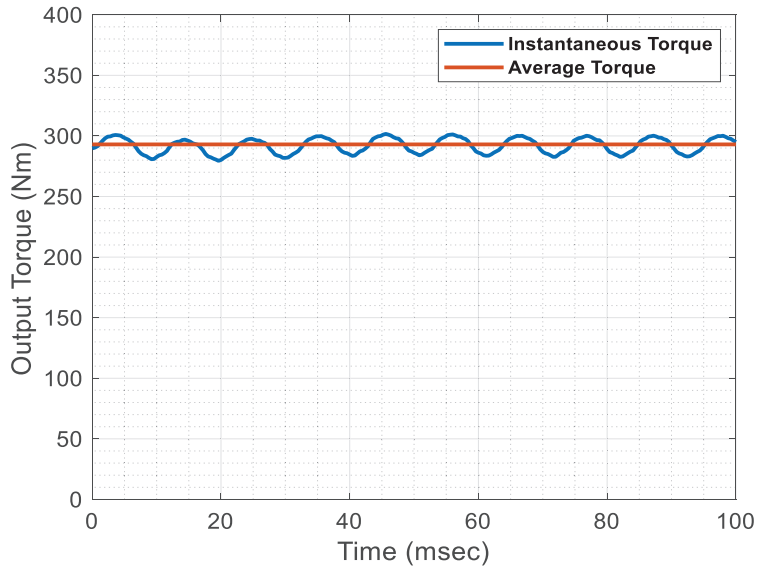


Figure 3. Computed electromechanical torque at rated operating conditions.

At 160 rpm, voltage measurements were recorded. Figure 4 displays the Back-EMF comparison between FEA simulations and experimental data, showing excellent agreement. The peak induced voltage was 321 V (simulated) and 320 V (measured), showing two complete sinusoidal cycles.

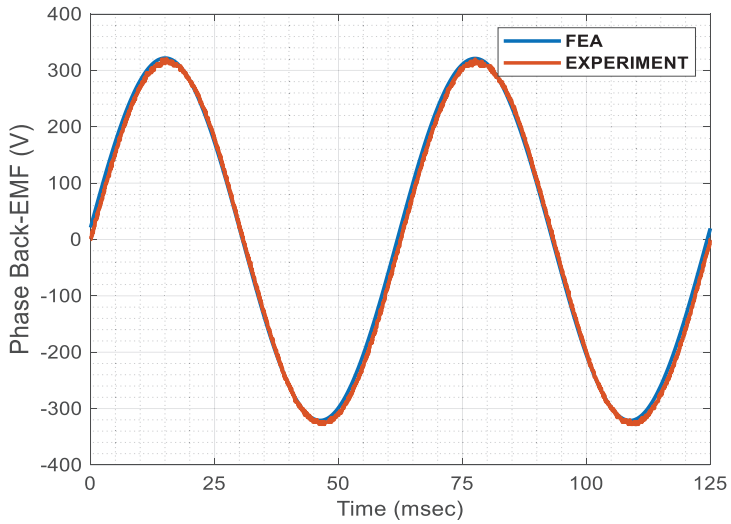


Figure 4. Comparison between simulated and measured Electromotive Force (Back-EMF) under nominal speed.

To assess the accuracy of the Finite Element Analysis (FEA) model and its agreement with experimental observations, a quantitative comparison was conducted using representative electrical and mechanical parameters under identical operating conditions. Table 3 summarizes the comparison, reporting absolute values obtained via simulation and experiment, along with the corresponding percentage error. The error for each parameter was calculated using the standard relative percentage formula.

Table 3. Comparison of simulated and experimental results.

Parameters	FEA	Experiment	Error
Phase voltage	225 V	240 V	6.25%
Electromotive force	220 V	221 V	0.45%
Phase current	10.4 A	10.8 V	3.70%
Torque	300 Nm	305 Nm	1.64%
Output power	5.02 kW	5.09 kW	1.37%
Input power	6 kW	6.05 kW	0.82%
Efficiency	83.66%	84.13%	0.56%

The results show that the maximum deviation occurs in the phase voltage (6.25%), while the electromotive force demonstrates excellent agreement (0.45%). The torque, output power, and efficiency also exhibit low errors, all within 2%. The average error across all parameters is 2.4%, which confirms a high level of consistency between simulated and experimental results.

The discrepancies can be attributed to practical factors not fully captured by the simulation, including switching noise, measurement inaccuracies, and ambient conditions. Despite these differences, the simulation results closely follow the physical behavior of the motor, validating the underlying modeling assumptions, boundary conditions, and mesh configuration used in the FEA process. These findings support the reliability of the proposed modeling framework and its ability to reflect the dynamic and steady-state behavior of the PMSM under real-world elevator operation scenarios.

The distribution of the magnetic flux density in the PMSM under nominal operating conditions is shown in Figure 5. It is observed that the intensity of the magnetic field does not exceed 2.172 Tesla. Moreover, there is a uniform distribution of the field, indicating minimal saturation and high motor efficiency.

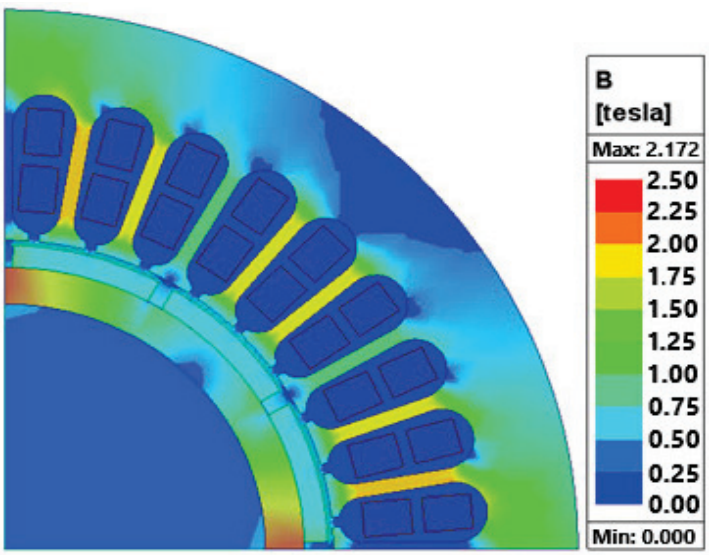


Figure 5. Magnetic flux density distribution of the PMSM at nominal operating conditions.

Motor current is a fundamental parameter as it directly influences the generation of the electromagnetic field, torque production, and associated losses. Using the nominal operating current, as determined by the motor's technical specifications, ensures a realistic simulation of standard operating conditions and facilitates a reliable comparison between FEA results and experimental measurements. Figure 6 shows the comparative analysis of the nominal motor operating current signals under full load conditions, again demonstrating strong convergence between simulated and measured data.

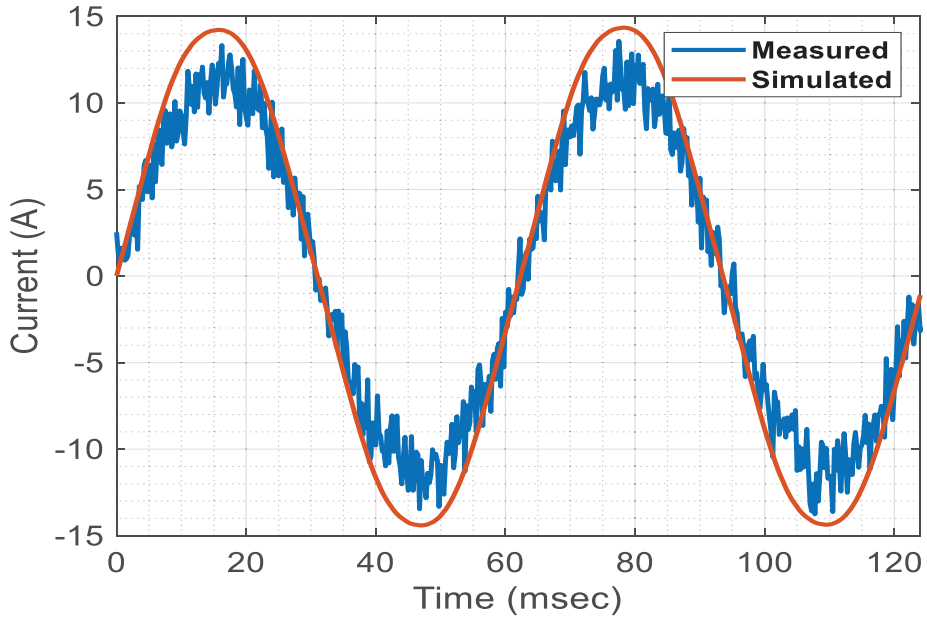


Figure 6. Comparison between measured and simulated current signals under nominal load.

2.3. Mathematical Model

Permanent Magnet Synchronous Motors (PMSMs) are widely adopted in elevator systems due to their high efficiency and reliability. However, operational stresses can lead to various electrical and mechanical faults, necessitating robust condition monitoring and predictive maintenance strategies. While traditional techniques, like the Extended Kalman Filter (EKF), offer real-time fault estimation, their computational demands limit real-time applicability. In this study, Finite Impulse Response (FIR) and Butterworth filters are proposed as alternatives for real-time fault detection and signal processing, and their performance is compared to EKF.

To analyze the operation of the PMSM, the system is represented in the d-q reference frame. The voltage equations for a PMSM under healthy conditions are given by:

$$\upsilon_d = Ri_d + L_d \frac{di_d}{d_t} - \omega L_q i_q \tag{1}$$

$$\upsilon_q = Ri_q + L_q \frac{di_q}{d_t} - \omega L_d i_d + \omega \lambda_m \tag{2}$$

where υ_d, υ_q are the direct and quadrature axis voltages, $i_{d,}, i_d$ are the corresponding currents, L_d, L_d are the inductances, ω is the angular speed, and λ_m is the permanent magnet flux linkage.

The electromechanical torque of the PMSM is given by:

$$T_e = \frac{3}{2} p\left[\lambda_m i_q + (L_d - L_q)\right] i_d i_q \tag{3}$$

where p is the number of pole pairs.

In the presence of faults, the above equations must be modified to account for changes in electrical and magnetic parameters.

When an Inter-Turn Short Circuit (ITSC) occurs in one of the stator windings, an additional fault current i_f is introduced, modifying the voltage equations:

$$\upsilon_d = R_s i_d + L_d \frac{di_d}{d_t} - \omega L_q i_q + M_f \frac{di_f}{d_t} \tag{4}$$

$$\upsilon_q = R_s i_q + L_q \frac{di_q}{d_t} - \omega L_d i_d + \omega \lambda_m + M_f \frac{di_f}{d_t} \tag{5}$$

where M_f is the mutual inductance between the faulty and healthy windings.

The fault current dynamics are given by:

$$\frac{di_f}{d_t} = -\frac{R_f}{L_f} i_f + \frac{1}{L_f} \upsilon_f \tag{6}$$

where R_f and L_f are the resistance and self-inductance of the shorted winding.

Permanent magnet degradation over time leads to a reduction in λ_m, expressed as:

$$\lambda_m(t) = \lambda_m(0) e^{-aT} \tag{7}$$

where T is the temperature and a is the demagnetization coefficient.

Bearing failures introduce vibrations that can be detected through Fast Fourier Transform (FFT). The characteristic fault frequencies f_{BPFO} (Ball Pass Frequency Outer Race) and f_{BPFI} (Ball Pass Frequency Inner Race) at which a rolling element (ball) passes over a specific point on the outer or inner race of the bearing are:

$$f_{BPFO} = \frac{n_b}{2} f_r \left(1 - \frac{d}{D} \cos\theta\right) \tag{8}$$

$$f_{BPFI} = \frac{n_b}{2} f_r \left(1 + \frac{d}{D} \cos\theta\right) \tag{9}$$

where n_b is the number of rolling elements, f_r is the shaft rotational frequency, d is the ball diameter, D is the pitch diameter, and θ is the contact angle.

A 4th-order low-pass Butterworth filter eliminates high-frequency noise in the current and vibration signals and can be described by Equation (10):

$$|H(f)| = \frac{1}{\sqrt{1 + \left(\frac{f}{f_c}\right)^{2n}}} \tag{10}$$

where f_c is the cutoff frequency and n is the filter order.

FIR filtering is applied to analyze Total Harmonic Distortion (THD) in stator currents:

$$THD = \frac{\sum_{n=2}^{\infty} I_n^2}{I_1^2} \times 100\% \tag{11}$$

A THD threshold of 5% indicates a potential fault.

EKF estimates system states dynamically. The discrete-time state-space model is:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + w_k \\ y_k &= Cx_k + \upsilon_k \end{aligned} \tag{12}$$

where $x_k = \left[i_d, i_q, \lambda_m, \omega\right]^T$ is the state vector, $u_k = \left[\upsilon_d, \upsilon_q\right]^T$ is the control input, and w_k, υ_k are the process and measurement noise.

The state estimation update follows:

$$\begin{aligned} \hat{x}_{k+1} &= A\hat{x}_k + Bu_k \\ K_k &= P_k C^T \left(CP_k C^T + R\right)^{-1} \\ \hat{x}_{k+1} &= \hat{x}_{k+1} + K_k(y_k - C\hat{x}_{k+1}) \end{aligned} \tag{13}$$

where K_k is the Kalman gain, P_k is the error covariance matrix, and R is the measurement noise covariance.

The filtering methods classify faults (shown in Table 4) as follows:

Table 4. Comparison of predictive maintenance strategy.

Method	Complexity	Real-Time Processing	Accuracy	Computational Cost
EKF	High	Medium	High	High
FIR+Butterworth	Low	High	Medium-High	Low

2.4. Harmonic Analysis for Fault Detection

Signal processing is a primary tool in the modern diagnostics of electrical and mechanical faults in motors. By analyzing waveforms corresponding to key motor characteristics, it becomes possible to extract critical information regarding the operating condition of the machine. These techniques are characterized by low computational cost, high speed, and efficiency, enabling the detection of anomalies that are not apparent through simple observation [32]. In the present application, delivering high quality torque without ripple is a fundamental design parameter. Numerous studies have concentrated on minimizing torque ripple by refining the machine's design, which in turn reduces excessive noise and additional harmonics that may lead to system losses and undesirable vibrations. As such, two main research directions have emerged: one focused on developing appropriate control strategies, and the other on optimizing the electromagnetic circuit of the electrical machine [33].

In elevator applications, torque ripple must be minimized to ensure smooth and accurate operation. A comparative investigation between a Flux Switching Permanent Magnet (FSPM) motor and a PMSM emphasized the importance of motor configuration and spatial harmonic reduction. While the FSPM motor provides enhanced performance and lower cost than a conventional PMSM, it exhibits a higher torque ripple of 9.6% compared to 5.9% for the PMSM [34]. Another research approach addressed torque ripple reduction through the optimization of pole/slot combinations. By adjusting geometric features such as tooth height and magnet angles, the torque ripple was reduced from 9.12% (10 p/12 s) to 4.03% using an optimized topology of 16 poles and 18 slots, thus achieving both high torque and efficiency [35].

Harmonic analysis, particularly the transformation of signals from the time domain to the frequency domain, remains the most widely adopted and reliable method for assessing the overall operational condition of electric machines. Figure 7a,b illustrate the analysis of the electromagnetic torque signal in both logarithmic and linear scales within the frequency domain, while Table 5 presents the characteristic parameters of the FFT analysis. Based on the graphs and the data provided, a fundamental frequency component is identified at 16 Hz, which corresponds to the electrical frequency of the studied PMSM motor.

Moreover, the spectral amplitude at this fundamental frequency was found to be 3.89 Nm, indicating a very weak oscillatory torque component. The majority of the signal

energy is concentrated at DC (Peak Amplitude = 294 Nm), suggesting that the torque remains nearly constant and closely aligned with the rated motor torque. According to the literature and based on comparisons with similar industrial applications, the torque ripple is evaluated as satisfactory, with potential for further reduction through improved machine design. The Total Harmonic Distortion (THD) is measured at 22.12%, indicating the presence of harmonics, but still within acceptable limits for unfiltered signals, as is the case in this study. Although the Noise Floor Index (0.0221 Nm) is particularly low, the presence of multiple lateral components (20) is also noted.

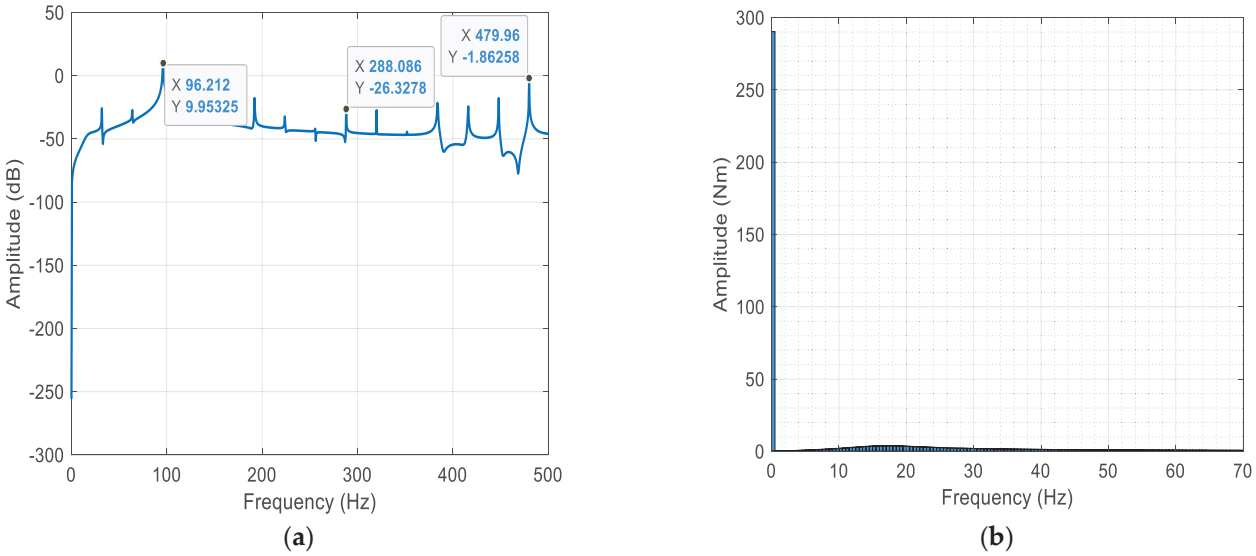


Figure 7. Frequency-domain analysis of the torque signal in the PMSM under nominal operation: (**a**) Logarithmic scale; (**b**) Linear scale.

Table 5. Key parameters from the FFT analysis.

Parameters	Value	Units
Fundamental frequency	16	Hz
Amplitude of FFT f_el	3.89	Nm
Max peak amplitude (DC)	296	Nm
THD	22.12	%
TR	6.13	%
Noise floor	0.0221	Nm
Sidebands near f_el	20	-
Dominant harmonics count	226	-

2.5. Techniques for Diagnosing Electrical Faults

In the domain of direct electrical fault diagnosis, efforts have focused on identifying appropriate and modern techniques with high accuracy. Two of the most widely recognized traditional methodologies, known for their efficiency and reliability, are Motor Current Signature Analysis (MCSA) and Park Vector Approach (PVA). MCSA is a non-intrusive method based on spectral analysis of the motor supply currents. It detects possible short-circuits between coils and does not require access to the inside of the motor, producing results during machine operation [35]. A similar function is performed by the park vector approach, which transforms the three-phase system into a two-phase system, enabling temporal and geometric analysis of the signal trajectory. This makes it easier to identify indicators of possible distortions [36,37].

The two characteristic techniques are discussed below to draw conclusions in the comparative investigation of the original current signals in Figure 8, where the time-

domain waveform of the healthy condition demonstrates smooth periodicity and stability without strong peaks, while the signal oscillations appear symmetrical and confined within a low range of values. By applying the MCSA method in both logarithmic and linear scales (shown in Figure 9a,b), we observe that, in the frequency domain, the healthy condition reveals sharp peaks at the fundamental operating frequencies, while higher-frequency components are absent, indicating reduced noise.

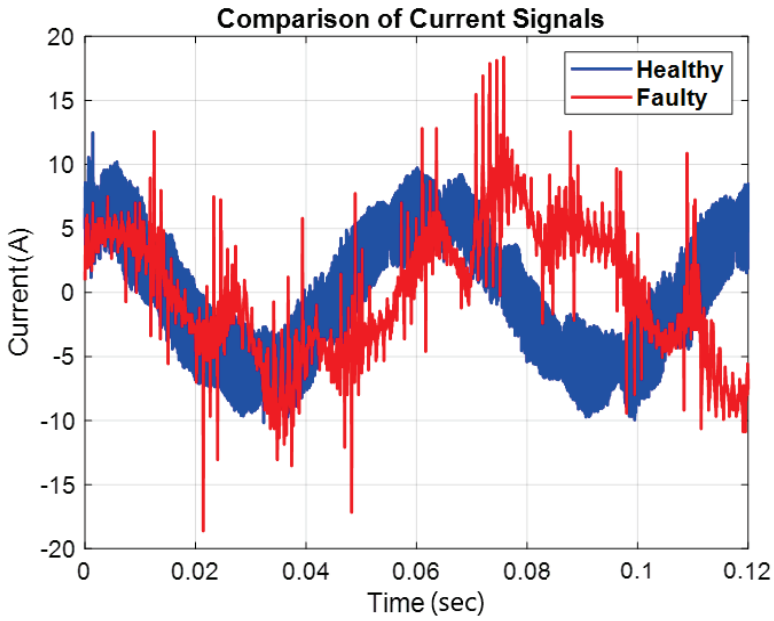


Figure 8. Comparison of the original measured current signal at half nominal load.

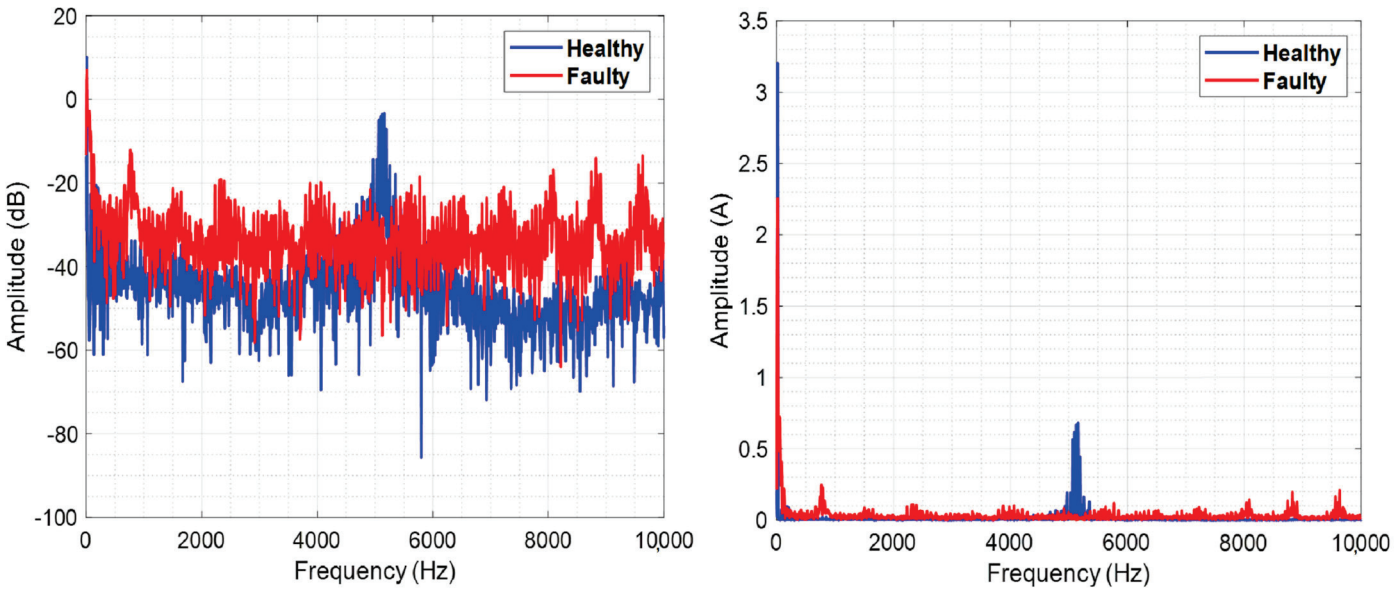


Figure 9. Frequency domain analysis of current signals under healthy and faulty conditions: (**a**) Logarithmic scale; (**b**) Linear scale.

Moreover, in the logarithmic scale, the healthy condition exhibits significantly lower noise levels over a wider frequency band. This is typical of a motor in good condition, where no unwanted frequency disturbances or energy beyond the operating limits is observed. The healthy operation curve (blue line) at low frequencies is smooth, with energy

mainly concentrated at fundamental frequencies, indicating that the motor is operating without parasitic effects or excessive loads at low frequencies, as detailed in Table 6.

Table 6. Current amplitude analysis across different frequencies.

Frequency (Hz)	Healthy (dB)	Faulty (dB)
50	−48.41	−6.33
100	−42.29	−19.73
150	−54.49	−30.23
200	−48.82	−22.33
250	−55.14	−26.78
300	−27.31	−27.34
350	−43.24	−33.88
400	−28.95	−33.45
450	−45.92	−27.78
500	−40.51	−46.77
550	−42.76	−32.12
600	−41.47	−24.16
650	−45.18	−42.70
700	−41.24	−38.51
750	−40.69	−20.88
800	−40.66	−17.69
850	−33.85	−23.94
900	−41.70	−27.10
950	−44.33	−35.60
1000	−45.94	−33.47

The application of PVA and Extended PVA (EPVA) revealed both key quantitative and visual differences between current signals under healthy and faulty conditions, as shown in Figure 10a,b. The healthy signal is characterized by a circular trajectory, stable amplitude, symmetry, and very good balance. In contrast, the faulty signal exhibits an irregular and distorted trajectory, dense clustering points indicating instability, as well as a large dispersion and displacement. Table 7 contributes to the analysis of the statistical variations in the signals, where the Variance Zero-seq and Variance a, b parameters indicate internal failures from coil short-circuits, as well as asymmetry in the current phases.

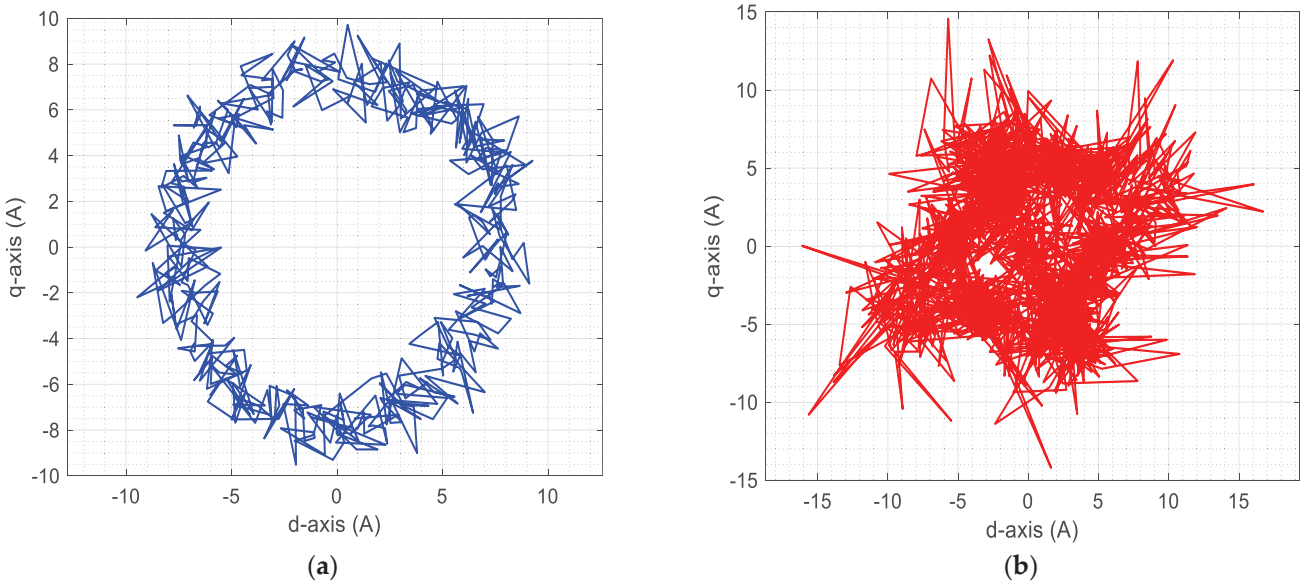


Figure 10. Park vector representation of a PMSM under different operating states: (**a**) Healthy; (**b**) Faulty.

Table 7. Extended park vector approach parameters for each operating condition.

Parameters	Healthy	Faulty
Mean a/b/zero-seq	0.0048/0.022/0.0682	−0.012/−0.16/−0.21
Variance a/b	30.37/29.82	21.33/18.99
Variance zero-seq	0.3927	5.32
Range a/b	18.74/10.24	32.83/28.79
Range zero-seq	3.4506	15.3687
Mean distance from origin	7.7252	6.3895
Energy	7725.5731	5712.2464

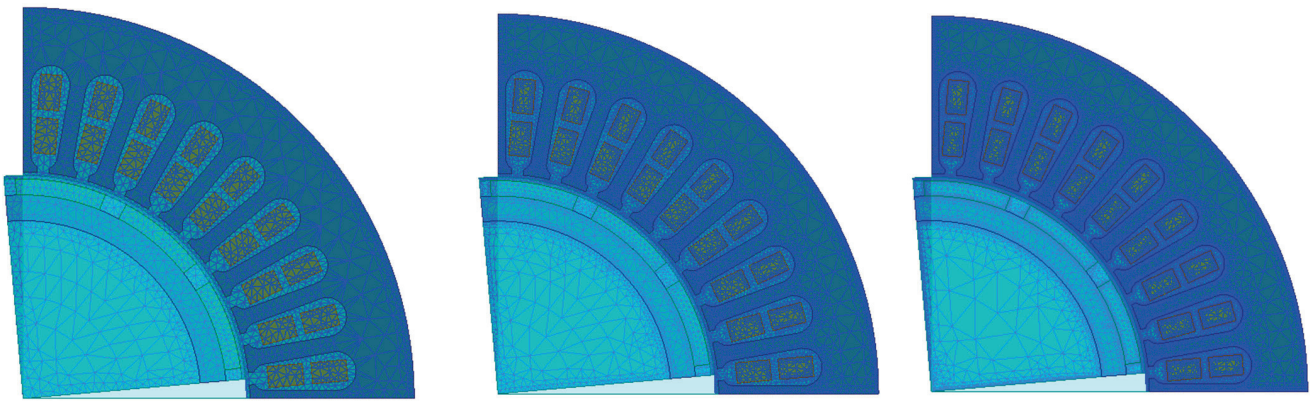
3. Fault Tolerance at Steady State

Passenger safety is a critical issue in elevator systems. The occurrence of various types of failure in PMSMs can affect the overall operation of the system. Many research efforts have focused on proper inverter control to ensure uninterrupted operation of the machine, even in the event of a simple short-circuit.

3.1. Mech Sensitivity Analysis in a PMSM

The occurrence of electrical faults (open-phase and short-circuit) can cause abrupt and localized changes in the electromagnetic field, creating flow asymmetries in critical design regions such as the airgap, thereby making the system highly sensitive to numerical deviations. A similar technique is applied in modern generators, where gradual improvement is adopted until numerical results are stabilized. Special emphasis is given to the airgap, as proper separation, uniform node subdivision, and the selection of a dense grid contribute to accurate current distribution [38].

Our research focused on mech sensitivity analysis to evaluate the numerical stability and independence of results with respect to mech discretization. As shown in Figure 11, three cases with different mesh densities were considered: coarse (2 mm), medium (1 mm), and fine (0.5 mm). Figure 12 presents the main electromagnetic parameters, such as average torque, magnetic flux, and maximum magnetic field value, as a function of mesh variation.



(**a**) (**b**) (**c**)

Figure 11. Mesh sensitivity analysis: (**a**) Coarse mesh (2 mm); (**b**) Medium mesh (1 mm); (**c**) Fine mesh (0.5 mm).

In the coarse mesh case, sparce discretization is observed in critical regions such as the airgap, stator tooth edges, and areas near the magnets. While suitable for initial estimates, it underestimates magnetic field peaks. The medium mesh offers denser discretization throughout the motor, with clear improvements in the gap and magnet regions, achieving a good balance of reliability and computational efficiency. As shown in Figure 11c, the

fine mesh is characterized by very small elements across the entire model. It is ideal for accurately capturing field distributions and losses.

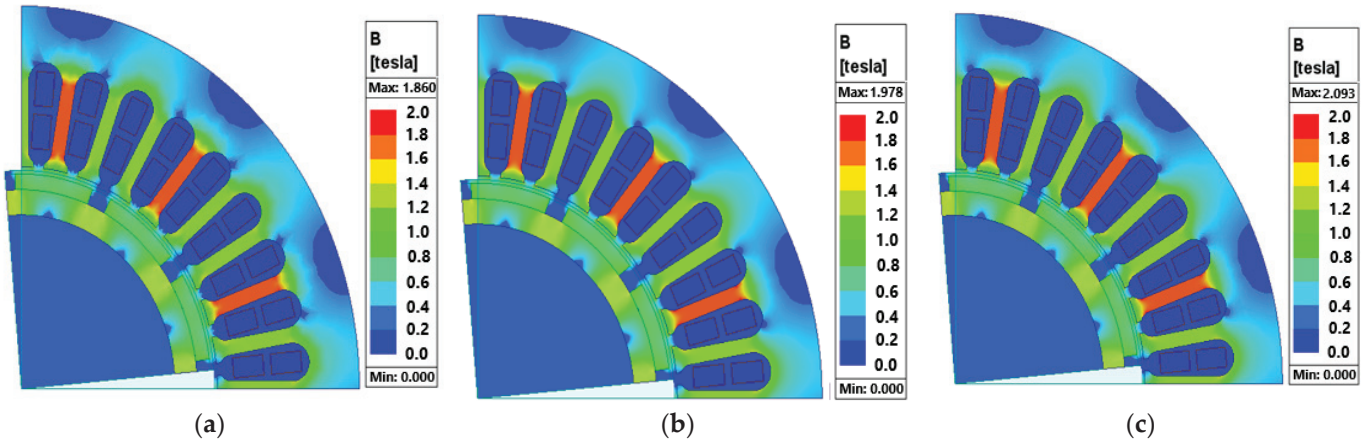


Figure 12. Magnetic field analysis in a PMSM: (**a**) Coarse mesh (2 mm); (**b**) Medium mesh (1 mm); (**c**) Fine mesh (0.5 mm).

Utilizing the Finite Element Method (FEM) enables a more comprehensive analysis of a magnetic field in each PMSM case studied, as depicted in Figure 12a–c. Despite the higher computational cost, the fine mesh provides clean, detailed results and captures flow transients, making it ideal for saturation analysis and fault conditions. Visually, the general magnetic field distribution remains stable across mesh variations, with magnetic axes, stator teeth, and saturation regions appearing at consistent locations and extents. The main difference lies in the maximum value of magnetic flux density (B), where improved peak clarity from the finer mesh provides a more complete physical picture.

Figure 13a,b show the primary variations in electromagnetic quantities as a function of mesh density. The average torque exhibited minimal changes (less than 0.3%) across different discretizations, suggesting numerical convergence. The magnetic flux showed a small but expected decrease, as the finer grid more accurately accounts for local flux leakages.

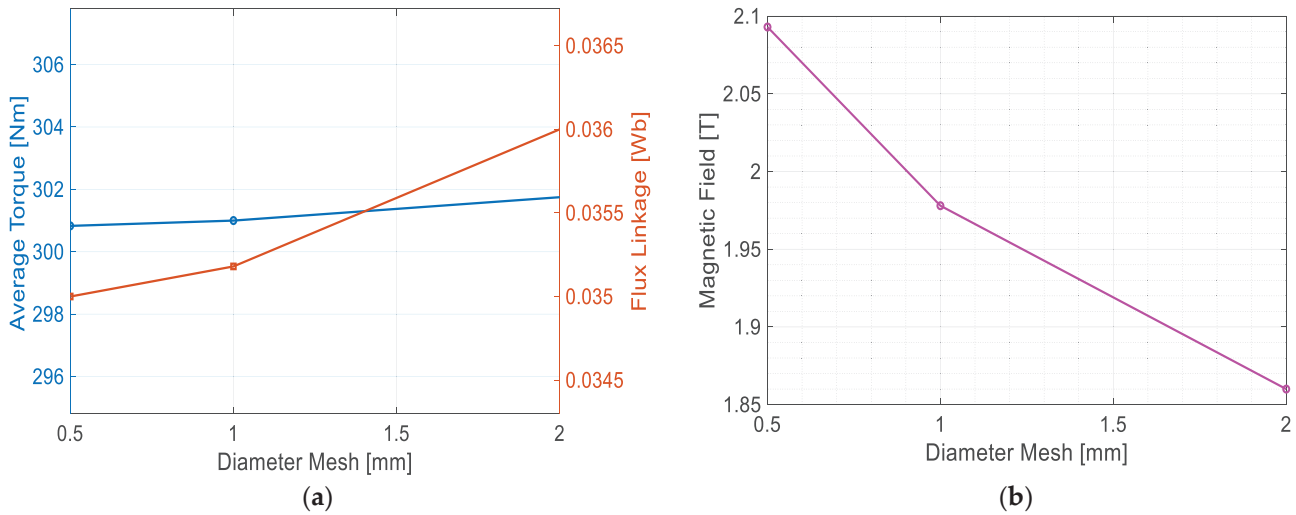


Figure 13. Comparative results of mesh sensitivity analysis in a PMSM: (**a**) Average torque and flux linkage; (**b**) Magnetic field intensity.

3.2. Elevator System Fault-Tolerance Capabilities

To ensure the correct operation of the drive system, the performance of the motor under both healthy and faulty conditions was investigated. The analysis included vibration

and current signals under short-circuit conditions. Several system loading experiments were conducted, in accordance with IEC 60034-2-1 standards [39], at 25%, 50%, 75%, and 100% load levels. Table 8 presents the nominal characteristics of the tested elevator. Figure 14a,b show the comparison between measured and simulated current waveforms for both healthy and faulty motor conditions, using switching frequencies of f_1 = 12 kHz and f_2 = 5 kHz. The results demonstrate good agreement between experimental and simulation data.

Table 8. Nominal characteristics of the elevator system.

Parameters	Symbol	Value
Mass nominal load	Q	450 kg
Mass of the chamber	P	495 kg
Mass of the counterweight	G	720 kg
Force power	F	42.75 N
Nominal velocity cabin	v	1 m/s
Nominal acceleration	a	0.19 m/s^2

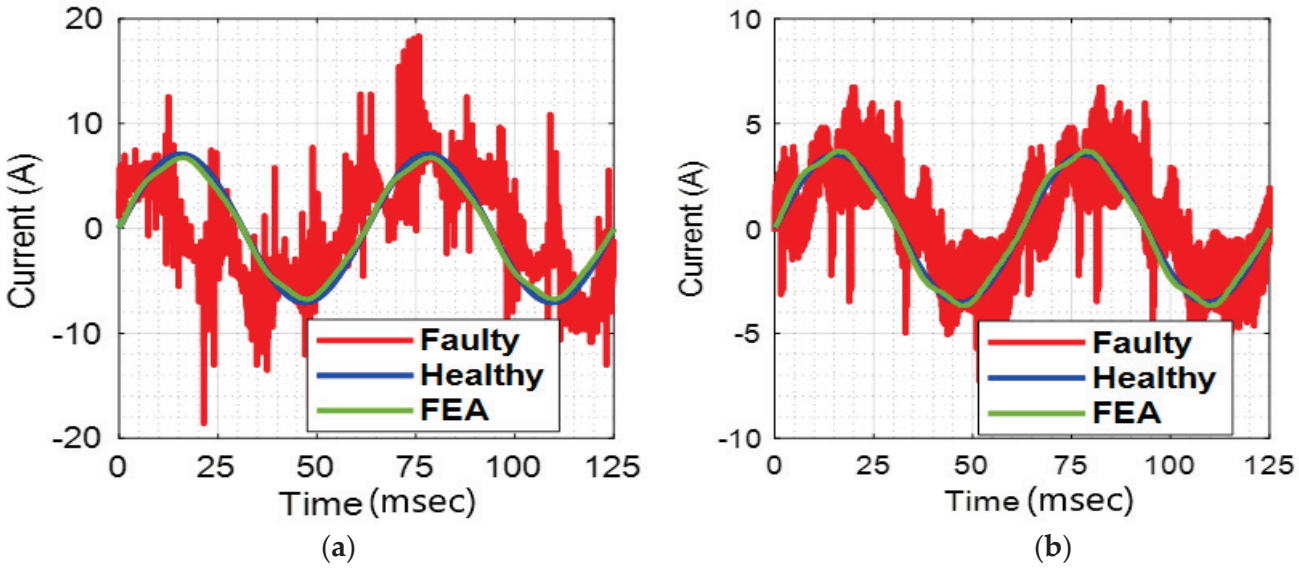


Figure 14. Measured and simulated current signal: (**a**) f_1 = 12 kHz; (**b**) f_2 = 5 kHz.

The comparative current plots in Figure 15a,b illustrate the limits and loads the motor can handle under various operating conditions. The highest current intensity occurs at full load during ascent and under no load during descent. For half the rated load, approximately equal current and power levels are observed in both directions, indicating that the motor primarily covers mechanical and core losses. Additionally, at load levels around zero active force, such as 200 Kg to 300 Kg, the current intensity does not exceed 5.5 A (rms) during either ascent or descent, confirming that the elevator operates safely under these conditions.

The maximum effective mass acting on the elevator motor is 225 kg, independent of the direction of movement. This value results from adding the nominal load (450 kg) and the empty cabin weight (495 kg), then subtracting the counterweight, which equals the cabin plus half the nominal load (720 kg). This configuration reflects the worst-case loading scenario. By multiplying this mass by the nominal acceleration of 0.19 m/s^2, the total force exerted on the motor during acceleration is calculated to be approximately 42.75 N.

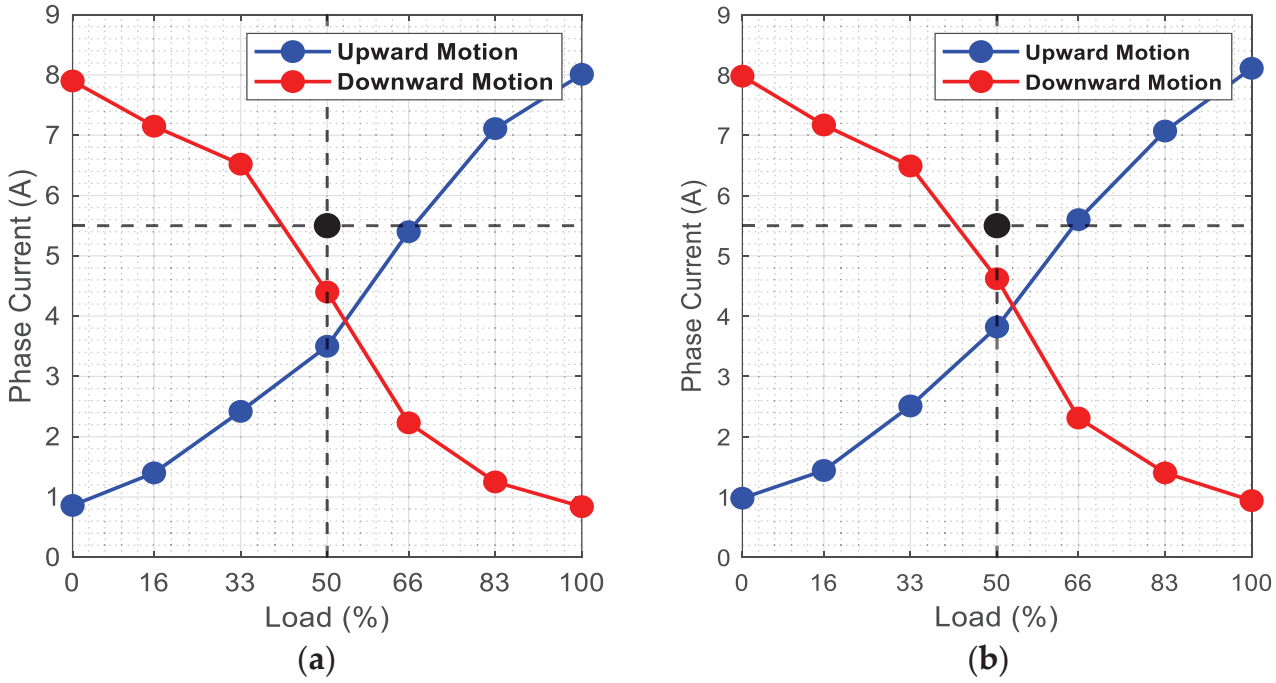


Figure 15. Fault-tolerant current-loading graphs: (**a**) Experiment; (**b**) FEA.

3.3. Condition Monitoring and Fault Detection in a PMSM

The proposed methodology is based on continuous monitoring and data acquisition for both the motor and inverter, which are crucial for the uninterrupted operation of the elevator. Through the digital platform and experimental setup, a dynamic PMSM model was developed (shown in Figure 16) with validated accuracy capturing key characteristics of both the inverter and the motor. Utilizing vector control in the Simulink/Matlab R2023a environment enabled effective PMSM operation modeling [40]. Figure 17a,b illustrate the inverter output current waveforms and motor electrical currents compared with simulation results, demonstrating convergence.

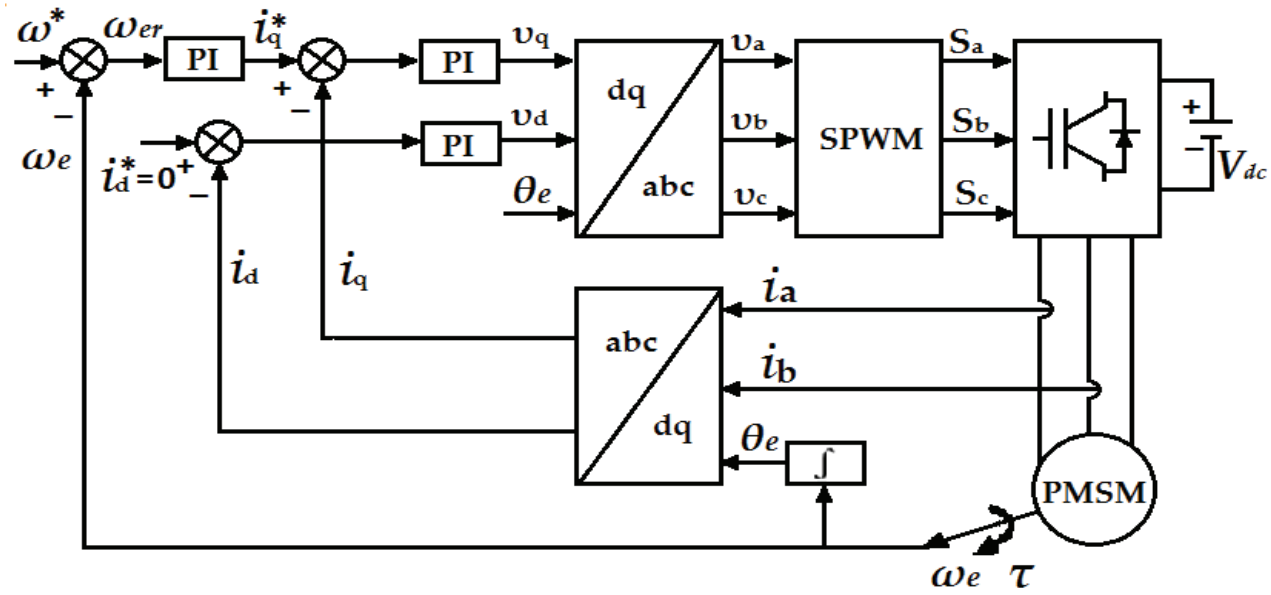


Figure 16. Typical Field Oriented Control (FOC) scheme for a PMSM.

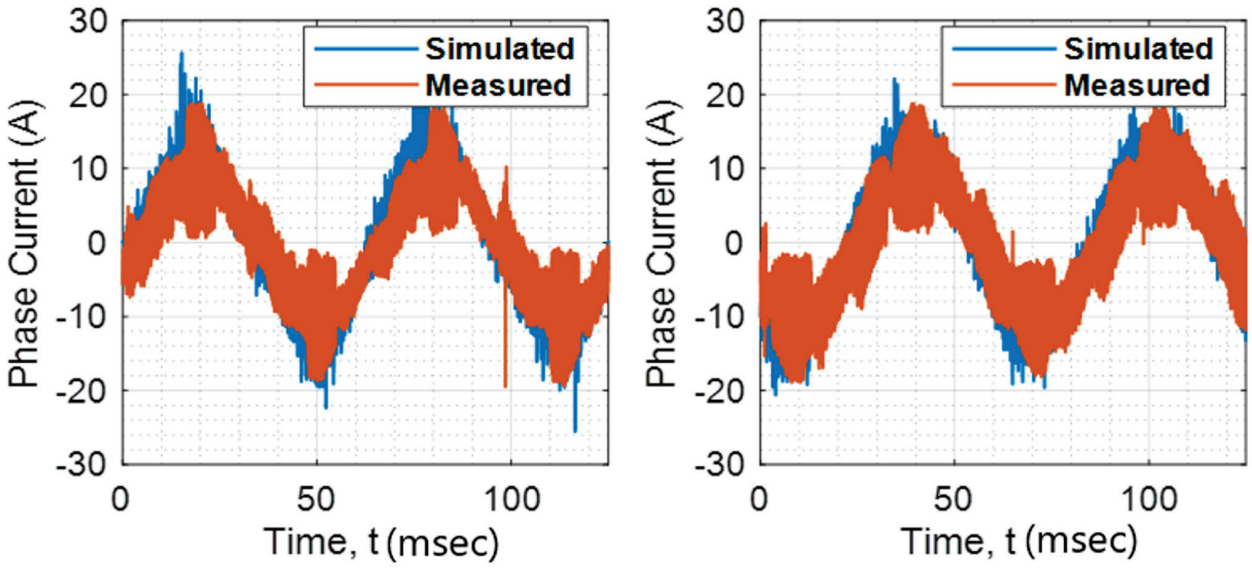


Figure 17. Current waveforms under nominal operation: (**a**) Inverter Output; (**b**) PMSM Input.

Continuous heavy loading in elevators can lead to significant mechanical issues affecting the drive system. Increased pulley friction and high inertia are major causes of eccentricity and bearing wear. The designed digital platform is based on ISO 20816-3 standards and uses vibration signal processing. According to Figure 18, the motor's operating condition, estimated health status, and high RMS speed values during operation are presented. To facilitate diagnosis, a color-coded scheme distinguishes different operational states.

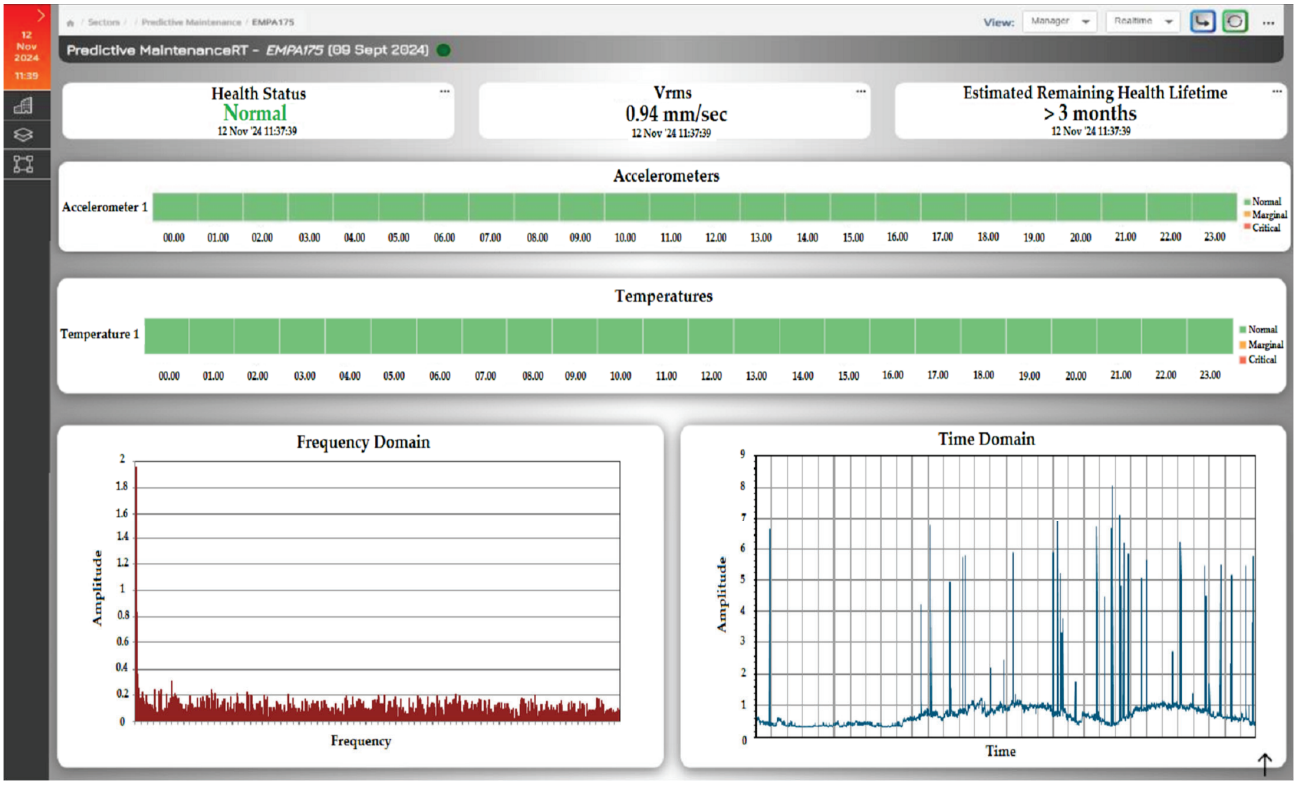


Figure 18. Dashboard visualization of vibration signals and predictive maintenance in a PMSM.

Based on the predictive maintenance results, the "healthy" state is classified as normal, indicating acceptable equipment operation. The vibration velocity measurement of 0.94 mm/s (V_{rms}) classifies the motor as "Newly Commissioned Machinery" requiring no immediate intervention. Considering temperature variation, the estimated remaining service life is over three months, a highly positive indicator of machine functionality.

To test the model's accuracy, statistical faulty data from PMSMs were used. Figures 19a,b and 20a,b show healthy and faulty vibration signals along with FFT analyses, which help identify the frequencies associated with faults. The faulty signal has lower amplitude compared to the healthy one but displays strong motor vibration fluctuations. The RMS value of 1.42 mm/s corresponds to "Unrestricted Operation", a Marginal condition requiring maintenance within one month.

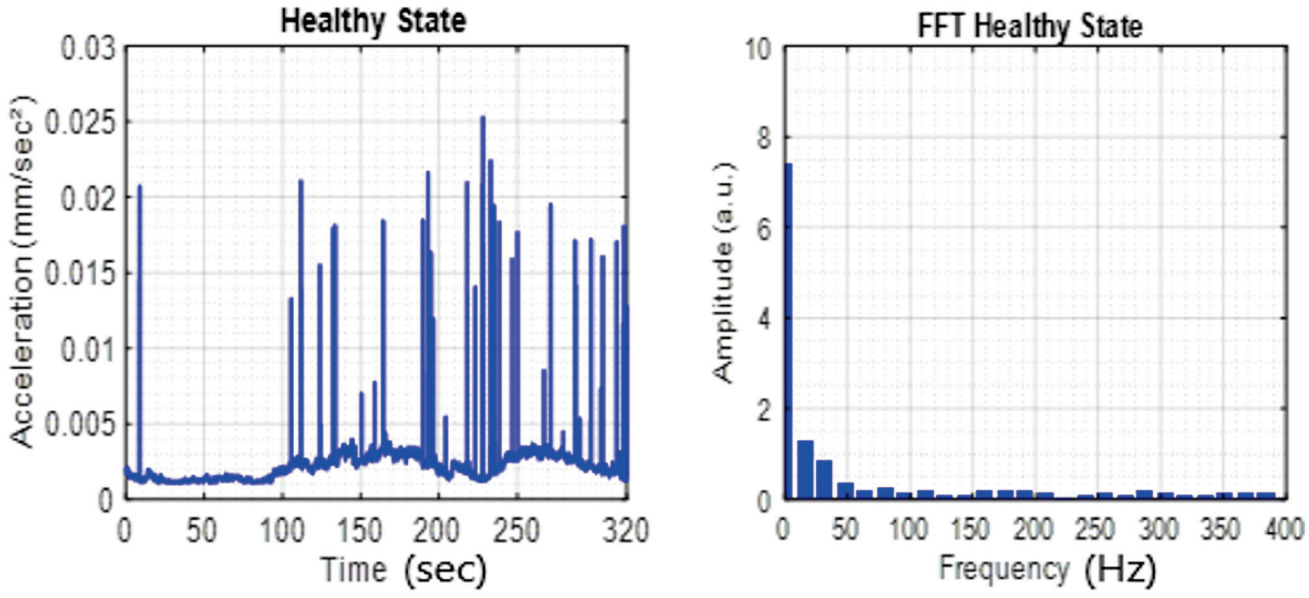


Figure 19. Healthy vibration signal: (**a**) Time domain; (**b**) Frequency domain.

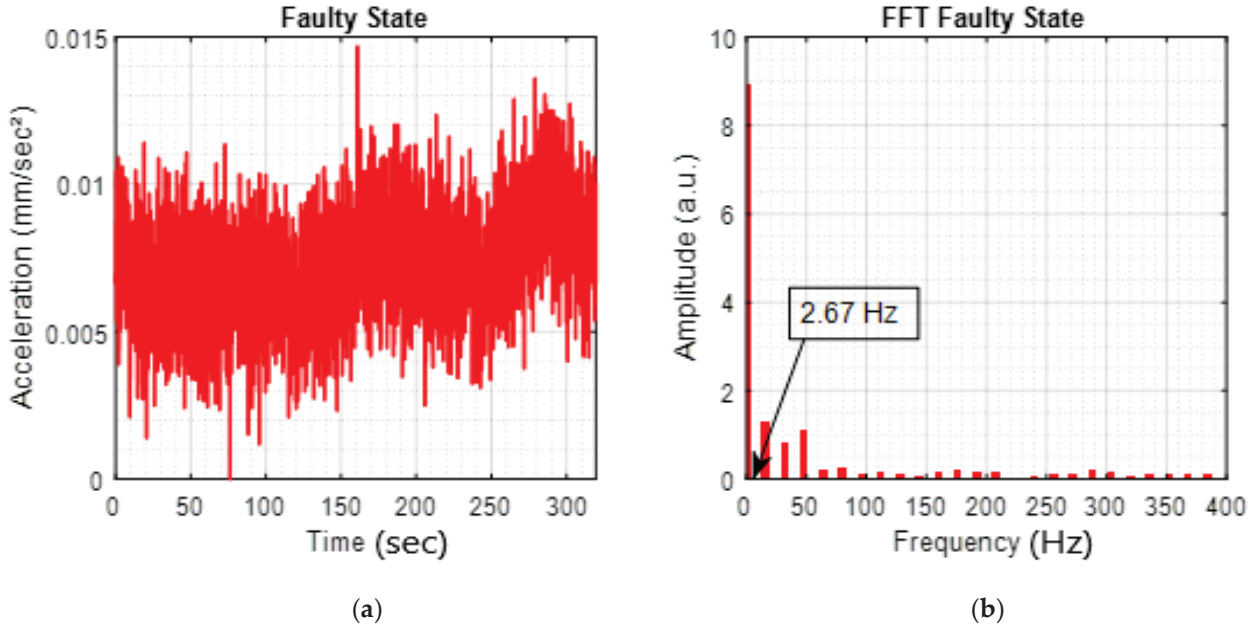


Figure 20. Faulty vibration signal: (**a**) Time domain; (**b**) Frequency domain.

Harmonic analysis of the faulty signal provides key insights into fault occurrence. A distinct peak at 2.67 Hz is associated with eccentricity, indicating bearing wear, supported by additional peaks at 32 Hz (2d harmonic), 48 Hz (3rd harmonic), and 96 Hz (6th harmonic). Notably, the absence of high frequencies (>200 Hz) suggests that the fault is at an early stage and corrective action should be taken to prevent severe secondary damage.

3.4. Proposed Methodology for Elevator Fault Prediction

The proposed methodology concerns elevator fault prediction focusing on the electric motor. It incorporates sensor data, preprocessing, feature extraction, and classification based on machine-learning model development to support an effective decision-making mechanism. The process is illustrated in Figure 21 and includes four interconnected modules, ensuring comprehensive monitoring and fault prediction.

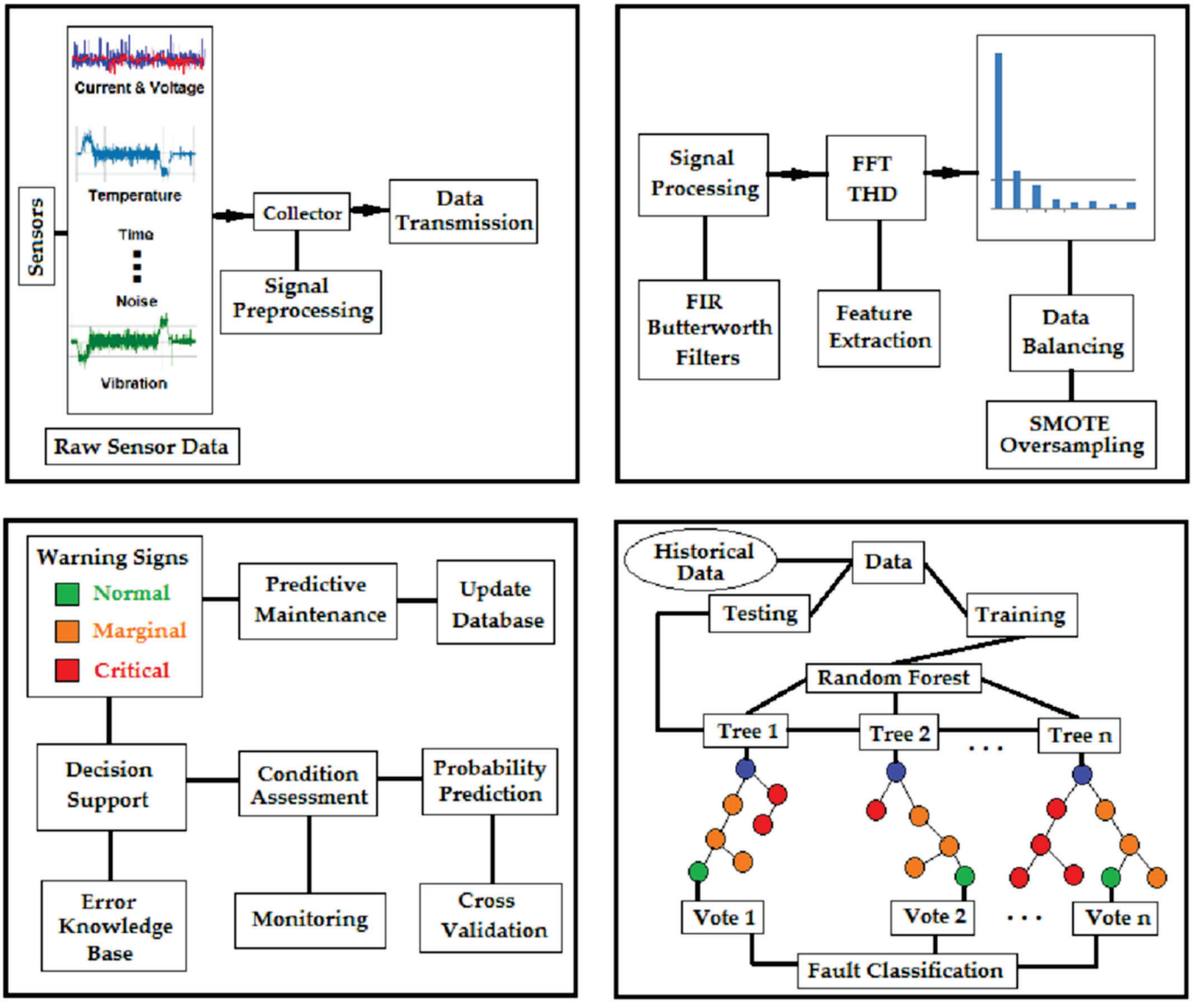


Figure 21. Flowchart of the fault diagnosis methodology.

The first stage of the methodology involves collecting data from multiple sensors placed at different locations on the machine. Once collected, the data are transmitted to a central processing unit via a secure communication system. Preprocessing of the signal is then applied using appropriate Finite Impulse Response (FIR) and Butterworth filters to eliminate high-frequency noise and any fluctuations. The preprocessing stage utilizes

techniques such as Fast Fourier Transform (FFT), Total Harmonic Distortion (THD), and statistical analysis to extract and differentiate key characteristics of healthy and faulty motor operation.

The Butterworth filter, used for signal preprocessing, was designed to attenuate high-frequency components, while preserving essential low-frequency information. Its mathematical formulation is presented in Equation (10). Application of the Butterworth filter, as illustrated in Figure 22, highlights the smooth sinusoidal waveform of the healthy current. The filter response confirms that the current is confined to low frequencies, indicating stable and normal operation. Conversely, the faulty signal exhibits strong deviations and distortions, pointing to abnormal motor winding behavior, reflected by the presence of higher frequencies.

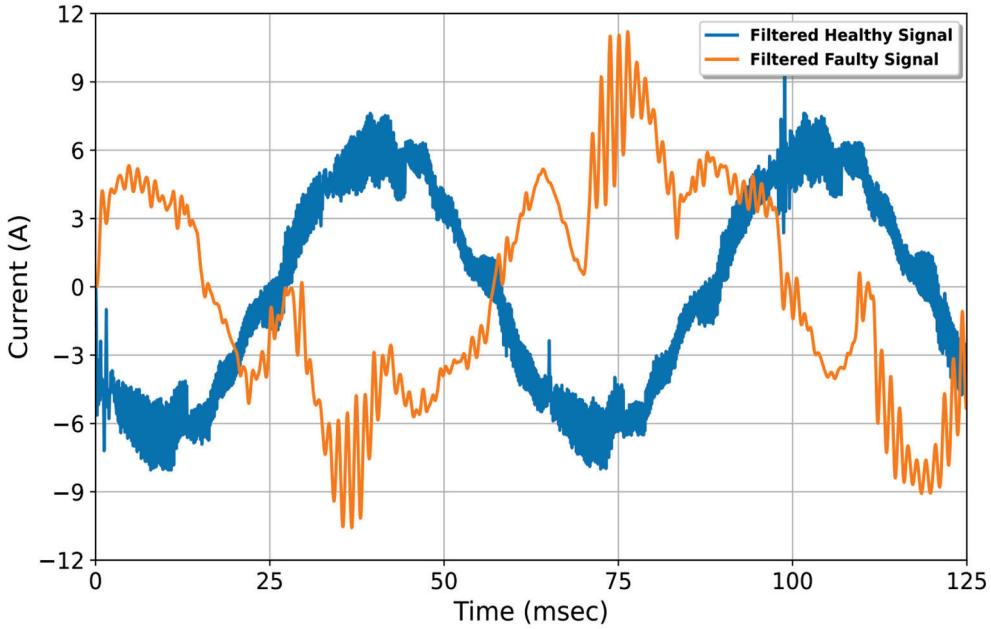


Figure 22. Healthy and faulty current signals after applying the Butterworth filter.

A machine-learning algorithm was developed using the Random Forest (RF) technique. This choice was based on the ability of RF to handle high dimensional data and deliver reliable predictions even under strong data noise. The process includes three phases: data segmentation, model training using selected features, and a voting mechanism to classify the input data. The final stage involves fault prediction, where each data segment processed by the RF algorithm outputs a failure probability, enabling continuous monitoring of the motor's operational status and generating appropriate preventive maintenance alerts.

The accuracy of the model's predictions is significantly influenced by several factors, including the quantity of available data across all operating states, the effectiveness of signal filtering, and the quality of model training. The reliability of the proposed methodology in accurately classifying current signals and predicting motor health conditions depends on the volume and diversity of the training data, which must adequately represent both normal and faulty operating scenarios. To support this, the monitoring platform was designed for continuous data acquisition at varying current levels under both healthy and degraded conditions. Furthermore, voltage measurements were also collected in parallel to capture potential variations that could influence current behavior, thereby enabling a more robust and holistic analysis.

To train and evaluate the proposed Random Forest classifier, the collected dataset was divided into training and testing sets using an 80/20 stratified split. This ensured that all three machine states (Normal, Marginal, and Critical), were proportionally represented

in both sets. The 80/20 split was chosen to provide sufficient training samples while preserving a reliable testing subset for performance validation. To improve robustness and avoid overfitting, five-fold cross-validation was performed during training. Performance metrics were averaged across folds and used for final reporting.

Although the dataset was obtained from a single elevator installation, the system was evaluated under varying operational scenarios, including ascending and descending motion at 25%, 50%, 75%, and 100% load levels, with injected electrical and mechanical faults. This variety of conditions ensures a rich data structure but also motivates future work involving datasets from different elevators and environments to enhance generalization capabilities.

The input feature vector included both time-domain and frequency-domain characteristics extracted from current and vibration signals. These features comprised time-domain attributes such as Root Mean Square (RMS), peak-to-peak value, skewness, kurtosis, and signal energy and frequency-domain features including Total Harmonic Distortion (THD), dominant FFT peaks, and sideband counts, as well as statistical metrics such as park vector variances and zero-sequence indicators.

All features were normalized prior to training to ensure scale invariance. The Random Forest model was built using 100 estimators with Gini impurity as the split criterion and a maximum depth of 15, empirically selected to balance complexity and accuracy. To address class imbalance, especially for the Critical class, class weighting was applied during training to reduce the penalty of misclassifications in underrepresented categories. Outlier filtering was also applied to remove artifacts caused by transient anomalies.

The classification performance was evaluated using standard metrics, such as Precision, Recall, F1-Score, and Accuracy. This categorization of current signals provides a clearer representation of the machine's operational condition, offering accurate monitoring with clear distinctions between Normal, Marginal, and Critical. The results on the machine-learning algorithm are summarized in Table 9, showing 94% accuracy, which reflects its effectiveness in correctly classifying the signals. The Marginal class achieved perfect recall, ensuring early-stage degradation is reliably detected. Both Normal and Critical states demonstrate high precision, indicating the model's ability to minimize false positives and support effective predictive maintenance. As a result, it detects conditions requiring immediate attention while enabling planned maintenance, reducing false alarms and enhancing system reliability.

Table 9. Results of the Random Forest algorithm.

Class	Precision (%)	Recall (%)	F1-Score (%)
Critical	97	90	93
Marginal	91	100	95
Normal	96	93	95
Accuracy	94	-	-
Macro Avg	95	94	94
Weighted Avg	95	94	94

To benchmark the RF model, alternative classifiers were trained using the same dataset and features: Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and Decision Tree (DT). Table 10 shows the comparative performance using the same five-fold cross-validation procedure.

Among all classifiers, Random Forest consistently achieved the best results in terms of accuracy and F1-score, demonstrating superior robustness under noisy and complex signal conditions.

Table 10. Comparative performance of machine-learning models.

Classifier	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Random Forest	95	94	94	94
Support Vector Machine	89	88	88	89
k-Nearest Neighbors	86	87	86	87
Decision Tree	84	83	83	85

Total Harmonic Distribution (THD) is a reliable indicator for monitoring the condition of an electric machine. Figure 23 presents a relationship diagram between current values and THD (%), showing how different machine operating states correspond to variations in THD. A correlation is observed between increasing current and THD, consistent with studied load scenarios. Using Motor Current Signal Analysis (MCSA) reveals the filter's behavior, where healthy current is confined to low frequencies, while faulty signals reflect high-frequency content.

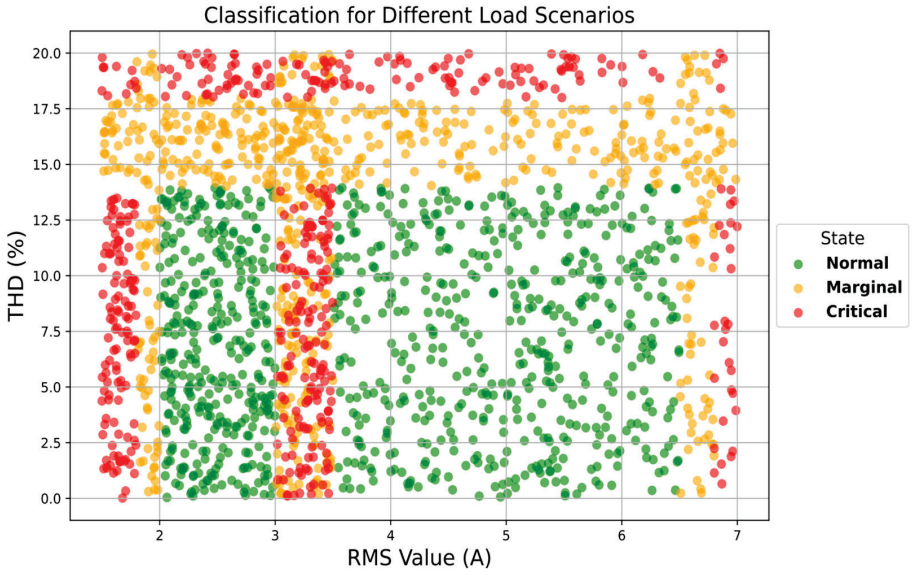


Figure 23. Harmonic distortion analysis and current values under different load conditions.

The predictive maintenance outcomes are illustrated in Figure 24.

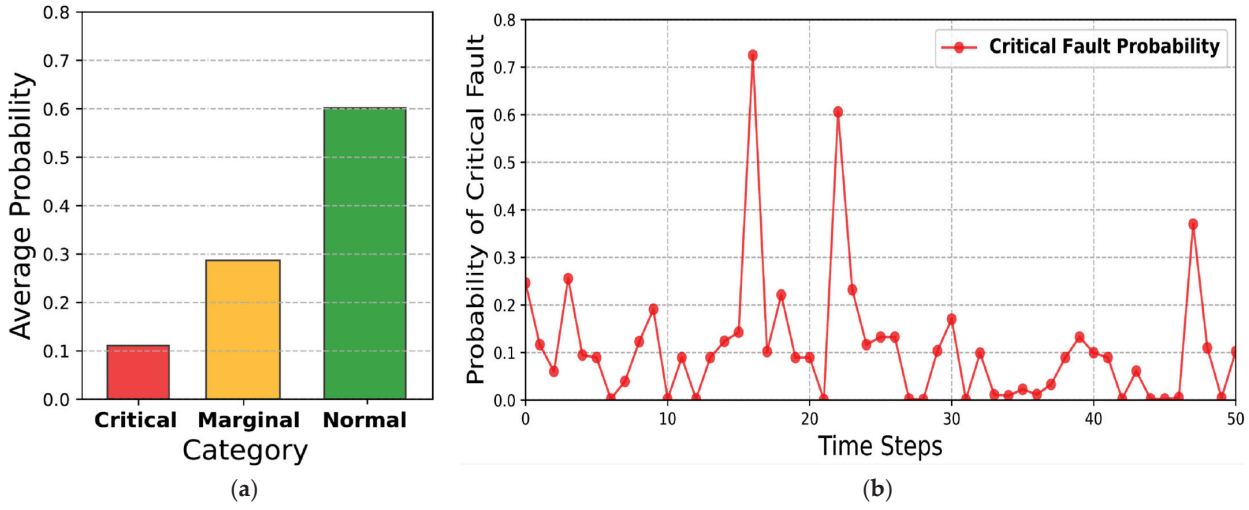


Figure 24. Predictive maintenance results: (**a**) Average predicted probabilities for each category; (**b**) Time-series of fault probability.

Based on Figure 24a, the Normal category shows the highest probability, indicating that most data are classified as normal by the model. The Marginal state has a moderate fault probability (~0.3), while the Critical state has the lowest probability, which reflects strong equipment reliability. In Figure 24b, most values cluster near zero, indicating a low risk of critical failure over time. However, the presence of occasional spikes highlights moments when failure probability increases significantly, serving as key indicators for predictive control interventions at those points.

4. Discussion

The proposed methodology demonstrated promising results in fault prediction for PMSM-based elevator systems. Nevertheless, several avenues remain open for further enhancement and exploration to improve the robustness, generalizability, and adaptability of the predictive maintenance framework. The following future directions are proposed as potential extensions of this study:

- Environmental Influence: Investigating the impact of environmental conditions such as ambient temperature, humidity, and dust accumulation on the electrical signals and vibration profiles could refine prediction accuracy under varying operating conditions.
- Data Generalization Across Systems: Incorporating data from multiple elevator installations operating under different loads and environments would enable the development of a generalized prediction model with better scalability across real-world applications.
- Advanced Machine-Learning Architectures: A comparative study involving other advanced prediction models, such as Artificial Neural Networks (ANNs), Long Short-Term Memory (LSTM), Gated Recurrent Units (GRUs), and attention-based mechanisms, particularly for time-series data, could reveal higher performance alternatives to Random Forest.
- Correlation of Electrical and Mechanical Signatures: Establishing a predictive correlation between electrical indicators (e.g., current distortions, THD) and mechanical stress or vibration signatures under dual fault conditions can enhance the accuracy and early detection capability of the model.
- Adaptive and Online Learning Framework: Implementing a dynamically updating model through local online retraining using real-time data streams would allow the system to adapt to operational drifts, sensor degradations, or changes in usage patterns without requiring complete retraining from scratch.
- Impact of Phase-to-Phase Faults: A dedicated analysis of the thermal and mechanical impact of inter-turn short circuits or phase-to-phase faults on motor windings could offer valuable insights into insulation health and long-term system stability.
- Modular Converter Redundancy: Designing and simulating a Modular Multilevel Converter (MMC) topology with fault-tolerant features such as sub-module bypassing and redundancy in power electronic elements (IGBT/MOSFET) would allow for self-healing capabilities and increased system uptime.
- Field Control Reconfiguration under Faults: Integrating an adaptive field-oriented control system capable of readjusting the d–q axis voltage vector in real-time when a fault occurs in a specific phase could maintain torque production and ensure partial functionality.
- Sensitivity Analysis on Design Topologies: A detailed sensitivity analysis involving various pole–slot combinations could lead to a reduction of spatial harmonics, minimizing torque ripple, acoustic noise, and improving overall efficiency.
- Topology Optimization Using Evolutionary Algorithms: Optimization of rotor and magnet geometries using single- and multi-objective genetic algorithms would enhance torque density and thermal efficiency while balancing multiple design trade-offs such as cost, weight, and fault resilience.

These proposed directions reflect the multi-disciplinary potential of the current research, encompassing aspects of mechanical design, control systems, signal processing, and artificial intelligence. By incorporating these enhancements, the methodology could evolve into a comprehensive and adaptive diagnostic platform, suitable not only for elevators but also for a wide range of industrial motor-driven applications.

5. Conclusions

This research demonstrates that the combination of a well-designed Permanent Magnet Synchronous Motor (PMSM) and the strategic application of machine-learning techniques can form the basis for safe, efficient, and intelligent elevator operation through predictive maintenance. A specialized PMSM design was developed, optimized for low-speed, high-torque, and high-efficiency performance, aligning with the stringent requirements of elevator systems. The electromagnetic simulations clearly highlighted the superior behavior of the PM rotor configuration in terms of fault tolerance, particularly under conditions typically encountered in real elevator operations. Mesh sensitivity analysis validated the numerical convergence and the accuracy of simulation outcomes, with fine meshes capturing magnetic saturation peaks more effectively. Nonetheless, the medium mesh (1 mm) was identified as a practical compromise, offering both computational efficiency and sufficient accuracy under normal and faulty conditions.

The integration of the motor into an experimental lifting system provided real-world validation of the proposed fault prediction framework. The data acquisition architecture, incorporating multiple sensors, enabled continuous monitoring of the motor's operational health. A robust signal processing pipeline was applied, including FIR and Butterworth filtering, Fast Fourier Transform (FFT), and Total Harmonic Distortion (THD) analysis. These techniques enhanced the visibility of fault signatures and improved the separability of healthy and faulty conditions. Using a Random Forest (RF) model, a fault classification algorithm was developed and achieved an accuracy of 94%, with high precision in distinguishing between Normal, Marginal, and Critical states. This high classification performance underpins the system's potential to minimize false positives, support proactive maintenance decisions, and ensure uninterrupted system performance. Moreover, the methodology's ability to incorporate both real-time and historical data into its decision-making mechanism marks a significant step toward intelligent condition monitoring. The system is capable of evolving into a comprehensive predictive maintenance platform applicable across a wide range of motor-driven applications beyond elevators.

In conclusion, the developed approach not only strengthens operational safety but also reduces unplanned downtime and maintenance costs, thus aligning with modern trends in smart building automation and Industry 4.0 frameworks. Compared to recent works in the literature [1–3], the present study introduces a comprehensive approach combining predictive diagnostics, ML-based classification, and physical motor design validation. Unlike Chen et al. [1], who focused on door fault detection using a single-layer ML, our method integrates PU learning and GAN-based data augmentation for enhanced classification robustness. The Soyaslan et al. [2] and Yetis et al. [3] target motor optimization lacks integration with real time monitoring or learning mechanisms. Our framework bridges this gap by enabling simultaneous motor performance evaluation and fault prediction under realistic elevator conditions.

Author Contributions: Conceptualization, V.I.V.; methodology, V.I.V. and T.S.K.; software, V.I.V.; validation, V.I.V. and T.S.K.; formal analysis, V.I.V.; investigation, V.I.V. and T.S.K.; data curation, V.I.V.; writing—original draft preparation, V.I.V. and T.S.K.; writing—review and editing, T.S.K.; visualization, V.I.V. and T.S.K.; supervision, T.S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PMSM	Permanent Magnet Synchronous Motor
EKF	Extended Kalman Filter
FFT	Fast Fourier Transform
STFT	Short-Time Fourier Transform
CWT	Continuous Wavelet Transform
FSPM	Flux Switching Permanent Magnet
FIR	Finite Impulse Response
ML	Machine Learning
THD	Total Harmonic Distribution
RF	Random Forest
SVM	Support Vector Machine
k-NN	k-Nearest Neighbors
DT	Decision Tree
PVA	Park Vector Approach
EPVA	Extended Park Vector Approach
MCSA	Motor Current Signal Analysis
ITSC	Inter Turn Short Circuit

References

1. Usman, A.; Saxena, A. Technical Roadmaps of Electric Motor Technology for Next Generation Electric Vehicles. *Machines* **2025**, *13*, 156. [CrossRef]
2. Kefalas, K.; Karlis, A.D. Supercapacitors based energy saving mode of electromechanical elevator's operation. In Proceedings of the International Conference on Electrical Machines (ICEM), Lausanne, Switzerland, 4–7 September 2016; pp. 1208–1214.
3. Sirsi, R.; Kamath, D. Energy Efficient Elevators and Technologies. In Proceedings of the IEEE India Council International Conference (INDICON), Roorkee, India, 15–17 December 2017.
4. Anand, R.; Mahesh, M. Elevator drives energy analysis with duty loads and behavior in dynamic conditions. In Proceedings of the International Conference on Emerging Technologies and Innovative Business Practices for the Transportation of Societies (EmergiTech), Balaclava, Mauritius, 3–6 August 2016.
5. Blanusa, B.; Knezevic, B.; Erceg, B.; Lekic, O.; Mrsic, P. Elevator drives: Present trends and perspectives. In Proceedings of the International Conference on Electromechanical and Power Systems (SIELMEN), Lasi, Romania, 11–13 October 2017; pp. 1–9. [CrossRef]
6. Bakhtiarzadeh, H.; Polat, A.; Ergene, L.T. Design and Analysis of a Permanent Magnet Synchronous Motor for Elevator Applications. In Proceedings of the International Conference on Optimization of Electrical and Electronic Equipment (OPTIM) & Int. Aegean Conference on Electrical Machines and Power Electronics (ACEMP), Brasov, Romania, 25–27 May 2017; pp. 293–298.
7. Avsar, Y.; Fenercioglu, A.; Soyaslan, M. Design Optimization of PM Synchronous Motor: Rail Mounted Belt Drive Elevator Systems. *IEEE Trans. Ind. Appl.* **2024**, *60*, 301–311. [CrossRef]
8. Yetis, H.; Mese, E.; Biyikli, M. Design and Comparison of Ferrite Based IPM and NdFeB Based SPM Synchronous Motors for Elevator Systems. In Proceedings of the International Conference on Electrical Machines (ICEM), Alexandroupoli, Greece, 3–6 September 2018; pp. 635–641.

9. Mirnikjoo, S.A.; Sangdeli, S.M.K.; Abdollahi, S.E. Multi-objective Optimization of a Permanent Magnet Synchronous Motor for Gearless Elevator. In Proceedings of the Power Electronic Drive Systems and Technologies Conference (PEDSTC), Tabriz, Iran, 2–4 February 2021. [CrossRef]
10. Chasiotis, I.D.; Karnavas, Y.L.; Dimadis, S.I. A Multi-Criteria Design Method of Line-Start PMSM with Enhanced Start-Up Capability for Gantry Crane Lifting Systems. In Proceedings of the IEEE Workshop on Electrical Machines, Design, Control and Diagnosis (WEMDCD), Athens, Greece, 22–23 April 2019; pp. 9–14.
11. Vlachou, V.I.; Karakatsanis, T.S.; Efstathiou, D.E. Design, Analysis and Application of Control Techniques for Driving a Permanent Magnet Synchronous Motor in an Elevator System. *Machines* **2024**, *12*, 560. [CrossRef]
12. Li, H.; Zhu, Z.Q.; Azar, Z.; Clark, R.; Wu, Z. Fault Detection of Permanent Magnet Synchronous Machines: An Overview. *Energies* **2025**, *18*, 534. [CrossRef]
13. Duan, Y.; Toliyat, H. A review of condition monitoring and fault diagnosis for permanent magnet machines. In Proceedings of the IEEE Power and Energy Society General Meeting, San Diego, CA, USA, 22–26 July 2012.
14. Orloeska-Kowalska, T.; Wolkiewicz, M.; Pietrzak, P.; Skowron, M.; Ewert, P.; Tarchala, G. Fault Diagnosis and Fault-Tolerant Control of PMSM Drives-State of the Art and Future Challenges. *IEEE Access* **2022**, *10*, 59974–60024. [CrossRef]
15. Zhan, T.; Yang, Y.; He, Y.; Feng, K. Motor Fault Diagnosis Method Based on Multimodal and Spatial Attention Mechanism. In Proceedings of the Global Reliability and Prognostics and Health Management Conference (PHM-Beijing), Beijing, China, 11–13 October 2024. [CrossRef]
16. Kim, M.; Park, J.; Moon, H.J.; Ko, S. EKF/MM-based Fault Diagnosis Algorithm of PMSM in the Presence of Sensor, Open and Short-Circuit Faults. In Proceedings of the International Conference on Control, Automation and Diagnosis (ICCAD), Paris, France, 15–17 May 2024. [CrossRef]
17. Lenz, R.; Kugi, A.; Kemmetmuller, W. Online Fault Diagnosis for Multiple Open-Circuit Faults in Multiphase Drives with Current Harmonics. In Proceedings of the International Conference on Electrical Machines (ICEM), Torino, Italy, 1–4 September 2024. [CrossRef]
18. Gyftakis, K.N.; Panagiotou, P.A.; Spyrakis, D. Detection of simultaneous mechanical faults in 6-kV pumping induction motors using combined MCSA and stray flux methods. *IET Electr. Power* **2021**, *15*, 643–651. [CrossRef]
19. Yang, J.; Dou, M.; Dai, Z.; Zhao, D.; Zhang, Z. Modeling and fault diagnosis of inter-turn short circuit for five-phase PMSM based on Particle Swarm Optimization. In Proceedings of the IEEE Applied Power Electronics Conference and Exposition (APEC), Long Beach, CA, USA, 20–24 March 2016. [CrossRef]
20. Huang, X.; Zhang, J.; Xu, Z.; Xu, S. Fault Diagnosis of Motor Bearings with Multiple Time-Frequency Extraction Method Under Variable Speed Conditions. In Proceedings of the International Power Electronics and Motion Control Conference (IPEMC2024-ECCE Assia), Chengdu, China, 17–20 May 2024; pp. 2468–2473.
21. Milfont, L.D.; Ferreira, G.T.C.; Giesbrecht, M. Fault Diagnosis in PMSM Motors: A Review of Techniques Based on Mathematical Models. In Proceedings of the Annual Conference of the IEEE Industrial Electronics Society (IECON 2024), Chicago, IL, USA, 3–6 November 2024; pp. 1–6.
22. Niu, G.; Dong, X.; Chen, Y. Motor Fault Diagnostics Based on Current Signatures: A Review. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 3520919. [CrossRef]
23. Vlachou, E.I.; Vlachou, V.I.; Efstathiou, D.E.; Karakatsanis, T.S. Overview of IoT Security Challenges and Sensors Specifications in PMSM for Elevator Applications. *Machines* **2024**, *12*, 839. [CrossRef]
24. Sergakis, A.; Salinas, M.; Gkiolekas, N.; Gyftakis, K.N. A Review of Condition Monitoring of Permanent Magnet Synchronous Machines: Techniques, Challenges and Future Directions. *Energies* **2025**, *18*, 1177. [CrossRef]
25. Creux, J.; Obeid, N.H.; Boileau, T.; Meibody-Tabar, F. Comparative Behavior Study of Various PMSM Topologies Under Demagnetization Conditions. In Proceedings of the International Conference on Electrical Machines (ICEM), Torino, Italy, 1–4 September 2024. [CrossRef]
26. Shih, K.J.; Hsieh, M.F.; Chen, B.J.; Huang, S.F. Machine Learning for Inter-Turn Short-Circuit Fault Diagnosis in Permanent Magnet Synchronous Motors. *IEEE Trans. Magn.* **2022**, *58*, 8204307. [CrossRef]
27. Huang, W.; Chen, H.; Zhao, Q. Fault Diagnosis of Inter-Turn Fault in Permanent Magnet-Synchronous Motors Based on Cycle-Generative Adversarial Networks and Deep Autoencoder. *Appl. Sci.* **2024**, *14*, 2139. [CrossRef]
28. Li, H.; Shen, J.; Shi, C.; Shi, T. Hybrid Learning Model-Based Inter-turn Short Circuit Fault Diagnosis of PMSM. In Proceedings of the IEEE Transportation Electrification Conference and Expo, Asia-Pacific (ITEC Asia-Pacific), Chiang Mai, Thailand, 28 November–1 December 2023; pp. 1–6. [CrossRef]
29. Yang, Y.; Haque, M.M.M.; Bai, D.; Tang, W. Fault Diagnosis of Electric Motors Using Deep Learning Algorithms and Its Application: A Review. *Energies* **2021**, *14*, 7071. [CrossRef]
30. Moysiadis, D.A.; Karatzinis, G.D.; Boutalis, Y.S.; Karnavas, Y.L. A Study of Noise Effect in Electrical Machines Bearing Fault Detection and Diagnosis Considering Different Representative Feature Models. *Machines* **2023**, *11*, 1029. [CrossRef]

31. *ISO 20816-3*; Mechanical Vibration-Measurements and Evaluation of Machine Vibration-Part 3: Industrial Machines with Nominal Power Above 15 kW and Nominal Speeds Between 120 r/min and 30000 r/min When Measured In Situ. International Organization for Standardization: Geneva, Switzerland, 2022.
32. Choi, S.; Pazouki, E.; Baek, J.; Bahrami, H.R. Iterative Condition Monitoring and Fault Diagnosis Scheme of Electric Motor for Harch Industrial Application. *IEEE Trans. Ind. Electron.* **2015**, *62*, 1760–1769. [CrossRef]
33. Jedryczka, C.; Danielczyk, D.; Szelag, W. Torque Ripple Minimization of the Permanent Magnet Synchronous Machine by Modulation of the Phase Currents. *Sensors* **2020**, *20*, 2406. [CrossRef] [PubMed]
34. Hwang, H.; Kim, D.; Lee, C.; Jin, H. Torque ripple reduction in a flux-switching permanent magnet machine targeted at elevator door applications by minimizing space harmonics. In Proceedings of the IEEE Energy Conversion Congress and Exposition (ECCE), Milwaukee, WI, USA, 18–22 September 2016. [CrossRef]
35. Khorasani, M.; Mirimani, S.M.; Gholamian, S.A. Optimal Design of a Permanent Magnet Synchronous Motor for Elevator Applications by Reducing Torque Ripple. In Proceedings of the International Conference on Electrical Machines and Drives (ICEMD), Tehran, Iran, 20–21 December 2023. [CrossRef]
36. Halder, S.; Bhat, S.; Bhaumik, C.; Rakshit, R. Stator Inter-Turn Fault Diagnosis in Permanent Magnet Synchronous Motor. In Proceedings of the International Conference on Smart Technologies for Power, Energy and Control (STPEC), Nagpur, India, 25–26 September 2020. [CrossRef]
37. Gubarevych, O.; Gerlici, J.; Kravchenko, O.; Melkonova, I.; Melnyk, O. Use of Park's Vector Method for Monitoring the Rotor Condition of an Induction Motor as a Part of the Built-In Diagnostic System of Electric Drives of Transport. *Energies* **2023**, *16*, 5109. [CrossRef]
38. Nusso, S.; Galea, M.; Bolognesi, P.; Vakil, G.; Gerada, C.; Brown, N. A Methodology to Remove Stator Skew in Small-Medium Size Synchronous Generators via Innovative Damper Cage Designs. *IEEE Trans. Ind. Electron.* **2019**, *66*, 4296–4307. [CrossRef]
39. *IEC 60034-2-1*; Rotating Electrical Machines-Part 2-1: Standard Methods for Determining Losses and Efficiency from Tests (Excluding Machines for Traction Vehicles). International Electrotechnical Commission: Geneva, Switzerland, 2014.
40. Mohamed, N.; Aymen, F.; Mouna, B.H.; Lassaad, S. Modeling and simulation of vector control for a Permanent Magnet Synchronous Motor in electric vehicle. In Proceedings of the International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Alkhobar, Saudi Arabia, 6–8 December 2021. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Analysis of Electromagnetic Characteristics of Outer Rotor Type BLDC Motor Based on Analytical Method and Optimal Design Using NSGA-II

Tae-Seong Kim [1], Jun-Won Yang [1], Kyung-Hun Shin [2], Gang-Hyeon Jang [3], Cheol Han [4] and Jang-Young Choi [1,*]

[1] Department of Electrical Engineering, Chungnam National University, Daejeon 34134, Republic of Korea; myxotjd@o.cnu.ac.kr (T.-S.K.); dws5523@o.cnu.ac.kr (J.-W.Y.)
[2] Department of Electrical Engineering, Changwon National University, Changwon 51140, Republic of Korea; kshin@changwon.ac.kr
[3] Convergence Technology Lab, Korea Electric Power Research Institute, Daejeon 34056, Republic of Korea; gh.jang@kepco.co.kr
[4] Advance E&E BLDC Motor Team Hanonsystems, Daejeon 34325, Republic of Korea; chan4@hanonsystems.com
* Correspondence: choi_jy@cnu.ac.kr

Abstract: This study investigates the electromagnetic analysis and optimal design of outer rotor type brushless DC (BLDC) motors for fan filter applications. The primary objective is to develop a method that integrates three-dimensional (3D) structural effects with efficient two-dimensional (2D) equivalent analysis. This study proposes a 2D equivalent analysis method that addresses the unique features of outer rotor type BLDC motors, particularly the permanent magnet (PM) overhang structure. This approach transforms the operating point on the B–H curve to facilitate accurate modeling in a 2D framework, overcoming traditional analysis limitations. An analytical method using spatial harmonics is introduced to derive essential electromagnetic quantities, namely flux linkage and back electromotive force (EMF). The method compensates for slot effects using the Carter coefficient, ensuring precise evaluation of circuit parameters and electromagnetic losses. To optimize motor performance, a multi-objective optimization technique is implemented using the Non-dominated Sorting Genetic Algorithm-II (NSGA-II), aiming to maximize both efficiency and power density. The research validates the proposed analytical approach against the finite element analysis method (FEM) results to confirm its accuracy.

Keywords: overhang structure; analytical method; finite element analysis method; NSGA-II; Pareto front

1. Introduction

The transition in power source operation methods aimed at reducing greenhouse gas emissions is a key factor driving new developments in industrial sectors. Consequently, industrial components are being electrified. Brushless DC (BLDC) motors have become increasingly prominent among industrial electric machines owing to their benefits over mechanically commutated DC motors, including greater operational reliability and service life, decreased need for maintenance, and enhanced performance consistency [1–7].

Outer rotor type BLDC motors have a rotor that is positioned outside the stator, providing more space to accommodate magnets than inner rotor type BLDC motors. This allows for the use of ferrite permanent magnets (PMs), which have the advantages of a

lower cost and abundant reserves, as opposed to rare-earth PMs, which have significant drawbacks such as supply monopolies and high prices [8].

Outer rotor type BLDC motors are commonly designed with a PM overhang to suppress leakage flux originating from the end-region of the stator coils. In addition, the rotor yoke can be accommodated on the outer surface, allowing for a reduction in the rotor yoke thickness. However, because the motor has an inherently three-dimensional (3D) structure that cannot be fully represented in two dimensions, conventional two-dimensional (2D) analysis, which assumes that the axial lengths of the stator and rotor are identical, fails to provide accurate predictions. Owing to this limitation, 3D analysis is essential; however, it has the drawback of a significantly increased analysis time. Therefore, an equivalent analysis method that can account for the 3D structure within a 2D analysis is required [9–14].

The initial design and analysis methods for permanent magnet synchronous motors (PMSMs) primarily include the finite element analysis method (FEM), which utilizes numerical analysis, and the analytical method, which employs the space harmonic method. The FEM performs simulations using specialized software and provides highly accurate results. However, its accuracy and efficiency depend on the specific software used and the experience level of the designer. In contrast, the analytical method analyzes the electromagnetic characteristics of PMSMs based on Maxwell's equations. This method employs Fourier-series-based magnetization modeling and derives partial differential equations from Maxwell's equations and the magnetic vector potential. By applying boundary conditions across the different domains of the machine, the resulting solution provides the flux density variation in the air gap. The analytical approach offers relatively high accuracy and significantly faster analysis times compared with the FEM, making it an efficient alternative for electromagnetic analysis [15,16].

By design, outer rotor type machines necessarily exhibit an increased rotor yoke volume and, therefore, a distinctly 3D flux path that cannot be captured by purely 2D approaches. To date, very few studies combine an analytic treatment of this extra yoke thickness with a full characteristic analysis of performance and saturation. Our work fills this gap by introducing a B–H operating-point transformation to compensate for the additional magnetic energy stored in the 3D yoke and validating its accuracy through both 3D FEM comparison and experimental data. This dual analytic–experimental framework is, to our knowledge, novel and essential for faithfully modeling outer rotor type topologies.

For most PMSMs, an optimal design is essential for improving various objective functions, including efficiency, power density, torque ripple, cogging torque, and PM usage. Among these, multi-objective optimization, which involves multiple objective functions, typically selects parameters that exhibit an inverse relationship. The optimal design methods for PMSMs can be broadly categorized into surrogate model-based and probabilistic optimization algorithms. Surrogate model-based optimization constructs an approximation model that reduces the analysis time compared with directly linking the algorithm with the simulation tool. The time required for the design process can be significantly reduced by utilizing an approximation model. However, because surrogate models are approximations of the actual objective function, discrepancies between the model and real motor characteristics are inevitable [17]. Probabilistic optimization algorithms require longer analysis times but possess global optimization capabilities, allowing them to find optimal solutions even for nonlinear characteristics. Because electrical machines inherently exhibit nonlinear behavior, employing global optimization techniques is necessary to achieve a valid and effective optimal design [18,19].

This paper proceeds as follows. First, the 2D equivalent model of the PM overhang structure is formed by adjusting the operating point of the PM on the B–H curve. In

addition, to prevent rotor yoke flux saturation in the 2D model, the yoke thickness is increased accordingly. Second, to simplify the analysis model, a slotless assumption is made, and an analytical method utilizing the space harmonic method is employed to derive the rotor flux linkage, back electromotive force (EMF), and stator armature reaction. The slot effect is compensated for using the Carter coefficient. Third, the efficiency and power density, which exhibit an inverse relationship, are selected as the objective functions. The optimization is then performed using a genetic algorithm, which is a probabilistic optimization technique. Finally, the optimized 2D equivalent model is converted into a 3D model with the same performance, followed by prototyping and experimental validation. The experimental prototype was manufactured by Shinsung ENG (Gwacheon-si, Republic of Korea) and is currently utilized in industrial applications in Korea. The comparative 3D finite element analysis was performed using Ansys Electronics Desktop (EM) 2023 R2 (Ansys Inc., Canonsburg, PA, USA). This study confirms the effectiveness of the proposed approach through a comparison between experimental outcomes and the predictions of the optimized model.

2. Two-Dimensional Equivalent Analysis Method Using Operating Point Conversion

The magnetic flux generated by the PMs in a PMSM is linked to the stator windings and flows into the core. The PM overhang structure plays a role in increasing this flux linkage. With increased flux linkage, flux saturation within the stator teeth intensifies. The flux density is higher in the presence of the overhang structure. This phenomenon is illustrated in Figure 1.

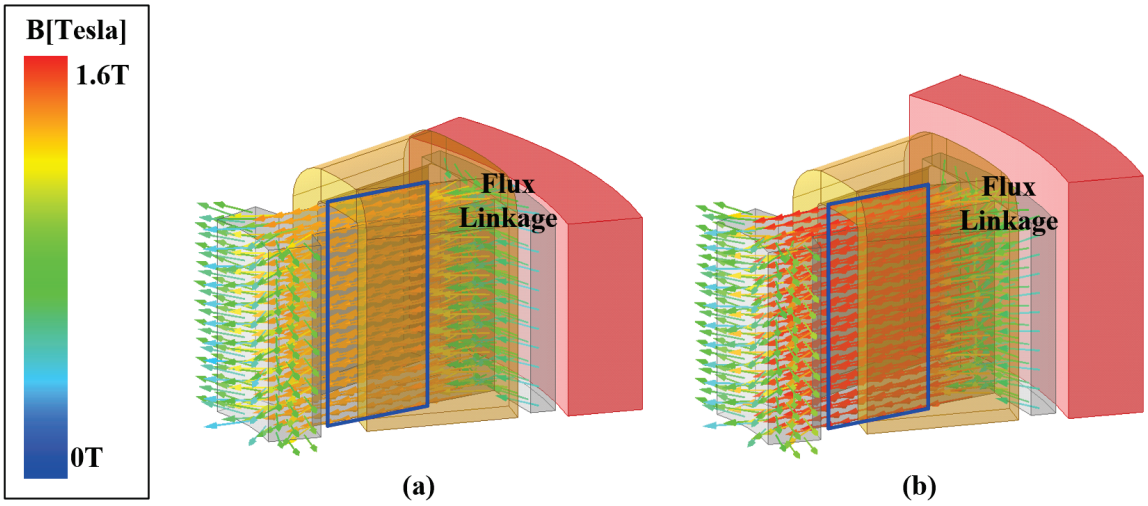


Figure 1. Stator core flux linkage: (**a**) general model and (**b**) model with overhang structure applied.

Therefore, the overhang structure is commonly employed in outer rotor type motors because it enables accurate analysis through 3D simulations. However, 3D FEM requires a long analysis time. To overcome this limitation, this study proposes a 2D equivalent analysis method that transforms the operating point of the PM on the B–H curve. Figure 2 illustrates the process of the proposed analysis method based on this operating point conversion.

The proposed method considers the PM overhang structure in 2D analysis by increasing the energy product of the PM material in the 2D-equivalent model by an amount that corresponds to the overhang structure. This ensures that the magnetic field energy in the proposed 2D analysis is equivalent to that in the 3D analysis. This process is represented as follows:

$$W_m = \frac{1}{2} B_m H_m V = \frac{1}{2} B'_m H'_m V' \quad (1)$$

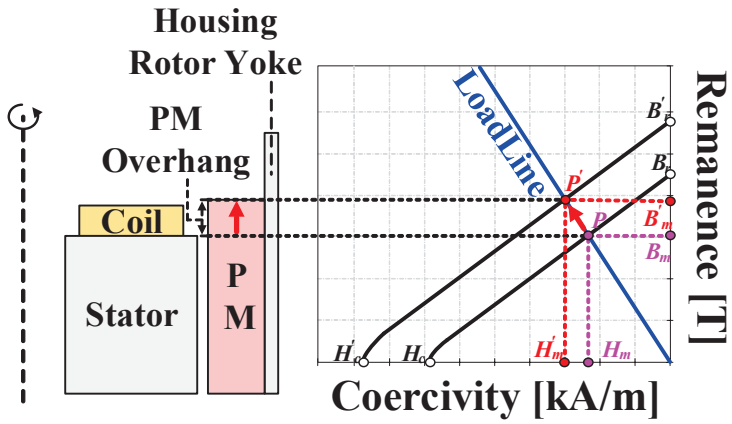


Figure 2. Operating point conversion process on the B–H Curve.

B_m and H_m represent the magnetic flux density and coercive force at the operating point of the permanent magnet, respectively, and V represents the volume of the permanent magnet. The left side is the three-dimensional analysis model, and the right side is the magnetic field energy of the proposed analysis model.

The material values of the proposed analysis model are used to derive the material properties of the 3D analysis model. For this purpose, the load line gradient must be determined, which is expressed as:

$$PC = \frac{B_m}{\mu_0 |H_m|} = \frac{1}{f_{LKG}} \frac{l_{PM}}{g'} \frac{A_g}{A_m} \tag{2}$$

where PC, μ_0, f_{LKG}, l_{PM}, g', A_g, and Am represent the load line gradient, permeability of vacuum, leakage factor, PM thickness, effective air gap length, flux-conducting area of the air gap, and cross-sectional area of the PM, respectively. In addition, the flux density at the operating point of the PM is derived as follows:

$$B_m = \frac{PC}{\mu_{rec} + PC} B_r \tag{3}$$

where μ_{rec} and $\boldsymbol{B_r}$ represent the relative permeability and remanence, respectively. Consequently, the flux density and coercive force at the operating point of the PM of the proposed analysis model are derived as follows:

$$B'_m = \sqrt{-\mu PC \frac{V}{V'} B_m H_m} \tag{4}$$

$$H'_m = -\frac{B'_m}{\mu_0 PC} \tag{5}$$

The remanence can be obtained by an inverse calculation using Equation (3). The PM data derived using the proposed analysis method are listed in Table 1.

Table 1. PM material data of the 3D analysis model and the proposed analysis model.

3D Analysis Model			Proposed Model		
Parameter	**Value**	**Unit**	**Parameter**	**Value**	**Unit**
B_m	0.3726	T	B'_m	0.4328	T
B_r	0.42	T	B'_r	0.4868	T
H_m	−270	kA/m	H'_m	−313	kA/m
H_c	−305	kA/m	H'_c	−353	kA/m

Figure 3a depicts the flux density saturation of the integrated rotor yoke during the no-load analysis of the 3D model. The flux density saturation is high in the radial direction of the core along the ***z*-axis**, whereas the axial direction of the core remains mostly unsaturated. This is because the actual rotor flux distribution is more concentrated in the radial direction than in the axial direction, as shown in Figure 3b.

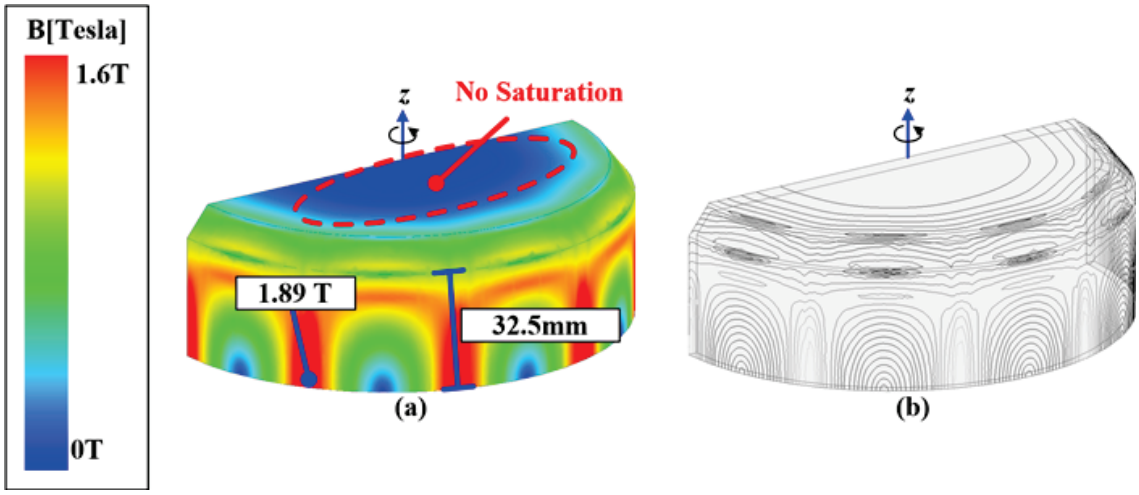


Figure 3. Integrated rotor yoke: (**a**) flux density saturation and (**b**) flux line distribution.

Therefore, the rotor yoke thickness of the proposed analysis model is determined based on the axial extent of the core, where the flux density saturation is high. Using this method, when the integrated rotor is equivalently modeled in a 2D analysis, the flux density saturation closely matches that of the 3D analysis model, as shown in Figure 4**.** In contrast, the conventional 2D model, which does not consider the increased rotor thickness, exhibits a higher level of rotor flux saturation compared to the 3D model. Furthermore, the effectiveness of the equivalent modeling technique is verified by comparing the rotor thickness and flux density shown in Table 2.

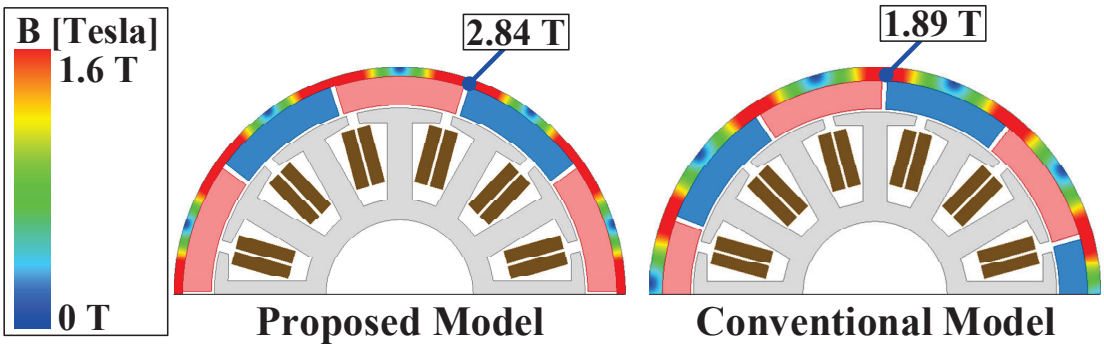


Figure 4. Flux density saturation of the proposed and conventional models.

Table 2. Three-dimensional model and rotor specifications of the proposed model.

3D Analysis Model			Proposed Model		
Parameter	**Value**	**Unit**	**Parameter**	**Value**	**Unit**
Rotor yoke	2.5	mm	Rotor yoke	3.8	mm
Flux density	1.89	T	Flux density	1.89	T

3. Electromagnetic Characteristic Analysis of a BLDC Motor

3.1. Electromagnetic Characteristic Analysis of the Motor Using Analytical Methods

3.1.1. No-Load Flux Linkage and Back EMF

Figure 5 shows the shape of the BLDC motor for fan filter applications, along with a simplified analysis model for the convenience of the analytical method.

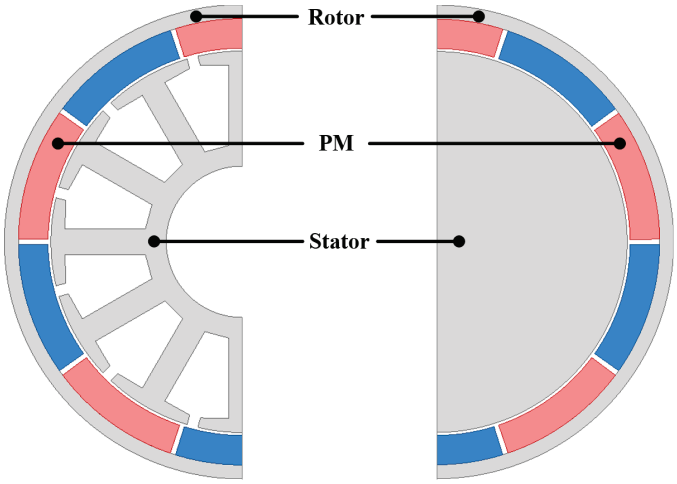


Figure 5. A simplified analysis model of an outer rotor type BLDC motor.

The following assumptions are necessary to evaluate the electromagnetic performance of the simplified model using the analytical method:

- The magnetic effects caused by the stator slots are corrected using the Carter factor.
- The stator and rotor cores are considered to have infinite magnetic permeability.
- The permeability of the PM is considered equal to the permeability of air.
- The stator current is modeled using a surface current density function.

Equation (6) represents the magnetization modeling of a PM, which is represented using a Fourier series for both the radial and tangential directions. Here, q corresponds to np, where n is the harmonic order and p represents the pole-pair number. Equations (7) and (8) represent the governing equations in the air gap (I) and PM (II) regions, respectively, as shown in Equation (7):

$$M = \sum_{n=1}^{\infty} \left\{ M_{rn} \cos(q\theta) \vec{i_r} + M_{\theta n} \sin(q\theta) \vec{i_\theta} \right\} \tag{6}$$

$$\nabla^2 A_I = 0 \tag{7}$$

$$\nabla^2 A_{II} = -\mu_0 (\nabla \times M) \tag{8}$$

Region I has no magnetization components; therefore, it can be expressed by the Laplace equation, whereas Region II contains magnetization components and is described by the Poisson equation. Solving the equations for each region leads to the derivation of the magnetic vector potential, as shown in Equations (9) and (10). Using the magnetic flux density expression in Equation (11), this formulation is further refined into Equations (12)–(15).

$$A_I = \sum_{n=1}^{\infty} \left\{ C_1 r^q + D_1 r^{-q} \right\} \sin(q\theta) \tag{9}$$

$$A_{II} = \sum_{n=1}^{\infty} \left\{ C_2 r^q + D_2 r^{-q} + \frac{-j\mu_0 r^q M_n}{(1-q^2)} \right\} \sin(q\theta) \tag{10}$$

$$\nabla \times \mathbf{A} = \mathbf{B} \tag{11}$$

$$B_{rn}^{I} = \frac{q}{r}\left[C_1 r^q + D_1 r^{-q}\right]\cos(q\theta) \tag{12}$$

$$B_{rn}^{II} = \frac{q}{r}\left[C_2 r^q + D_2 r^{-q} + \frac{\mu_0 r^q M_n}{q^2-1}\right]\cos(q\theta) \tag{13}$$

$$B_{\theta n}^{I} = -q\left[C_1 r^{q-1} - D_1 r^{-(q+1)}\right]\sin(q\theta) \tag{14}$$

$$B_{\theta n}^{II} = -q\left[C_2 r^{q-1} - D_2 r^{-(q+1)} + \frac{\mu_0 M_n}{q^2-1}\right]\sin(q\theta) \tag{15}$$

The boundary conditions between each region in Figure 6 can be identified from Equations (16)–(18). The unknown coefficients can be determined by substituting the magnetic flux density equations for each region.

$$r = R_I \rightarrow B_{\theta n}^{I} = 0 \tag{16}$$

$$r = R_{PM} \rightarrow B_{rn}^{I} = B_{rn}^{II}, B_{\theta n}^{I} + \mu_0 M_\theta = B_{\theta n}^{II} \tag{17}$$

$$r = R_r \rightarrow B_{\theta n}^{II} = \mu_0 M_\theta \tag{18}$$

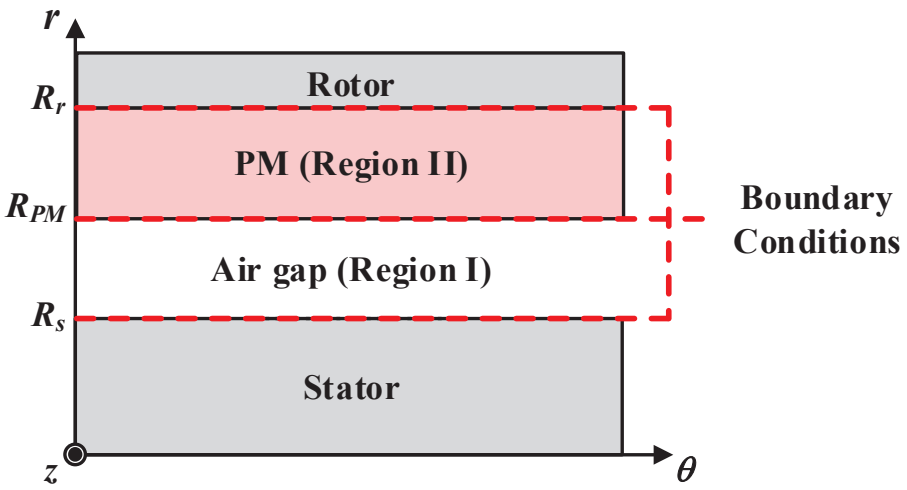


Figure 6. Shape equivalence and boundary conditions for analytical methods.

The flux linkage is calculated through the integration of the flux density across the air gap over one coil pitch. The EMF generated by the PM can be expressed based on Faraday's law. As shown in Figure 7, the derived results closely match the FEM and experimental results, thereby validating the proposed 2D equivalent and analytical methods.

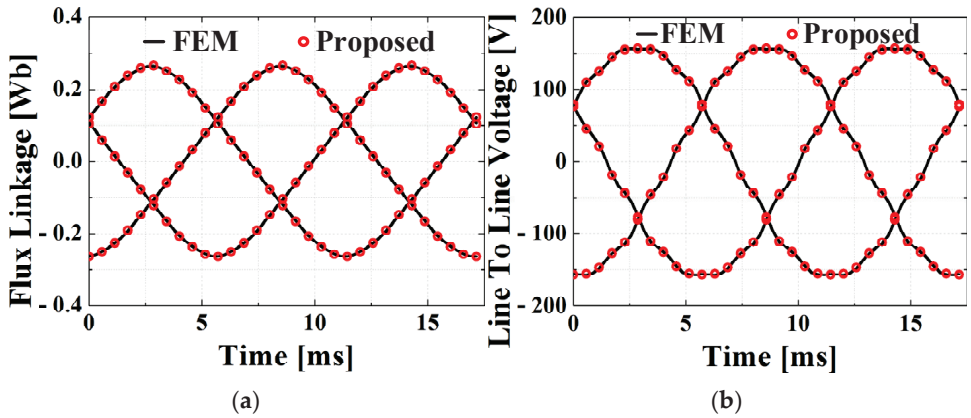


Figure 7. By PMs: (**a**) flux linkage and (**b**) line-to-line EMF.

3.1.2. Stator Armature Reaction

The magnetic field characteristics caused by the armature reaction of the stator winding are essential for predicting the inductance, which is a key parameter in the equivalent circuit of an electric motor. To derive these characteristics analytically, the magnetization components of the PM are assumed to be absent. Equation (19) expresses the current density distribution on the stator surface using a Fourier series. As with the expressions obtained for the PM case, the magnetic induction due to the armature effect is derived based on the magnetic vector potential, as presented in Equations (20) and (21).

$$J = \sum_{n=1}^{\infty} [J_{an} \cos(n\theta) + J_{bn} \sin(n\theta)] \tag{19}$$

$$B_{r\mathrm{I}}^{I} = -\frac{n}{r} \left[\left(A_{1\mathrm{I}} r^{n} + B_{1\mathrm{I}} r^{-n} \right) \sin(n\theta) - \left(C_{1\mathrm{I}} r^{n} + D_{1\mathrm{I}} r^{-n} \right) \cos(n\theta) \right] \tag{20}$$

$$B_{\theta\mathrm{I}}^{I} = -n \left[\left(A_{1\mathrm{I}} r^{n-1} - B_{1\mathrm{I}} r^{-n-1} \right) \cos(n\theta) + \left(C_{1\mathrm{I}} r^{n-1} - D_{1\mathrm{I}} r^{-n-1} \right) \sin(n\theta) \right] \tag{21}$$

The boundary conditions between the regions can be identified from Equations (22) and (23). The unknown coefficients can be determined by substituting the magnetic flux density equations for each region.

$$r = R_s \rightarrow B_{\theta 1}^{I} = -\mu_0 J \tag{22}$$

$$r = R_i \rightarrow B_{\theta 1}^{I} = 0 \tag{23}$$

3.1.3. Carter Coefficient

The electric motor has slots that result in leakage flux. As the equivalent air gap increases, the alteration in the magnetic flux distribution across the gap leads to variations in electromagnetic characteristics such as the back EMF. The slot effect can be considered by applying the Carter coefficient to the magnetic flux in the air gap, as presented in Equations (24)–(26).

$$\gamma = \frac{\tau}{g} \left(\frac{B_{so}\tau}{\tau} - \frac{4g}{\pi r} \log \left(1 + \frac{\pi B_{so}}{4g} \right) \right) \tag{24}$$

$$K_c = \frac{\tau}{\tau - \gamma g} \tag{25}$$

$$g_{eff} = K_c g \tag{26}$$

In the above equations, B_{so}, τ, g, and g_{eff} represent the slot opening, slot pitch angle, and the nominal and effective air gap lengths, respectively. The equivalent air gap length is corrected by applying the Carter factor to compensate for the influence of stator slotting.

3.2. Circuit Constant

The flux linkage for an individual coil is calculated by integrating the flux density across the air gap over one coil pitch. The analysis model consists of a 10-pole, 12-slot concentrated winding with a coil pitch of 1. The air gap flux density and flux linkage are expressed with respect to the rotor angle in the following equations.

$$B(r, \boldsymbol{\theta}) = \sum_{n} B_n \cos(\boldsymbol{\theta} - w_r t) \tag{27}$$

$$\boldsymbol{\lambda}_{ph} = N_c R_s L_{stk} B_n \int_{pos}^{neg} \cos q(\boldsymbol{\theta} - w_r t) \tag{28}$$

Here, N_c represents the number of turns in one phase, R_s represents the stator outer diameter, L_{stk} represents the axial length, and B_n represents the magnetic flux density.

The back EMF is obtained by differentiating the flux linkage of the winding with respect to time and is expressed as follows:

$$E_{ph} = -\frac{d\boldsymbol{\lambda}_{ph}}{dt} = -N_c R_s L_{stk} B_n w_r \int_{pos}^{neg} \sin q(\boldsymbol{\theta} - w_r t) \tag{29}$$

The method for calculating the conductor length is shown in Figure 8, and the resistance per phase of the winding is determined using the following equations:

$$l_{endturn} = 2\pi r \tag{30}$$

$$l_a = N_{ph} N_{turn} (2 \cdot l_{axial} + l_{endturn}) \tag{31}$$

$$R_{ph} = \rho_c \frac{l_a}{A_c N_{sn} N_{parrallel}^2} \tag{32}$$

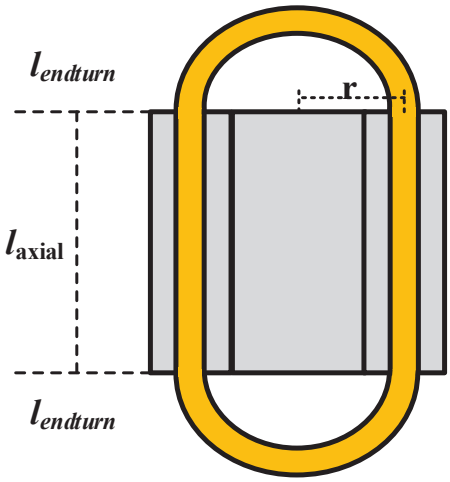


Figure 8. Winding length calculation method.

$l_{endturn}$ indicates that the end turns at both ends of the stator are assumed to be circles, where N_{ph} represents the equivalent number of slots, N_{turn} is the number of turns, ρ_c denotes the resistivity, A_c represents the area of one conductor's cross section, N_{sn} is the number of strands per conductor, $N_{parallel}$ is the number of parallel circuits, and l_a is the total length of the equivalent coil.

The inductance, which represents the flux linkage per unit current, can be determined using the armature reaction field. It consists of self-inductance and mutual inductance and can be expressed by the following equation:

$$L_{g_self} = \frac{\lambda_A}{I_A} = \sum_{n=1}^{coil} \int_{\theta_{n_pos}}^{\theta_{n_neg}} B_{\mathrm{aI}} d\theta \tag{33}$$

$$L_{g_mutual} = \frac{\lambda_A}{I_B} = \sum_{n=1}^{coil} \int_{\theta_{n_pos}}^{\theta_{n_neg}} B_{\mathrm{bI}} d\theta \tag{34}$$

3.3. Electromagnetic Loss

3.3.1. Core Loss

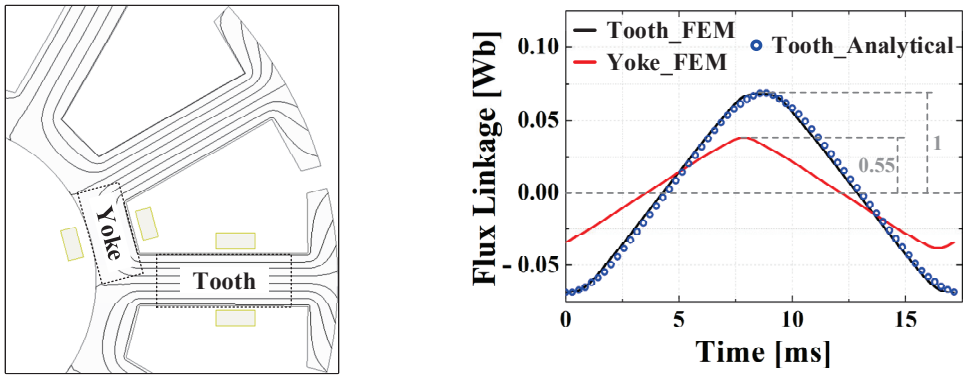
The core loss constitutes a significant portion of the electromagnetic losses in electric motors and must be carefully considered. The core loss consists of the hysteresis loss, which arises from variations in magnetic field intensity and flux reversal within the steel; the classical eddy current loss, resulting from the alternating magnetic flux and induced voltage in the steel sheets; and the abnormal eddy loss, which is attributed to the magnetic

domain thickness and externally applied mechanical stress on the electrical steel sheets. The core loss can be estimated using the Steinmetz equation, derived from the experimentally obtained coefficients for each type of core loss.

$$P_{core} = P_h + P_c + P_e = K_h f(B_m)^2 + K_c(fB_m)^2 + K_e(fB_m)^{1.5} \tag{35}$$

where K_h, K_c, and K_e correspond to the hysteresis loss, classical eddy current loss, and excess loss coefficients, respectively, which are derived from the experimental data provided by the silicon steel manufacturer. f and B_m denote the frequency of the magnetic flux density and the m-th harmonic, respectively. In the motor, the stator core is formed by stacking thin sheets of steel; thus, the eddy current losses are minimal, and the hysteresis loss constitutes the dominant portion of the core loss. The total core loss in the stator core can be derived by calculating the core loss in a single tooth and yoke and then multiplying it by the number of slots. The core loss in the tooth is derived by evaluating the integral of the air gap flux density over the angular section occupied by the tooth and then multiplying it by the lamination length to obtain the flux crossing the tooth. An FEM-based approach is required to derive the flux crossing the yoke analytically.

The flux crossing the teeth and yoke has a consistent ratio depending on the pole slot combination of the motor. In this study, this ratio (TY-ratio) is determined based on the ratio of peak magnetic flux magnitudes through the teeth and yoke. The model used in this study, which has 10 poles and 12 slots, shows that the flux density in the teeth is higher than that in the yoke, as shown in Figure 9.



(**a**) (**b**)

Figure 9. Initial model: (**a**) flux diagram in the stator core and (**b**) comparison of the results of deriving flux linkage using analytical method and FEM.

In this model, the flux crossing the stator yoke is obtained by multiplying the flux crossing the stator tooth by the TY ratio using an analytical method. The flux crossing the stator tooth and yoke, depending on the rotor position, can be derived using the following equations:

$$B_t = \sum_{n=1}^{\infty} \frac{R_s}{T_w} \int_0^{\theta_t} B_n \cos q(\theta - w_r t) d\theta \tag{36}$$

$$B_y = TYratio \cdot \sum_{n=1}^{\infty} \frac{R_s}{T_w} \int_0^{\theta_t} B_n \cos q(\theta - w_r t) d\theta \tag{37}$$

where R_s and T_w represent the stator radius and tooth thickness, respectively, while θ_t is the angle subtended by a single stator tooth. The total core loss in the stator core can be calculated based on the derived flux crossing the teeth and yoke. As shown in Table 3, the method for deriving the core loss in this section is valid.

Table 3. Core loss analysis results.

Parameter	Analytical	FEA
Core loss [W]	3.50	3.70

3.3.2. Copper Loss

When a current is applied to the stator winding of a PM motor, heat is generated owing to the resistance, which is referred to as copper loss. The copper loss can be broadly classified into AC and DC copper losses. AC copper loss occurs because of the skin effect. As the frequency of the current increases, the current in the center of the winding is offset, and the current density on the surface of the winding increases. This results in an increase in the effective resistance. However, because the model in this study does not operate at high rotational speed and the strand conductor diameter is small, this effect is not considered. The DC copper loss is determined using the RMS value of the current flowing through the winding and the phase resistance of the winding, as shown below. The analytical results are presented in Table 4.

$$P_{copper} = 3 \cdot I_{rms}^2 \cdot R_{ph} \tag{38}$$

where I_{rms} denotes the effective value of the phase current, and R_{ph} represents the resistance of a single phase winding

Table 4. Copper loss analysis results.

Parameter	Analytical
Copper loss [W]	7.78

4. Optimal Design Using Multi-Objective Genetic Algorithm

4.1. Initial Design Specifications of the Outer Rotor BLDC Motor

The torque of the motor consists of magnetic and reluctance torques. In the case of an SPMSM without saliency, only magnetic torque is utilized. The magnetic torque is defined as the product of magnetic flux loading and electric current loading, where the magnetic flux loading is affected by the characteristics of the PMs. The corresponding design variables of the rotor and PMs are shown in Figure 10a. In this case, T_r and T_m represent the thicknesses of the rotor yoke and PM, respectively, whereas θ_τ and θ_{em} indicate the pole pitch and pole arc ratio, respectively. R_{ro} denotes the outer radius of the rotor. Once the specific magnetic loading is determined, the specific electric loading must be designed. The maximum phase current required to satisfy the required torque and number of turns in the stator winding are derived using the parameters influenced by the selected specific magnetic loading, as shown in Equations (38) and (39). In these equations, E_{max} and Φ_{max} represent the back electromotive force and flux linkage under no-load conditions, respectively, while we denotes the electrical angular velocity. I_{max} refers to the maximum phase current, and T_{req} represents the required torque. The maximum values of the back EMF and applied power are expressed as follows:

$$E_{\max} = \phi_{\max} w_e N_c \tag{39}$$

$$I_{\max} = \frac{T_{req}}{1.5 p N_c \phi_{\max}} \tag{40}$$

The slot area is determined by calculating the stator tooth and yoke thickness based on the flux density saturation limit of the teeth. The corresponding design variables of the stator are presented in Figure 10b, where T_t and T_y represent the thicknesses of the stator tooth and yoke, respectively. *Hs0*, *Hs1*, and *Hs2* denote the lengths of shoe 1, shoe 2, and

the tooth, respectively. *Bs1*, *Bs2*, and θ_{so} indicate the upper and lower widths of the slot and slot opening angle, respectively. In addition, R_{si} and R_{so} refer to the inner and outer radii of the stator, respectively.

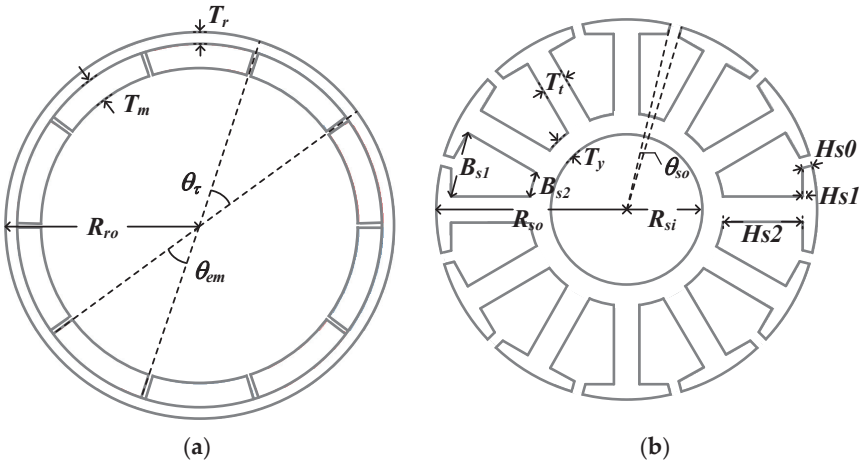


Figure 10. Outer rotor type motor design: (**a**) rotor and (**b**) stator.

4.2. Optimal Design Process of Outer Rotor Type BLDC Motor

In the input stage, the outer diameter of the motor, stator flux density saturation, current density, fill factor, and rated power are considered constraints. The material of each structure, pole slot combination, air gap length, and coil pitch are treated as fixed variables. These selected constraints and fixed variables are presented as initial conditions.

In this optimization process, a population consisting of random combinations of design variables within the specified rotor shape variable range of the Non-dominated Sorting Genetic Algorithm-II (NSGA-II) is generated. Subsequently, the performance of each design variable combination is evaluated using a simulation model, and the objective function values are calculated. In this study, the power density and efficiency are used as objective functions. Individuals with better performance according to the evaluation function are chosen, and next-generation populations are created via genetic operations such as crossover and mutation. The algorithm proceeds through a fixed number of evolutionary cycles, progressively identifying better solutions in each cycle. Ultimately, a Pareto front, which is a set of Pareto optimal solutions showing various trade-off relationships, is derived.

In the modeling stage, the stator design process is performed according to the rotor specifications proposed by the NSGA-II. The performance of the designed motor model is evaluated, and the results are fed back into the NSGA-II. Assessment is conducted using the fitness, which is determined by the values of the objective functions. If the requirements are not met, a penalty is applied to the fitness, prompting the NSGA-II to recognize the model as unsuitable.

In this study, the multi-objective optimization was performed using the Non-dominated Sorting Genetic Algorithm II (NSGA-II). The population size for each generation was set to 100, and the optimization process was carried out for 20 generations, which served as the stopping criterion. The crossover rate was fixed at 0.9, while the mutation rate was set to 0.01. These parameter values were selected based on commonly adopted practices in multi-objective optimization studies and were found to provide a good balance between convergence speed and solution diversity in preliminary tests. All optimization procedures were implemented using MATLAB R2023b, and the same random seed was used for reproducibility.

In the optimization stage, the optimal model is selected from the Pareto front derived in the previous steps. The Pareto front represents the set of all optimal solutions, and the optimal point is selected based on the priority and importance of the design objectives [20].

In our approach, we first define all design constraints (outer diameter, stator flux-density saturation limit, current density limit, slot fill factor, rated power) and fixed variables (material selections, pole slot combination, air gap length, coil pitch). NSGA-II then generates an initial population by randomly sampling design variables—rotor yoke thickness, magnet overhang, pole–arc ratio, axial length, etc.—within their allowable ranges. Each candidate is converted to the 2D equivalent model (Section 2) and analyzed (Section 3) to compute its power density and efficiency and to check constraint compliance. After evaluation, non-dominated sorting assigns Pareto fronts, and crowding-distance operations preserve diversity. Selected individuals undergo crossover and mutation to form the next generation. This loop continues until a preset number of generations is reached or no further improvement occurs. Finally, the converged Pareto front is presented, and the optimal design is chosen based on the relative importance of efficiency versus power density. The aforementioned optimization process is shown in Figure 11.

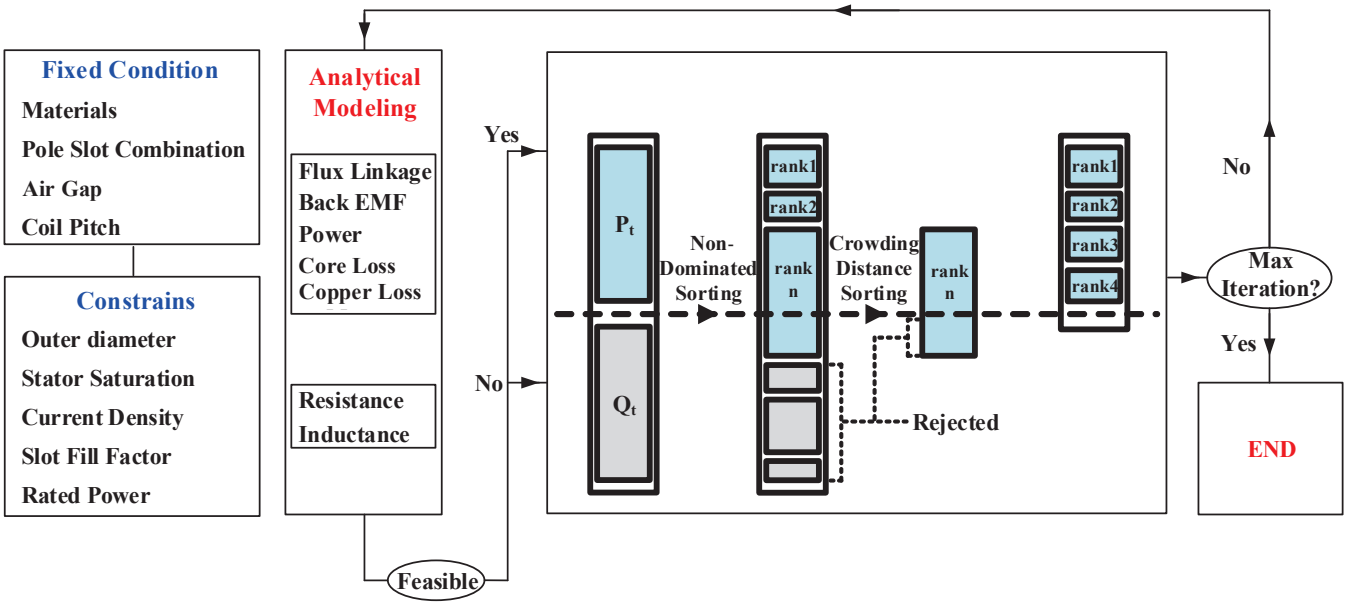
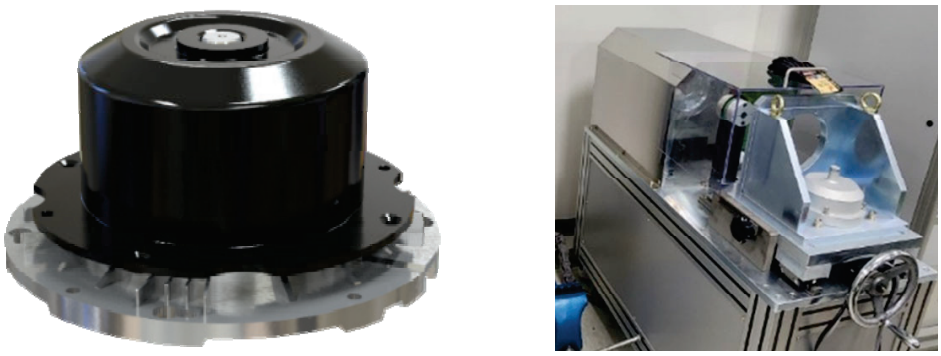


Figure 11. Optimal design flow chart of outer rotor type BLDC motor.

4.3. Optimized Design Strategy for Outer Rotor Type BLDC Motors

Figure 12 shows the manufactured model of the outer rotor type BLDC motor and configuration of the back EMF experimental setup. The experimental setup shown in Figure 12b consists of the manufactured outer rotor BLDC motor, a programmable DC power supply, and an inverter drive. Back EMF was measured under no-load conditions using an oscilloscope across two phases while the rotor was driven by an outer rotor type motor at a rated speed. The motor's mechanical speed was maintained at 1500 rpm and monitored using an optical encoder. Figure 13 compares the electromagnetic analysis results of the 2D equivalent analysis method and space harmonic method, proposed in this study through operating point transformation with the FEM analysis results and experimental data. In comparison to the experimental data, both the FEM and analytical method yield very similar results, suggesting the validity of the proposed approach.



(**a**) (**b**)

Figure 12. Outer rotor type BLDC motor: (**a**) manufactured model and (**b**) experimental set.

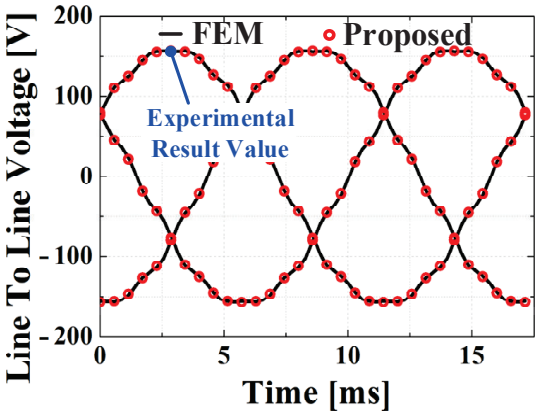


Figure 13. Comparison of the proposed method and experimental results using back EMF.

In addition to its accuracy, the proposed analytical method significantly reduces the computational burden. In our test environment (Intel i7 CPU, 128 GB RAM), a single evaluation using the analytical method took approximately 10 s, while a 2D FEM simulation under similar conditions required about 15 min. This corresponds to a reduction in computation time of over 95%. This improvement becomes especially important during the multi-objective optimization process using NSGA-II, where hundreds of design candidates are evaluated per generation. As such, the proposed method enables fast and practical optimization without compromising accuracy.

In this study, a multi-objective optimization technique was applied to optimize the efficiency and power density simultaneously, with the aim of finding a balance between these two performance metrics. As a result of deriving the Pareto front with the efficiency and power density as objective functions, it can be observed from Figure 14 that the two objective functions have an inverse relationship.

Efficiency and power density are considered key performance indicators in motor design. During the process of optimizing these two objective functions simultaneously, the weight setting plays a significant role in the final design results. The higher the power density, the smaller the size that is required to achieve the same power. This refers to the amount of power generated per unit volume of the motor. As motor power increases and the housing size decreases, the power density increases. Therefore, in this study, a higher weight was assigned to the power density to optimize the motor. L_{stk}, R_{ro}, and Thick PM represent the axial length, external radius of the motor, and PM thickness, respectively. Figure 15 compares the shapes of each initial model and the optimal model.

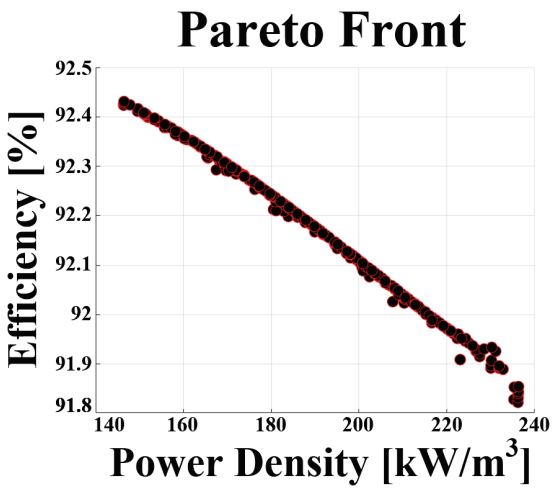


Figure 14. Optimal design results derived from Pareto front.

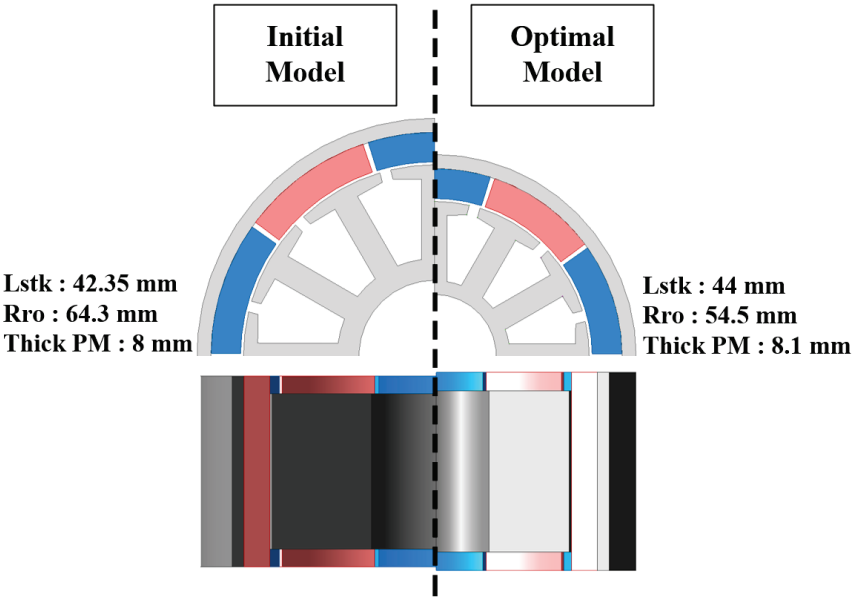


Figure 15. Comparison of initial and optimal design shapes.

Figure 16 shows the electromagnetic analysis results for the initial and optimal design models. While the efficiency remained at the same level, the power density improved by approximately 30.8% compared with the initial model. The material usage of the initial and optimal models is shown in Table 5. To verify the validity of the proposed optimal design model, a comparative analysis of the analytical method and FEM results was conducted. The results are presented in Table 6. Based on key performance indicators such as efficiency and power, the error between the two methods was found to be within 1%. Specifically, the power predicted by the analytical method was 104.23 W, whereas that calculated by the FEM was 104.87 W, with a difference of only approximately 0.61%. In terms of efficiency, the analytical method predicted 91.99% and the FEM predicted 91.24%, showing a minor difference of approximately 0.82%. In conclusion, the high similarity between the proposed analytical method and FEM supports the validity and practicality of the proposed optimal design model. This method is expected to be useful for the future design and optimization of related fields.

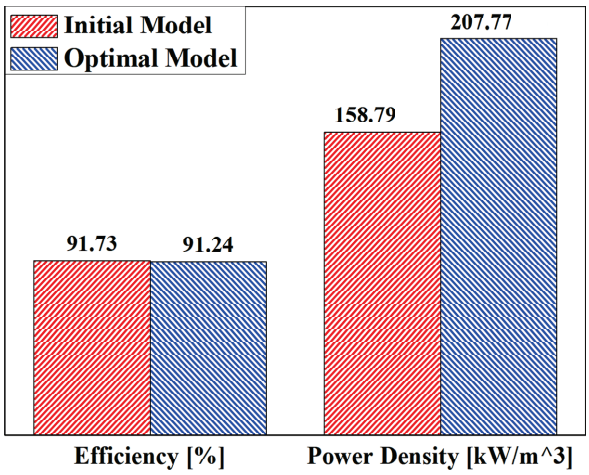


Figure 16. Comparison of initial and optimal design electromagnetic analysis results.

Table 5. Comparison of material usage between initial and optimal models.

Material	Unit	Initial Model Usage	Optimal Model Usage
Core (35H440)	kg	2.74	2.74
PM (FB-5BE)	kg	0.58	0.5

Table 6. Comparison of electromagnetic analysis results between the proposed method and the optimal model of FEM.

Parameter	Unit	Proposed	FEM	Error
L_{stk}	mm	44		-
R_{ro}	mm	54.5		-
Thick PM	mm	8.1		-
I_{max}	A	0.559		-
Power	W	104.23	104.87	0.61%
Efficiency	%	91.99	91.24	0.82%

5. Conclusions

This study has proposed an analytical method using 2D equivalent analysis and the space harmonic method, considering the 3D structure of outer rotor type BLDC motors, as well as an optimal design method using the NSGA-II.

To account for the commonly applied 3D structure of the PM overhang in outer rotor type BLDC motors, a 2D equivalent analysis method using an operating point transformation was introduced. In addition, to model the integrated rotor yoke equivalently, the distribution of the actual rotor flux was considered, and the rotor yoke for the 2D model was selected. This approach addresses the issue of long analysis times in 3D analysis.

Next, the electromagnetic characteristics of the outer rotor type BLDC motor were analyzed using the space harmonic method. The flux linkage, back EMF, and air gap flux density owing to the stator–armature reaction were derived.

Finally, multi-objective optimization design was carried out using the NSGA-II. The power density and efficiency were set as objective functions, and the Pareto front was derived to select the optimal model.

The experimental results of the manufactured model were compared with the FEM analysis results to validate the proposed method. The proposed method was confirmed to be highly accurate. This implies that the proposed methods for analysis and optimization are effective tools in the development of outer rotor type BLDC motors.

Author Contributions: Conceptualization, T.-S.K., J.-W.Y., K.-H.S., G.-H.J. and J.-Y.C.; methodology, T.-S.K.; validation, T.-S.K.; writing, T.-S.K.; writing—review and editing, J.-W.Y., K.-H.S., G.-H.J., C.H. and J.-Y.C.; supervision, K.-H.S. and J.-Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korean government (MOTIE) (20223A10100040, development of design technology for converting retired thermal plants into synchronous condenser plants).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The author Cheol Han was employed by the Advance E&E BLDC Motor Team Hanonsystems. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Shao, J. An Improved Microcontroller-Based Sensorless Brushless DC (BLDC) Motor Drive for Automotive Applications. *IEEE Trans. Ind. Appl.* **2006**, *42*, 1216–1221. [CrossRef]
2. Pillay, P.; Krishnan, R. Application characteristics of permanent magnet synchronous and brushless DC motors for servo drives. *IEEE Trans. Ind. Appl.* **1991**, *27*, 986–996. [CrossRef]
3. Park, S.J.; Park, H.W.; Lee, M.H.; Harashima, F. A new approach for minimum-torque-ripple maximum-efficiency control of BLDC motor. *IEEE Trans. Ind. Electron.* **2000**, *47*, 109–114. [CrossRef]
4. Damiano, A.; Floris, A.; Fois, G.; Marongju, I.; Porru, M.; Serpi, A. Design of a High-Speed Ferrite-Based Brushless DC Machine for Electric Vehicles. *IEEE Trans. Ind. Appl.* **2017**, *53*, 4279–4287. [CrossRef]
5. Bianchi, N.; Bolognani, S.; Luise, F. Analysis and Design of a PM Brushless Motor for High-Speed Operations. *IEEE Trans. Energy Convers.* **2005**, *20*, 629–637. [CrossRef]
6. Kim, H.K.; Hur, J. Dynamic Characteristic Analysis of Irreversible Demagnetization in SPM- and IPM-Type BLDC Motors. *IEEE Trans. Ind. Appl.* **2012**, *53*, 982–990. [CrossRef]
7. Kim, T.-H.; Ehsani, M. Sensorless Control of the BLDC Motors From Near-Zero to High Speeds. *IEEE Trans. Power Electron.* **2004**, *19*, 1635–1645. [CrossRef]
8. Liu, H.-C.; Kim, H.W.; Jang, H.K.; Jang, I.-S.; Lee, J. Ferrite PM Optimization of SPM BLDC Motor for Oil-Pump Applications According to Magnetization Direction. *IEEE Trans. Appl. Supercond.* **2020**, *30*, 1–5. [CrossRef]
9. Azzouzi, J.; Barakat, G.; Dakyo, B. Quasi-3-D Analytical Modeling of the Magnetic Field of an Axial Flux Permanenet-Magnet Synchronous Machine. In Proceedings of the IEEE International Electric Machines and Drives Conference, IEMDC'03, Madison, WI, USA, 1–4 June 2003; IEEE: Pictasaway, NJ, USA, 2003; Volume 3, pp. 1941–1947. [CrossRef]
10. Kim, K.-C.; Koo, D.-H.; Lee, J. The study on the overhang coefficient for permanent magnet machine by experimental design method. *IEEE Trans. Magn.* **2007**, *43*, 2483–2485. [CrossRef]
11. Ko, Y.-Y.; Song, J.-Y.; Seo, M.-K.; Han, W.; Kim, Y.-J.; Jung, S.-Y. Analytical method for overhang effect of surface-mounted permanent-magnet motor using conformal mapping. *IEEE Trans. Magn.* **2018**, *54*, 1–5. [CrossRef]
12. Woo, D.-K.; Lim, D.-K.; Yeo, H.-K.; Ro, J.-S.; Jung, H.-K. A 2-D finite-element analysis for a permanent magnet synchronous motor taking an overhang effect into consideration. *IEEE Trans. Magn.* **2013**, *49*, 4894–4899. [CrossRef]
13. Lee, T.-Y.; Seo, M.-K.; Kim, Y.-J.; Jung, S.-Y. Motor Design and Characteristics Comparison of Outer-Rotor-Type BLDC Motor and BLAC Motor Based on Numerical Analysis. *IEEE Trans. Appl. Supercond.* **2016**, *26*, 1–6. [CrossRef]
14. Shin, H.-S.; Shin, K.-H.; Jang, G.-H.; Cho, S.-K.; Jung, K.-H.; Choi, J.-Y. Experimental Verification and 2D Equivalent Analysis Techniques of BLDC Motor With Permanent Magnet Overhang and Housing-Integrated Rotor Core. *IEEE Trans. Appl. Supercond.* **2020**, *30*, 1–5. [CrossRef]
15. Kim, K.-H.; Park, H.-I.; Jang, S.-M.; Choi, J.-Y. Comparison of Characteristics of Double-Sided Permanent-Magnet Synchronous Motor/Generator According to Magnetization Patterns for Flywheel Energy Storage System Using an Analytical Method. *IEEE Trans. Magn.* **2015**, *51*, 1–4. [CrossRef]
16. Lee, H.K. Characteristic Analysis and Analytical Study of the Permanent Magnet Synchronous Motor According to Rotor Eccentricity. Master's Thesis, Chungnam National University, Daejeon, Republic of Korea, 2020.
17. Lim, D.-K.; Woo, D.-K.; Kim, I.-W.; Ro, J.-S.; Jung, H.-K. Cogging Torque Minimization of a Dual-Type Axial-Flux Permanent Magnet Motor Using a Novel Optimization Algorithm. *IEEE Trans. Magn.* **2013**, *49*, 5106–5111. [CrossRef]
18. Kim, W.H. Multi-Objective Optimal design of SPMSM for electric compressor using analytical method and NSGA-II algorithm. Master's Thesis, Chungnam National University, Daejeon, Republic of Korea, 2022.

19. Jo, S.T. Optimal Design of High-Speed Permanent Magnet Synchronous Generator for Supercritical Organic Rankine Cycle Power System. Master's Thesis, Chungnam National University, Daejeon, Republic of Korea, 2023.
20. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2002**, *6*, 182–197. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Review

Review of Automated Operations in Drilling and Mining

Athanasios Kokkinis [1], Theodore Frantzis [1,*], Konstantinos Skordis [1], George Nikolakopoulos [1,2] and Panagiotis Koustoumpardis [1]

[1] Robotics Group, Department of Mechanical Engineering and Aeronautics, University of Patras, 26504 Patras, Greece; up1072366@ac.upatras.gr (A.K.); up1097472@ac.upatras.gr (K.S.); geonik@ltu.se (G.N.); koust@upatras.gr (P.K.)

[2] Robotics and AI Team, Luleå University of Technology, 97187 Luleå, Sweden

* Correspondence: theodfrantzis@ac.upatras.gr

Abstract: Current advances and trends in the fields of mechanical, material, and software engineering have allowed mining technology to undergo a significant transformation. Aiming to maximize the efficiency and safety of the mining process, several enabling technologies, such as the recent advances in artificial intelligence, IoT, sensor fusion, computational modeling, and advanced robotics, are being progressively adopted in mining machine manufacturing while replacing conventional parts and approaches that used to be the norm in the rock ore extraction industry. This article aims to provide an overview of research trends and state-of-the-art technologies in face exploration and drilling operations in order to define the vision toward the realization of fully autonomous mining exploration machines of the future, capable of operating without any external infrastructure. As the trend of mining at large depths is increasing and as the re-opening of abandoned mines is gaining more interest, near-to-face mining exploration approaches for identifying new ore bodies need to undergo significant revision. This article aims to contribute to future developments in the use of fully autonomous and cooperative smaller mining exploration machines.

Keywords: drilling robots; mining; mining sensors; autonomous mining

1. Introduction

Rock ores have an integral role in product manufacturing. It is widely accepted that modern society's needs and ore extraction requirements are highly correlated. It is without a doubt that traditional methods, which require the participation of human personnel, will not be sufficient to satisfy production demand, since they have a high downtime and may cause harm to the personnel in hazardous environments [1,2]. It is also important to consider that the inefficiency of human-curated processes may cause harm to the environment. Lastly, it is becoming increasingly more important to reduce the overall environmental impact, whenever possible [3,4].

The aforementioned problems underline a need for a radical transformation in the way the mining industry undergoes exploration and production drilling, as well as extraction processes. This transformation can be observed through an already developing trend, as there is an increase in automation and the use of robots in these fields. Regarding software implementations, from human-assisted monitoring systems to remotely operated or fully autonomous mining operations, the boundaries of automation have substantially expanded within the mining engineering field. This article reviews automation in drilling and mining, focusing on trends and state-of-the-art technologies for face exploration and production drilling. In other words, this review focuses on fully autonomous, smaller, and cooperative mining machines.

The mapping of knowledge domains for automated operations in drilling and mining referenced in this article can be examined in Table 1. The rest of this article is structured as follows. Section 2 will briefly describe the current state of the mining industry, focusing on

the current challenges that it faces and how fully automated mining can be the main route to overcoming those challenges. Section 3 will provide an overview of drilling approaches, focusing on conventional methods used in the mining industry. Section 4 will present the state of the art in automated mining and intelligent systems, summarizing current advancements and research trends. In Section 5, the technological infrastructure utilized in automated mining will be presented, including various sensors and software crucial to these processes. Section 6 will touch upon potential advancements in intelligent mining systems, such as the integration of artificial intelligence (AI) and neural networks, aiming to enhance safety, efficiency, and precision in automated operations. Finally, Section 7 will conclude the article by sharing additional research visions, particularly emphasizing the development of infrastructure-less systems adaptable to various deep mine and abandoned mine situations.

Table 1. The mapping of knowledge domains for automated operations in drilling and mining.

Type	Topic	References
Drilling approaches	Cable tool drilling	[5,6]
	Auger drilling	[5–7]
	Rotary drilling	[5,6]
	Diamond drilling	[5,6]
	Directional drilling	[8]
Automated mining intelligent machines	Rotary Steerable Drilling Systems	[9–11]
	Longwall	[12–15]
	Automated hauling	[16]
	Charging robots	[16,17]
	Intelligent Rock-Drilling Jumbo and DTH Drill	[16]
Sensors in automated mining	LIBS	[18–20]
	LIDAR	[21–25]
	TFS	[26,27]
Localization systems	Light-based localization systems	[16,28,29]
	Daisy chaining of fiducial markers	[30,31]
	Magnetic induction	[21,32,33]
Communication systems	WiFi	[21,24]
	LoRaWAN	[21,34,35]
Digital model	DT	[36,37]
Software in intelligent systems	e-Drilling's software suite	[38,39]
	Draco	[40–42]
	Forestall	[43,44]
	TIMining Aware	[44,45]
Collaborative robotics	Caterpillar's MineStar Command for Underground Systems (AHS)	[46–49]
	EH-RemoteHeadControl v2 by Elgór-Hanses S.A.	[15]
	UGVs	[50]
	UAVs	[50–54]
Research about intelligent mining machines	Drill piston control	[55]
	Photogrammetry	[25,56–59]
	Neural networks	[60,61]
	VR simulation for multi-agent system action planning	[57,62]

Table 1. *Cont.*

Type	Topic	References
Research about intelligent mining machines	AI-assisted DTs	[63–65]
	Optimized haulage scheduling using RL	[66]
	Controlling equipment using DRL	[67,68]
	Using DRL in mine production scheduling	[69]
	Development and research for environmental monitoring	[70,71]
	Monitoring of the drilling rig	[72]
	Intelligent control of drilling	[72]
	MWD and machine learning for rock monitoring	[72]
	AI and big data technology in ventilation systems	[72]
	Intelligent rock stress monitoring	[72]
	Quantum computing for open-pit profile optimization	[73–75]

2. Problem Statement and Motivation

Assessing the current state of the mining industry, it can be observed that an important part of the operations is conducted sub-optimally, while the need for resource extraction is constantly increasing. In the meantime, the quality of ore grades located on the surface is decreasing, which leads to the extraction of ores from deeper depths [69]. Furthermore, the increasing focus on net-zero emission policies encourages mining companies to adopt more environmentally friendly policies that require the acquisition of different equipment and raises the need for more effective ways of mining [66,76–79]. Society's interest in worker health and safety also requires companies to protect their workers from extreme mining conditions, such as dust, dirt, poorly lit areas, and areas with a high potential for rock tumbling [78].

The current issues that the rock ore mining industry is facing can be resolved or at least mitigated by fully automating the processes of mining and drilling while utilizing intelligent systems [77,79]. Intelligent systems can increase efficiency if they are used for the scheduling and monitoring of the operation of fully autonomous machines such as hauling fleets. This efficiency can be translated into a reduction in machine wait time, a reduction in equipment energy consumption and resource usage, and the ability to adapt to random events that would normally cause disruptions in the ore extraction process [78]. Additionally, fully automating the mining industry can aid in the protection of workers, who can observe machines operating in hazardous areas such as blasting zones with the use of proper sensors from virtually identical environments (DTs) [77].

The fully automated and intelligent approaches discussed in this paper cannot be implemented immediately. There are plenty of specifications that need to be met in order for the proposed technologies to be mature enough to be implemented in raw material production. For instance, the intelligent methods that are currently being researched are characterized by high computational intensity and increasing learning time depending on the project's complexity [66]. Additionally, both fully automated processes and intelligent systems need to be robust, flexible, and adaptable, incorporating uncertainties that are a result of insufficient data or unpredictable events that might occur during the mining operation [78]. Furthermore, for intelligent systems to be implemented, maintenance and operational costs must be considered, along with the large variety of smart sensors needed for the collection of proper data. Finally, the reliability of these sensors under difficult conditions, such as environments with high dust concentration, humidity, and a lack of proper lighting, should be considered [77].

3. Overview of Drilling Approaches

Drilling has been the subject of significant change throughout history, providing many different approaches when it comes to mining. To date, the approaches described below have appeared in the mining industry.

3.1. Cable Tool Drilling

Cable tool drilling was the first developed drilling technique, and it was used to extract salt from salt quarries [5]. Its modern implementation follows the same principle: a drilling bit is lifted through a mechanism to a great height and is then allowed to fall down the borehole to break the rock formation [5,6], and after that, a *bailer* collects the fragments [6].

3.2. Auger Drilling

Auger drilling utilizes a screw-like drill bit to break the formation [5]. This method does not usually include the use of *mud*; instead, it uses the helical path of its bit, called a *flight*, to force the fragments to the surface [6,7].

3.3. Rotary Drilling

Rotary drilling is a technique that utilizes the rotational motion of the drill bit to deepen the borehole [5]. It uses a motor and a long shaft to rotate the drill bit and a mining fluid, called *mud*, to force the fragments to the surface, and it then sieves through the fluid to separate the solid materials [6].

3.4. Diamond Drilling

Diamond drilling applies the same principle of operation that the rotary drilling technique uses. The main difference between the two is that diamond drilling relies on the abrasive wear its drill bit causes to produce an uninterrupted cylindrical output of whatever ore it is drilling. This is achievable because its bit is actually a metallic matrix with embedded synthetic industrial diamonds with high hardness [5,6].

3.5. Directional Drilling

Directional drilling allows the boring procedure to be performed at a controlled angle and normally features a down-the-hole motor. The system may or may not have the ability to change its angle multiple times [8]. Normally, this can be accomplished in a multitude of ways, such as the following:

- Whipstock: A special geometry that allows the drill bit to slide off its surface. The geometry's angle is what directs the drilling bit [8].
- Bent-house PDM: The motor's housing is at an angle, allowing the motor to drill at a specific, predefined trajectory [8].
- Best sub: In this configuration, there are hubs that connect two different bent subs at different angles. It provides a greater degree of freedom when drilling [8].

4. Automated Mining and Intelligent Machines

4.1. Rotary Steerable Drilling Systems

The evolution of directional drilling systems led to *Rotary Steerable Drilling Systems* (*RSDSs*). These systems can change the drill bit's trajectory mid-operation, allowing the implementation of complex trajectory planning and exploration, as reported in [9,10]. These systems are primarily down-the-hole motor systems that can be divided based on the method they employ to affect the bit's direction and can be roughly divided into the following types:

- Push-the-bit systems: Push-the-bit systems have a pump that pushes actuators fixated on the system's external housing. These pneumatic systems utilize the polar array of these actuators, normally pads, to push the bit to the desired direction [10], as can be

seen in Figure 1. The most famous implementations are the Schlumberger PowerDrive and Baker-Hughes Autotrak systems, as in [9,11].

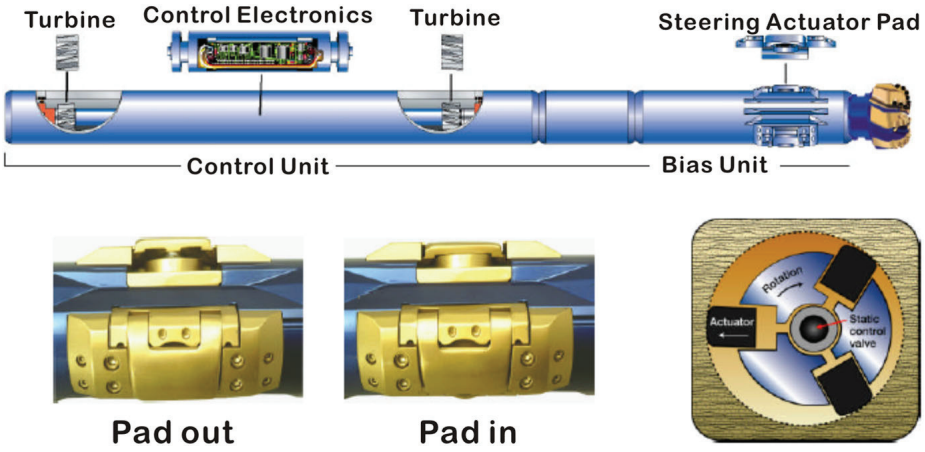


Figure 1. The Schlumberger PowerDrive system. Figure taken from [8].

- Point-the-bit systems: These systems vary from the push-the-bit systems in that the actuators reside on the inside of the configuration [10,11]. The most well-explored example of this is the Halliburton Sperry-sun Geo-Pilot system [9]. In these cases, the control of the direction of the bit is possible by using two eccentric rings, changing its angle directly using motors, as shown in Figure 2.

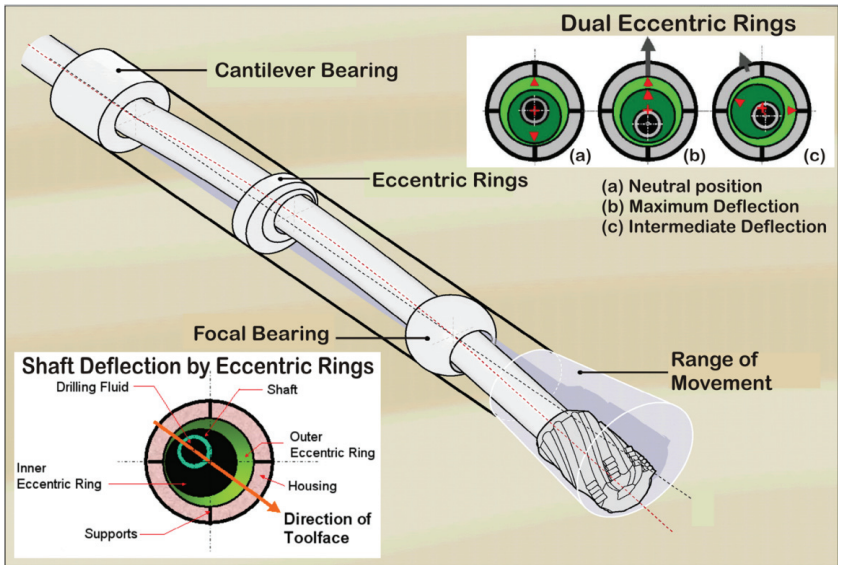


Figure 2. The Halliburton Sperry-sun Geo-Pilot system. Figure taken from [8].

- Hybrid systems: Hybrid systems are, in essence, a combination of the previous two systems. They have multiple actuators that assist them in better controlling every aspect of the drilling procedure. In the Schlumberger PowerDrive Archer system, the bit's shaft uses a universal joint and internal actuators to push to orient it at the desired angle [9,11].

4.2. Longwall

Longwall systems (Figure 3) are large robotic systems that combine the drilling and hauling operations through the use of a conveyor belt system and several horizontal motors with drill bits that perform the drilling [12,13]. Famur's MIKRUS longwall system features a cutting and loading head, which has two cutting drums that can break coal and a conveyor

belt that can retrieve it from the mine [14]. It also has a multitude of sensors that allow the process to be even more automated [15].

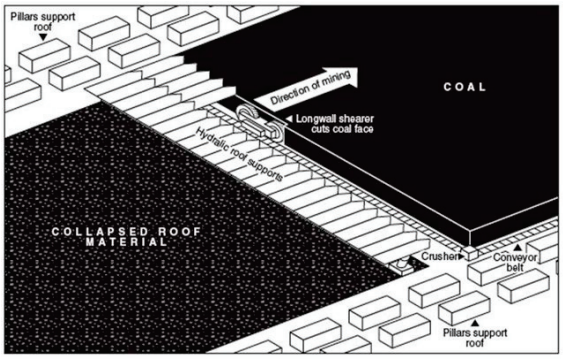


Figure 3. A longwall system with most of its parts. Figure taken from [12].

4.3. Automated Hauling

Automated hauling is the process in which the drilled formations are reclaimed automatically. The state-of-the-art choice for such a process is a robotic hauler, also called a *scraper*.

Scrapers are intelligent systems that can transport ore fragments from the drilling point to a preset point. They utilize sensors to understand their position and orientation in the world and use weight sensors while digging the ore fragments to provide data about the extracted ore [16]. Additionally, they can be operated autonomously or remotely by a human. Such a system is shown in Figure 4.



Figure 4. Autonomous scraper with human-control capability. Figure taken from [16].

These robots are designed with a dual capability to handle various challenges in complex environments. They can operate autonomously, adjusting to issues like limited signal propagation in underground areas, or be controlled by human operators when direct guidance is needed.

Aside from scrapers, there are also automatic mining trucks that are used for the same purpose as the scrapers, but they have a much larger capacity than the scrapers and are used to retrieve extracted ore [16].

4.4. Charging Robots

In mining, a controlled explosion can be used to break the rock formation. This process may prove dangerous due to the possible cave-in of the mine. Instead of using humans to transport and charge explosives, robots are used to carry, secure, and charge them [16,17].

4.5. Intelligent Rock-Drilling Jumbo and DTH Drill

Rock drilling is a process that is implemented during excavation procedures or during the production of ore through blasting. Depending on the method of drilling, different

machines are used. Such machines are the hydraulic rock-drilling jumbo and the down-the-hole (DTH) drill. Intelligent and unmanned versions of these machines allow for efficient ore mining and accurate blasting. This efficiency stems from systems that incorporate smart blockage prevention mechanisms for the drilling pipes, alongside a fully automated drill-pipe bank that oversees the management of the drilling tubes based on the drill's wear status. The accuracy in blasting is supported by systems that utilize anti-deviation control technology that controls the drilling pipes in real time [16].

By observing the trends mentioned in this section on the automation process of mining operations, it can be noticed that there are different automated systems that can partially fulfill the tasks needed for mining without the need for a human presence on site. Given the fact that the collapse of rocks in mines is one of the common causes of fatal accidents [1], it seems that removing human workers from these sites can lead to an important increase in the safety of the overall mining industry.

5. Technological Infrastructure in Automated Mining

5.1. Sensors Used in Automated Mining

Aside from the sensors used by specific robotic systems to perform tasks, there are sensors that orchestrate the coordinated action of several systems and provide analytics regarding the mining progress.

5.1.1. LIBS

Laser-Induced Breakdown Spectroscopy is a method of examining the contents of the core extracted by a mining machine. It utilizes a laser to emit photons onto the cylinder's surface, which creates plasma that excites electrons, and a spectrometer measures the response photons. This technology was recently adapted as an integrated system in down-the-hole equipment after being exclusively used in the laboratory [18–20].

5.1.2. LIDAR

Lidar technology has always been an important technology when the mapping of an area is required. There are currently systems that employ this technology to map and navigate entire caves and mines through the application of SLAM algorithms [21–24], such as automatic loading trucks [25]. An example of the sensor's output is shown in Figure 5.

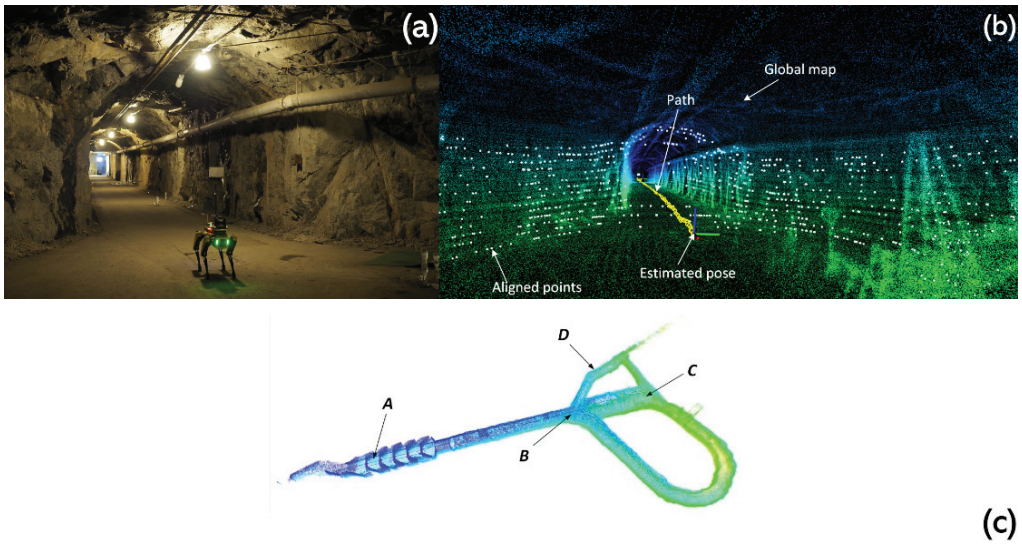


Figure 5. Obtaining a point cloud map using a LIDAR sensor using a quadruped robot. (**a**) A visual representation of the experimental environment (mine). (**b**) The estimated pose in the global point cloud map as perceived by the sensors on the robot. The estimated pose is depicted in the form of XYZ-axis (X-axis in red, Y-axis in green and Z-axis in blue). (**c**) The created point cloud map. Figures taken from [80].

5.1.3. Time-of-Flight Camera

A time-of-flight camera (TFC) or sensor (TFS) uses human-made signals of light, which can be created by either a laser or an LED [26,27]. The laser TFC is a scanner-less LiDAR type of sensor. The main difference that separates these sensors from LiDARs is that only one pulse is needed to capture the surrounding scene instead of many consecutive scans. Real-world applications of TFCs show their practicality. Such an example is their use in the local planning of a group of quadrupedal robots for the exploration of mines, as described in [51].

5.2. Communication and Localization Systems Used in Automated Mining

Sensors provide useful data that can later be utilized for the localization of the robot and the workers. Successful localization and sensor information require a robust and effective network to transmit the information to all the machines and workers, both in the mine and at the control center.

5.2.1. Light-Based Localization Systems

Currently, there are navigation platforms that allow for the precise localization of each robot in a 3D-digitized environment [28]. This can be implemented by using multiple navigation stations with lasers mounted on top of them, which allow them to perceive their pose through rotary encoders and predefined installation location coordinates [16] or through alternative advanced positioning methods [29].

5.2.2. Daisy Chaining of Fiducial Markers

It is evident that localization in the mining environment may prove difficult due to the fact that technologies and methods, such as GPS and motion-based sensor tracking, may prove inaccurate or unavailable in deep mines. High-end systems often employ additional robots whose job is to place easily recognizable constructs so that the robots can extrapolate their positions [30,31].

5.2.3. Magnetic Induction

In order to solve the problem that the underground environment poses during the localization process of robotic systems and personnel, non-RF methods have been employed, which include the use of magnetic induction methods [32]. Whether they use the Earth's magnetic field, which varies because of the mine's unique local structure, or the superposition effect of artificial magnetic fields through coils [21], it is possible for the sensor to extrapolate its position extremely accurately [33] if the coils' positions are known or if the local magnetic field is already mapped.

5.2.4. WiFi and LoRaWAN

WiFi is a major RF technology that can be implemented in mines in the form of a large network of signal repeaters to help the propagation of information in complex tunnels [21,24]. Another major RF technology is LoRaWAN (Long-Range WAN), which operates at a lower frequency (that translates to a larger wavelength), which enables stations to receive and transmit data through soil and minerals at a greater distance [21,34,35], albeit with lower bandwidth.

5.3. Simulations Using Digital Models

Digital twins (DTs) are digitized models of the actual electromechanical elements and mimic their behavior throughout the process. They provide the operators with a more comprehensive visualization of the physical world's state than traditional graphs by using multiple sensors and reducing resource expenditure while allowing the implementation of more complex and targeted software for decision-making. The cost of implementing DTs comes in the form of increased hardware demands and the need for high computational power, which is reflected by the quality of the sensing systems [36].

Ref. [37] proposes the six-layer digital twin architecture shown in Figure 6. By combining IoT sensors, cloud networking, and simulation software, it is possible to transform the physical world into a digitized model and apply secure and automated data-driven decision-making to carry out equipment maintenance and control through the different data analysis layers. Furthermore, this model allows streamlined information communication to every member of the mining crew.

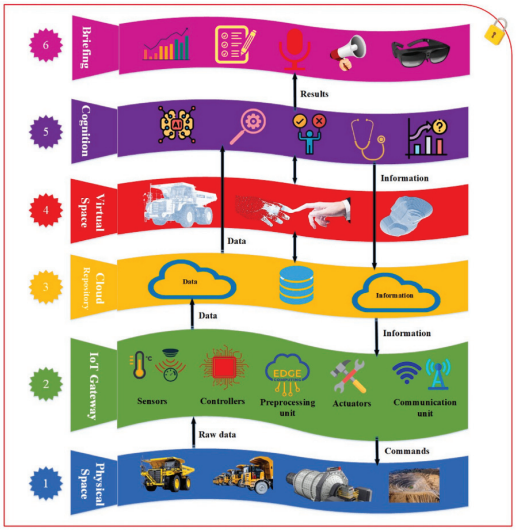


Figure 6. The proposed DT-based architecture for the mining industry. Figure taken from [37].

5.4. Software Used in Intelligent Systems

While the systems' hardware may be able to perform various tasks, without the direction of specialized software, it would not be possible to automate them.

5.4.1. e-Drilling's Software Suite

e-Drilling, located in Stavanger, Norway offers a software suite that allows the constant evaluation of the state of the drill and the well's borehole. It falls under the general digital twin tools that mimic the movement and state of the actual machine through a simulated model. This software also allows for better visualization and control of the process by the workers through its various modules [38,39]. For example, the wellAhead module is used to simulate a digital twin of the well, and the wellSIM module handles the mechanical loads and drilling parameter simulation of the drill itself. The wellAhead module's digital twin is depicted in Figure 7.

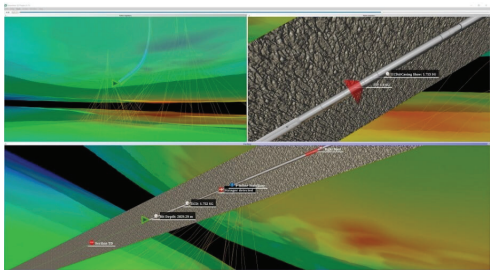


Figure 7. e-Drilling's wellAhead module's digital twin in action. Figure taken from [81].

5.4.2. Draco

Draco is an open-source library that compresses and decompresses three-dimensional geometrical meshes and point clouds [42]. By compressing the data, it is possible to upload them wirelessly with restricted available bandwidth from separate actors with limited

computational power to a main computational system, which will then only send back motion commands [40–42].

5.4.3. Forestall

Forestall uses proven predictive algorithms for the mining and oil industries and, coupled with cloud-based machine health and predictive maintenance products, offers optimized asset performance with minimized downtime. Platform-agnostic health monitoring ensures comprehensive insights into critical machinery's condition [43,44].

5.4.4. TIMining Aware

TIMining Aware creates real-time mine and block model visualization. The live mine plan compliance feature ensures that operations align with predefined strategies and regulatory requirements, and global accessibility across multiple devices fosters collaboration and productivity among users [44,45].

5.5. Collaborative Robotics in Mining

The above technologies have all been utilized in individual units and in networks of robots. The true extent of their usefulness can only be expressed through the coordination of many different systems for the betterment of the mining process. Currently, in the mining industry, several types of robots are being used. As mentioned above, autonomous hauling trucks as a part of an Autonomous Haulage System (AHS), especially in open-pit mines [82], drilling robots, charging robots, and automated conveyor belts are some of the categories that they can be separated into [83,84].

Examples of commercially available collaborative systems include the AutoMine drill and machine fleet, which allow the simultaneous control of several pieces of mining equipment [47,61,85].

Another example is Caterpillar's MineStar Command for Underground Systems, which automates multiple hauling trucks while allowing human operators to watch from afar [46–49].

An additional commercial application is EH-RemoteHeadControl v2 by Elgór-Hanses S.A., which requires the collaboration of a roadheader, a power supply unit, and some of the many sensors mentioned above for the creation of a DT around the shearer so that the roadheader can be operated remotely and safely for the creation of drifts in the mine [15].

Furthermore, research trends include examining the benefits of automated systems such as AHSs in open pits and exploring how decentralized AHSs can be applied in underground mines [86]. Also, the usage of Unmanned Ground Vehicles (UGVs) and Unmanned Aerial Vehicles (UAVs) is considered beneficial for mine inspections and rescue missions [50–54].

Research on systems of collaborative robots such as AHSs shows that they allow more efficient battery and fuel consumption in autonomous vehicles in comparison to manually driven ones while removing the operators of the hauling tracks from the mines and assigning them the role of observers. Furthermore, there is no longer a delay when changing shifts, since the trucks operate continuously [82,86,87].

Research on the use of UGVs and UAVs for inspection and rescue missions aims to increase the safety of inspectors from dangers that they may face, such as high concentrations of toxic gases or weak mine ceilings that might lead to the need for a rescue mission or even fatal accidents [52,88]. In rescue missions, where time is of the essence, the use of robots aims to reduce the focus on the safety of the rescue team, thus providing trapped miners with a faster and more efficient rescue plan [50,54].

The current infrastructure shows that it is possible to digitize mines and deploy robots that can collaborate on a peer-to-peer basis, while the human factor can be integrated with the role of the observer.

6. Possible Evolution of Intelligent Mining Machines

Aside from the aforementioned and already implemented systems, there are also numerous others that are currently being tested or researched. Some examples are listed below:

- Drill piston control for better bit-pointing accuracy: This approach promises better results when it comes to pointing the bit of an RSDS. It uses multiple pistons, valves, water, and stabilization clamp pads to better control the orientation and force applied by the bit [55].
- Photogrammetry in the mining industry: While Virtual Reality technology is being used today for minor tasks, it is being researched and worked upon extensively due to the presently unexplored advantages described in [56,57], especially with the improvement in modern cameras. Mining professionals are already adopting technologies from research, especially within the area of the digitization and mapping of mines with high-resolution cameras. Additionally, large-area scans for open-pit sites, which once required the use of special fiducial markers (ground control points) [58], are no longer needed with the introduction of Real-Time Kinematics and Post-Process Kinematics technologies [59]. Furthermore, there is ongoing research on multi-camera systems that can produce rock-mass digital models with a camera rig and can also support UAVs [25].
- Neural networks for augmented personnel safety: Neural networks can be used for a diversity of tasks. Research has shown that one of their uses can be the monitoring of the activity of personnel inside the mine to decrease accidents and exposure to hazardous areas [60]. One such example is the use of a YOLOv8 model (Figure 8) to determine the location of humans and hazardous areas [61]. YOLOv8 is a convolutional neural network that makes predictions of bounding boxes and class probabilities all at once.

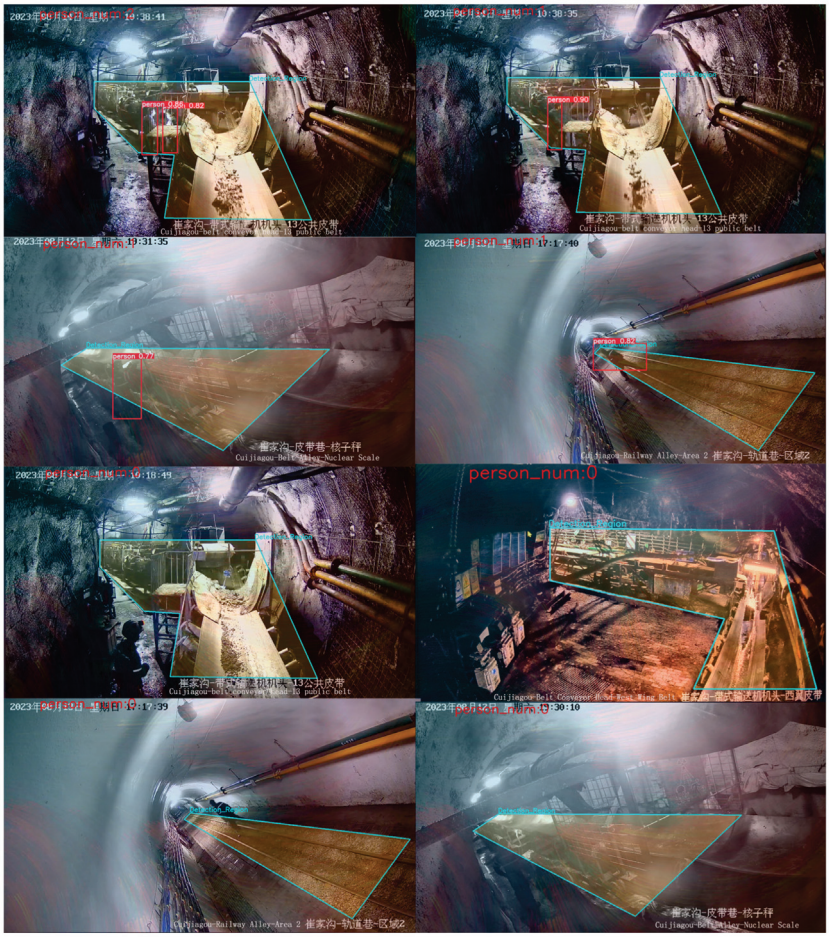


Figure 8. A YOLOv8 model detects personnel and hazardous areas. Figure taken from [89].

- Virtual Reality simulation for multi-agent system action planning: Virtual Reality's importance is already underlined, but the range of its applications cannot be stressed enough. Currently, there are ready-to-use software solutions that allow for the simulation of multiple robots in a virtual mining environment, expressed through the game development engine Unity3D [62]. Simulators of this kind may prove essential in mining procedure planning, especially in more complex systems with multiple interactions [57].
- AI-assisted DTs: Advancements that are currently being made in the field of artificial intelligence suggest that AI will be further integrated into FEA simulation software to provide better insights regarding mining data [63,64]. Surveys report that AI will be refined enough to be widely accepted in the industry by 2030 [65], providing a better bridge of communication between the real and digital worlds.
- Optimized haulage scheduling using Reinforcement Learning: Reinforcement Learning (RL) methods can be used to schedule the operation of fleets of autonomous haulers in real time utilizing model-free Q-Learning algorithms. In a simulated environment, this method resulted in a reduction in the wait time of each hauler and reduced the emission of greenhouse gases that resulted from the extensive fuel consumption of the machines. This reduction in fuel consumption was as high as 30% in comparison to fixed-schedule approaches while increasing productivity by 50% [66].
- Controlling equipment using Deep Reinforcement Learning: Deep Reinforcement Learning (DRL) can be used to control underground, individually operating machines as a substitute for human operators. Machines such as Load-Haul-Dump machines can be controlled with the use of a DRL framework by implementing a soft actor-critic (SAC) algorithm in a simulated environment, lowering energy consumption and increasing productivity [68]. In more detail, in [68], the energy consumption was found to be 21% lower for the autonomous vehicle in comparison with the energy consumed when a similar vehicle was operated by a human in the real world, and productivity was increased by 7%. Another example of operating multiple machines using RL, implementing proximal policy optimization (PPO) and SAC, can be examined in [67].
- Using DRL in mine production scheduling: DRL methods can even be utilized for the design of underground mine production layouts [69]. Utilizing DRL in the pre-production stages of the mining process allows for a better evaluation of the mineral grade since it utilizes uncertainty to create a stochastic design in comparison to deterministic designs. Deterministic designs are created by the more conventional geostatistical methods, such as mineable shape optimizers, which tend to lead to the under-performance of mining projects. A case study of underground stoping for gold production in [69] resulted in 8.3% higher expected profit and 3.4% more gold mined than the baseline that was provided by a mineable shape optimizer.
- Development and research for environmental monitoring: To ensure that the mining industry does not damage the environment, tools, indices, and methodologies are being developed to ensure the minimization of pollution [70,71]. One such example can be seen in Figure 9.
- Monitoring of the drilling rig: Research on monitoring drilling bit wear has been conducted to predict and analyze the drilling bit's condition. These studies have shown that sound and vibration signals, as well as examining the geological and mechanical properties of rocks, can be used to survey the drilling bit's wear condition [72].
- Intelligent control of the drilling: Computer vision and Deep Neural Networks for image classification are used to control the drilling rig and improve procedures such as drilling path planning and intelligent hole positioning [72].
- Measurement while drilling (MWD) and machine learning for rock monitoring: MWD is a technology that provides real-time measurements of drilling parameters, e.g., torque, rotation speed, and drilling speed. These parameters are used to determine the magnitude of the rock stress caused by the drilling operation with the aid of digital panoramic imaging technology for boreholes. Additionally, intelligent rock

property sensing can be achieved by implementing machine learning techniques with the data provided by MWD [72].

- AI and big data technology in ventilation systems: Current research on ventilation systems focuses on acquiring ventilation parameters, such as wind volume and pressure, temperature, humidity, and the detection/concentration of toxic gases. To achieve this, research focuses on the proper placement of the minimum number of sensors that provide reliable data about those parameters. After obtaining these data, diagnostic methods are proposed based on learning algorithms and big data analysis, as well as airflow control methods [72].
- Intelligent rock stress monitoring: Rock stress during drilling operations causes the deformation of the surrounding rock. Therefore, real-time monitoring methods have been developed based on machine vision and learning algorithms that provide early warning and predictions about surrounding rock deformation and damage [72].
- Quantum computing for open-pit profile optimization: There are currently studies that show that it is possible to implement quantum computing to determine the best possible parameters for ore extraction in open-pit mines. This could potentially help with the planning of the drill site [73–75].

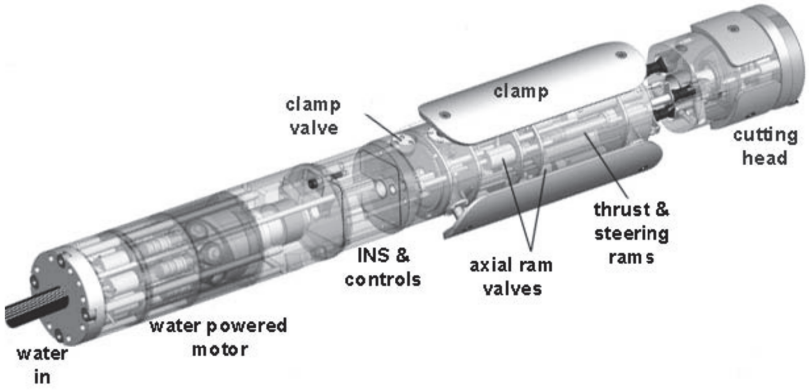


Figure 9. Automatic drill operating through pistons with water. Figure taken from [55].

There seems to be a trend of using digital technology to further improve the efficiency and capabilities of mining robotic solutions. This also means that there is a constant need for additional data-gathering sensors and systems that will help tune the function of the software to match the physical needs of the mine.

7. Future Trends and Visions in the Robotization and Digitalization of the Mining Industry

Emerging technologies seem to be heavily dependent on the current state of mining technology, which does not necessarily mean that they will be able to satisfy the current needs of the mining industry. Case studies seem to refer to technologies that follow implementations that require pre-existing communication systems. It is also apparent that modern practices tend to favor non-modular approaches that often rely on specific proprietary equipment and software to operate. While today's systems seem to be adequate for standard mines, they are not expected to work in environments that have been abandoned before the installation of such infrastructure.

While this approach may be sufficient for most mining industries, it is often not flexible enough to cover all of the possible mine configurations. To better generalize automation solutions, creating a unified framework is essential. This framework needs to categorize the different layers and parts of the data handling and the physical aspects of these systems in order to encourage the standardization of the software and methodology used.

Additionally, it would be interesting to test rig-less and infrastructure-less methods of automation, since installing these technologies may not be available in old, abandoned

mines. For example, it is not realistic to expect that 5G and UHF communication systems will be present in every mine. This points to the fact that decentralized approaches may need to be investigated in a more thorough way. However, the limitations of this approach should be taken into account, such as bandwidth limitations and latency in the communication between different machines and systems. As the mining industry is currently focusing on extraction technologies at large depths and revisiting the extractions at abandoned mines, there are major trends in miniaturizing mining machines, revisiting mining plans, e.g., the volume of drifts, and moving toward a future with completely autonomous mining operations.

Even with a technology standard, it is imperative that safety and health considerations be taken into account before sending crew members into mines to install the necessary infrastructure. To that end, it would be beneficial to implement systems that allow for the better inspection of hazardous elements without the necessary control or supervision from a human operator. To achieve this, the introduction of robotic systems that monitor the necessary non-automated operations while also being able to contribute to reducing the personnel's exposure time to potentially adversarial conditions is envisioned. It is evident that such automation for mining safety, while still an immature field, is something that can be expanded on.

Lastly, it seems that while intelligent systems are being studied and developed to reduce environmental strain from mining operations, said systems are not yet mature enough to minimize the environmental impact. Thus, it should be noted that there are still not enough solutions when it comes to automated waste management, and this is something that has to be addressed in the near future. Currently, practices such as the use of waste neutralization tanks and ground storage are used to combat solid mining waste (slag, tailings, etc.), but even they may pose problems with specific types of waste [3]. These intelligent systems will require the corresponding machines to operate properly. The machines need to contain adequate sensors and the proper computational power. They require an intelligent and generative design.

8. Conclusions

There are a number of new technologies developed for the automation of the mining process, which is a testament to the ever-increasing need for better efficiency. The inclusion of automation systems is what allows modern manufacturing needs to be met while reducing possible human harm and human error. From partially automated to fully automated systems, this integration reshapes the foundation of the mining industry and engineering. The use of many different types of autonomous and intelligent machines that can communicate on a peer-to-peer basis in situations where pre-existing communication networks are weak, faulty, or nonexistent can raise the efficiency of the ore extraction and increase the safety of the workers in the mines. The use of automated systems for mining operations could improve workers' health and increase their safety. This might lead to a reduction in needed unspecialized personnel and an increase in personnel with higher educational backgrounds. The digitization of the mining process and the use of more sophisticated digital models can also enhance the remote monitoring of mining operations. Automated mining is still a field with a lot of progress to be made. As time goes by, with continued intellectual and financial investments in intelligent systems design, it will become even clearer that these systems should be embraced and implemented as the new standard in the mining industry.

Finally, it should be noted that this article focuses on the automation and robotization of drilling and mining operations, and as such, the focus has been on these areas. Other important aspects of mining, e.g., the effectiveness of the selected methods and environmental impact, have not been addressed, as they are considered out of scope.

Author Contributions: Conceptualization, A.K. and T.F.; writing—original draft preparation, A.K., T.F. and K.S.; writing—review and editing, A.K., T.F., K.S., G.N. and P.K.; supervision, G.N. and P.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was funded by the Horizon Europe project PERSEPHONE [grant number 101138451].

Data Availability Statement: Data are contained within the article.

Acknowledgments: We would like to express our gratitude toward Achilleas Kousios for providing us with his insights during the editing process of this paper.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

3D	Three-dimensional
5G	Fifth-generation mobile network
AHS	Autonomous Haulage System
AI	Artificial intelligence
DRL	Deep Reinforcement Learning
DT	Digital twin
DTH	Down the hole
FEA	Finite Element Analysis
GPS	Global Positioning System
IoT	Internet of Things
LED	Light-Emitting Diode
LIBS	Laser-Induced Breakdown Spectroscopy
LiDAR	Light Detection and Ranging
LoRaWAN	Long-Range Wide-Area Network
MWD	Measurement while drilling
PPO	Proximal policy optimization
RF	Radio Frequency
RL	Reinforcement Learning
RSDS	Rotary Steerable Drilling System
SAC	Soft actor-critic
SLAM	Simultaneous Localization and Mapping
TFC	Time-of-flight camera
TFS	Time-of-flight sensor
UAV	Unmanned Aerial Vehicle
UGV	Unmanned Ground Vehicle
UHF	Ultra-High Frequency
YOLO	You Only Look Once

References

1. Yinka-Banjo, C.; Bagula, A.; Osunmakinde, I.O. Autonimous multi-robot behaviours for safety inspection under the constraints of underground mine terrains. *Ubiquitous Comput. Commun. J.* **2012**, *7*, 1316–1328.
2. Mueller, C.; Assibey-Bonsu, W.; Baafi, E.; Dauber, C.; Doran, C.; Jerzy Jaszczuk, M.; Nagovitsyn, O. *Mining Goes Digital: Proceedings of the 39th International Symposium'Application of Computers and Operations Research in the Mineral Industry' (APCOM 2019), June 4–6. 2019, Wroclaw, Poland*, 1st ed.; CRC Press: London, UK, 2019. [CrossRef]
3. Kalisz, S.; Kibort, K.; Mioduska, J.; Lieder, M.; Małachowska, A. Waste Management in the Mining Industry of Metals Ores, Coal, Oil and Natural Gas—A Review. *J. Environ. Manag.* **2022**, *304*, 114239. [CrossRef] [PubMed]
4. Agboola, O.; Babatunde, D.E.; Isaac Fayomi, O.S.; Sadiku, E.R.; Popoola, P.; Moropeng, L.; Yahaya, A.; Mamudu, O.A. A Review on the Impact of Mining Operation: Monitoring, Assessment and Management. *Results Eng.* **2020**, *8*, 100181. [CrossRef]
5. Akpedeye, K.U. Advancement on Drilling Technology in Petroleum Industry. Bachelor's Thesis, University of Applied Sciences, Riihimäki, Finland, 2010.
6. Todd, D.K.; Mays, L.W. *Groundwater Hydrology*; John Wiley & Sons: Hoboken, NJ, USA, 2004.

7. Zacny, K.; Paulsen, G.; McKay, C.; Glass, B.; Dave, A.; Davila, A.; Marinova, M.; Mellerowicz, B.; Heldmann, J.; Stoker, C.; et al. Reaching 1m Deep on Mars: The Icebreaker Drill. *Astrobiology* **2013**, *13*, 1166–1198. [CrossRef] [PubMed]
8. Tianshou, M.; Chen, P.; Zhao, J. Overview on Vertical and Directional Drilling Technologies for the Exploration and Exploitation of Deep Petroleum Resources. *Geomech. Geophys. Geo-Energy Geo-Resour.* **2016**, *2*, 365–395. [CrossRef]
9. Li, F.; Ma, X.; Tan, Y. Review of the Development of Rotary Steerable Systems. *J. Phys. Conf. Ser.* **2020**, *1617*, 012085. [CrossRef]
10. Sugiura, J. Systematic Testing with Push- and Point-the-Bit Rotary-Steerable Systems Leads to the Optimal BHA Design for Stability, Steerability and Borehole Quality. In Proceedings of the 2008 AADE Fluids Conference and Exhibition held at the Wyndam Greenspoint Hotel, Houston, TX, USA, 8–9 April 2008.
11. Ivanova, T.; Božek, P.; Korshunov, A.; Koretskiy, V. Analysis of Distortion Mechanisms Used in Rotary Steerable Systems. *Multidiscip. Asp. Prod. Eng.* **2020**, *3*, 331–346. [CrossRef]
12. Saki, S. Gob Ventilation Borehole Design and Performance Optimization for Longwall Coal Mining Using Computational Fluid Dynamics. Ph.D. Thesis, Colorado School of Mines, Golden, CO, USA, 2016. [CrossRef]
13. Peng, S. *Longwall Mining*, 3rd ed.; CRC Press: Boca Raton, FL, USA, 2019.
14. Famur. 'Machinery & Equipment for Underground Mining', 6 October 2024. Available online: https://famur.com/en/machinery-equipment-for-underground-mining/ (accessed on 6 October 2024).
15. Bołoz, Ł.; Biały, W. Automation and Robotization of Underground Mining in Poland. *Appl. Sci.* **2020**, *10*, 7221. [CrossRef]
16. Li, J.; Zhan, K. Intelligent Mining Technology for an Underground Metal Mine Based on Unmanned Equipment. *Engineering* **2018**, *4*, 381–391. [CrossRef]
17. Bonchis, A.; Duff, E.; Roberts, J.; Bosse, M. Robotic Explosive Charging in Mining and Construction Applications. *IEEE Trans. Autom. Sci. Eng.* **2014**, *11*, 245–250. [CrossRef]
18. Fayyaz, A.; Ali, R.; Waqas, M.; Liaqat, U.; Ahmad, R.; Umar, Z.A.; Baig, M.A. Analysis of Rare Earth Ores Using Laser-Induced Breakdown Spectroscopy and Laser Ablation Time-of-Flight Mass Spectrometry. *Minerals* **2023**, *13*, 787. [CrossRef]
19. Bhatt, C.R.; Jain, J.C.; Goueguel, C.L.; McIntyre, D.L.; Singh, J.P. Determination of Rare Earth Elements in Geological Samples Using Laser-Induced Breakdown Spectroscopy (LIBS). *Appl. Spectrosc.* **2018**, *72*, 114–121. [CrossRef] [PubMed]
20. Bellie, V.; Gokulraju, R.; Rajasekar, C.; Vinoth, S.; Mohankumar, V.; Gunapriya, B. Laser Induced Breakdown Spectroscopy for New Product Development in Mining Industry. *Mater. Today Proc.* **2021**, *45*, 8157–8161. [CrossRef]
21. Zeeshan, M.; Chavda, M.; Ehshan, K.M.; Nayek, R.; Malik, S. A Review on Non-RF Underground Positioning Techniques for Mining Applications. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 9510217. [CrossRef]
22. Dima, E.; Sjöström, M. Camera and Lidar-Based View Generation for Augmented Remote Operation in Mining Applications. *IEEE Access* **2021**, *9*, 82199–82212. [CrossRef]
23. Wang, G.; Wu, J.; Xu, T.; Tian, B. 3D Vehicle Detection With RSU LiDAR for Autonomous Mine. *IEEE Trans. Veh. Technol.* **2021**, *70*, 344–355. [CrossRef]
24. Stathoulopoulos, N.; Pagliari, E.; Davoli, L.; Nikolakopoulos, G. Redundant and Loosely Coupled LiDAR-Wi-Fi Integration for Robust Global Localization in Autonomous Mobile Robotics. In Proceedings of the 2023 21st International Conference on Advanced Robotics (ICAR), Abu Dhabi, United Arab Emirates, 5–8 December 2023; pp. 121–127. [CrossRef]
25. Janiszewski, M.; Prittinen, M.; Uotinen, L.; Torkan, M.; Rinne, M. Rapid Photogrammetric Method for Rock Mass Characterization in Underground Excavations. In Proceedings of the ISRM IV Nordic Symposium on Rock Mechanics and Rock Engineering, Reykjavik, Iceland, 24–26 May 2023. [CrossRef]
26. Hansard, M.; Lee, S.; Choi, O.; Horaud, R.P. *Time-of-Flight Cameras: Principles, Methods and Applications*; Springer Science & Business Media: Berlin, Germany 2012.
27. Horaud, R.; Hansard, M.; Evangelidis, G.; Ménier, C. An Overview of Depth Cameras and Range Scanners Based on Time-of-Flight Technologies. *Mach. Vis. Appl.* **2016**, *27*, 1005–1020. [CrossRef]
28. Kanellakis, C.; Nikolakopoulos, G. Evaluation of Visual Localization Systems in Underground Mining. In Proceedings of the 2016 24th Mediterranean Conference on Control and Automation (MED), Athens, Greece, 21–24 June 2016; pp. 539–544. [CrossRef]
29. Keskin, M.F.; Sezer, A.D.; Gezici, S. Localization via Visible Light Systems. *Proc. IEEE* **2018**, *106*, 1063–1088. [CrossRef]
30. Capozzi, R.; Wilde, M.; Kish, B. Daisy Chain Navigation and Communication in Underground Environments. In Proceedings of the 2021 IEEE Aerospace Conference (50100), Big Sky, MT, USA, 6–13 March 2021; pp. 1–9. [CrossRef]
31. Tardioli, D.; Riazuelo, L.; Sicignano, D.; Rizzo, C.; Lera, F.; Villarroel, J.L.; Montano, L. Ground Robotics in Tunnels: Keys and Lessons Learned after 10 Years of Research and Experiments. *J. Field Robot.* **2019**, *36*, 1074–1101. [CrossRef]
32. Pasku, V.; De Angelis, A.; De Angelis, G.; Arumugam, D.D.; Dionigi, M.; Carbone, P.; Moschitta, A.; Ricketts, D.S. Magnetic Field-Based Positioning Systems. *IEEE Commun. Surv. Tutor.* **2017**, *19*, 2003–2017. [CrossRef]
33. Lin, S.C.; Alshehri, A.; Wang, P.; Akyildiz, I. Magnetic Induction-Based Localization in Randomly-Deployed Wireless Underground Sensor Networks. *IEEE Internet Things J.* **2017**, *4*, 1454–1465. [CrossRef]
34. Seguel, F.; Palacios-Játiva, P.; Azurdia-Meza, C.A.; Krommenacker, N.; Charpentier, P.; Soto, I. Underground Mine Positioning: A Review. *IEEE Sens. J.* **2022**, *22*, 4755–4771. [CrossRef]
35. Branch, P.; Cricenti, T. A LoRa Relay Based System for Detonating Explosives in Underground Mines. In Proceedings of the 2020 IEEE International Conference on Industrial Technology (ICIT), Buenos Aires, Argentina, 26–28 February 2020; pp. 259–264. [CrossRef]

36. Botín-Sanabria, D.M.; Mihaita, A.S.; Peimbert-García, R.E.; Ramírez-Moreno, M.A.; Ramírez-Mendoza, R.A.; Lozoya-Santos, J.d.J. Digital Twin Technology Challenges and Applications: A Comprehensive Review. *Remote Sens.* **2022**, *14*, 1335. [CrossRef]
37. Hazrathosseini, A.; Moradi Afrapoli, A. The Advent of Digital Twins in Surface Mining: Its Time Has Finally Arrived. *Resour. Policy* **2023**, *80*, 103155. [CrossRef]
38. Josef, N. Using Digital Technologies to Automat and Optimize Drilling Parameters in Real-Time, Its Impact on Value Creation, and Work Process. Master's Thesis, University of Stavanger, Stavanger, Norway, 2023.
39. eDrilling. eDrilling Products, 6 Octomber 2024. Available online: https://www.edrilling.no/products (accessed on 6 October 2024).
40. Jansen van Rensburg, B.; Puech, W.; Pedeboy, J.P. The First Draco 3D Object Crypto-Compression Scheme. *IEEE Access* **2022**, *10*, 10566–10574. [CrossRef]
41. Source, G.O. Google Open Source. Introducing Draco: Compression for 3D Graphics. Google Open Source Blog(blog), 6 October 2024. Available online: https://opensource.googleblog.com/2017/01/introducing-draco-compression-for-3d.html (accessed on 6 October 2024).
42. Google/Draco. C++. 2016.Reprint, Google, 5 October 2024. Available online: https://github.com/google/draco (accessed on 6 October 2024).
43. PETRA. FORESTALL®, 16 July 2024. Available online: https://www.petradatascience.com/forestall/ (accessed on 6 October 2024).
44. Hazrathosseini, A.; Afrapoli, A.M. Intelligent Fleet Management Systems in Surface Mining: Status, Threats, and Opportunities. *Min. Metall. Explor.* **2023**, *40*, 2087–2106. [CrossRef]
45. TIMining. TIMining AWARE—TIMining. TIMining AWARE, 6 October 2024. Available online: https://www.timining.com/en/timining-aware/ (accessed on 6 October 2024).
46. Cat Mining. Cat® MineStar™ Command for Underground, 17 October 2016. Available online: https://www.youtube.com/watch?v=WU6u5XTMKNI (accessed on 6 October 2024).
47. Gustafson, A.; Paraszczak, J.; Tuleau, J.; Schunnesson, H. Impact of Technical and Operational Factors on Effectiveness of Automatic Load-Haul-Dump Machines. *Min. Technol.* **2017**, *126*, 185–190. [CrossRef]
48. George, B.K. Optimization of Quarry Operations and Maintenance Schedules. Ph.D. Thesis, Virginia Tech, Blacksburg, VA, USA, 2023.
49. George, B.; Nojabaei, B. Data Analyses of Quarry Operations and Maintenance Schedules: A Production Optimization Study. *Mining* **2023**, *3*, 347–366. [CrossRef]
50. Dinelli, C.; Racette, J.; Escarcega, M.; Lotero, S.; Gordon, J.; Montoya, J.; Dunaway, C.; Androulakis, V.; Khaniani, H.; Shao, S.; et al. Configurations and Applications of Multi-Agent Hybrid Drone/Unmanned Ground Vehicle for Underground Environments: A Review. *Drones* **2023**, *7*, 136. [CrossRef]
51. Miller, I.D.; Cladera, F.; Cowley, A.; Shivakumar, S.S.; Lee, E.S.; Jarin-Lipschitz, L.; Bhat, A.; Rodrigues, N.; Zhou, A.; Cohen, A.; et al. Mine Tunnel Exploration Using Multiple Quadrupedal Robots. *arXiv* **2020**, arXiv.1909.09662. [CrossRef]
52. Nordström, S.; Stathoulopoulos, N.; Dahlquist, N.; Lindqvist, B.; Tevetzidis, I.; Kanellakis, C.; Nikolakopoulos, G. Safety Inspections and Gas Monitoring in Hazardous Mining Areas Shortly After Blasting Using Autonomous UAVs. *Authorea* **2024**. [CrossRef]
53. Lindqvist, B.; Stathoulopoulos, N.; Karlsson, S.; Bai, Y.; Haluska, J.; Tevetzidis, I.; Koval, A.; Kanellakis, C.; Nikolakopoulos, G.; Patel, A. Deployment of Autonomous Uavs in Underground Mines: Field Evaluations and Use-Case Demonstrations. *SSRN Electron. J.* **2023**. [CrossRef]
54. Zimroz, P.; Trybała, P.; Wróblewski, A.; Góralczyk, M.; Szrek, J.; Wójcik, A.; Zimroz, R. Application of UAV in Search and Rescue Actions in Underground Mine—A Specific Sound Detection in Noisy Acoustic Signal. *Energies* **2021**, *14*, 3725. [CrossRef]
55. Wendt, M.; Einicke, G. Development of a Water-Hydraulic Self-Propelled Robotic Drill for Underground Mining. In *Field and Service Robotics*; Springer Tracts in Advanced Robotics; Springer: Berlin/Heidelberg, Germany, 2005; Volume 25, pp. 355–366. [CrossRef]
56. Akbulut, N.; Anani, A. Application of Virtual Reality in the Mining Industry—Where Are We Now? In Proceedings of the SME Annual Conference and Expo, Denver, CO, USA, 26 February–1 March 2023.
57. Xie, J.; Yang, Z.; Wang, X.; Zeng, Q.; Li, J.; Li, B. A Virtual Reality Collaborative Planning Simulator and Its Method for Three Machines in a Fully Mechanized Coal Mining Face. *Arab. J. Sci. Eng.* **2018**, *43*, 4835–4854. [CrossRef]
58. Patikova, A. Digital Photogrammetry In The Practice Of Open Pit Mining. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2004**, *34*.
59. Jaud, M.; Bertin, S.; Beauverger, M.; Augereau, E.; Delacourt, C. RTK GNSS-Assisted Terrestrial SfM Photogrammetry without GCP: Application to Coastal Morphodynamics Monitoring. *Remote Sens.* **2020**, *12*, 1889. [CrossRef]
60. Akhmetov, T. Industrial Safety Using Augmented Reality and Artificial Intelligence. Ph.D. Thesis, Nazarbayev University, Astana, Kazakhstan, 2023.
61. Automation.AutoMine®, 16 July 2024. Available online: https://www.rocktechnology.sandvik/En/Products/Technology/Automation/Automine/ (accessed on 6 October 2024).

62. Shi, H.; Xie, J.; Wang, X.; Li, J.; Ge, X. An Operation Optimization Method of a Fully Mechanized Coal Mining Face Based on Semi-Physical Virtual Simulation. *Int. J. Coal Sci. Technol.* **2019**, *7*, 147–163. [CrossRef]
63. Zayed, S.M.; Attiya, G.M.; El-Sayed, A.; Hemdan, E.E.D. A Review Study on Digital Twins with Artificial Intelligence and Internet of Things: Concepts, Opportunities, Challenges, Tools and Future Scope. *Multimed. Tools Appl.* **2023**, *82*, 47081–47107. [CrossRef]
64. Huang, Z.; Shen, Y.; Li, J.; Fey, M.; Brecher, C. A Survey on AI-Driven Digital Twins in Industry 4.0: Smart Manufacturing and Advanced Robotics. *Sensors* **2021**, *21*, 6340. [CrossRef]
65. Rosen, R.; Hartmann, D.; Van der Auweraer, H.; Hermann, M.; Wolfrum, P. *Simulation & DigitalTwin*; Siemens Corporation: Princeton, NJ, USA, 2020.
66. Huo, D.; Sari, Y.A.; Kealey, R.; Zhang, Q. Reinforcement Learning-Based Fleet Dispatching for Greenhouse Gas Emission Reduction in Open-Pit Mining Operations. *Resour. Conserv. Recycl.* **2023**, *188*, 106664. [CrossRef]
67. Fidêncio, A.X.; Glasmachers, T.; Naro, D. Application of Reinforcement Learning to a Mining System. In Proceedings of the 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI), Herl'any, Slovakia, 21–23 January 2021; pp. 000111–000118. [CrossRef]
68. Backman, S.; Lindmark, D.; Bodin, K.; Servin, M.; Mörk, J.; Löfgren, H. Continuous Control of an Underground Loader Using Deep Reinforcement Learning. *Machines* **2021**, *9*, 216. [CrossRef]
69. Noriega, R.; Boisvert, J. Optimization of Underground Mining Production Layouts Considering Geological Uncertainty Using Deep Reinforcement Learning. *Eng. Appl. Artif. Intell.* **2025**, *139*, 109493. [CrossRef]
70. Song, W.; Song, W.; Gu, H.; Li, F. Progress in the Remote Sensing Monitoring of the Ecological Environment in Mining Areas. *Int. J. Environ. Res. Public Health* **2020**, *17*, 1846. [CrossRef] [PubMed]
71. Myaskov, A.; Temkin, I.; Deryabin, S.; Marinova, D. Factors and Objectives of Sustainable Development at the Implementation of Digital Technologies and Automated Systems in the Mining Industry. *E3S Web Conf.* **2020**, *174*, 04023. [CrossRef]
72. Liu, X.; Zhang, X.; Wang, L.; Qu, F.; Shao, A.; Zhao, L.; Wang, H.; Yue, X.; Li, Y.; Yan, W.; et al. Research Progress and Prospects of Intelligent Technology in Underground Mining of Hard Rock Mines. *Green Smart Min. Eng.* **2024**, *1*, 12–26. [CrossRef]
73. Hindy, Y.; Pointing, J.; Tolunay, M.; Venkatarao, S.; Motta, M.; Latone, J. A Quantum Computational Approach to the Open-Pit Mining Problem. *arXiv* **2021**, arXiv:2107.11345.
74. Paradezhenko, G.; Pervishko, A.; Yudin, D. Quantum-Assisted Open-Pit Optimization. *JETP Lett.* **2024**, *119*, 470–478. [CrossRef]
75. Hindy, Y.; Pointing, J.; Tolunay, M.; Venkatarao, S.; Motta, M.; Latone, J.A. Application of the Variational Quantum Eigensolver to the Ultimate Pit Problem. In Proceedings of the 2023 IEEE International Conference on Quantum Computing and Engineering (QCE), Bellevue, WA, USA, 17–22 September 2023; Volume 1, pp. 681–691. [CrossRef]
76. Zhironkina, O.; Zhironkin, S. Technological and Intellectual Transition to Mining 4.0: A Review. *Energies* **2023**, *16*, 1427. [CrossRef]
77. Zhironkin, S.; Ezdina, N. Review of Transition from Mining 4.0 to Mining 5.0 Innovative Technologies. *Appl. Sci.* **2023**, *13*, 4917. [CrossRef]
78. Clausen, E.; Sörensen, A.; Nienhaus, K. Mining 4.0. In *Handbook Industry 4.0: Law, Technology, Society*; Frenz, W., Ed.; Springer: Berlin/Heidelberg, Germany, 2022; pp. 785–792. [CrossRef]
79. Faz-Mendoza, A.; Gamboa-Rosales, N.K.; Medina-Rodríguez, C.E.; Casas-Valadez, M.A.; Castorena-Robles, A.; López-Robles, J.R. Intelligent Processes in the Context of Mining 4.0: Trends, Research Challenges and Opportunities. In Proceedings of the 2020 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 8–9 November 2020; pp. 480–484. [CrossRef]
80. Stathoulopoulos, N.; Koval, A.; Nikolakopoulos, G. A Comparative Field Study of Global Pose Estimation Algorithms in Subterranean Environments. *Int. J. Control Autom. Syst.* **2024**, *22*, 690–704. [CrossRef]
81. Automatic Realtime Monitoring Enhances Safety and Reduces Cost of Drilling Operations—eDrilling. 2024. Available online: https://www.edrilling.no/blogs/post/automatic-realtime-monitoring-enhances-safety-and-reduces-cost-of-drilling-operations (accessed on 6 October 2024).
82. Cerna, G.E.P.; Hernández, J.R.C.; Herazo, J.C.M.; Castillo, A.P. Evaluation of the Overall Effectiveness (OEE) of Autonomous Transportation System (AHS) Equipment and Its Impact on Mine Design. Open Pit Mine Case Study. *Procedia Comput. Sci.* **2023**, *224*, 468–473. [CrossRef]
83. Frank, M. *A Step Towards the Design of Collaborative Autonomous Machines: A Study on Construction and Mining Equipment*; Blekinge Institute of Technology: Karlskrona, Sweden, 2019.
84. Ding, X.; Guo, J.; Ren, Z.; Deng, P. State-of-the-Art in Perception Technologies for Collaborative Robots. *IEEE Sens. J.* **2022**, *22*, 17635–17645. [CrossRef]
85. Patil, S.D.; Mitra, A.; Tuggali Katarikonda, K.; Wansink, J.D. Predictive Asset Availability Optimization for Underground Trucks and Loaders in the Mining Industry. *Opsearch* **2021**, *58*, 751–772. [CrossRef]
86. Matsui, K.; Adan, F.; Angeloudis, P. Decentralised Motion Planning for Autonomous Mining Haulage Trucks Using Prioritised Multi-Agent MPPI. *IFAC-PapersOnLine* **2024**, *58*, 154–161. [CrossRef]
87. Ishimoto, H.; Hamada, T. Safety Concept and Architecture for Autonomous Haulage System in Mining. In Proceedings of the 37th International Symposium on Automation and Robotics in Construction, Kitakyushu, Japan, 27–28 October 2020. [CrossRef]

88. Yinka-Banjo, C.; Osunmakinde, I.O.; Bagula, A. Multi-Robot Systems: A Cooperative Behaviour and Performance Evaluations for Mine Safety Inspections. In Proceedings of the Proceedings of the Southern African Institute for Computer Scientist and Information Technologists Annual Conference 2014 on SAICSIT 2014 Empowered by Technology, New York, NY, USA, 29 September–1 October 2014; SAICSIT '14, pp. 241–250. [CrossRef]
89. Ni, Y.; Huo, J.; Hou, Y.; Wang, J.; Guo, P. Detection of Underground Dangerous Area Based on Improving YOLOV8. *Electronics* **2024**, *13*, 623. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Comparative Analysis of Dry, Minimum Quantity Lubrication, and Nano-Reinforced Minimum Quantity Lubrication Environments on the Machining Performance of AZ91D Magnesium Alloy

Berat Baris Buldum [1], Kamil Leksycki [2,*] and Suleyman Cinar Cagan [1]

[1] Department of Mechanical Engineering, Mersin University, Ciftlikkoy, Mersin 33343, Türkiye; barisbuldum@mersin.edu.tr (B.B.B.); cinarcagan@mersin.edu.tr (S.C.C.)
[2] Faculty of Mechanical Engineering, University of Zielona Góra, Prof. Z. Szafrana Street 4, 65-516 Zielona Gora, Poland
* Correspondence: k.leksycki@iim.uz.zgora.pl

Abstract: This study investigates the machining performance of AZ91D magnesium alloy under three different cooling environments: dry, minimum quantity lubrication (MQL), and nano-reinforced MQL (NanoMQL) with multi-walled carbon nanotubes. Turning experiments were conducted on a CNC lathe with systematically varied cutting parameters, including cutting speed (150–450 m/min), feed rate (0.05–0.2 mm/rev), and depth of cut (0.5–2 mm). The machining performance was evaluated through cutting force measurements, surface roughness analysis, and tool wear examination using SEM. The results demonstrate that the NanoMQL environment significantly outperforms both dry and conventional MQL conditions, providing a 42.2% improvement in surface quality compared to dry machining and a 33.6% improvement over conventional MQL. Cutting forces were predominantly influenced by the depth of cut and the feed rate, while cutting speed showed variable effects. SEM analysis revealed that the NanoMQL environment substantially reduced built-up edge formation and flank wear, particularly under aggressive cutting conditions. The superior performance of the NanoMQL environment is attributed to the enhanced thermal conductivity and lubrication properties of carbon nanotubes, which form a protective tribofilm at the tool–workpiece interface. This study provides valuable insights for optimizing the machining parameters of AZ91D magnesium alloy in industrial applications, particularly where high surface quality and tool longevity are required.

Keywords: magnesium alloy; turning; tribology; surface roughness; wear; minimum quantity lubrication; carbon nanotubes; sustainable machining

1. Introduction

Magnesium alloys, particularly AZ91D, have become critical materials in the automotive, aerospace, and electronics industries due to their low density, high specific strength, excellent castability, and corrosion resistance [1–3]. These properties make AZ91D especially suitable for lightweight applications where weight reduction and durability are essential, such as automotive components, portable electronic devices, and aerospace structures [4]. However, despite these advantages, the machinability of magnesium alloys presents significant challenges, primarily due to their high thermal conductivity, low melting point, and tendency to form built-up edges (BUEs) during machining [5]. These issues

often result in poor surface quality, increased tool wear, and reduced process efficiency, limiting their broader adoption in precision manufacturing [6].

To address these challenges, researchers have explored various cooling and lubrication techniques to improve the machinability of magnesium alloys. Conventional flood cooling has been widely used to reduce cutting temperatures and improve surface quality. However, this method is associated with high environmental and economic costs due to the large volumes of cutting fluid required and the subsequent disposal challenges [7,8]. As a result, alternative methods such as minimum quantity lubrication (MQL) have gained attention. MQL involves the application of a minimal amount of cutting fluid directly at the tool–workpiece interface, significantly reducing fluid consumption while maintaining effective cooling and lubrication [9]. Studies have shown that MQL can improve surface quality and reduce tool wear compared to dry machining, making it a more sustainable alternative to flood cooling [10,11].

Although the present study focuses on the turning of AZ91D magnesium alloy, several researchers have also explored its machinability in milling operations. For instance, Jouini et al. [12] reported that both dry and cryogenic high-speed milling of AZ91D can produce mirror-like surface finishes, with cryogenic cooling notably extending tool life and further enhancing surface quality. Similarly, Zagórski and Korpysa [13] systematically examined how milling parameters affect surface roughness, finding that feed per tooth is the most critical factor, while higher cutting speeds tend to reduce 3D surface roughness. Marakini et al. [14] investigated high-speed face milling of AZ91 alloy and observed that both uncoated and PVD-coated carbide inserts significantly improve surface roughness and hardness, with PVD-coated tools yielding defect-free surfaces and higher compressive residual stresses. These studies underscore the importance of optimizing cutting parameters and lubrication strategies to achieve superior surface integrity in AZ91D milling. Nevertheless, it is important to recognize that the mechanisms and outcomes of turning may differ from those in milling, due to the inherent differences between these machining processes.

Building on the success of MQL, recent advancements have focused on enhancing its performance through the incorporation of nanoparticles into the cutting fluid. NanoMQL utilizes nanoparticles such as multi-walled carbon nanotubes (MWCNTs) to improve the thermal conductivity and lubrication properties of the cutting fluid [15]. These nanoparticles form a protective tribofilm at the tool–workpiece interface, reducing friction, adhesion, and heat generation during machining [16]. Experimental studies have demonstrated that NanoMQL can significantly enhance surface quality, reduce cutting forces, and extend tool life compared to both dry and conventional MQL environments [17,18].

Despite the widespread use of AZ91D magnesium alloy in lightweight applications, its machinability under advanced cooling and lubrication techniques such as NanoMQL remains underexplored. Previous studies have primarily focused on conventional machining methods, with limited attention given to the effects of nano-reinforced cutting fluids on the machining performance of AZ91D [19]. This gap in the literature highlights the need for a comprehensive investigation into the potential benefits of NanoMQL for machining AZ91D, particularly regarding cutting forces, surface roughness, and tool wear.

Cutting forces are a critical parameter in machining, as they directly influence tool wear, surface quality, and energy consumption. Studies have shown that cutting forces are primarily affected by cutting parameters such as cutting speed, feed rate, and depth of cut, as well as the cooling and lubrication environment [9,20]. For magnesium alloys, the high thermal conductivity and low melting point exacerbate the effects of cutting forces, leading to increased tool wear and poor surface finish [21]. By reducing friction and heat generation, NanoMQL has the potential to mitigate these issues and improve the overall machinability of AZ91D [12,22].

Surface roughness is another key performance indicator in machining, as it directly impacts the functional performance of machined components, including their fatigue resistance, corrosion susceptibility, and aesthetic appeal [23]. Achieving a high-quality surface finish is particularly challenging for magnesium alloys due to their tendency to form BUE and adhere to the cutting tool [24]. Previous studies have demonstrated that NanoMQL can significantly improve surface roughness by forming a thin protective film at the tool–workpiece interface, reducing adhesion and wear [25]. However, the specific effects of NanoMQL on the surface quality of AZ91D under varying cutting parameters remain largely unexplored.

Tool wear is a major concern in machining, as it affects tool life, machining accuracy, and overall process efficiency. The wear mechanisms in magnesium alloy machining are influenced by both mechanical and thermal factors, with flank wear and crater wear being the most common types observed [26]. The use of NanoMQL has been shown to reduce tool wear by enhancing heat dissipation and minimizing adhesion at the tool–workpiece interface [27]. However, the effectiveness of NanoMQL in mitigating tool wear during the machining of AZ91D has not been thoroughly investigated, particularly under aggressive cutting conditions [28].

This study aims to address these research gaps by conducting a comprehensive analysis of the machining performance of AZ91D magnesium alloy under dry, MQL, and NanoMQL environments. The primary objectives are to evaluate the effects of these cooling and lubrication methods on cutting forces, surface roughness, and tool wear, and to identify the optimal machining parameters for achieving superior performance. By systematically varying cutting speed, feed rate, and depth of cut, this study provides valuable insights into the interactions between machining parameters and cooling environments. The findings are expected to contribute to the development of more efficient and sustainable machining processes for magnesium alloys, aligning with the growing emphasis on environmentally friendly manufacturing practices.

Although several studies have examined the machinability of magnesium alloys under dry and MQL conditions, there is a lack of systematic research directly comparing dry, MQL, and NanoMQL environments for AZ91D magnesium alloy using identical experimental parameters. In particular, the tribological and performance benefits of NanoMQL with multi-walled carbon nanotubes for this alloy have not been comprehensively reported. Therefore, this study not only aims to fill these gaps but also provides new insights into the optimization of sustainable machining processes for AZ91D.

2. Materials and Methods

2.1. Materials

AZ91D is the most widely used magnesium alloy grade (designated by the aluminum and zinc content). Aluminum serves as the primary alloying element while zinc improves strength and corrosion resistance in this magnesium-based alloy. It is employed in applications requiring lightweight components with good castability and mechanical properties. These alloys are commonly used in automotive components, portable electronic housings, power tool casings, and aerospace parts.

AZ91D magnesium alloy samples were used for experimentation. The chemical composition of this alloy consists of Al 9.21%, Zn 0.45%, Mn 0.17%, Fe 0.0018%, Be 0.00084%, Si 0.016%, Cu 0.002%, and Ni 0.00085%, and the balance is Mg element. The Brinell hardness value of the material was measured as 63 BSD. The yield strength of the AZ91D magnesium alloy was determined to be 160 MPa, and the tensile strength as 230 MPa. The material exhibits a fatigue strength of 90–110 MPa, elongation of 3%, thermal conductivity of 72 W/mK, and a melting temperature of 578 °C.

2.2. Experiments

The experiments were conducted on a Johnford TC-35 industrial CNC lathe (Johnford Machine Tool Company, Taichung, Taiwan) with a maximum power capacity of 10 kW and spindle speed of 3500 rpm. This Fanuc-operated machine features a 250 mm *x*-axis travel, 600 mm *z*-axis travel, 250 mm hydraulic chuck diameter, and a 12-position turret. Prior to testing, workpiece preparation included drilling Ø 5 mm center holes at both ends, with one end machined in the traveling center-mirror axis configuration to prevent chuck ejection during testing. Surface preparation involved removing 0.5–1.0 mm of material to eliminate potential axial misalignment, post-heat treatment oxide layers, and decarburized regions. The cutting operation employed TCMT16T304 cutting inserts with STGCR1616h11 tool holders (Sandvik Coromant, Sandviken, Sweden). Cooling was provided by a minimum quantity lubrication (MQL) system delivering 0.0036 mL of cutting oil per injection at approximately 5 bar pressure, with two injections occurring each second. Figure 1 presents a schematic representation of the MQL system's arrangement during the turning process.



Figure 1. Schematic illustration of the experimental setup.

The metalworking process utilized CUTTEX SYN-10, a biodegradable cutting oil manufactured by Belgin Oil Company (Gebze, Turkiye). This synthetic vegetable-based lubricant, enhanced with performance additives, is applied directly at the tool–material interface. The product contains no chlorine or heavy metals, making it environmentally friendly, while its light coloration facilitates workpiece visibility. Its formulation includes anti-wear and anti-corrosion additives that extend cutting tool lifespan, while its synthetic base provides longevity between oil changes and its high flash point reduces oil mist formation and fire hazards. For the NanoMQL system, this same base oil was modified with 1% by weight of MWCNTs sourced from Nanografi (Ankara, Turkiye). These 92% pure nanoparticles, measuring 8–10 nm, were thoroughly incorporated through a com-

bined process of mechanical mixing and ultrasonic homogenization, with each step lasting one hour.

Experimental investigations were performed across three distinct cooling environments (dry, MQL, and NanoMQL) with systematically varied cutting parameters. The cutting speed was tested at three levels (150, 300, and 450 m/min), combined with three feed rates (0.05, 0.1, and 0.2 mm/rev) and three cutting depths (0.5, 1, and 2 mm), as detailed in Table 1. Throughout all trials, ISO standard TCMT16T304 cutting inserts were utilized in conjunction with ISO standard STGCR1616h11 tool holders.

Table 1. Parameters and levels for machining experiments.

Parameters	Units	Levels		
		1	**2**	**3**
Cutting Speed	m/min	150	300	450
Feed rate	mm/rev	0.05	0.1	0.2
Depth of Cut	mm	0.5	1	2
Cooling environments	-	Dry	MQL	NanoMQL

2.3. Experimental Measurement Procedures

The experimental analysis incorporated multiple measurement techniques to comprehensively evaluate machining performance. Cutting force data were captured at 20 mm intervals using a KISTLER 9257B piezoelectric dynamometer (Kistler Group, Winterthur, Switzerland) capable of simultaneously measuring main cutting force (Fc), feed force (Ff), and radial force (Fr). The force signals underwent amplification via a KISTLER Type 5019 multichannel charge amplifier before transmission to a computer through an RS-232C connection, where DynoWare Type 2825A1-2 software generated the corresponding force graphics.

Surface quality assessment was conducted using a portable Mitutoyo Surftest SJ 201 roughness meter (Mitutoyo Corporation, Kawasaki, Japan), with data collected from three equidistant positions (120° apart) around each workpiece to ensure representative sampling. This evaluation is critical as surface characteristics significantly influence component performance attributes including fatigue resistance, corrosion susceptibility, and operational longevity. Manufacturing engineers recognize that surface irregularities often serve as initiation sites for fractures and corrosion processes, though achieving optimal surface finish values typically involves challenging and cost-intensive processes.

A detailed examination was carried out using the Scanning Electron Microscope (SEM) (JEOL JSM-6060LV, JEOL Ltd., Tokyo, Japan) facilities at Gazi University, Faculty of Technology, Department of Metallurgy and Materials Engineering, to detect deterioration of the cutting tools.

The flank wear (VB) was measured from SEM images using the scale bar as reference, following the procedure described in ISO 3685 for turning operations [29]. The reliability of this method is supported by recent studies, which show that automated image analysis algorithms and manual measurements (e.g., using ImageJ) yield very similar results, with relative errors as low as 2.5% for common magnifications [30]. The ImageJ software used for measurements was obtained from the official website (https://imagej.nih.gov/ij/, accessed on 12 May 2025) [31].

3. Results

3.1. Analysis of Machining Forces

Table 2 and Figure 2 present the cutting force data obtained from machining AZ91D magnesium alloy in a dry environment. Analysis of these results reveals that cutting

forces increase significantly with greater cutting depths and feed rates, while cutting speed variations produce less pronounced effects on force magnitudes. The minimum cutting force was recorded at parameters of 300 m/min cutting speed, 0.05 mm/rev feed rate, and 0.5 mm cutting depth (60.86 N resultant force). Conversely, maximum cutting force occurred at 450 m/min cutting speed, 0.2 mm/rev feed rate, and 2 mm cutting depth (234.42 N resultant force).

Table 2. Force values measured during the dry-cutting process of AZ91D magnesium alloy.

Exp No.	Cutting Speed	Feed Rate	Depth of Cut	F_x (N)	F_y (N)	F_z (N)	F_{result} (N)
1	150	0.05	0.5	28.36	23.75	51.84	63.685
2	150	0.05	1	52.22	23.64	69.41	90.020
3	150	0.05	2	91.93	20.44	102.85	139.453
4	150	0.1	0.5	30.23	23.74	59.38	70.735
5	150	0.1	1	57.37	24.41	85.82	106.077
6	150	0.1	2	100.97	23.51	135.7	170.769
7	150	0.2	0.5	31.36	27.14	72.78	83.767
8	150	0.2	1	67.35	24.59	115.64	136.064
9	150	0.2	2	98.67	25.02	183.17	209.554
10	300	0.05	0.5	28.7	19.95	49.82	60.858
11	300	0.05	1	58.9	26.27	71.04	95.948
12	300	0.05	2	89.29	21.92	101.17	136.706
13	300	0.1	0.5	31.04	27.38	61.04	73.750
14	300	0.1	1	64.42	24.65	88.94	112.552
15	300	0.1	2	105.51	22.37	136.72	174.141
16	300	0.2	0.5	30.1	24.91	74.49	84.115
17	300	0.2	1	61.92	32.95	114.06	133.901
18	300	0.2	2	116.78	27.82	187.79	222.882
19	450	0.05	0.5	30.46	24.2	50.42	63.684
20	450	0.05	1	48.83	19.42	66.75	84.953
21	450	0.05	2	98.91	21.44	102.89	144.323
22	450	0.1	0.5	31.87	25.43	60.83	73.230
23	450	0.1	1	54.74	23.74	84.2	103.197
24	450	0.1	2	107.99	22.99	136.02	175.191
25	450	0.2	0.5	41.12	30.4	71.89	88.222
26	450	0.2	1	59.36	29.23	111.94	130.033
27	450	0.2	2	131.53	30.6	191.62	234.424

Previous studies on dry machining of AZ91D magnesium alloys have consistently reported substantial force increases with greater cutting depths and feed rates [32]. Regarding cutting speed effects, the literature presents mixed findings, with some researchers documenting force reductions at higher speeds [33], while others report minimal influence [34]. Our experimental results similarly show inconsistent cutting speed effects, with forces decreasing in some trials and increasing in others as cutting speed rises.

This variable response to cutting speed can be attributed to two competing mechanisms: thermal softening of the workpiece material at higher speeds, which tends to reduce cutting forces, and material adhesion to the cutting tool edge, which increases cutting resistance and consequently elevates forces [35]. The specific balance between these phenomena appears to vary across different parameter combinations in the current study.

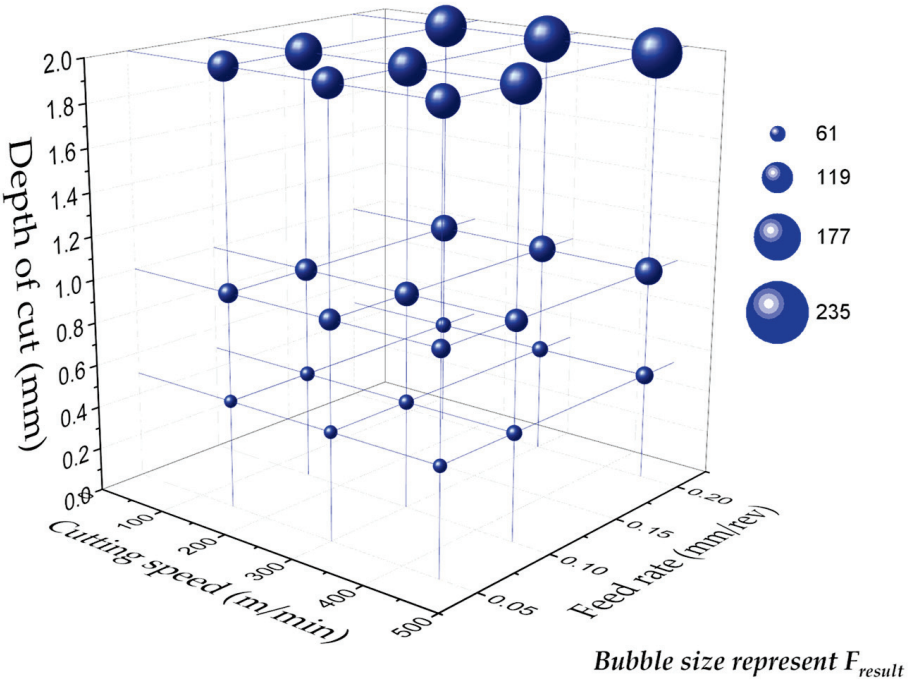


Figure 2. Force values required for machining AZ91D magnesium alloy in a dry environment.

3.2. Surface Roughness Evaluation

Table 3 and the graphics are shown in Figures 3–5 present the surface roughness values obtained from machining AZ91D magnesium alloy under dry, MQL, and NanoMQL environments. Analysis of these results reveals significant variations in surface quality across different cutting parameters and cooling conditions.

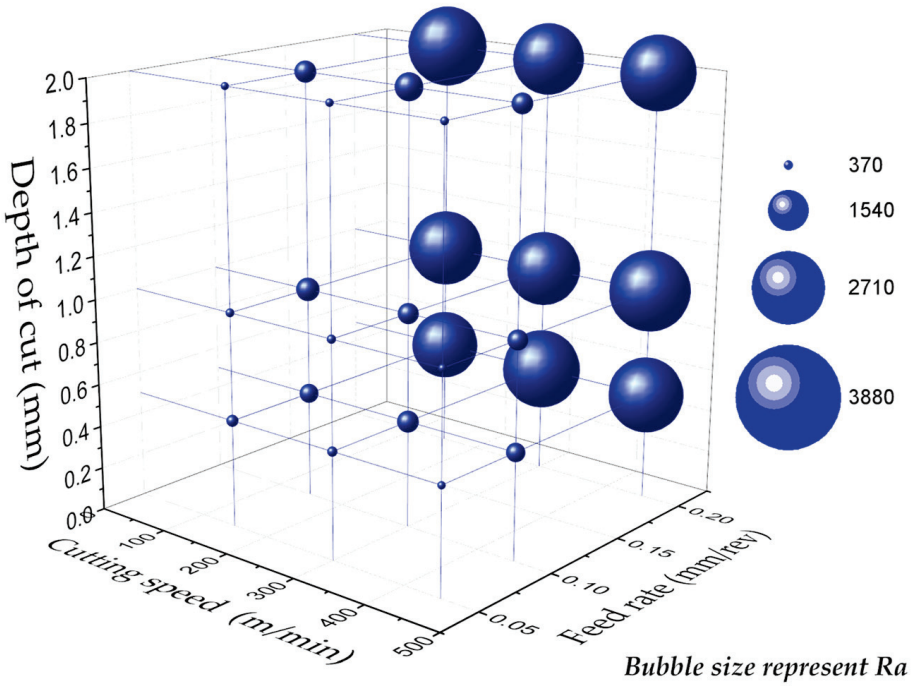


Figure 3. Surface roughness graph obtained by machining AZ91D magnesium alloy in a dry environment.

Table 3. Surface roughness (Ra) values obtained by machining AZ91D magnesium alloy under dry, MQL, and NanoMQL environments.

Cutting Speed	Feed Rate	Depth of Cut	Ra (Dry, µm)	Ra (MQL, µm)	Ra (NanoMQL, µm)
150	0.05	0.5	0.587	0.322	1.910
150	0.05	1	0.405	0.356	2.660
150	0.05	2	0.399	0.360	2.386
150	0.1	0.5	0.933	0.966	4.393
150	0.1	1	1.137	0.986	3.784
150	0.1	2	1.007	0.940	4.018
150	0.2	0.5	3.353	3.042	8.098
150	0.2	1	3.703	3.128	8.424
150	0.2	2	3.897	3.084	9.965
300	0.05	0.5	0.492	0.348	2.535
300	0.05	1	0.421	0.423	2.578
300	0.05	2	0.382	0.389	2.208
300	0.1	0.5	1.01	0.954	3.766
300	0.1	1	0.964	0.958	3.827
300	0.1	2	1.251	1.121	4.308
300	0.2	0.5	3.833	3.053	8.463
300	0.2	1	3.557	3.168	8.429
300	0.2	2	3.386	2.924	8.809
450	0.05	0.5	0.391	0.380	2.110
450	0.05	1	0.39	0.397	2.600
450	0.05	2	0.377	0.387	2.160
450	0.1	0.5	0.885	0.893	3.744
450	0.1	1	0.946	0.964	3.032
450	0.1	2	0.921	0.842	3.959
450	0.2	0.5	3.571	3.062	8.190
450	0.2	1	3.815	3.210	7.929
450	0.2	2	3.498	2.993	8.728

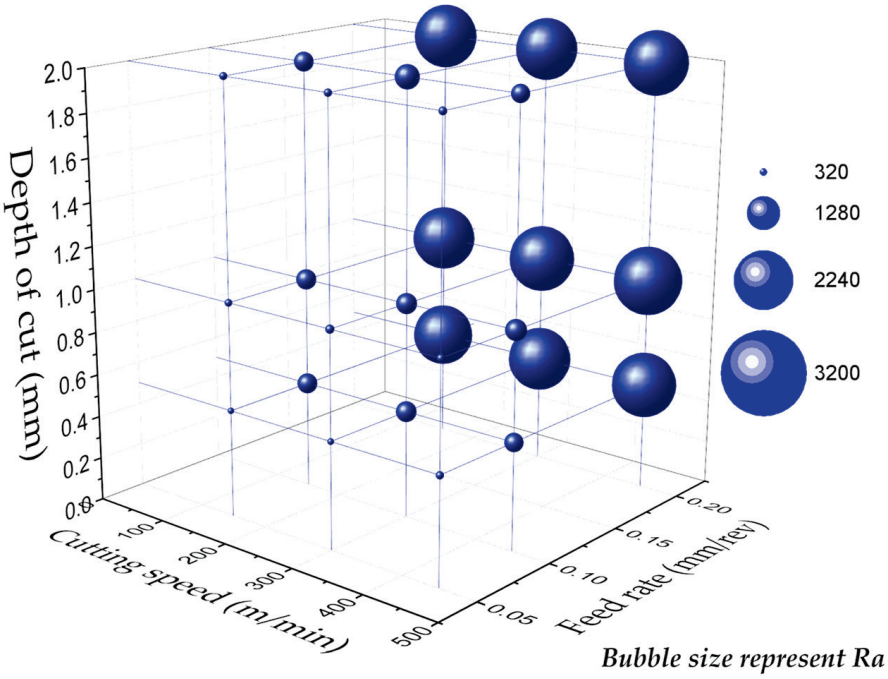


Figure 4. Surface roughness graph obtained by machining AZ91D magnesium alloy in MQL environment.

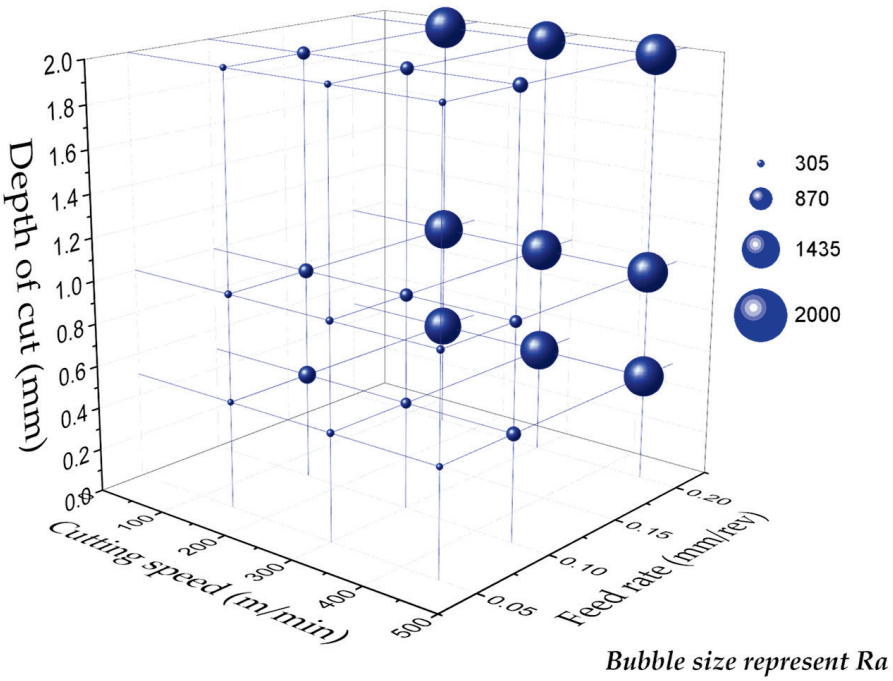


Figure 5. Surface roughness graph obtained by machining AZ91D magnesium alloy in multi-walled carbon-nanotube-reinforced MQL environment.

In dry machining conditions, surface roughness decreases with decreasing feed rate, which aligns with established machining theory [36]. Generally, increasing cutting speed leads to improved surface finish, with the best surface quality (Ra = 0.377 μm) achieved at 450 m/min cutting speed, 0.05 mm/rev feed rate, and 2 mm cutting depth. The poorest surface finish (Ra = 3.897 μm) occurred at 150 m/min cutting speed, 0.2 mm/rev feed rate, and 2 mm cutting depth. This relationship between cutting speed and surface quality can be attributed to reduced built-up edge formation at higher speeds [37].

Figure 3 illustrates the three-dimensional representation of surface roughness values obtained in the dry machining environment. The bubble chart clearly demonstrates the significant influence of feed rate on surface roughness, with larger bubbles (indicating higher Ra values) predominantly appearing at the 0.2 mm/rev feed rate position. The visualization confirms that the combination of high feed rate and low cutting speed produces the roughest surfaces, while the combination of high cutting speed and low feed rate yields the smoothest surfaces. The size gradient of the bubbles effectively represents the magnitude of surface roughness across the experimental parameter space.

When applying minimum quantity lubrication, surface roughness similarly increases with higher feed rates. The effect of cutting speed shows a non-linear pattern, with roughness increasing as speed rises from 150 to 300 m/min, then decreasing as speed further increases to 450 m/min. The optimal surface quality (Ra = 0.322 μm) was obtained at 150 m/min cutting speed, 0.05 mm/rev feed rate, and 0.5 mm cutting depth. The roughest surface (Ra = 3.210 μm) resulted from parameters of 450 m/min cutting speed, 0.2 mm/rev feed rate, and 1 mm cutting depth. Compared to dry machining, MQL generally provides superior surface finish, with an average improvement of 12.9% [10].

Figure 4 presents the three-dimensional bubble chart for surface roughness values in the MQL environment. Similar to the dry machining results, the chart shows a clear correlation between feed rate and surface roughness, with the largest bubbles concentrated at the 0.2 mm/rev feed rate. However, the overall bubble sizes are noticeably smaller compared to the dry environment chart, visually confirming the improved surface finish achieved with MQL. The distribution pattern also reveals the non-linear effect of cutting

speed, with medium cutting speeds (300 m/min) showing slightly larger bubbles than both lower and higher speeds at equivalent feed rates and cutting depths.

The nano-reinforced MQL environment demonstrates that surface roughness increases with higher feed rates, consistent with the other environments. Cutting speed shows a different pattern, with roughness decreasing as speed increases from 150 to 300 m/min, then remains relatively stable between 300 and 450 m/min. The finest surface quality (Ra = 0.308 μm) was achieved at 450 m/min cutting speed, 0.05 mm/rev feed rate, and 2 mm cutting depth, while the poorest finish (Ra = 2.001 μm) occurred at 150 m/min cutting speed, 0.2 mm/rev feed rate, and 2 mm cutting depth. Notably, NanoMQL provides significantly better surface quality compared to both dry and conventional MQL environments, with improvements of 42.2% and 33.6%, respectively [8].

Figure 5 displays the surface roughness graph obtained by machining AZ91D magnesium alloy in a multi-walled carbon-nanotube-reinforced MQL environment. The most striking feature of this visualization is the substantially smaller bubble sizes compared to both dry and conventional MQL charts, particularly at the 0.2 mm/rev feed rate position. This visual representation effectively illustrates the superior surface finish achieved with NanoMQL across all parameter combinations. The bubble distribution pattern confirms that, while feed rate remains the dominant factor affecting surface roughness, its negative impact is significantly mitigated by the nano-enhanced lubricant. Additionally, the chart shows more consistent bubble sizes across different cutting speeds, suggesting that NanoMQL provides more stable machining conditions regardless of speed variations.

The superior performance of NanoMQL can be attributed to the enhanced thermal conductivity and lubrication properties of carbon nanotubes, which effectively reduce friction at the tool–workpiece interface and improve heat dissipation from the cutting zone [27]. This is particularly beneficial for magnesium alloys like AZ91D, which are prone to built-up edge formation and poor surface finish due to their high thermal conductivity and low melting point [38,39]. The nanoparticles in the cutting fluid create a thin protective film between the tool and workpiece, reducing adhesion and improving machining stability [16].

3.3. Tool Wear Analysis

The wear mechanisms occurring during the machining of AZ91D magnesium alloy were examined using SEM micrographs. These images provide critical insights into the tool degradation patterns under different cooling conditions. Figures 6 and 7 present SEM images of tool wear after machining in dry, MQL, and NanoMQL environments at different cutting parameters.

Figure 6 shows the SEM images obtained from experiments conducted at moderate cutting conditions (cutting speed: 150 m/min, feed rate: 0.05 mm/rev, and depth of cut: 0.5 mm). Under these conditions, flank wear was observed in all three environments, though with varying severity. The dry machining condition (Figure 6a) exhibited the most pronounced flank wear, followed by the MQL environment (Figure 6b), while the NanoMQL environment (Figure 6c) showed the least wear. This pattern can be attributed to the superior cooling and lubrication properties of the nano-enhanced cutting fluid, which effectively reduces friction and heat generation at the tool–workpiece interface [40].

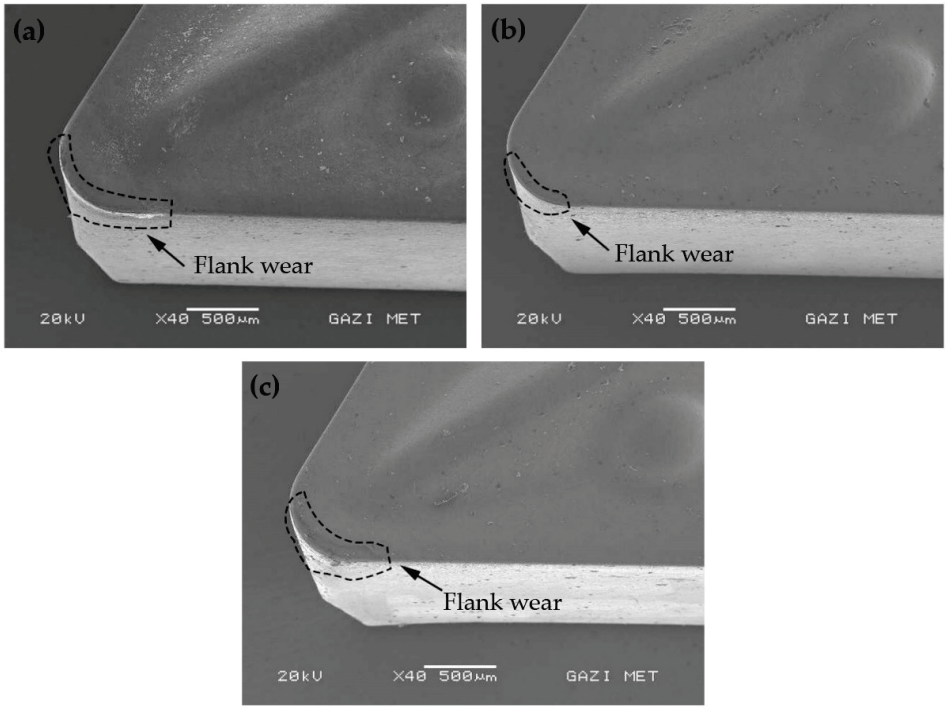


Figure 6. SEM images obtained from experiments performed in different environments: (**a**) dry, (**b**) MQL, and (**c**) NanoMQL. Turning parameters: cutting speed 150 m/min, feed rate 0.05 mm/rev, and depth of cut 0.5 mm.

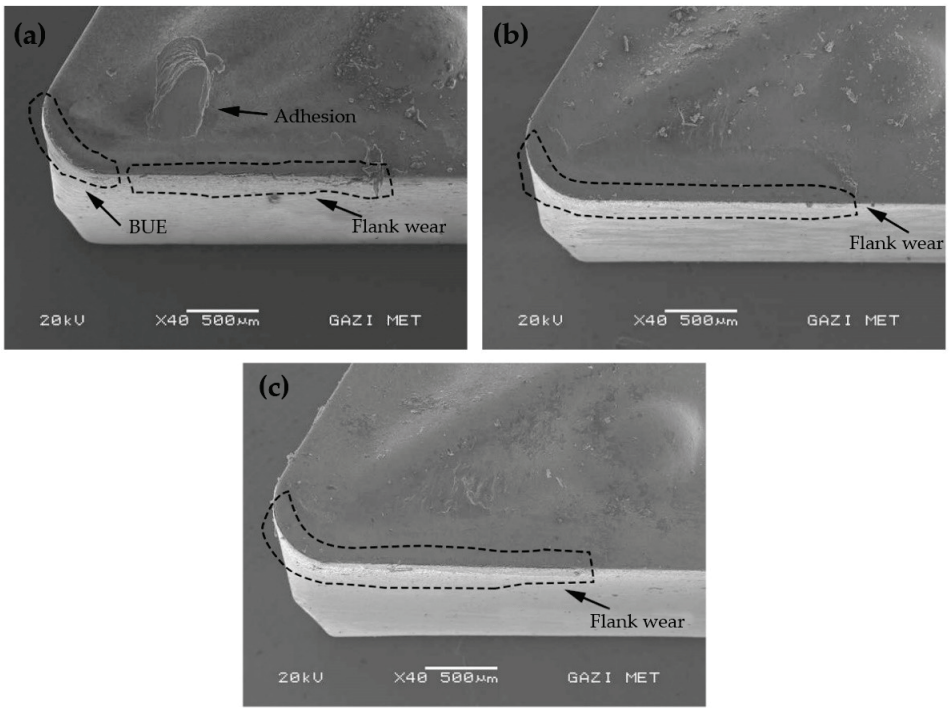


Figure 7. SEM images obtained from experiments performed in different environments: (**a**) dry, (**b**) MQL, and (**c**) NanoMQL. Turning parameters: cutting speed 450 m/min, feed rate 0.2 mm/rev, and depth of cut 2 mm.

Figure 7 presents SEM images from experiments performed at more aggressive cutting parameters (cutting speed: 450 m/min, feed rate: 0.2 mm/rev, and depth of cut: 2 mm). Under these severe conditions, significant differences in wear mechanisms were observed. In the dry environment (Figure 7a), both built-up edge (BUE) formation and adhesion were

evident alongside extensive flank wear. The presence of BUE is particularly detrimental as it alters the effective tool geometry and can lead to poor surface finish [2]. In contrast, the MQL environment (Figure 7b) showed reduced BUE formation but still exhibited considerable flank wear. The NanoMQL environment (Figure 7c) demonstrated the least severe wear pattern with minimal adhesion and more uniform flank wear distribution.

The depth of cut was found to have the most significant impact on tool wear, while cutting speed and feed rate had comparatively less influence within the parameter ranges studied. This finding aligns with previous research on magnesium alloy machining, which indicates that the depth of cut directly affects the contact area between the tool and workpiece, thereby increasing mechanical and thermal loads on the cutting edge [41].

The superior performance of the NanoMQL environment can be attributed to the multi-walled carbon nanotubes' ability to form a protective tribofilm at the tool–workpiece interface. This film reduces direct contact between the tool and workpiece materials, thereby minimizing adhesion and friction [42]. Additionally, the enhanced thermal conductivity of the nano-enhanced fluid facilitates more efficient heat dissipation from the cutting zone, preserving the tool's hardness and wear resistance properties [43].

These results are consistent with findings by Wang et al. [11], who reported that nano-enhanced cutting fluids significantly reduced tool wear when machining magnesium alloys due to the rolling and sliding effects of nanoparticles at the contact interface. Similarly, Sharma et al. [44] observed that carbon-nanotube-enhanced cutting fluids provided superior performance in terms of tool life and wear resistance compared to conventional cooling methods when machining lightweight alloys.

As shown in Table 4, the average flank wear values measured under each lubrication condition clearly demonstrate the superior performance of NanoMQL compared to both dry and conventional MQL environments. These findings are in agreement with those of previous studies, confirming that the use of nano-reinforced lubricants can significantly enhance tool life during the machining of magnesium alloys. The substantial reduction in flank wear observed with NanoMQL can be attributed to several synergistic effects provided by the nano-additives. The presence of MWCNTs in the lubricant improves the formation and stability of a lubricating film at the tool–workpiece interface, effectively reducing direct metal-to-metal contact and, consequently, friction and adhesive wear. Furthermore, the high thermal conductivity of MWCNTs enables more efficient heat dissipation from the cutting zone, lowering the thermal load on the cutting edge and minimizing thermal softening and diffusion wear mechanisms. In addition, these nano-additives can act as nanosized ball bearings, further reducing friction and promoting smoother chip flow. Collectively, these effects not only decrease the rate of tool wear but also contribute to improved surface finish and dimensional accuracy of the machined parts. Similar mechanisms and benefits of nano-reinforced lubricants have been reported in the literature for various machining operations involving difficult-to-cut materials [28,45]. Therefore, the superior performance of NanoMQL observed in this study is consistent with the growing body of evidence supporting the adoption of nanofluid-based lubrication strategies in advanced manufacturing.

Table 4. Average flank wear (VB) under different lubrication conditions.

Lubrication Condition	Average Flank Wear (VB, μm)
Dry	163 ± 4
MQL	94 ± 7
NanoMQL	58 ± 3

The superior performance of NanoMQL in turning operations can be attributed to the presence of multi-walled carbon nanotubes in the lubricant. These nanoparticles enhance the lubricating film strength, reduce friction at the tool–workpiece interface, and improve heat dissipation during turning. As a result, the tool experiences lower thermal and mechanical loads, leading to reduced wear. This mechanism is consistent with previous findings in the literature [46,47]. The effectiveness of different cooling and lubrication strategies in turning operations is primarily related to their ability to reduce friction and dissipate heat at the tool–workpiece interface. Under dry conditions, there is no external medium to carry away heat or reduce friction, resulting in higher tool temperatures and accelerated wear. MQL introduces a small amount of lubricant in the form of a fine mist, which partially reduces friction and provides some cooling, leading to moderate improvements in tool life and surface quality.

NanoMQL, in which multi-walled carbon nanotubes are dispersed in the lubricant, offers additional benefits. The MWCNTs act as nanosized ball bearings, further reducing friction at the interface. They can also form a protective tribo-film on the tool surface, which minimizes direct contact between the tool and the workpiece, thereby reducing adhesive and abrasive wear. Moreover, the high thermal conductivity of MWCNTs enhances heat dissipation from the cutting zone, keeping the tool temperature lower and further extending tool life. These mechanisms have been reported in several recent studies [42,48,49].

4. Conclusions

This comprehensive investigation into the machining of AZ91D magnesium alloy under dry, MQL, and NanoMQL environments has yielded several significant findings with important implications for industrial applications.

The cutting force analysis revealed that the depth of cutting and the feed rate are the primary factors influencing force magnitudes during machining, with cutting forces increasing substantially as these parameters increase. Cutting speed demonstrated a more complex relationship with cutting forces, showing variable effects that can be attributed to the competing mechanisms of thermal softening and material adhesion. The minimum cutting force (60.86 N) was achieved at a moderate cutting speed (300 m/min) with low feed rate and depth of cutting; meanwhile, the maximum force (234.42 N) occurred at high values of all three parameters.

Surface roughness measurements demonstrated the superior performance of the NanoMQL environment, which provided a 42.2% improvement over dry machining and a 33.6% improvement over conventional MQL. The feed rate consistently emerged as the most influential parameter affecting surface quality across all environments, with higher feed rates producing rougher surfaces. The finest surface finish (Ra = 0.308 μm) was achieved in the NanoMQL environment at a high cutting speed (450 m/min), a low feed rate (0.05 mm/rev), and a large depth of cut (2 mm). This exceptional performance can be attributed to the enhanced lubrication and cooling properties of the carbon-nanotube-reinforced cutting fluid.

Tool wear analysis through SEM imaging revealed significant differences in wear mechanisms across the three environments. The NanoMQL environment consistently demonstrated reduced flank wear and minimal built-up edge formation compared to both dry and MQL conditions. Compared to dry cutting, NanoMQL reduced flank wear by approximately 64.4%. Compared to conventional MQL, the reduction was about 38.3%. This was particularly evident under aggressive cutting parameters, where the dry environment exhibited extensive adhesion and BUE formation. The protective tribofilm formed by carbon nanotubes at the tool–workpiece interface effectively minimized direct contact

between materials, reducing friction and adhesion while enhancing heat dissipation from the cutting zone.

These findings collectively establish that nano-reinforced MQL represents a superior approach for machining AZ91D magnesium alloy, particularly in applications requiring high surface quality and extended tool life. The optimal machining parameters identified in this study (high cutting speed, low feed rate, and moderate–high depth of cut with NanoMQL) provide valuable guidelines for industrial implementation. Furthermore, the environmentally friendly nature of the biodegradable cutting fluid used in this study, combined with the minimal quantity application, aligns with sustainable manufacturing principles.

Future research should explore the long-term effects of carbon nanotube concentration on machining performance and investigate the potential for further optimization through the combination of nano-enhanced cutting fluids with advanced tool coatings. Additionally, the economic implications of implementing NanoMQL systems in industrial settings should be evaluated to provide a comprehensive cost–benefit analysis.

Author Contributions: Conceptualization, S.C.C. and B.B.B.; methodology, S.C.C. and B.B.B.; validation, S.C.C., B.B.B. and K.L.; formal analysis, S.C.C. and B.B.B.; investigation, S.C.C. and B.B.B.; data curation, S.C.C.; writing—original draft preparation, S.C.C. and K.L.; writing—review and editing, S.C.C. and K.L.; visualization, S.C.C. and K.L.; supervision, B.B.B.; project administration, S.C.C. and B.B.B.; funding acquisition, S.C.C. and B.B.B. All authors have read and agreed to the published version of the manuscript.

Funding: This project with the code 2019-2-TP3-3557 is supported financially by the Mersin University Scientific Research Projects Unit.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yang, Y.; Xiong, X.; Chen, J.; Peng, X.; Chen, D.; Pan, F. Research advances of magnesium and magnesium alloys worldwide in 2022. *J. Magnes. Alloys* **2023**, *11*, 2611. [CrossRef]
2. Risaliti, E.; Del Pero, F.; Arcidiacono, G.; Citti, P. Optimizing Lightweight Material Selection in Automotive Engineering: A Hybrid Methodology Incorporating Ashby's Method and VIKOR Analysis. *Machines* **2025**, *13*, 63. [CrossRef]
3. Tan, J.; Ramakrishna, S. Applications of magnesium and its alloys: A review. *Appl. Sci.* **2021**, *11*, 6861. [CrossRef]
4. Maqbool, A.; Khan, N.Z.; Siddiquee, A.N. Towards Mg based light materials of future: Properties, applications, problems, and their mitigation. *J. Manuf. Sci. Eng.* **2022**, *144*, 030801. [CrossRef]
5. Yang, J.; Zhu, Z.; Han, S.; Gu, Y.; Zhu, Z.; Zhang, H. Evolution, limitations, advantages, and future challenges of magnesium alloys as materials for aerospace applications. *J. Alloys Compd.* **2024**, *1008*, 176707. [CrossRef]
6. Manjunath, K.; Tewary, S.; Khatri, N.; Cheng, K. Monitoring and predicting the surface generation and surface roughness in ultraprecision machining: A critical review. *Machines* **2021**, *9*, 369. [CrossRef]
7. Tiwari, A.; Singh, D.; Mishra, S. A review on minimum quantity lubrication in machining of different alloys and superalloys using nanofluids. *J. Braz. Soc. Mech. Sci. Eng.* **2024**, *46*, 112. [CrossRef]
8. Sharma, A.K.; Tiwari, A.K.; Dixit, A.R. Effects of Minimum Quantity Lubrication (MQL) in machining processes using conventional and nanofluid based cutting fluids: A comprehensive review. *J. Clean. Prod.* **2016**, *127*, 1–18. [CrossRef]
9. Li, D.; Zhang, T.; Zheng, T.; Zhao, N.; Li, Z. A comprehensive review of minimum quantity lubrication (MQL) machining technology and cutting performance. *Int. J. Adv. Manuf. Technol.* **2024**, *133*, 2681. [CrossRef]
10. Pu, Z.; Outeiro, J.C.; Batista, A.C.; Dillon, O.W., Jr.; Puleo, D.A.; Jawahir, I.S. Enhanced surface integrity of AZ31B Mg alloy by cryogenic machining towards improved functional performance of machined components. *Int. J. Mach. Tools Manuf.* **2012**, *56*, 17. [CrossRef]
11. Wang, Y.; Li, C.; Zhang, Y.; Yang, M.; Li, B.; Jia, D.; Hou, Y.; Mao, C. Experimental evaluation of the lubrication properties of the wheel/workpiece interface in minimum quantity lubrication (MQL) grinding using different types of vegetable oils. *J. Clean. Prod.* **2016**, *127*, 487. [CrossRef]

12. Jouini, N.; Ruslan, M.S.M.; Ghani, J.A.; Che Haron, C.H. Sustainable high-speed milling of magnesium alloy AZ91D in dry and cryogenic conditions. *Sustainability* **2023**, *15*, 3760. [CrossRef]
13. Zagórski, I.; Korpysa, J. Surface quality in milling of AZ91D magnesium alloy. *Adv. Sci. Technol. Res. J.* **2019**, *13*, 119. [CrossRef]
14. Marakini, V.; Pai, S.P.; Bhat, U.K.; Thakur, D.S.; Achar, B.P. High-speed face milling of AZ91 Mg alloy: Surface integrity investigations. *Int. J. Lightweight Mater. Manuf.* **2022**, *5*, 528. [CrossRef]
15. Cagan, S.C.; Buldum, B.B. A study on the machinability of environmentally friendly turning of titanium grade 2 alloy. *J. Tribol.* **2024**, *146*, 064201. [CrossRef]
16. Korkmaz, M.E.; Gupta, M.K. Nano lubricants in machining and tribology applications: A state of the art review on challenges and future trend. *J. Mol. Liq.* **2024**, *407*, 125261. [CrossRef]
17. Tomac, N.; Tannessen, K.; Rasch, F.O. Machinability of particulate aluminium matrix composites. *CIRP Ann.* **1992**, *41*, 55. [CrossRef]
18. Chu, A.; Li, C.; Zhou, Z.; Liu, B.; Zhang, Y.; Yang, M.; Gao, T.; Liu, M.; Zhang, N.; Dambatta, Y.S.; et al. Nanofluids minimal quantity lubrication machining: From mechanisms to application. *Lubricants* **2023**, *11*, 422. [CrossRef]
19. Viswanathan, R.; Ramesh, S.; Subburam, V. Measurement and optimization of performance characteristics in turning of Mg alloy under dry and MQL conditions. *Measurement* **2018**, *120*, 107. [CrossRef]
20. Ji, C.; Sheng, R.; Wu, H.; Zhou, Z.; Yan, X.; Dong, L.; Li, C. Bibliometric analysis and research trends in minimum quantity lubrication for reducing cutting forces. *Int. J. Adv. Manuf. Technol.* **2024**, *135*, 4995–5033. [CrossRef]
21. Dixit, U.S.; Sarma, D.; Davim, J.P. *Environmentally Friendly Machining*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.
22. Kumar, A.; Gill, S.; Singh, G.; Sharma, S.; Dwivedi, S.; Mohammed, K.; Sharma, K.; Kozak, D.; Hunjet, A.; Abbas, M. Experimental investigation into machining performance of magnesium alloy AZ91D under dry, minimum quantity lubrication, and nano minimum quantity lubrication environments. *High Temp. Mater. Process.* **2024**, *43*, 20220328. [CrossRef]
23. Cica, D.; Tesic, S.; Markovic, M.; Sredanovic, B.; Borojevic, S.; Zeljkovic, M.; Kramar, D.; Pušavec, F. Multi-Objective Optimization of Milling Ti-6Al-4V Alloy for Improved Surface Integrity and Sustainability Performance. *Machines* **2025**, *13*, 221. [CrossRef]
24. Jawahir, I.S.; Attia, H.; Biermann, D.; Duflou, J.; Klocke, F.; Meyer, D.; Newman, S.T.; Pusavec, F.; Putz, M.; Rech, J.; et al. Cryogenic manufacturing processes. *CIRP Ann.* **2016**, *65*, 713. [CrossRef]
25. Srikant, R.R.; Rao, D.N.; Subrahmanyam, M.S.; Krishna, V.P. Applicability of cutting fluids with nanoparticle inclusion as coolants in machining. *Proc. Inst. Mech. Eng. Part J J. Eng. Tribol.* **2009**, *223*, 221. [CrossRef]
26. Trent, E.M.; Wright, P.K. *Metal Cutting*; Butterworth-Heinemann: Oxford, UK, 2000.
27. Singh, V.; Sharma, A.K.; Sahu, R.K.; Katiyar, J.K. State of the art on sustainable manufacturing using mono/hybrid nano-cutting fluids with minimum quantity lubrication. *Mater. Manuf. Process.* **2022**, *37*, 603. [CrossRef]
28. Said, Z.; Gupta, M.; Hegab, H.; Arora, N.; Khan, A.M.; Jamil, M.; Bellos, E. A comprehensive review on minimum quantity lubrication (MQL) in machining processes using nano-cutting fluids. *Int. J. Adv. Manuf. Technol.* **2019**, *105*, 2057. [CrossRef]
29. *ISO 3685:1993*; Tool Life Testing with Single-Point Turning Tools. International Organization for Standardisation: Geneva, Switzerland, 1993.
30. Sousa, V.F.; Gil, J.; Silva, T.E.; de Jesus, A.M.; Silva, F.J.; Tavares, J.M.R. An Image Analysis Algorithm for Measuring Flank Wear in Coated End-Mills. *Comput. Mater. Contin.* **2025**, *83*, 177. [CrossRef]
31. ImageJ Software, Image Processing and Analysis in Java. Available online: https://imagej.net/ij/ (accessed on 12 May 2025).
32. Sun, S.; Brandt, M.; Dargusch, M.S. Characteristics of cutting forces and chip formation in machining of titanium alloys. *Int. J. Mach. Tools Manuf.* **2009**, *49*, 561. [CrossRef]
33. Nouari, M.; List, G.; Girot, F.; Coupard, D. Experimental analysis and optimisation of tool wear in dry machining of aluminium alloys. *Wear* **2003**, *255*, 1359. [CrossRef]
34. Mabrouki, T.; Girardin, F.; Asad, M.; Rigal, J.F. Numerical and experimental study of dry cutting for an aeronautic aluminium alloy (A2024-T351). *Int. J. Mach. Tools Manuf.* **2008**, *48*, 1187. [CrossRef]
35. Dandekar, C.R.; Shin, Y.C.; Barnes, J. Machinability improvement of titanium alloy (Ti–6Al–4V) via LAM and hybrid machining. *Int. J. Mach. Tools Manuf.* **2010**, *50*, 174. [CrossRef]
36. Davim, J.P. Design of optimisation of cutting parameters for turning metal matrix composites based on the orthogonal arrays. *J. Mater. Process. Technol.* **2003**, *132*, 340. [CrossRef]
37. Shaw, M.C.; Cookson, J. *Metal Cutting Principles*; Oxford University Press: New York, NY, USA, 2005.
38. Sallakhniknezhad, R.; Ahmadian, H.; Zhou, T.; Weijia, G.; Anantharajan, S.K.; Sadoun, A.M.; Abdelfattah, W.M.; Fathy, A. Recent Advances and Applications of Carbon Nanotubes (CNTs) in Machining Processes: A Review. *J. Manuf. Mater. Process.* **2024**, *8*, 282. [CrossRef]
39. Carou, D.; Rubio, E.M.; Davim, J.P. Machinability of magnesium and its alloys: A review. *Tradit. Mach. Process.* **2015**, 133–152. [CrossRef]

40. Haghnazari, S.; Abedini, V. Effects of hybrid Al_2O_3–CuO nanofluids on surface roughness and machining forces during turning AISI 4340. *SN Appl. Sci.* **2021**, *3*, 203. [CrossRef]
41. Akyüz, B. Wear and machinability properties of AS series magnesium alloys. *Proc. Inst. Mech. Eng. Part B J. Eng. Manuf.* **2016**, *230*, 701. [CrossRef]
42. Manikanta, J.E.; Abdullah, M.; Ambhore, N.; Kotteda, T.K. Analysis of machining performance in turning with trihybrid nanofluids and minimum quantity lubrication. *Sci. Rep.* **2025**, *15*, 12194. [CrossRef]
43. Vardhanapu, M.; Chaganti, P.K.; Tarigopula, P. Characterization and machine learning-based parameter estimation in MQL machining of a superalloy for developed green nano-metalworking fluids. *J. Braz. Soc. Mech. Sci. Eng.* **2023**, *45*, 154. [CrossRef]
44. Sharma, V.S.; Dogra, M.; Suri, N. Cooling techniques for improved productivity in turning. *Int. J. Mach. Tools Manuf.* **2009**, *49*, 435. [CrossRef]
45. Aslan, A.; Salur, E. Applications of nanofluids in minimum quantity lubrication machining. *Nanomater. Manuf. Process.* **2022**, 53–84. [CrossRef]
46. Kanth, V.K.; Sreeramulu, D.; Srikiran, S.; Kumar, M.P.; Jagdeesh, K.E.; Govindh, B. Experimental investigation of cutting parameters using nano lubrication on turning AISI 1040 steel. *Mater. Today Proc.* **2019**, *18*, 2095. [CrossRef]
47. Patole, P.B.; Kulkarni, V.V.; Bhatwadekar, S.G. MQL Machining with nano fluid: A review. *Manuf. Rev.* **2021**, *8*, 13. [CrossRef]
48. Altaf, S.F.; Parray, M.A.; Khan, M.J.; Bhat, F.A. Machining with minimum quantity lubrication and nano-fluid MQL: A review. *Tribol. Online* **2024**, *19*, 209. [CrossRef]
49. Xu, M.; Ali, S.; Kurniawan, R.; Gautam, R.K.S.; Sundaresan, T.K.; Ahmad, K. Nanoparticle-based lubrication during machining: Synthesis, application, and future scope—A critical review. *Int. J. Adv. Manuf. Technol.* **2025**, *136*, 4141. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Balancing Productivity and Sustainability in EDM: A Comprehensive Analysis of Energy Consumption and Electrode Degradation

Sunil Kumar Maurya [1], Gianni Campatelli [1,*], Massimo Veracini [2], Massimo Arcioni [2] and Dario Clori [2]

[1] Industrial Engineering Department, Università degli Studi di Firenze, 50139 Firenze, Italy; sunilkumar.maurya@unifi.it
[2] Baker Hughes SRL, 50127 Firenze, Italy; massimo.veracini@bakerhughes.com (M.V.); massimo.arcioni@bakerhughes.com (M.A.); dario.clori@bakerhughes.com (D.C.)
* Correspondence: gianni.campatelli@unifi.it; Tel.: +39-3204644763

Abstract: Die-sinking Electrical Discharge Machining (EDM) is a manufacturing process for fabricating complex geometries in challenging applications. However, its energy-intensive nature and complex parameter interactions pose challenges in balancing productivity, sustainability, and electrode wear. This study presents a comprehensive analysis of energy consumption and electrode degradation in EDM. Utilizing an advanced experimental setup with real-time energy monitoring, this study investigated the trade-off between machining parameters, energy efficiency, and electrode wear. The study employed a simple and standardized electrode geometry and varied EDM parameters, such as discharge current and pulse duration. The obtained results clearly demonstrated that optimizing EDM machining parameters, particularly discharge current, significantly influenced machining efficiency and electrode wear. Specifically, employing high-current settings of 140 A substantially reduced the total machining time from approximately 33 h (at conservative settings of 40 A) down to around 3.5 h, achieving nearly a tenfold improvement. Moreover, it also led to a reduction in specific energy consumption (SEC), decreasing from 0.81 Wh/mm^3 at the low current (40 A) to 0.19 Wh/mm^3 at the higher current (140 A), underscoring a definitive inverse relationship between discharge current and energy consumption. The study outcomes provide practical guidelines for enhancing the operational efficiency and sustainability of EDM in advanced manufacturing sectors.

Keywords: electrical discharge machining (EDM); energy consumption; electrode degradation; machining parameters; material removal rate (MRR); specific energy consumption (SEC)

1. Introduction

Die-sinking Electrical Discharge Machining (EDM) is widely acknowledged as an advanced and key manufacturing process, particularly for fabricating complex geometries with materials that are traditionally difficult to machine. This technique operates through a controlled sequence of electrical discharges, facilitated by a dielectric fluid, with the aim of achieving a very high surface finish [1]. As presented in Figure 1, the erosion process replicates the contour of the electrode. This achieves high fidelity in the reproduction of complex shapes, a feature that is critically important for high-precision sectors, such as aerospace, automotive, and tooling, where conventional machining methods are often inadequate [2,3]. Products produced using such techniques are molds and dies or fully machined impellers for the energy sector. Despite its significant advantages, die-sinking

EDM faces challenges owing to its inherent energy-intensive nature and the complexities associated with its operational parameters. Key machining factors—namely, electrical current, applied voltage, and pulse duration—exhibit a multifaceted correlation that directly affects the material removal rate (MRR), overall machining time, and surface integrity of the workpiece [4]. Consequently, achieving optimal process performance requires the fine-tuning of these parameters to enhance productivity while concurrently mitigating environmental and economic costs [5].

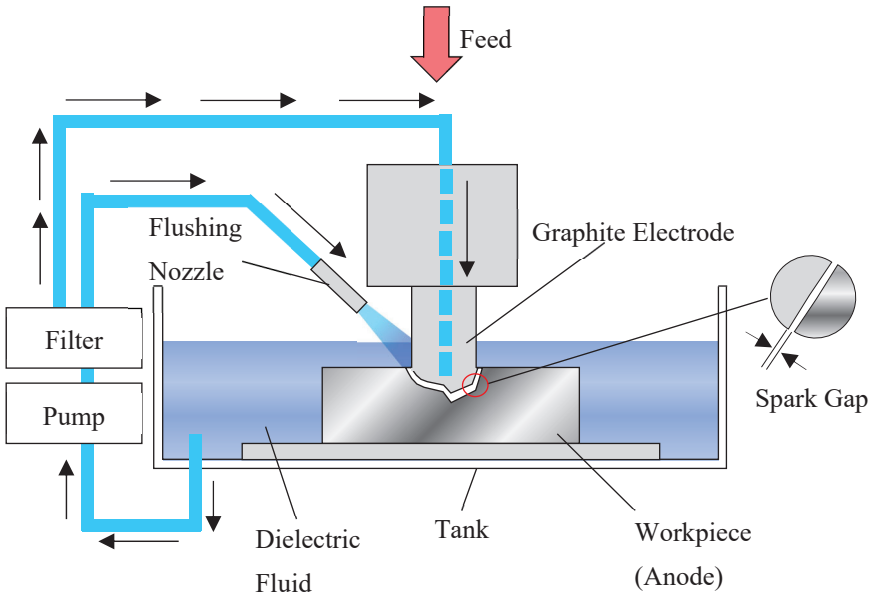


Figure 1. Schematic diagram of die-sinking EDM.

The motivation for investigating energy consumption and electrode degradation in EDM arises primarily from the significant operational costs and environmental concerns associated with energy-intensive manufacturing processes, as well as the economic impact of frequent electrode replacement or reconditioning. With industries increasingly focusing on sustainability and efficient resource utilization, understanding the interplay between EDM parameters, energy usage, and electrode wear becomes essential to improve machining efficiency, reduce manufacturing costs, and minimize the environmental footprint.

In response to these challenges, considerable research efforts have been directed toward developing advanced solutions. Innovations have focused on designing and implementing high-durability electrode materials and geometries that significantly enhance machining precision and efficiency [6]. Simultaneously, the energy consumption of die-sinking EDM machines has emerged as a critical metric influenced by multiple factors, including the nature of the discharge process, auxiliary systems, and machine design. Therefore, the accurate measurement and control of energy consumption are essential not only for improving machining efficiency but also for reducing operational costs and minimizing environmental impacts. This growing emphasis on the environmental footprint of manufacturing processes has led to the development of energetic models that provide a robust framework for evaluating overall energy utilization in machining processes. These models systematically analyze the power absorbed by all the machine subsystems. They employ both system sizing (e.g., subsystem nominal power) and direct measurement techniques to accurately determine total power consumption during machining operations [3,6,7].

This quantitative approach facilitates a comprehensive understanding of energy distribution throughout the process, enabling precise assessments that are critical for optimizing the process performance. Moreover, by incorporating the energy demands of auxiliary systems—such as chillers, which often impose a significant energy burden on micro-EDM

machines—these models empower manufacturers to identify and target specific areas for energy consumption optimization [8,9]. In an automotive case study, the researchers used line balancing to reduce waiting time and improve assembly efficiency. While effective for task allocation in linear workflows, their study did not address machining systems. This research adapts lean principles to EDM, where optimization must also consider thermal effects, energy demand, and tool wear [10]. Complementing these energetic models, discharge-based models offer a focused perspective on the dynamics of the electrical discharge process by correlating essential machining parameters—such as current, voltage, and pulse timing—with the material removal rate (MRR), thus providing estimates of the energy required for effective material removal [4,6]. This study successfully applied Manufacturing Cycle Efficiency (MCE) to reduce time wastage and improve productivity in automotive wiring harness assembly. However, their work focused on mechanical operations without considering energy-intensive processes. The present study extends MCE analysis to EDM, integrating energy consumption and electrode wear—factors critical for optimizing subtractive manufacturing systems [11]. This targeted analysis enhances the ability to fine-tune machining operations by directly linking the process parameters with the energy efficiency outcomes. Further insights are gained by isolating the energy consumed during active discharge from that used by auxiliary systems, which yields valuable data to assess and improve the intrinsic performance of the process itself [12].

Advancements in measurement techniques have led to the integration of real-time monitoring systems into modern machining environments. These systems combine advanced sensors with sophisticated data acquisition techniques to continuously track energy consumption across all machine subsystems, enabling the prompt detection of inefficiencies and allowing operators to dynamically adjust process parameters to maintain optimal energy use [7,8]. In addition to these systems, acoustic emission (AE) monitoring techniques have emerged as non-invasive methods for assessing energy fluxes during machining operations. By analyzing the vibroacoustic signals generated during material removal, AE monitoring provides a real-time evaluation of the energy expended, ensuring that the process maintains optimal performance without disrupting the operational workflow [13].

An important concept in the evaluation of energy efficiency is the specific energy consumption (SEC) metric, which is the energy consumed to remove a standard volume of material, crucial for assessing the overall efficiency of an EDM process. Hence, the SEC provides a valuable comparative framework for evaluating the performance of different machining setups. This metric is particularly beneficial for contrasting the energy performance of conventional EDM technologies with that of more advanced systems, thereby driving further process improvements and optimization [14,15].

At the heart of these energy considerations lies the complex relationship between the material removal rate (MRR) and machining parameters in die-sinking EDM, a relationship characterized by nonlinear interactions and trade-offs. A thorough understanding of these correlations is critical for optimizing the process to achieve high productivity and energy efficiency while maintaining surface quality. In advanced machining processes, several key electrical parameters play a decisive role in influencing both the MRR and overall process performance. One of the most influential of these parameters is the peak current (I_p)—a peak in current intensity that amplifies the energy delivered during each electrical discharge, thereby accelerating the material removal process. However, this benefit is counterbalanced by elevated thermal loads, which can adversely affect the surface finish and accelerate electrode wear [5,16]. This relationship between current and MRR is particularly significant in applications such as mold making and aerospace manufacturing, where the demand for rapid material removal often takes precedence over the need for a high-quality surface finish [17].

Complementing the role of current, voltage (V_x) is fundamental in regulating the energy transfer during each discharge. Maintaining an optimal voltage level is essential for ensuring stable sparking conditions that promote uniform material erosion; conversely, excessive voltage can lead to arcing and unstable discharge conditions, which detrimentally affect both the MRR and the quality of the machined surface [17,18]. Furthermore, the pulse timing parameters, specifically the pulse-on time (T_{on}) and pulse-off time (T_{off}), further refine the efficiency of material removal. Extended pulse-on times facilitate greater energy transfer per discharge, thereby enhancing the MRR, but the increased energy transfer also elevates the risk of thermal damage to both the workpiece and the electrode, potentially compromising surface integrity [16]. In contrast, shorter pulse durations focus energy delivery more precisely, reducing the heat-affected zones, although this approach may limit the overall material removal rate. Additionally, the pulse-off time is critical, as it governs the recovery period for the dielectric fluid and ensures the effective removal of debris; insufficient off time may result in erratic discharges, diminishing process efficiency and intensifying electrode wear [17].

Ultimately, the intricate correlation between these electrical parameters underscores the necessity for a holistic optimization strategy, where adjustments made to enhance one parameter often require compensatory modifications in others to sustain process stability and efficiency. For instance, increasing the current to boost the MRR may necessitate a reduction in pulse duration or voltage to mitigate the risks associated with thermal damage and excessive electrode wear [19]. Through such a balanced and integrated approach, manufacturers can effectively tailor die-sinking EDM processes to meet specific operational demands while achieving high levels of precision and energy efficiency.

Numerous external studies have explored the energy aspects of EDM, such as energetic modeling and parameter optimization, to improve efficiency and electrode lifespan. However, these studies have often isolated theoretical modeling from real-world applications, rarely integrating real-time monitoring data or considering the compounded impact of auxiliary machine systems. The current research bridges these critical gaps by conducting empirical investigations that combine advanced real-time monitoring systems with systematic experimental analyses. This integrated approach not only expands upon previous external findings but also provides practical insights into parameter optimization, specifically addressing simultaneous considerations of machining productivity, energy efficiency, and electrode wear. A key novelty is the use of a simplified, standardized graphite electrode across all trials, which ensures consistent volume removal and enables accurate comparisons of specific energy consumption under varying process settings. This allows for a deeper understanding of how parameter changes impact both energy use and electrode degradation—delivering actionable insights for optimizing EDM processes in both performance and sustainability.

Therefore, the primary objective of this research is to empirically quantify the effects of key EDM machining parameters—particularly discharge current and pulse duration—on energy consumption, machining efficiency, and electrode wear. It is hypothesized that optimized combinations of these parameters significantly lower specific energy consumption (SEC) and machining time without causing unacceptable electrode degradation. The experimental validation of these hypotheses, detailed in the following section, directly addresses the limitations identified in prior research.

2. Materials and Methods

In this study, an Agie Charmilles FORM 3000 VHP, manufactured by GF machining solutions, Milan Italy, presented in Figure 2, was used to conduct the experiments and collect data in a plant owned by Baker Hughes, a company that operates worldwide in

the energy sector. The machine has been equipped with a Modbus-capable energy meter manufactured by Ceam Group, Empoli, Italy, used for acquiring energy consumption data. The current was recorded using Class 1, 100 A:5 A open-core clamp-on current transformers. The open-core design enables easy installation by clamping around the power cables without the need for disconnection or interruption of the power supply. The schematics in Figure 3 shows the data collection process using the RS485 Modbus protocol wirelessly over WiFi.



Figure 2. Agie Charmilles Form 3000 VHP EDM Machine.

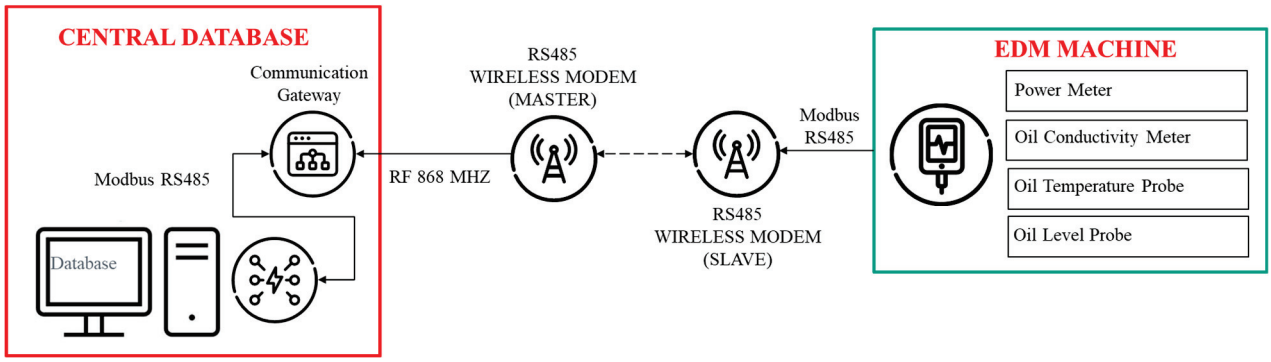


Figure 3. Data acquisition setup schematics.

From the schematics, it can be seen that the machine, as shown, is fitted with a power meter, an oil conductivity meter, an oil temperature probe, and an oil level probe. However, for this study, we will only discuss the results from the power meter and PLC data. This information is first sent to a WiFi modem using the RS485 protocol, which is then transmitted to another control or master WiFi modem that receives it and sends it to a Communication Gateway. The Communication Gateway is located in the central database cluster. This cluster consists of a locally hosted database that stores all the information coming from the production machines. This information is useful for controlling the overall performance of all the plants and comparing the relative efficiency of the installed machines.

Although the data acquisition system employed high-resolution power meters and the real-time monitoring of auxiliary systems, several sources of error may still affect the measurements. Potential uncertainties arise from the dynamic fluctuations in discharge pulses, minor delays in power signal synchronization, and thermal drift in sensor calibration over extended machining cycles. Additionally, the indirect measurement of SEC relies

on the assumption of a constant material removal volume per trial, which may be affected by unmeasured electrode micro-chipping or flushing inefficiencies. While these factors were minimized through repeated trials and equipment calibration, they represent inherent limitations in high-resolution energy analysis in EDM.

In this study, three sets of experiments were performed to determine the optimal setup for the process. The setup, analysis, and results are presented in the following sections.

3. Experiment and Setup

3.1. Preliminary Experiment

The first experiment involved an in-depth analysis of the energy consumption characteristics of the target EDM machine, conducted via a series of cutting tests employing production electrodes. This phase of the study focused on the machining process during the roughing stage of impeller vane development, with details provided in Figure 4.

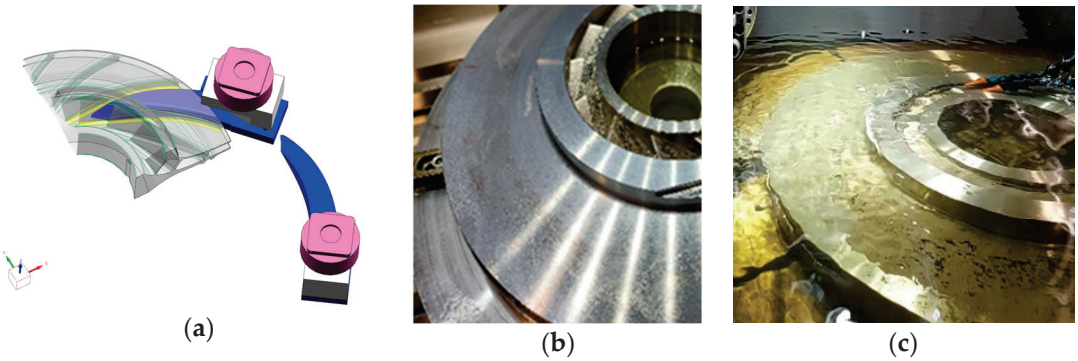


Figure 4. (**a**) Illustration of the impeller and electrode, (**b**) impeller stock, (**c**) impeller in an EDM machine.

A single-electrode geometry featuring an impeller vane was employed, and four different combinations of parameters were tested. The dielectric fluid used in all the experiments was standard-grade kerosene, which is the default working fluid in the production line. Kerosene is widely used in industrial EDM applications due to its high dielectric strength, moderate viscosity, and effective cooling and flushing characteristics. The machine's internal system maintained a continuous circulation of the fluid, ensuring stable flushing during machining. Although the exact flow rate is not directly adjustable or specified by the manufacturer, it is automatically controlled to meet the default preset conditions for each selected machining mode. The discharge current values selected for this study (40 A, 80 A, 100 A, and 140 A) correspond to the available preset configurations provided by the EDM machine manufacturer. These presets are pre-programmed into the machine and represent standard industrial settings optimized by the manufacturer for different machining objectives. Importantly, the only variable intentionally modified in this study was the discharge current; all other parameters—including pulse duration, duty cycle, and flushing conditions—were left at their default values as defined by the machine's internal control logic. The two preset types—High Material Removal Rate (High MRR) and Low Tool Wear—are built-in technology options that cannot be manually altered in terms of their internal settings, ensuring consistency and repeatability. This approach reflects realistic industrial usage scenarios, where operators rely on certified machine presets to balance performance and tool wear without manually tuning every discharge parameter. The preset and pulse current settings used in this experiment are given in Tables 1 and 2.

Table 1. Types of test cases performed.

Test N°	Ampere	Preset Type
1	80	2 (Low Tool Wear)
2	100	1 (High M.R.R.)
3	100	2 (Low Tool Wear)
4	80	1 (High M.R.R.)

Table 2. Preset specifications.

Preset 1 High MRR			Preset 2 Low Tool Wear		
Specifications	**Values**	**Units**	**Specifications**	**Values**	**Units**
T: Pulse-on time	154	µs	T: Pulse-on time	422	µs
P: Pulse-off time	87	µs	P: Pulse-off time	87	µs
I: Discharge current	user-defined	amps	I: Discharge current	user-defined	amps
U: Voltage	100	volts	U: Voltage	100	volts

Before the four tests from Table 1 were performed, the base energy consumption of the EDM machine while idle was measured, and after each test, the total time, mean machining power, total energy, energy consumed by the refrigeration unit, and fixed spark current were recorded. The plots for total system power vs. time are shown in Figures 5–8 for all four test cases. These plots are raw unlabeled data; hence, subsystem differences cannot be identified in these plots.

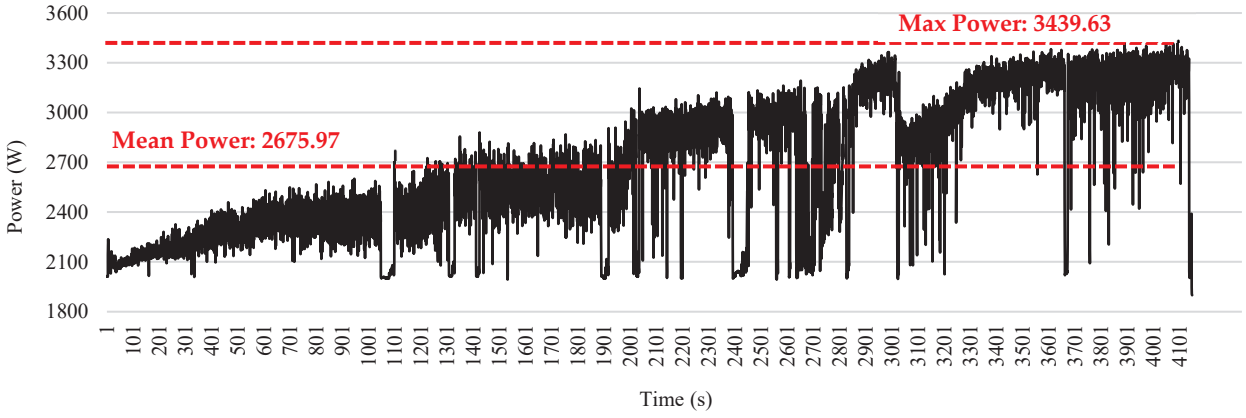


Figure 5. Test 1—Low Tool Wear at 80 amps.

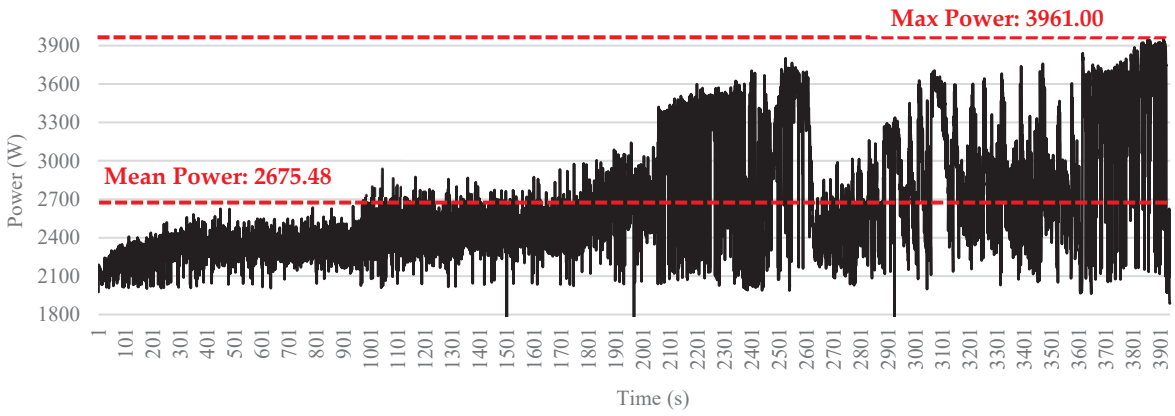


Figure 6. Test 2—High MRR at 100 amps.

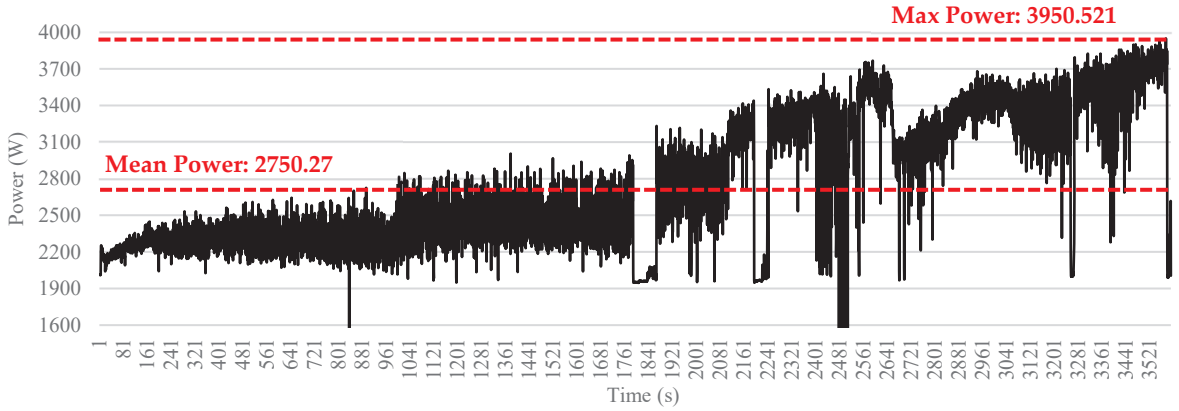


Figure 7. Test 3—Low tool Wear at 100 amps.

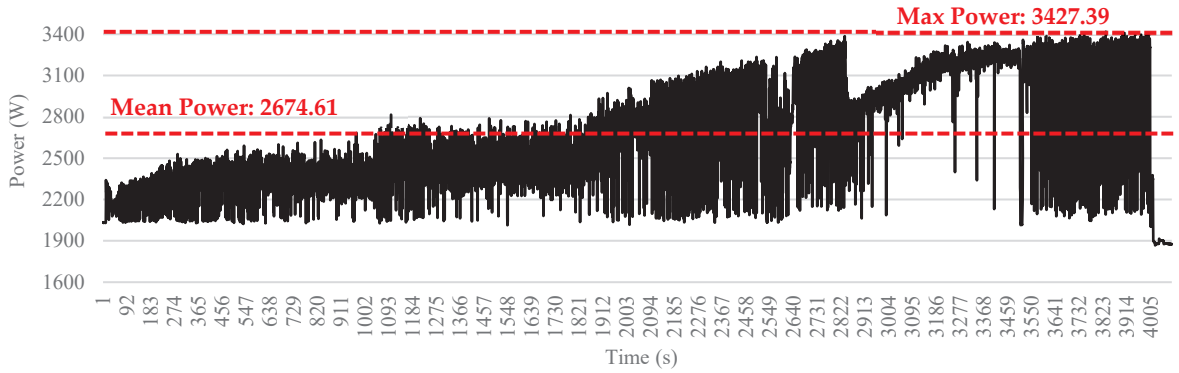


Figure 8. Test 4—High MRR at 80 amps.

Upon analyzing the results seen in Table 3, it became evident that Test 3, which used the preset "Low Tool Wear, 100 A", was the most energy-efficient and time-effective setting among all the tests. This test achieved the shortest machining time of approximately 3716 s and consumed the least amount of total energy, 10,220.4 kJ. Furthermore, this preset recorded the highest spark current value of 28.91 A, indicating a more aggressive electrical discharge rate. Despite the higher current, the overall system efficiency was enhanced, as indicated by the lowest energy-per-second ratio among all the tests. This suggests that a higher pulse current, when optimized correctly, can result in faster machining without significantly increasing energy costs.

In contrast, Test 1, corresponding to the "Low Tool Wear, 80 A" preset, was the least efficient of the four. It had the longest machining time of approximately 4303 s and consumed the highest total energy, 11,523.6 kJ. Although the spark current was considerably lower at 18.19 A, the longer operational time offset any potential energy savings, leading to an overall reduced efficiency. Notably, the refrigeration energy consumption in this case was also the highest, indicating longer support system usage owing to the extended machining duration.

Tests 2 and 4, which used the "High MRR" presets at 100 A and 80 A, respectively, showed similar performance. Their machining times and total energy consumption figures were slightly higher than those of Test 3 but lower than those of Test 1. Their energy efficiency (measured in joules per second of machining time) remained relatively constant at approximately 2676.24–2673.36 J/s, further highlighting that small differences in machining duration can significantly affect overall energy usage. This can also be seen in a detailed analysis of the last 100 s of the manufacturing cycle, when the power graph is normalized and overlapped with other types of presets, as reported in Figure 9.

Table 3. Test case results.

	Parameters	Total Time	Base Energy kJ	Machining Energy kJ	Total Energy kJ	External Aux System (Refrigeration) kJ	Fixed Spark Current A	Energy per Second J/s	Verdict
1	Low Tool Wear 80 A	01:11:43.314	6886.8	4636.8	11,523.6	12,909.9	18.19	2677.68	Worse Parameter
2	High MRR 100 A	01:07:48.747	6508.8	4377.6	10,890	12,206.2	17.78	2676.24	Neutral
3	Low tool Wear 100 A	01:01:56.287	5943.6	4284	10,220.4	11,148.8	28.91	2750.04	Best Parameter
4	High MRR 80 A	01:10:26.957	6764.4	4539.6	11,300.4	12,680.8	15.76	2673.36	Neutral

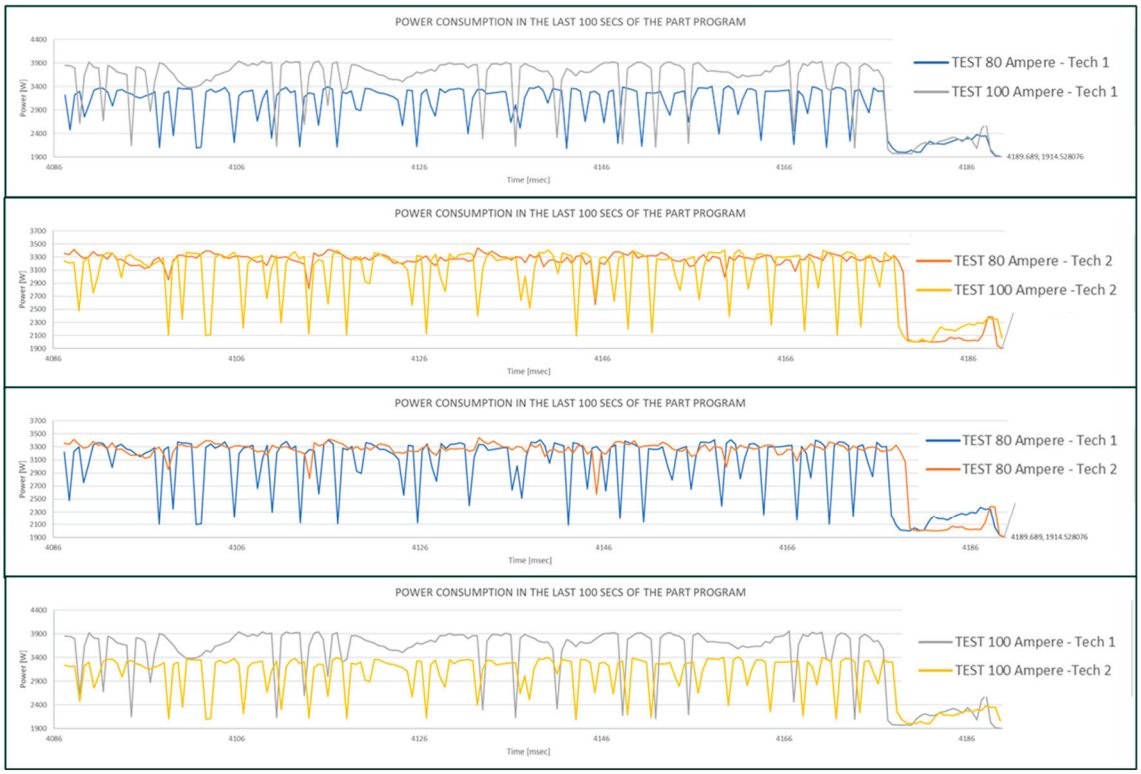


Figure 9. Power consumption of the four tests.

The energy consumed by the refrigeration unit in all the tests was substantial, accounting for more than the energy used in actual machining. However, Test 3 again stood out, with the lowest refrigeration energy demand (11,148.8 kJ), likely due to the reduced overall operation time. This reinforces the importance of minimizing machining duration, not only for power savings at the electrode but also to reduce the energy draw from auxiliary systems, which could constitute a significant share of the energy consumed.

In conclusion, this study demonstrates that the preset of Test 3 (Low Tool Wear, 100 A) offers the best compromise between machining speed and energy efficiency. It outperformed the other presets in every metric considered—shortest machining time, lowest total energy, and highest energy efficiency—making it the optimal choice for applications in which tool wear is manageable. However, Test 1 (Low Tool Wear, 80 A) is less favorable because of its longer machining time and higher total energy cost, despite its conservative current setting.

3.2. Standardized Electrode Shape Experiment

The results obtained from the initial experiment indicate the optimal solution for the impeller case study; however, the complex geometry of the electrode introduced significant challenges concerning the repeatability and reproducibility of this solution. The variability in electrode wear patterns, alterations in flushing fluid dynamics, and unpredictability of debris formation collectively influenced the total energy consumption and degradation characteristics of the electrode over the duration of machining. To address these uncertainties, a second experiment was conducted using a simplified electrode geometry. Specifically, a standardized graphite electrode was fabricated as a rectangular cuboid measuring 90 mm × 90 mm × 20 mm (length × breadth × height), as presented in Figure 10.

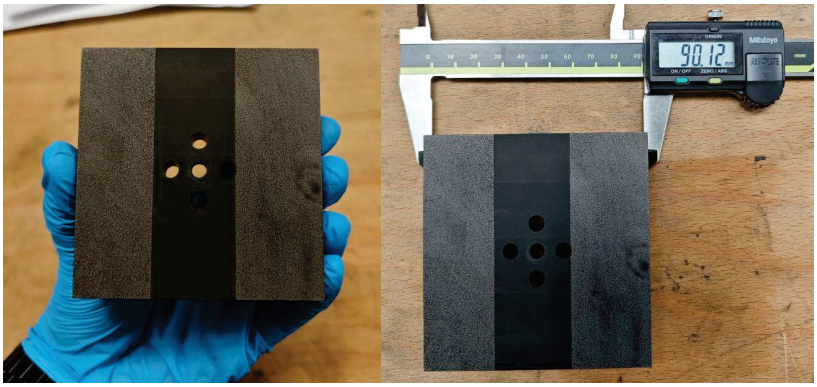


Figure 10. Standard electrode geometry used for tests.

A comprehensive set of 18 trials was performed, systematically varying the EDM parameters, including the discharge current, pulse duration, and other inherent machine "technology" settings or presets. Specifically, in this experiment, another preset was used in addition to presets 1 and 2 used in previous experiments. All three presets and their specifications are given in Tables 4 and 5. Each experimental run involved the erosion of a consistent metal volume of 5400 mm^3, defined by the intended contact area dimensions of 90 mm × 30 mm × 2 mm, as shown in Figure 11.

Table 4. EDM machine presets (technologies).

Tech 1 High MRR			**Tech 2 Low Tool Wear**			**Tech 3 50% & 50%**		
Pulse Time (T)	154	µs	Pulse Time (T)	422	µs	Pulse Time (T)	237	µs
Pulse Gap (P)	87	µs	Pulse Gap (P)	87	µs	Pulse Gap (P)	87	µs
Current	User-Defined	A	Current	User-Defined	A	Current	User-Defined	A
Voltage	100	V	Voltage	100	V	Voltage	100	V

Table 5. User-defined parameters for the experiment.

18 Tests Removing 5400 mm^3 (90 mm × 30 mm × 2 mm)	
40	1 (High M.R.R.)
60	2 (Low Tool Wear)
80	3 (50% & 50%)
100	
120	
140	

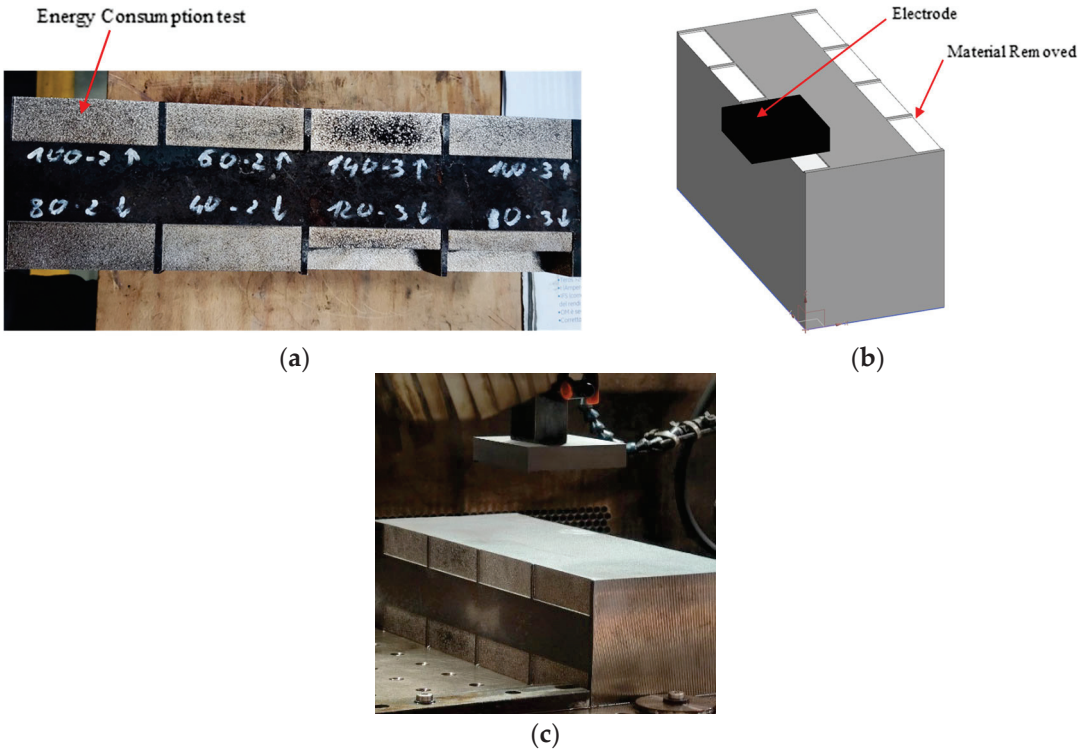


Figure 11. (**a**) Billet showing electrode indents after machining, (**b**) schematic of electrode with respect to raw material, and (**c**) raw material setup on EDM machine.

During these tests, the power sensors recorded the electrical load of the EDM machine, enabling the calculation of total energy consumed for each 5400 mm^3 removed. From this, the SEC for material removal was computed as a key metric for sustainability. In parallel, the graphite tool was closely examined for evaluating electrode wear after each test by measuring the electrode before and after machining.

From the initial 18 trials, trends have been highlighted. The results are reported in Table 6 in the Test ID column, and the test cases are named by the current value and type of preset used. For example, for TEST 140-3, the current used was 140 amperes, and the preset used was Tech 3. The data showed that aggressive EDM parameters that increased the material removal rate (MRR) dramatically shortened the machining time for a given volume. In contrast, more conservative settings (designed to minimize tool wear) led to much longer erosion times. The faster the removal, the lower the total energy consumed in completing the cut, even if the instantaneous power draw was higher. In fact, it was found that "higher machining time" and, hence, slower cutting, led to higher overall energy consumption, whereas a higher MRR resulted in lower specific energy consumption. This outcome was consistent across the trials and was a crucial insight, indicating that any strategy to reduce machining time (within practical limits) would directly save energy per part. The observed reduction in machining time with increasing discharge current, without a proportional increase in total energy consumption, can be attributed to the decreasing relative contribution of auxiliary system overheads. At lower currents, prolonged erosion times result in the extended operation of subsystems, such as dielectric circulation, machine control units, and servo motors. These auxiliary components contribute significantly to the total energy footprint during long cycles. In contrast, higher currents remove material more aggressively, thereby shortening the overall operation duration and effectively compressing the auxiliary energy load into a smaller time frame. This shift leads to lower specific energy consumption (SEC) values at higher-current settings, despite the rise in instantaneous power.

Table 6. Results for SEC tests on a simple geometry.

	Total Erosion Time	Delta	Energy Consumption	Delta	Average Energy Consumption Every Second	Delta	Removed Material	Material Removal Rate	Specific Energy Consumption
TEST ID	**[min]**	**[%]**	**[kWh]**	**[%]**	**[Wh]**	**[%]**	**[mm³]**	**[mm³/min]**	**[Wh/mm³]**
TEST 140-3	15.3	−88%	0.82	−81%	1.10	0%	5184	338.087	0.159
TEST 120-3	17.8	−86%	0.89	−80%	0.98	−10%	5688	319.285	0.156
TEST 140-2	19.1	−85%	1.00	−77%	1.12	2%	5400	283.394	0.186
TEST 140-1	22.0	−82%	1.02	−77%	0.88	−20%	5400	245.440	0.190
TEST 100-3	23.6	−81%	1.10	−75%	0.87	−20%	5184	219.542	0.213
TEST 100-2	24.1	−81%	1.11	−75%	0.98	−11%	5184	215.094	0.214
TEST 120-1	27.1	−78%	1.17	−73%	0.82	−25%	5400	199.015	0.217
TEST 120-2	24.2	−81%	1.18	−73%	1.03	−6%	5400	222.848	0.219
TEST 80-3	29.4	−77%	1.27	−71%	0.82	−25%	5688	193.518	0.224
TEST 80-2	30.3	−76%	1.30	−70%	0.92	−16%	5688	187.734	0.229
TEST 100-1	32.0	−75%	1.31	−70%	0.80	−27%	5400	168.763	0.243
TEST 80-1	44.0	−65%	1.75	−60%	0.77	−30%	5400	122.625	0.323
TEST 60-3	52.7	−58%	2.11	−52%	0.82	−25%	5382	102.190	0.391
TEST 60-1	73.5	−42%	2.69	−38%	0.78	−29%	5400	73.517	0.498
TEST 40-2	75.0	−40%	2.74	−37%	0.80	−27%	5688	75.849	0.482
TEST 40-1	125.6	0%	4.37	0%	0.74	−32%	5400	42.992	0.810

The SEC showed a distinct inverse correlation with the machining current. At higher currents (e.g., 140 A), the SEC dramatically decreased to approximately 0.19 Wh/mm^3 due to the significantly reduced erosion time and thus lower auxiliary system energy usage. Conversely, at lower currents (e.g., 40 A), the SEC increased to around 0.81 Wh/mm^3, reflecting extended machining durations that disproportionately amplified energy overheads. Test 140-3, characterized by high-current and balanced technology settings, was the most energy-efficient scenario, with a SEC of just 0.159 Wh/mm^3, whereas Test 40-1 showed the highest SEC of 0.810 Wh/mm^3. This nearly fivefold difference clearly emphasizes the importance of parameter optimization to minimize total energy consumption and increase sustainability. The observed SEC trends offer manufacturers practical guidance—strategically increasing discharge currents, within electrode wear limits, can markedly improve sustainability and reduce operational costs, particularly by cutting the auxiliary system energy demands associated with prolonged machining cycles.

An electrode wear analysis was also performed using the two extreme test cases from experiment 2 in Table 6 (Tech 3 was ignored, as in most cases, the results were not optimal). A comparison was made between Test 1 (higher current of 140 amperes, Technology 1) and Test 2 (much lower current of 40 amperes, Technology 2). As reported in Figure 12, both tests aimed to remove an identical volume of approximately 129,600 mm^3 from the workpiece. However, notable differences emerged between the two tests regarding erosion time, achieved volume removal, and electrode tool wear.

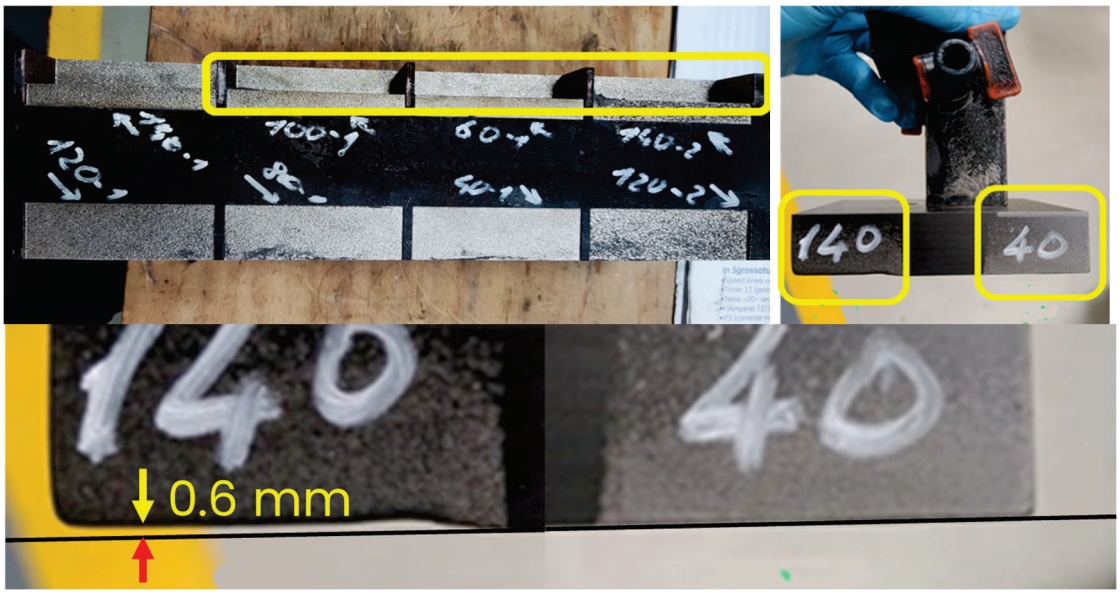


Figure 12. Comparison of TEST 140-1 and TEST 40-2.

Test 1 (140 A) was conducted aggressively, prioritizing high material removal rates, and the machining was completed in approximately 210 min. However, the actual volume removed was marginally below the intended target, specifically 127,980 mm^3, reflecting a shortfall of approximately −1.25%. The close-up images provided clearly show the measurable electrode wear—approximately 0.6 mm of material loss on the graphite electrode, as seen in Figure 12. Owing to this wear, the electrode's ability to achieve the full intended machining depth was slightly compromised. Therefore, a time adjustment was calculated to account for the shortfall, adding an additional 2.6 min (1.25%) of machining time for a corrected total duration of 212.6 min had the full volume been achieved.

Test 2, which employed a more conservative approach at 40 A (Technology 2), demonstrated a drastically different scenario. This low-current method significantly minimized electrode wear so effectively that the removed volume matched precisely the expected 129,600 mm^3 without any measurable deviation. However, this advantage came with an exceedingly prolonged erosion time of 2010 min (over 33 h), highlighting the practical trade-off between electrode longevity and productivity.

In terms of electrode degradation, a clear increase in volumetric wear was observed with increasing current. However, when normalized over machining time, the wear rate (expressed in mm^3/min) showed only a modest increase. For instance, at 40 A, the wear rate was approximately 0.018 mm^3/min, while at 140 A, it rose to 0.023 mm^3/min. This relatively low increase in wear rate compared to the large decrease in machining time suggests that higher-current presets provide a favorable trade-off, especially when the electrode replacement cost is not a limiting factor. This also reinforces the viability of high-current EDM as a productivity-oriented yet sustainable choice.

4. Conclusions

This research rigorously investigated the correlation between energy consumption and electrode wear in die-sinking Electrical Discharge Machining (EDM), addressing a notable gap in the current literature concerning comprehensive empirical evaluations of real-time energy usage integrated with precise electrode wear assessments. EDM has long been recognized as a crucial process for machining intricate geometries in challenging materials. However, prior studies have typically focused either solely on theoretical energy modeling or isolated analyses of individual machining parameters. These studies often

neglected a holistic examination of the combined influences of process settings, auxiliary system loads, and electrode deterioration.

The novelty of the present study lies in its methodical integration of empirical data collection techniques and robust analytical methodologies, facilitating a holistic appraisal of EDM operational dynamics. Thus, this study explicitly considered auxiliary energy demands, particularly from refrigeration systems. By doing so, it provided a comprehensive portrayal of the energy profile during EDM operations. This is a distinct advancement beyond the conventional fragmented analytical approaches prevalent in the existing literature.

The experimental outcomes yielded significant insights. They particularly highlighted that optimized high-current EDM parameters substantially improved both machining productivity and energy efficiency. This counters the prevalent notion that aggressive discharge settings invariably lead to undesirable electrode wear. Specifically, comparative analyses between high-current (140 A) aggressive machining and conservative low-current (40 A) settings highlighted a critical trade-off. The conservative parameters successfully minimized electrode wear to negligible amounts but at the cost of drastically increased machining times and correspondingly higher cumulative energy consumption. In contrast, the strategically optimized high-current setting notably reduced the overall machining time by up to an order of magnitude while maintaining acceptable electrode wear levels. Quantitative metrics reinforced this finding, demonstrating a lower specific energy consumption (SEC), a pivotal sustainability indicator, under high-current, optimized conditions.

These results have substantial implications for advanced manufacturing sectors, particularly the aerospace, automotive, and precision tooling industries, where operational efficiency, cost-effectiveness, and sustainability are paramount. This study empirically substantiates the viability and benefits of employing aggressive EDM settings. For example, for operations where electrode wear within ~0.6 mm is acceptable, using a 140 A discharge current with the High MRR preset is strongly advised, as it significantly reduces machining time and achieves the lowest specific energy consumption (0.159 Wh/mm^3) among the tested settings. On the other hand, for applications requiring minimal tool wear, the Low Tool Wear preset at 100 A offers a balanced compromise between machining speed and electrode longevity. Future studies can explore intermediate current levels (e.g., 60 A, 120 A) to identify more precise thresholds for optimization. Additionally, investigating alternative electrode materials such as copper–tungsten or composite electrodes may reveal further opportunities for enhancing wear resistance while maintaining energy efficiency. Expanding the analysis to include different workpiece materials and dielectric fluids could also help generalize these findings across a broader range of EDM applications.

For now, this study provides practical guidelines to enable process engineers to significantly enhance productivity and energy efficiency without incurring prohibitive electrode wear or environmental penalties.

Author Contributions: Conceptualization, S.K.M., G.C. and M.V.; methodology, S.K.M.; software, S.K.M.; validation, S.K.M., G.C., M.V. and M.A.; formal analysis, S.K.M.; investigation, S.K.M., M.V. and M.A.; resources, M.V., M.A. and D.C.; data curation, S.K.M.; writing—original draft preparation, S.K.M.; writing—review and editing, S.K.M. and G.C.; visualization, S.K.M.; supervision, G.C. and M.V.; project administration, G.C. and D.C.; funding acquisition, G.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data is unavailable publicly due to data privacy and the NDA clause of the Baker Hughes SRL company data privacy policy. Data can be provided to research paper reviewers upon request.

Acknowledgments: The authors gratefully acknowledge the support of Baker Hughes SLR, Florence, Italy, for providing access to the EDM machinery, graphite electrodes, and technical resources required to conduct the experiments. The valuable contribution of labor hours and in-kind support from the engineering team at Baker Hughes played a critical role in the successful execution of this research. Their collaboration and technical assistance are sincerely appreciated.

Conflicts of Interest: The authors Massimo Veracini, Massimo Arcioni and Dario Clori were employed by the company Baker Hughes SRL. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

EDM	Electrical Discharge Machining
SEC	Specific Energy Consumption
MRR	Material Removal Rate
MCE	Manufacturing Cycle Efficiency
AE	Acoustic Emission
PLC	Programmable Logic Controller
VHP	Very High Performance (as in FORM 3000 VHP)
RS485	Recommended Standard 485 (a serial communication protocol)

References

1. Peças, P.; Henriques, E. Intrinsic innovations of die sinking electrical discharge machining technology: Estimation of its impact. *Int. J. Adv. Manuf. Technol.* **2009**, *44*, 880–889. [CrossRef]
2. Ahmed, N. Die-sinking EDM performance under the effect of tool designs and tool materials. *Int. J. Interact. Des. Manuf. (IJIDeM)* **2024**, *18*, 1–22. [CrossRef]
3. Marrocco, V.; Modica, F.; Fassi, I.; Bianchi, G. Energetic consumption modeling of micro-EDM process. *Int. J. Adv. Manuf. Technol.* **2017**, *93*, 1843–1852. [CrossRef]
4. Tanjilul, M.; Kumar, A.S. Die-sinking of super dielectric based electrical discharge machining using 3D printed electrodes. *Procedia CIRP* **2020**, *95*, 471–475. [CrossRef]
5. Pujiyulianto, E.; Suyitno. Effect of pulse current in manufacturing of cardiovascular stent using EDM die-sinking. *Int. J. Adv. Manuf. Technol.* **2021**, *112*, 3031–3039. [CrossRef]
6. Yang, M.; Zhang, D.; Wu, B.; Zhang, Y. Energy Consumption Modeling for EDM Based on Material Removal Rate. *IEEE Access* **2020**, *8*, 173267–173275. [CrossRef]
7. Tristo, G.; Bissacco, G.; Lebar, A.; Valentinčič, J. Real time power consumption monitoring for energy efficiency analysis in micro EDM milling. *Int. J. Adv. Manuf. Technol.* **2015**, *78*, 1511–1521. [CrossRef]
8. Grigoriev, S.N.; Kozochkin, M.P.; Porvatov, A.N.; Okunkova, A.A.; Volosova, M.A.; Kuptsov, V.R.; Nadykto, A.B.; Pivkin, P.M. A new method to efficiently control energy use in Electrical Discharge Machining (EDM). In *Technologies for Optical Countermeasures XVIII and High-Power Lasers: Technology and Systems, Platforms, Effects V*; Titterton, D.H., Grasso, R.J., Richardson, M.A., Bohn, W.L., Ackermann, H., Eds.; SPIE: Bellingham, WA, USA, 2021; p. 7. [CrossRef]
9. Ahmed, N. Machining and wear rates in EDM of D2 steel: A comparative study of electrode designs and materials. *J. Mater. Res. Technol.* **2024**, *30*, 1978–1991. [CrossRef]
10. Jebbor, I.; Benmamoun, Z.; Hachimi, H.; Raouf, Y.; Haqqi, M.; Akikiz, M. Improvement of an Assembly Line in the Automotive Industry: A Case Study in Wiring Harness Assembly Line. *Adv. Transdiscipl. Eng.* **2023**, *35*, 62–71. [CrossRef]
11. Jebbor, I.; Benmamoun, Z.; Hachimi, H. Application of Manufacturing Cycle Efficiency to Increase Production Efficiency: Application in Automotive Industry. In Proceedings of the 2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology, IRASET 2024, Fez, Morocco, 16–17 May 2024. [CrossRef]
12. Zheng, J.; Jia, R.; Hu, X.; Qi, T.; Pan, Q.; Zhu, J.; Qiu, X.; Xia, Z.; Xv, B.; Deng, C.; et al. Regulating cutting fluid parameters for optimal energy and economic performance: Methods for efficient and Low-Energy electrical machining. *Energy Convers. Manag.* **2024**, *314*, 118707. [CrossRef]
13. Maradia, U.; Knaak, R.; Busco, W.D.; Boccadoro, M.; Wegener, K. A strategy for low electrode wear in meso–micro-EDM. *Precis. Eng.* **2015**, *42*, 302–310. [CrossRef]

14. Tran, V.T.; Le, M.H.; Vo, M.T.; Le, Q.T.; Hoang, V.H.; Tran, N.-T.; Nguyen, T.-A.; Nguyen, H.N.; Nguyen, V.T.T. Optimization design for die-sinking EDM process parameters employing effective intelligent method. *Cogent Eng.* **2023**, *10*, 2264060. [CrossRef]
15. Ming, W.; Guo, X.; Zhang, G.; Hu, S.; Liu, Z.; Xie, Z.; Zhang, S.; Duan, L. Optimization of process parameters and performance for machining Inconel 718 in renewable dielectrics. *Alex. Eng. J.* **2023**, *79*, 164–179. [CrossRef]
16. Mezoudj, M.; Belloufi, A.; Abdelkrim, M. Experimental Investigation on the Effect of Machining Parameters in Electric Discharge Machining Using AISI 1095-Treated Steel. *Int. J. Mod. Manuf. Technol.* **2019**, *XI*, 77–85.
17. Chidambaram, V.; Ramasamy, M.; Saravanan, N.K.; Devan, V. Investigations of process parameters on surface finish of aluminium component produced by die sink electric discharge machining process. *Indian J. Eng. Mater. Sci.* **2024**, *31*, 113–123. [CrossRef]
18. Rahman, M.M.; Khan, M.A.R.; Kadirgama, K.; Noor, M.M.; Bakar, R.A. Modeling of material removal on machining of Ti-6Al-4V through EDM using copper tungsten elec-trode and positive polarity. *World Acad. Sci. Eng. Technol.* **2010**, *71*, 576–581.
19. Singh, S.K.; Singh, A.K.; Kumar, J. Desirability Approach to Control Machining Parameters During Die-Sinking EDM of Inconel-686. *Acad. J. Manuf. Eng.* **2022**, *20*, 19–30.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

Experimental Design and Simulation of a Fly-Cutting Plant for Academic Environment Practices

Diego Fernando Ramírez-Jiménez [1,*] and Cristian Alejandro Torres Valencia [2]

[1] Programa de Ingeniería Electrónica, Universidad del Quindío, Armenia 630001, Colombia
[2] Faculty of Engineering, Electronic and Computer Science Department, Pontificia Universidad Javeriana, Cali 760031, Colombia; cristian.torres@javerianacali.edu.co
* Correspondence: dframirez@uniquindio.edu.co; Tel.: +57-300-662-7619

Abstract: Test plants or laboratory prototypes are essential for developing training activities in engineering. In the field of automation and control, simulators or high-fidelity equipment models commonly used in industrial processes are necessary. These tools allow engineering trainees to gain experience working with devices similar to those they will encounter in their professional contexts. This paper presents the design and simulation of a fly-cutting plant for academic use. A 3D model was developed in SketchUp, incorporating features typical of industrial plants. The system's simulation was carried out in MATLAB R2023b using mathematical modeling. The primary contribution of this work is the design of a low-cost, compact industrial prototype that includes a conveyor belt and a continuous cutting mechanism, enabling the understanding and operation of large-scale industrial processes. Performance tests were conducted using MATLAB, Simulink, and Code Composer Studio. Subsequently, operational and cutting tests were performed using classical control techniques. Additionally, the design features of the fly-cutting plant, which can be easily implemented for process control training activities, are detailed.

Keywords: three-dimensional (3D) modeling; academic environments; digital control; flying shear; industrial prototype; real-time control; fly-cutting plant

1. Introduction

In engineering, teaching processes are in constant evolution, raising new challenges that must be satisfied to provide high-quality education according to the particular context in which it is being applied. With the development of new technologies, educational environments have undergone a significant transformation in the digital era. This evolution can be traced back to the Gutenberg printing press, which emerged in 1450 as a revolutionary tool for the dissemination of knowledge on a mass scale [1]. In the context of academia, the advent of the digital era is largely associated with the rise of information and communications technology (ICT). Initially, ICT enabled the development of innovative tools to enhance the accessibility and advancement of educational materials. Moreover, the emergence of web-based platforms paved the way for the advancement of educational technologies and interaction tools. These developments promoted collaborative learning, multiple intelligences, and the integration of mobile devices, ultimately improving engineering trainees' comprehension and knowledge retention [1–3]. Another significant aspect of contemporary higher education is the implementation of learning management systems (LMSs) in virtual and distance learning environments. These systems leverage information and communications technology (ICT) tools to personalize learning processes, thus improving student motivation [4]. Additionally, the popularization of artificial intelligence (AI) has marked a significant evolution in the field

of learning management systems (LMSs), particularly in higher education institutions. AI not only enhances the capabilities of existing digital tools but also introduces innovative features such as adaptive learning systems, research support tools, and autonomous or assisted teaching mechanisms. These advancements personalize the learning journey and align it with the individual needs of each learner [5].

The training processes for engineers in automation and control should focus not only on the handling of technological tools, time management, and collaborative work, but also on fostering consistent engagement with the challenges faced by society and industry. Therefore, it is crucial to define the value of academic work in automation and control processes, helping students to develop fundamental skills, including those related to analog circuits, digital electronics, dynamic systems analysis, introductory control, and other activities that require precision and robustness [6]. To enhance the rigor of academic work, it is vital to strengthen core engineering domains while establishing laboratories equipped with test plants that replicate authentic industrial processes. This will facilitate the integration of diverse scenarios beyond the academic domain.

In recent years, test plants have gained significant interest due to two primary factors: First, the critical importance of automation and process control in the corporate landscape, coupled with the need for academic institutions to keep pace with these advancements [7]. Second, the high cost of laboratory equipment from companies such as Festo or Mitsubishi. As a result, many academic institutions have designed cost-effective, modular, and scalable test prototypes aimed at enhancing engineering education while also conducting research that has led to process improvements in industry. For example, in [8], applications for robotic manipulation and digital control were developed within an educational framework. Regarding fly-cutting plants and their role across different industrial sectors, numerous applications have been implemented using a wide range of system configurations or architectures. Table 1 provides a comparative analysis of existing fly-cutting designs used in both academic and industrial environments.

Table 1. Review of previous works carried out on fly-cutting systems.

Paper	Summary and Applications
FPGA implementation of general purpose real time controller-applied to flying shears in hot rolling mills [9].	The paper discusses FPGA-based controllers for industrial applications. It focuses on flying shear control in hot rolling mills by implementing speed, position, and torque control strategies for flying shear automation using a state estimator.
Optimizing design of control system for flying shear in tandem cold rolling line [10].	The rotary flying shear was used in tandem cold rolling lines. The key parameters for automatic cutting were identified, the control system was designed using SIMATIC TDC, and the algorithm of the automatic cutting process was optimized and improved to achieve a stable and efficient automatic cutting.
Study on the dynamics performance of the drive train of flying shear machine [11].	In this paper, a model is established with multiple backlashes, time-varying mesh stiffness, and error excitation to investigate the dynamic performance of the transmission system during the process. The study focuses on the transmission system's dynamic performance for flying shear continuous rolling mills, and it has academic applications in nonlinear dynamics research.

Table 1. *Cont.*

Paper	Summary and Applications
The wafer-level vacuum sealing and electrical interconnection using electroplated gold bumps planarized by single-point diamond fly cutting [12].	In this paper, a single-point diamond fly-cutting of electroplated Au bumps and Au-Au diffusion bonding was used to achieve a 100% yield of vacuum sealing and a high bonding shear strength. The research has applications in both the academic and industrial fields. It focuses on MEMS integration with LSI technology. The developed process is useful for inertial sensors and infrared imagers.
250 °C wafer-level vacuum sealing using electroplated copper bonding frame planarized by fly-cutting [13].	A wafer-level heterogeneous integration and vacuum packaging technology by thermo-compression bonding using an electroplated Cu sealing frame planarized by single-point diamond fly-cutting is reported. The technology is applicable in both academic and industrial settings. It enables wafer-level heterogeneous integration and vacuum packaging. A single-point diamond fly-cutting is applied for planarization. Mechanical stress application enhances grain boundary density.
Design of flying shear device and algorithm based on motioncontrol system [14].	The paper discusses flying shear device and algorithm design. It utilizes SIEMENS S7-300 PLC and SINAMIC S120 controllers, focuses on high precision and stability in motion control, and implements PID control for tension and speed regulation. It provides insights into flying shear device implementation. The design can aid learners in understanding motion control. Implementation of PID control for motor regulation and application of friction and inertia compensation in torque control.
The optimization of the dynamic regime of the asynchronous vectorial ac motor in flying shear [15].	The paper optimizes dynamic regimes of asynchronous motors in flying shear devices, focuses on minimizing energy consumption during cutting processes, describes complex motor operation with frequent speed changes, utilizes vector control for motor performance enhancement, and analyzes torque, current, and speed parameters for optimization.
Design of the flying shear servo control system [16].	The paper presents a flying shear servo control system. A universal servo controller replaces previous motor controllers. The system is used for steel plate shearing lines, and it optimizes motion curves for cutting processes.
Flying shear control system [17].	The research focuses on industrial applications in continuous rolling mills. It addresses energy saving and scrap minimization in manufacturing. The flying shear control system includes position control algorithms based on optimal positioning and simulation on a digital computer for verification.
Improving the operation of a flying dividing shear by using direct torque control [18].	The paper discusses direct torque control (DTC) for a flying dividing shear (FDS) in steel mills. DTC improves torque control for high-speed, dynamic applications like FDS. The control system utilizes a programmable logic controller for supervision.

Table 1. *Cont.*

Paper	Summary and Applications
Metallurgical investigation of flying shear blades at hot rolling rebar mills [19].	The paper investigates flying shear blades in rebar mills. The research focuses on industrial applications in rebar rolling mills. It aims to improve the performance and service life of the shear blade. Metallurgical analysis addresses premature failure of industrial components, along with chemical analysis using X-ray fluorescence (XRF) spectroscopy, hardness testing using a Rockwell C scale hardness machine, and metallographic techniques for surface preparation and analysis.
Reliability design of an electronic cam curve for flying shear machine in short materials cutting [20].	The structure and the production process for flying shear machines are introduced. Then, a quintic polynomial is applied to the design of an electronic cam system for the rotary knife axis in the cutting materials. The dimensionless equation for a quintic polynomial cam curve is deduced. Finally, the curve is plotted with the cam constructor integrated into the Siemens engineering development software SCOUT 18, and it is tested with a laboratory platform, which consists of a SIMOTION motion controller and SINAMICS S120 motor drivers. The results show that the running stability of the flying shear machine and the position control accuracy of the rotary knife can be effectively improved by using the curve designed in this paper.

Source: The authors.

The papers featured in this collection highlight a variety of insights related to plants and on-the-fly shearing systems, incorporating both theoretical and practical perspectives. For example, studies such as [11,15,20] explored academic subjects such as optimizing dynamic scenarios, analyzing transmission system performance, and creating electronic designs for cutting mechanisms. These studies achieved significant theoretical progress in dynamics and control. In contrast, research works including [14,16] focused on specific advancements in cutting plants, with the aim of enhancing system precision and stability through advanced controllers such as SIEMENS PLC and PID algorithms. Additionally, novel strategies for generalized control were introduced to cut costs and boost integration. The research in [9,10,18] explored the usage of technologies such as FPGA and DTC to increase the efficiency of speed, position, and torque control, consequently minimizing waste and optimizing energy use. Finally, studies such as [12,13,17,19] combined technological elements with material analysis to prolong the useful life of the components, implement vacuum sealing in MEMSs, and refine blade performance in industrial setups. These investigations contributed to both applied research and the development of practical solutions in industrial settings. Furthermore, other studies on flying shear plants, such as [21], introduced a three-dimensional model of a flying shear, focusing on structural and dynamic analysis of the shear process using MATLAB/SimMechanics and Solidworks/COSMOS motion. The results offer valuable theoretical insights into designing electronic control systems and optimizing parameters for similar flying shears.

In [22], the mechanism for a sheet shearing plant was designed with an analytical focus on the losses generated in transformer cores due to irregular shearing. These works represent some of the advancements in the modeling of shear systems conducted at universities and research institutes, most of which remain at the simulation level. Another area of focus in prototype development at the academic level has been the integration of hardware development, where devices are manipulated or controlled through remote applications in

addition to the use of simulations. In [23], a remote laboratory was developed for two test plants (an industrial robot and a mobile robot) using Java Servlet environments. In [24], a study was conducted on the implementation of a real-time control system for a 20 KVA power converter using the Texas Instruments LAUNCH-XL F28379D board. This research highlights the importance of employing dedicated hardware in control systems. Additional studies stress the value of using laboratory plants that replicate authentic industrial processes in an academic context. In [25], the authors emphasized the evolution of pedagogical methodologies and developed a state-space control system in a float-shield laboratory plant to replicate real-world environments in the classroom.

As evidenced in the literature reviewed, research on flying shear prototypes has primarily focused on simulations of estimated mathematical models. The methodology employed in this study makes a significant contribution by designing and simulating a fly-cutting plant that considers real-world behavior. This ensures the replicability of our work and the full implementability of the presented prototype in academic settings. Another notable contribution of this research is its integration of concepts from control theory, mechanical engineering, and physical principles. This interdisciplinary approach provides a detailed understanding of the operation of the proposed fly-cutting plant and facilitates research across new areas. Additionally, this study contributes to the field of electronic engineering by bridging the industrial sector in the pursuit of efficient solutions to enhance processes through new technologies, such as augmented reality laboratories and 3D printing. These technologies enable the testing of designed prototypes prior to their final implementation.

This research presents the design and simulation of a fly-cutting plant for implementing real-time control strategies in academic environments. The plant was designed using a three-dimensional modeling approach, enabling the creation of a digital prototype that could be subsequently implemented. The operational efficacy of the fly-cutting plant was validated through real-time simulations, with the plant represented in mathematical terms. This study makes additional contributions to the automation and control field within the Electronic Engineering Program at Universidad del Quindío, including the following:

- The design of a laboratory plant to emulate the continuous cutting process of moving parts, a procedure frequently performed in various industrial sectors.
- The definition of requirements and mechanisms is necessary to implement a fly-cutting plant on a laboratory scale.
- The development of applications using specialized tools for the design and simulation of real-time control systems, oriented toward the modeling and control of industrial processes at the laboratory scale. This contrasts to previous work, such as [26], which focused on robot manipulation.

This paper is structured into distinct sections. First, the methodology, including the materials and methods used for the design and simulation of the fly-cutting plant, is outlined. Next, the results obtained and their analysis in relation to the implemented control strategy are presented. Finally, a comprehensive summary of the main conclusions derived from this research is provided.

2. Materials and Methods

2.1. Materials

2.1.1. Fly-Cutting Prototype

Fly-cutting plants or machines (also known as flying shears) are mechanical systems equipped with a cutting blade and a conveyor belt to transport the material to be processed. These devices operate either in mechanical synchronism between the cutting blade and the moving material or by matching their speeds with the assistance of additional electronic

devices. Various manufacturers, such as RISHBIN, produce different types of fly-cutting machines. RISHBIN's catalog includes metal plate cutting machines classified by cutting type: rotary, crank–rotary lever, eccentric crank, and pendulum plants [27]. An illustrative example is provided in [28], where a crank-type cutting machine is analyzed for kinematics, profile error, acceleration, and velocities across different cutting phases. Schneider Electric offers an alternative classification system for RISHBIN flying shears, dividing shear systems into T1, T2, and T3 types [29]. Figure 1 illustrates the T1 fly-shearing plant, which features a cutting system where the saw's base supports the material and synchronizes its speed with the material while performing the cut.

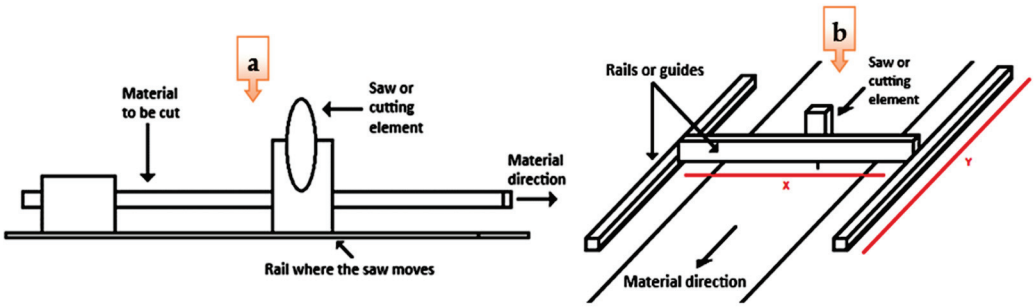


Figure 1. Industrial flying cutting plants: (**a**) T1 type and (**b**) T3 type. Source: The authors.

2.1.2. MATLAB PIDTuner

MATLAB (Matrix Laboratory) is a numerical computational tool that integrates capabilities for calculation, visualization, and programming within an interactive environment. It allows users to express problem solutions in a format similar to mathematical notation, with arrays serving as the fundamental data structure since they do not require explicit dimensioning. All data are represented with double precision, enabling more accurate calculations and easier interaction [30]. MATLAB provides a variety of toolboxes that offer specialized functions for diverse fields of knowledge. The Control Toolbox includes the PIDTuner application, which facilitates the automated adjustment of gains in designed controllers. The design process is graphical, enabling the evaluation of controller performance and robustness through associated graphical analyses. A key feature of PIDTuner is its ability to support controller design in both the time and frequency domains. Figure 2 illustrates the user interface of the MATLAB PIDTuner application [31].

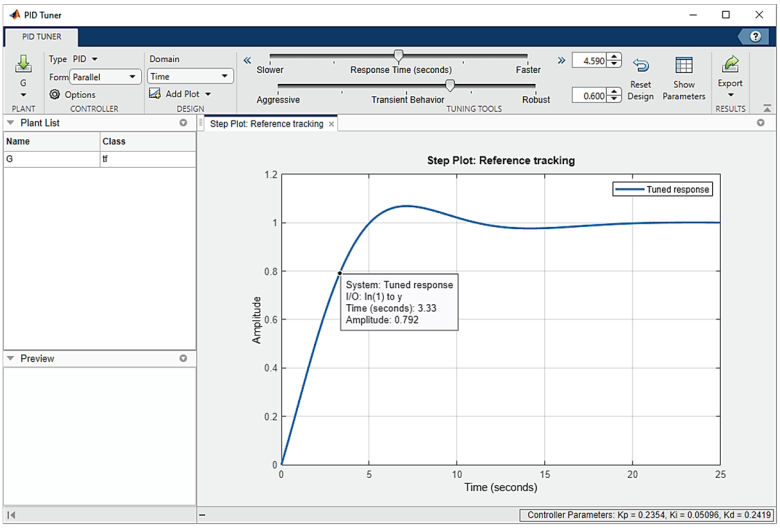


Figure 2. Graphical user interface for MATLAB PIDTuner. Source: [31].

2.1.3. SketchUp 3D Design

SketchUp is a versatile 3D modeling tool that supports the creation of three-dimensional designs, models, and extended reality experiences. Available on multiple platforms, including desktop, web, and mobile, it is an excellent choice for professionals across various disciplines, particularly in engineering. One of its most notable features is the integration of artificial intelligence (AI) functionalities, which optimize and accelerate workflows, streamlining the design process. Figure 3 illustrates the graphical user interface of SketchUp's desktop version.

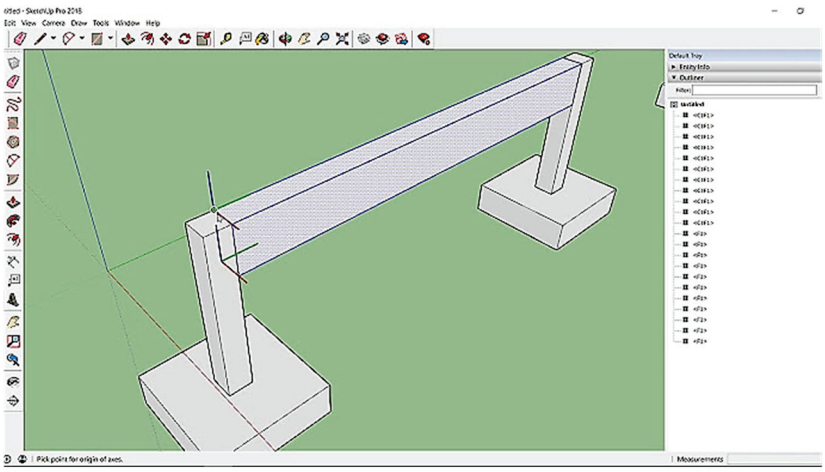


Figure 3. Graphical user interface for SketchUp desktop version. Source: [32].

2.2. Methods

2.2.1. Plant Structure

The structure of the designed fly-cutting plant is illustrated in Figure 4. The block diagram consists of two principal systems: The first is the mechanical system, which includes a conveyor belt and a cutting mechanism with three-dimensional movement along the X-, Y-, and Z-axes. The second is the electronic system, comprising a control board with peripherals and other control elements, a power circuit housing the main card for motor operation, and a switching circuit that allows the system to operate with internal devices or an external control device.

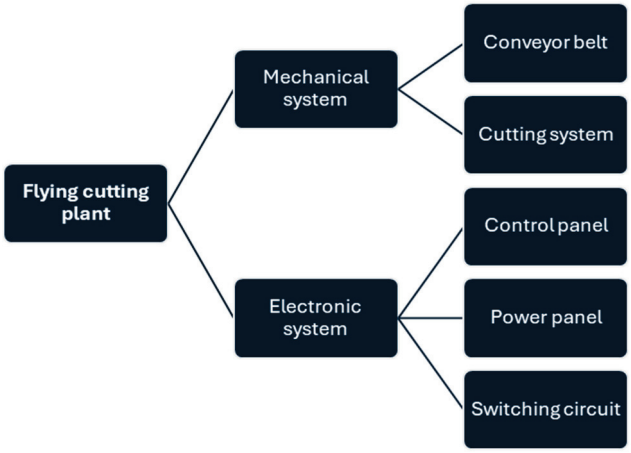


Figure 4. Block diagram of the structure of the fly-cutting plant. Source: The authors.

2.2.2. Mechanical System

The plant features a conveyor belt and a cutting mechanism that moves along the X-, Y-, and Z-axes. Each component was analyzed individually to meet the specified requirements. The following key elements were identified:

- Cutting system: This must be capable of creating markings or patterns on the conveyor belt.
- Conveyor belt: The system must be modular to facilitate easy relocation and stable to prevent vibrations from causing irregularities during the cutting process.
- Dimensions: The plant requires a fixed base on the floor with rollers positioned at a height of approximately 80 cm, corresponding to the hand height of an average operator.
- Ease of transport: As a structure designed for academic environments, it must be easy to move between spaces.
- Belt traction: The traction system comprises two rollers to ensure proper alignment and prevent the belt from deviating. One roller must be mechanically connected to a motor, controlled by electronic circuits.
- Location of the cutting system and circuit boards: The belt structure must include designated spaces to house the circuit boards and protect the electronic components.
- Tape material: The upper part of the tape must be smooth and allow for the erasure of the marks made with the cutting system.

The conveyor belt structure was designed using a SketchUp 3D model and will be constructed with 4 × 2 cm rectangular iron tubes of 0.8 mm gauge. One of the upper horizontal profiles, which supports the rollers, was designed to be fastened with bolts, allowing it to be disassembled if the conveyor belt or any component requires replacement. Figure 5 illustrates the designed structure of the conveyor belt.

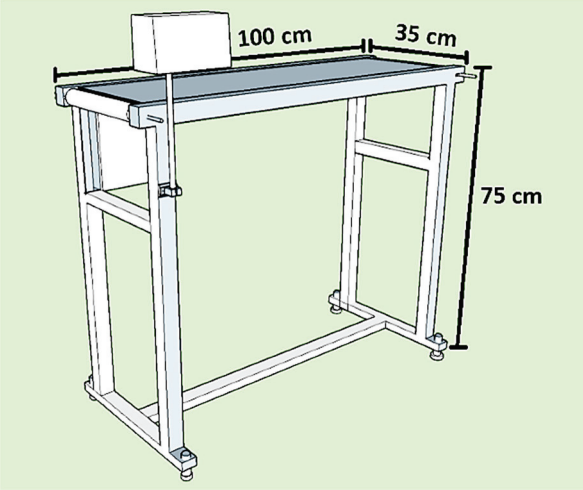


Figure 5. Structure designed for the conveyor belt. Source: The authors.

2.2.3. Roller Design

The rollers are essential for ensuring the proper rotation of the conveyor belt and were designed using SketchUp 3D Pro 2023 software to facilitate subsequent belt calibration and alignment. An 8 mm shaft-bearing system was used at one end of the belt to align the roller, which is driven by the motor through pulleys. This roller is a fixed drum roller that rotates in conjunction with the smooth shaft, enabling movement through the bearings. Figure 6 illustrates a representation of the fixed drum roller.

The opposite end of the belt features a freely rotating drum roller, supported by bearings that enable axial movement. The 8 mm threaded shaft allows clamping at both ends after the conveyor belt is calibrated and aligned. Belt alignment is achieved using a pair of grooves in each profile where the roller fits, clamping nuts for necessary adjustments, and tensioners to pull the shaft outward. Figure 7 illustrates a representation of the idler pulley.

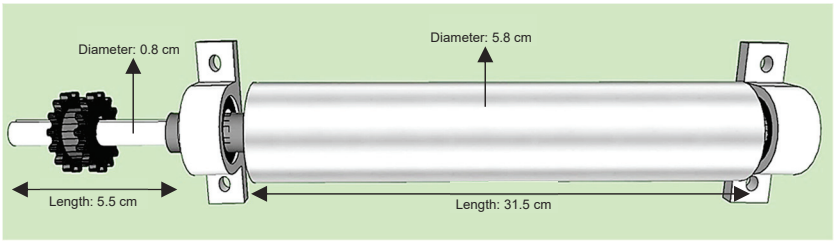


Figure 6. Fixed drum roller with bearings and pulley. Source: The authors.

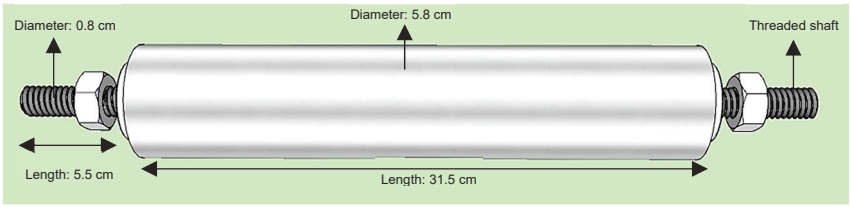


Figure 7. Free drum roller. Source: The authors.

2.2.4. Cutting System

The design and operation of the fly-cutting plant were validated by assessing its ability to execute a precise cutting path, characterized by uniformity and the absence of zigzagging patterns. To achieve this, it is crucial that the components are arranged correctly and that the motors or actuators operate with precision. Accordingly, a cutting system was designed consisting of a three-dimensional plane defined by the X, Y, and Z coordinates. The X- and Y-axes are driven by two motors, while an actuator or servo controls the movement along the Z-axis, executing the cutting path. Figure 8 provides an illustration of the cutting system.

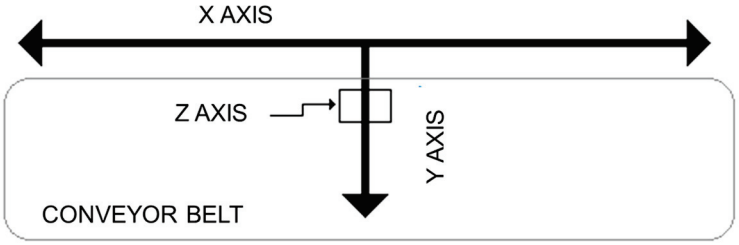


Figure 8. Motion coordinates for the cutting system. Source: The authors.

To develop the system depicted in Figure 8, it is essential to consider kinematic aspects within both the system and the designated conveyor belt area, as increasing the total area requires a proportional extension of the X- and Y-axes. This led to setting the conveyor belt's operational area at 80% of its total capacity as the upper limit, ensuring scalability for the cutting system and its potential application in other areas requiring Cartesian movement, such as CNC machines or modular laser cutting systems.

The linear motion of the cutting system is achieved using V-SLOT extruded aluminum profiles and high-precision bearings, commonly employed in the manufacturing of 3D printers, CNC machines, Cartesian robots, and similar devices. Finally, the Z-axis is actuated by a servomechanism adapted from those used in drawing machines, incorporating linear bearings and a base for fixing the cutting point. The design of the cutting system, created using SketchUp 3D modeling software, is illustrated in Figure 9.

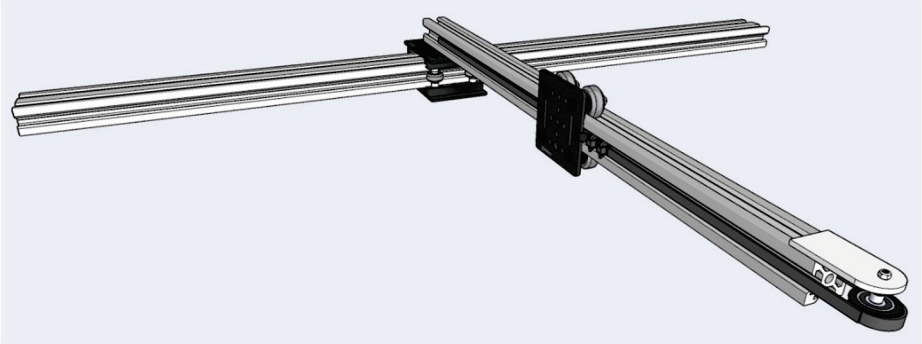


Figure 9. Design and 3D modeling of the cutting system using SketchUp software. Source: The authors.

3. Results

To develop the experimental framework, modifications were made to various operational parameters of the fly-cutting plant. The operational tests were conducted in the controlled environment of the Universidad del Quindío control laboratory, a humidity-free enclosure with an average temperature of 25 °C. Regarding the mechanical and operational configuration of the fly-cutting plant, the working area of 3000 cm² was reduced to 2200 cm² to avoid the need for a significant extension of the X- and Y-axis rails. The dimensions of the metal profiles and the configuration of various components provided a working area equivalent to 80% of the conveyor belt's total area.

For the operational tests of the conveyor belt motors and the cutting system, eight characterizations and input/output data collections were performed to gather consistent information about the behavior of these mechanisms and their respective technical specifications. The motion systems were tested at their maximum allowable input. Finally, the sensors and encoders used to measure motor speed, and the location of the cutting system were characterized based on the information provided by the manufacturers in the corresponding data sheets.

3.1. Mechanical Parameter Validation

The mechanical system of the proposed fly-cutting plant features finite movement along the axes, limited by the dimensions of the conveyor belt. Therefore, it is crucial to determine the operating limits of each component within the mechanical transmission system to validate them through experimental testing.

The belt drive consists of the rollers shown in Figures 6 and 7, along with a pulley system that connects one of the rollers to the motor driving the belt. Figure 10 illustrates the mechanical system used to move the conveyor belt.

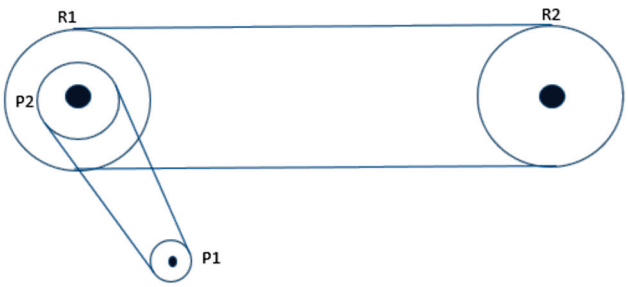


Figure 10. Pulleys' and rollers' relationship for the movement of the conveyor belt. Source: The authors.

Theoretically, the maximum angular speed achievable by the output of the conveyor belt's motor gearbox is 100 rpm, which is also the maximum speed of pulley P1. To determine the tangential speed of the rollers, we use the radius values of the pulleys and

rollers shown in Table 2 and apply the equations of uniform circular motion. The following relations are used:

$$w_1 = 100 \text{ rpm} \tag{1}$$

Table 2. Radius of pulleys and rollers on conveyor belts.

Device	Radius (cm)
Pulley (P1)	0.7
Pulley (P2)	1.3
Roller (R1)	2.165
Roller (R2)	2.165

Source: The authors.

Converting to radians per second, the angular speed of the pulley P1 is as follows:

$$w_1 = 10.47 \text{ rad/s} \tag{2}$$

The angular velocity is transmitted to pulley P2 linearly; therefore, the tangential speed v_1 at pulley P1 is equal to the tangential speed v_2 at pulley P2.

$$vt_1 = vt_2 \tag{3}$$

We know

$$vt = 2\pi * r/T \tag{4}$$

then

$$vt = w * r \tag{5}$$

Rewriting Equation (3), we have

$$w_1 * r_1 = w_2 * r_2 \tag{6}$$

Using Equation (6), we can know the angular speed at the pulley P2:

$$w_2 = (w_1 * r_1)/r_2 = (10.47 * 0.7)/1.3 = 5.64 \text{ rad/s} \tag{7}$$

As the pulley P2 is attached to the roller axis R1, both rotate at the same angular speed.

$$w_2 = w_roll1 = 5.64 \text{ rad/s} \tag{8}$$

Since both rollers have the same radius, their speeds are identical. Based on the angular speed of roller R1, the tangential speeds of both rollers, R1 and R2, can be determined, corresponding to the speed of the conveyor belt.

$$Vt_roll1 = w_roll1 * r = 5.64 * 2.165 = 12.21 \text{ cm/s} \tag{9}$$

After performing a theoretical analysis to determine the relationship between speeds in the mechanical system of the conveyor belt, experimental tests were conducted to validate the results. As shown in Figure 11, when maximum tension is applied to the motor, it achieves a speed of 93.39 rpm, corresponding to a linear speed of 11.40 cm/s on the conveyor belt.

Similarly to the conveyor belt motor, a theoretical analysis was conducted for the X-axis motion of the cutting system. To prevent damage or movement beyond the conveyor

belt's length, limits were established at each end of the cutting system, as shown in Figure 9. After these adjustments, the cutting system's movement along the X-axis was determined to be 62 cm. This is illustrated in Figure 12.

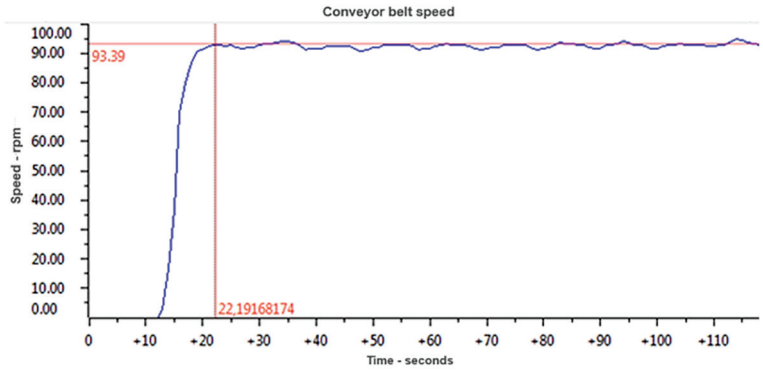


Figure 11. Maximum angular speed achieved by the conveyor belt motor. Source: The authors.

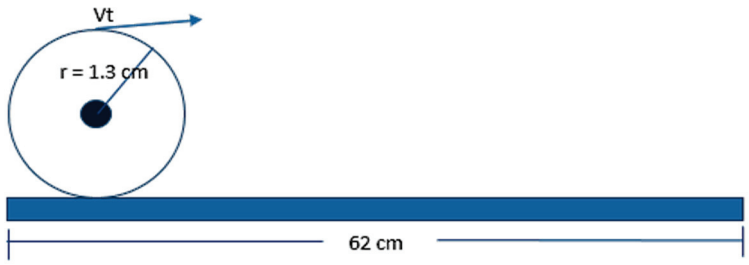


Figure 12. Movement of the pulley P2 on the X-axis. Source: The authors.

The movement of the cutting system along the X-axis must match the speed of the conveyor belt. To achieve this, the pulley attached to the motor must have the same characteristics as the fixed pulley of the drive roller (Figure 6). Another important aspect of the X-axis movement is the inclusion of a gearbox similar to that of the conveyor belt motor. This configuration allows for a theoretical maximum speed of 100 rpm. Therefore, the tangential speed of the X-axis for the cutting system can be expressed as follows:

$$w_x = 100 \text{ rpm} = 10.47 \text{ rad/s} \tag{10}$$

$$Vt_x = 10.47 \times 1.3 = 13.61 \text{ cm/s} \tag{11}$$

Based on the tangential speed of the X-axis motor, as previously established in Equation (11), the X-axis cutting system can traverse the entire workpiece length in approximately 4.55 s, representing the maximum allowable time for the system to complete a single cutting operation. Figure 13 illustrates the operational characteristics of the motor used in the X-axis cutting system. The maximum speed attained was 88.20 rpm, equivalent to a tangential speed of 12.0 cm/s. In comparison, the time required to travel the distance between the two endpoints was determined to be 4.71 s. This value was calculated by subtracting the number of start and end samples from the predefined sample time of 0.1 s used for simulation purposes.

For the angular velocity of the Y-axis cutting system, a maximum speed of 214 rpm was achieved using a square wave of 714 Hz. This result was obtained through experimentation with varying signal frequencies controlling the Y-axis motor driver. The displacement along the Y-axis of the cutting system is defined as 30.5 cm, corresponding to the width of the conveyor belt, and is ensured by two sensors that mark the beginning and end of the

movement. The maximum frequency enabling this displacement was 500 Hz, with a duty cycle of 71%, a peak voltage of 4.78 V, and a root-mean-square (RMS) voltage of 3.89 V.

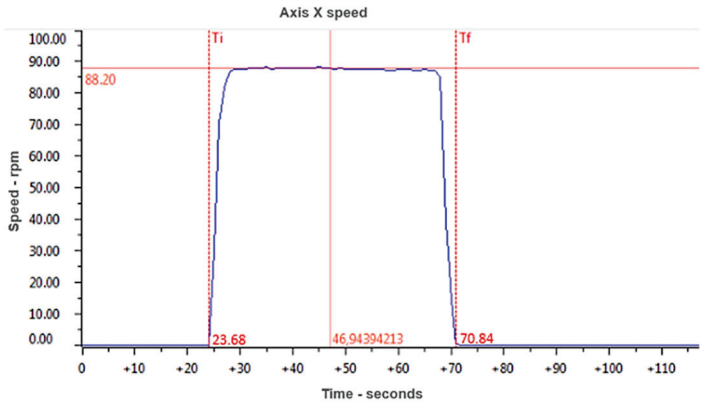


Figure 13. Maximum speed reached by the X-axis motor in the cutting system. Source: The authors.

For the Z-axis movement, a PWM signal with a frequency of 50 Hz was generated. The signal's high-level duration was adjusted to vary the angle of rotation. Figure 14 illustrates the configurations used for the simulations and motor response tests on the Y- and Z-axes of the cutting system.

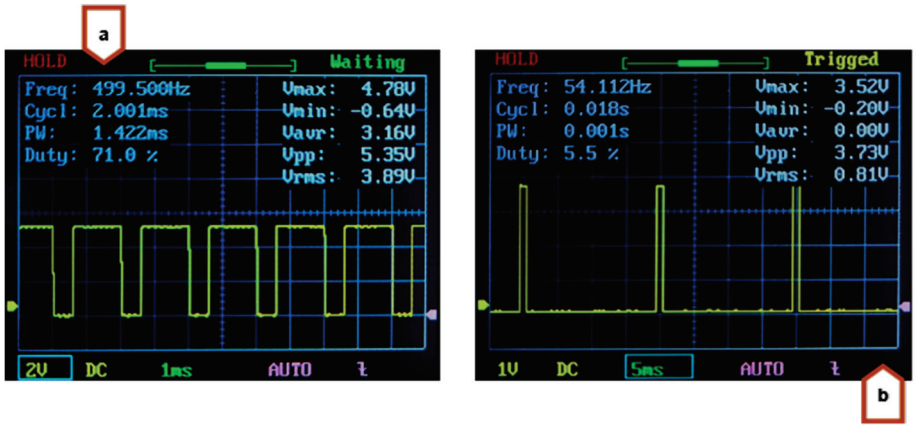


Figure 14. PWM signals for motor operation in the cutting system: (**a**) Motor for motion on the y-axis. (**b**) Motor for motion on the z-axis. Source: The authors.

The data illustrated in Figure 14b indicate that an angular displacement of 0° relative to the Z-axis servo motor positions the cutting blade on the workpiece. Additionally, it can be observed that an angular displacement of 50° is sufficient to return the blade to its initial position.

3.2. Electronic Circuits

The power circuits illustrated in Figure 15 were designed to facilitate easy connections to the two L298N drivers, which propel the geared motors controlling the conveyor belt and the X-axis of the cutting system. Additionally, a DRV8825 driver was incorporated to regulate the Y-axis stepper motor within the cutting system. Connection pins were added to enable direct linkage with a development PCB, as well as specific pins for the switching circuit.

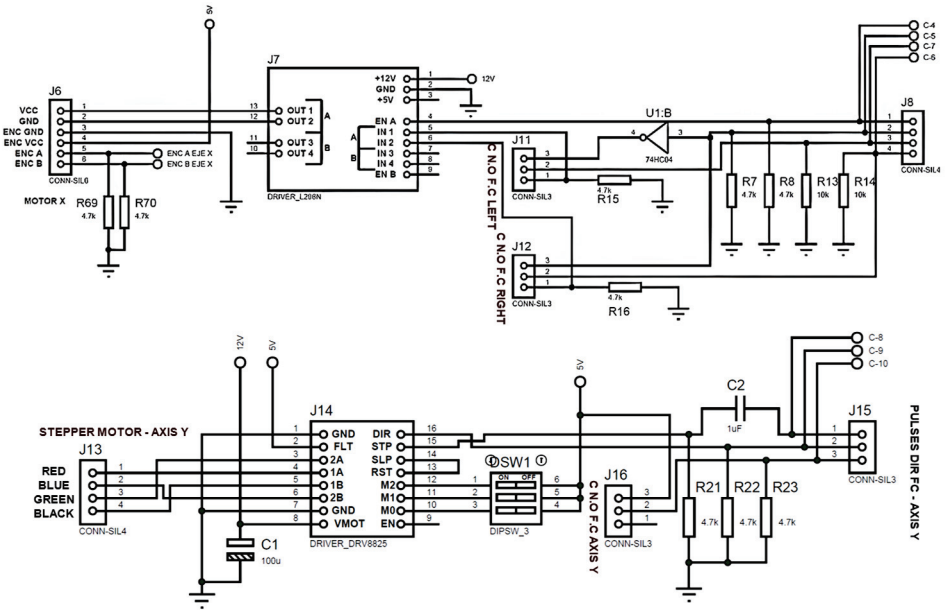


Figure 15. Power circuit for DRV8825 drivers (bottom) and L298N (top). Source: The authors.

The power circuit design includes terminal blocks, Molex, USB, and RJ-45 connectors for interconnecting with other components of the cutting plant, such as the power supply, sensors, actuators, control circuits, and programming interfaces. The electronic circuits were designed with scalability in mind, allowing for rapid and intuitive connections between devices necessary for integration during implementation. The design of the electronic circuits highlights three primary groups:

1. Mechanical System: Includes sensors (e.g., limit switches, encoders), DC motors, stepper motors, and servos.
2. Power Board: Contains drivers and signal conditioning for managing sensors, actuators, and embedded control systems.
3. Control Board: Facilitates user interaction with the fly-cutting plant, enabling different modes of operation, such as internal and external modes.

The circuit illustrated in Figure 15 (lower section) features a DIP switch that defines the motor's operational steps. Capacitor C2 and resistor R21 collectively form an RC circuit, also known as a high-pass filter. This configuration allows the passage of frequencies exceeding the cutoff frequency, mitigating potential interference from lower frequencies.

Optocoupler isolation can be employed for data transmission and reception between the control and power PCBs. However, due to the switching delays of the PC817 optocouplers, which range from 4 µs to 5 µs, a high-speed electronic circuit was designed using 6N137 optocouplers. These optocouplers, with a propagation delay of 48 ns to 50 ns, support transfer rates of up to 115,200 baud. The designed circuit is shown in Figure 16.

The designed power circuit also incorporates the following features:

- The option to use either an internal or external power supply.
- A latching circuit that keeps the start LED illuminated. This is initiated upon pressing the start button and deactivated upon pressing the stop button.
- The ability to select either digital pins or the Enhanced Quadrature Encoder (EQEP) module, commonly integrated into embedded systems, for the encoder readout pins. Finally, Figure 17 shows the 3D design of the power circuit.

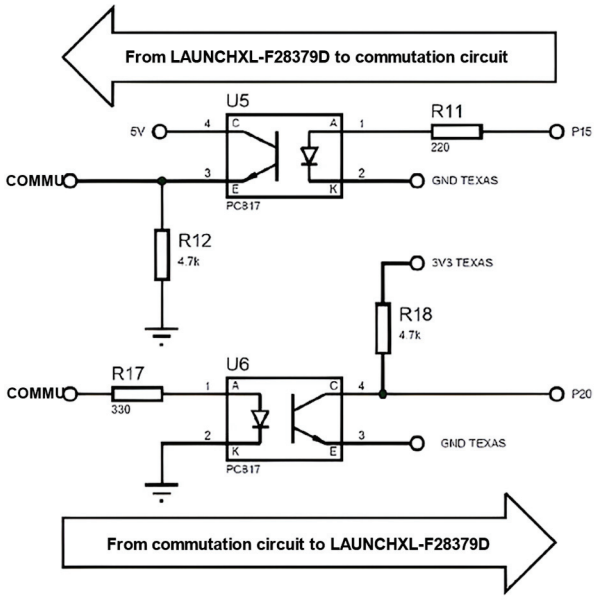


Figure 16. Tx and Rx circuits for high-speed optocouplers. Source: The authors.

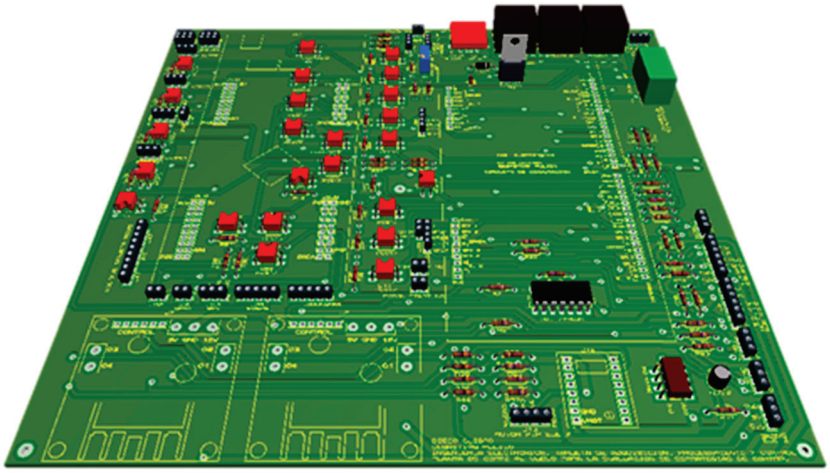


Figure 17. Three-dimensional (3D) design made in ARES for the power circuit. Source: The authors.

To perform signal switching and enable control from the processing device, it is necessary to consider whether the signal is read or written, as this determines the use of multiplexers or demultiplexers. The latter are constructed with analog switches. The 74HC157 integrated circuit is used for multiplexing signals, while the 74HC4066 integrated circuit is employed for demultiplexing. The electronic configuration for multiplexing and demultiplexing signals is illustrated in Figure 18.

A similar methodology is applied to all input and output signals used for commutation, including pulse-width modulation (PWM), gearmotors, steering, stepper motor control, servo duty cycle, encoder read pulses, and general-purpose input and output signals. Once the input and output signals for the commutation circuit are defined, its three-dimensional model is generated, as shown in Figure 19.

The primary function of the control circuit is to serve as a conduit between the user interface and the power circuit. To achieve this, the circuit was equipped with three RJ-45 connectors to facilitate signal transmission for sensor and actuator management, serial communication, general-purpose ports, and other applications. An Arduino Nano was integrated into the design to enable information visualization and the transmission of corresponding values to the power circuits.

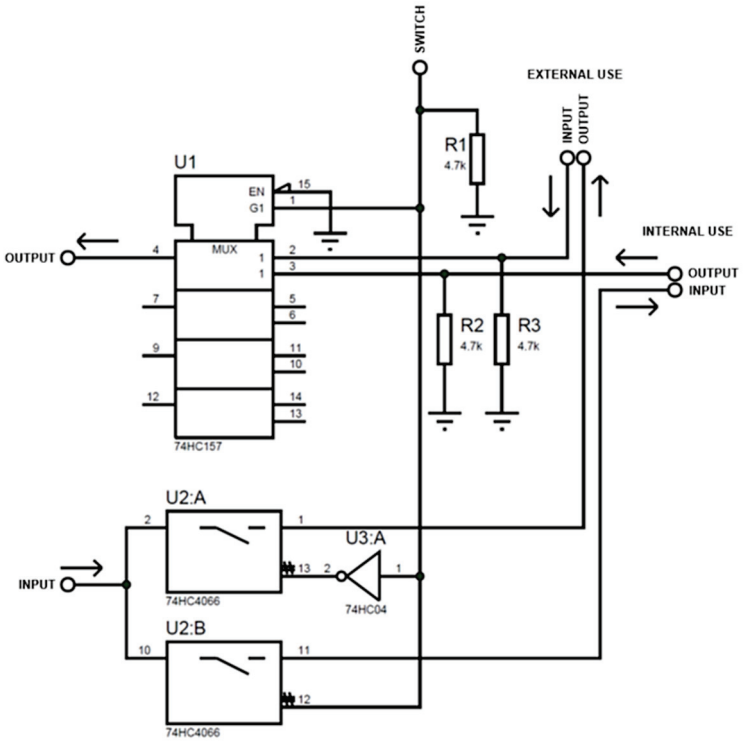


Figure 18. Proposed signal switching circuit. Source: The authors.

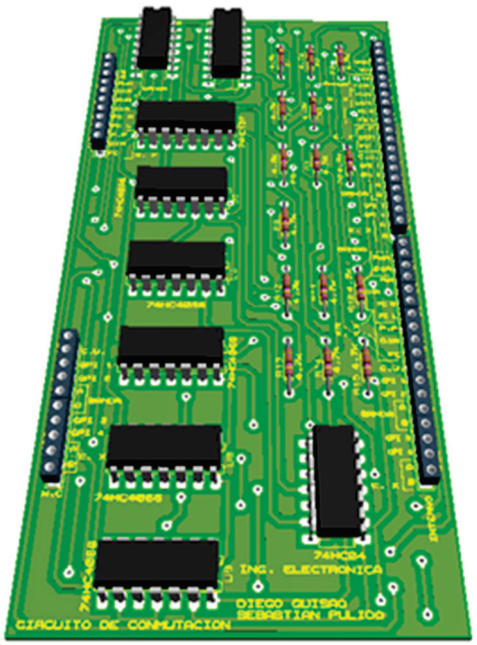


Figure 19. Three-dimensional (3D) model for the commutation circuit in Figure 20. Source: The authors.

The control circuit was configured to include a buzzer, jumpers for configuring pull-up or pull-down resistors, general-purpose ports, I2C communication pins for the LCD, and a power supply for additional connections. The design of the control circuit is illustrated in Figure 20.

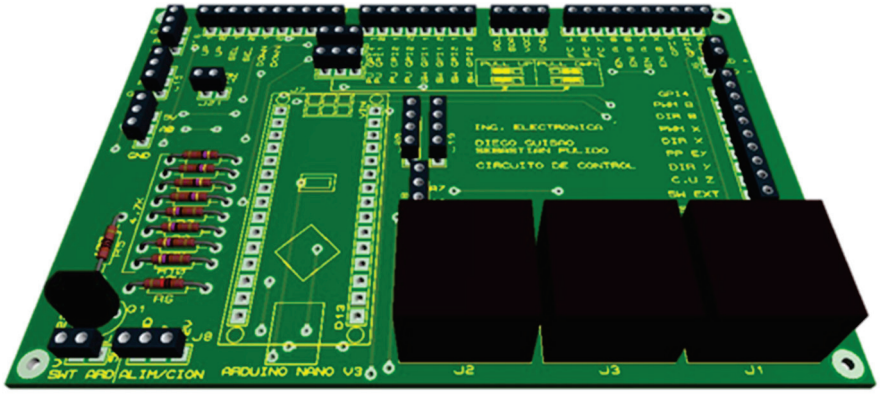


Figure 20. Proposed 3D model for the control circuit. Source: The authors.

3.3. System Programming and PID Controller

The integration of the designed plant's functionalities, including mechanical elements and electrical circuits, requires implementing a programming strategy for each module and card. Additionally, it is essential for all of these components to collectively establish communication with the plant. A variety of software tools can be used to program such systems, and the choice depends on the programmer's preferences and experience. In this case, several alternatives were considered, including Energia Nu designed in Austin, Texas, USA in 2012 by Robert Wessels as a Wiring/Arduino-based environment for Texas Instruments microcontrollers, Arduino that was developed in Ivrea, Italy in 2005 by a group of students and professors at the Interaction Design Institute Ivrea, and Code Composer Studio (CCS) developed in Dallas, Texas, USA in 1997 by Texas Instruments (TI).

Energia is an open-source integrated development environment (IDE) based on the Wiring and Arduino framework. It enables the development of applications using high-level programming, as described in [33]. Another explored option was CCS, an IDE featuring a C/C++ compiler, source code editor, debugger, and other tools for managing workspaces efficiently, making it a viable choice for developing various applications.

For this research, CCS was selected as the most suitable tool. If the designed plant is implemented in the future, CCS will allow the programmer to create microcontroller-based embedded systems for real-time applications, such as those used in Texas Instruments systems based on C2000 microcontrollers [34,35]. Figure 21 shows the CSS workspace in its desktop version.

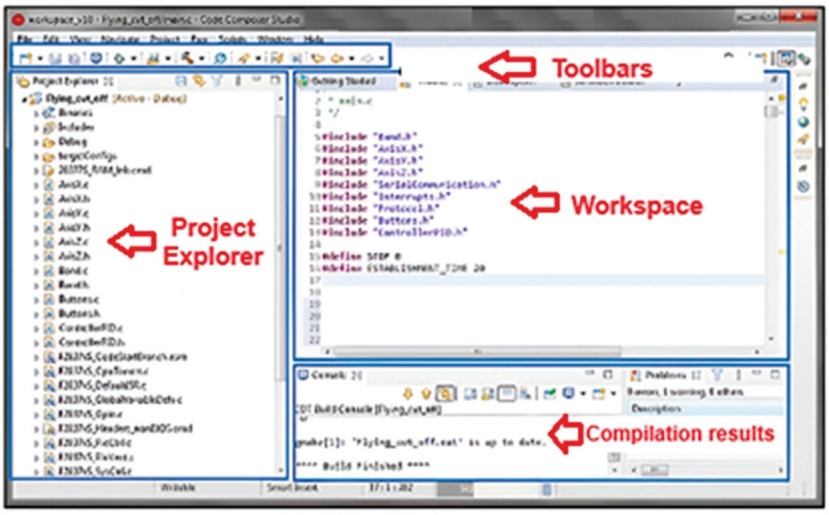


Figure 21. Workspace for the CCS desktop application. Source: The authors.

A review of the CCS editor's work environment shows that its programming structure is comparable to that of other development environments. Programming can be developed in the C language, as the default compiler used in CCS is ANSI C89. Since the objective of the designed plant was to integrate all constructed modules and devices, the programming was consolidated into a global project, resulting in the generation of numerous native files. These files contain the essential functions required for managing the sensors and actuators of the cutting plant, particularly for real-time control of the components. The files used for simulation development are presented in Appendix A. Figure 22 illustrates the defined programming structure for the proposed fly-cutting plant.

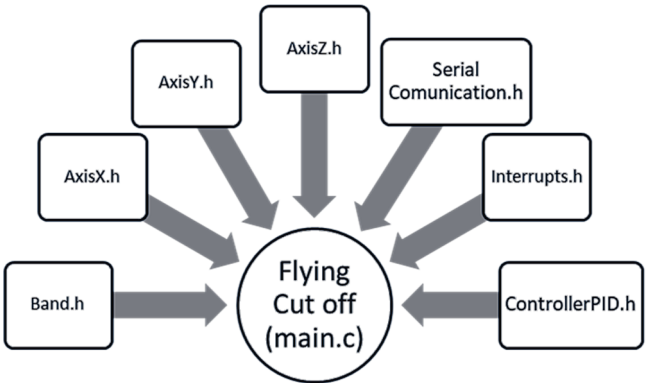


Figure 22. The programming structure developed for the simulation of the fly-cutting plant. Source: The authors.

Conversely, designing the controller for the fly-cutting plant requires an understanding of the plant's behavior and dynamics. Given the experimental nature of the design, factors such as friction, uncertainty, and various dynamic components were integrated into the identification process rather than being considered separately. The resulting model for the proposed system is as follows:

$$G(s) = 8368.1/(0.182\ s + 1) \tag{12}$$

The plant model was paired with MATLAB's PIDTuner to adjust the PID controller settings. The tuning process indicated that derivative action was unnecessary for the intended system, which is typical for this type of configuration. A key feature of the designed controller is its 0% overshoot and short settling time, both crucial for efficiently completing the trimming process. The parameters of the tuned controller are shown in Figure 23.

The implemented control loop is shown in Figure 24.

Since the PI controller is implemented on a microcontroller, the continuous model must be discretized. This requires determining a sampling time, which depends on the processing capacity of the digital device and the analysis of the open-loop and closed-loop poles. The following provides a brief overview of the method used to compute these poles and determine the plant's sampling time:

$$s(0.182s + 1) = 0, \qquad s_1 = 0, \quad s_2 = -5.4945 \tag{13}$$

The feedback system is described by

$$(8.6524 \times 10^{-5}\ s + 5.3054 \times 10^{-4})8368.1 + 0.182s^2 + s = 0 \tag{14}$$

$$s^2 + 9.4728s + 24.3935 = 0,\ \ s_3,4 = -4.7364 \pm 1.4006i \tag{15}$$

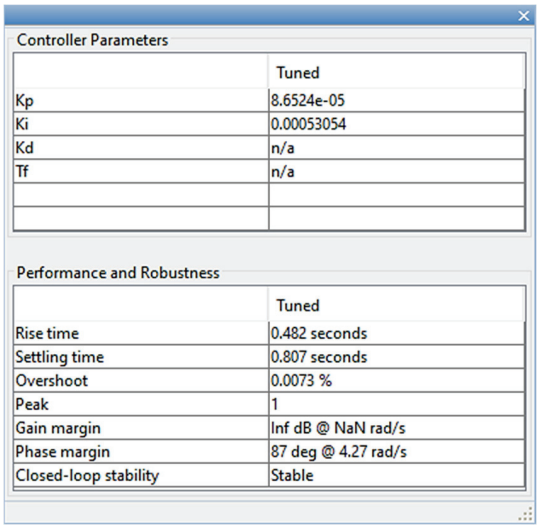


Figure 23. Controller parameters from PIDTuner. Source: The authors.

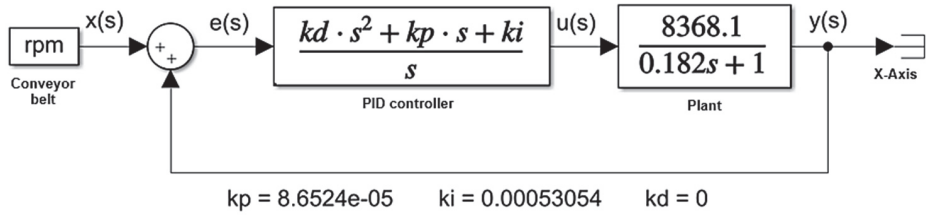


Figure 24. Continuous-time block diagram for the proposed controller. Source: The authors.

The highest frequency of the plant, denoted as ω_(max), is determined by the pole with the greatest magnitude, which is 5.4945 in this case. Based on this value, the maximum sampling time can be calculated.

$$f_m \geq (10\omega_max)/2\pi, \qquad Tm \leq 2\pi/(10\omega_max) \tag{16}$$

$$Tm \leq 0.114\ s$$

For the task of reading the encoders and implementing the control strategy, a sampling time of T_m = 0.1 s was selected. This sampling interval was then used to determine the discrete-time PI controller constants. Various controller tuning methods can be applied in this scenario, such as the Ziegler–Nichols approaches [34–36]. In this study, the Ziegler–Nichols rules, as outlined in Table 3, were adopted due to their compatibility with the sampling time T_m.

Table 3. PID controller equivalence between Z-domain and S-domain.

Gain	*Gc*(*s*)	*Gc*(*z*)	*Gc*(*z*) Including *Tm* and Delay
Proportional	k_p	K_p	K_p
Integral	$\frac{k_i}{s}$	$\frac{k_i z}{z-1}$	$\frac{k_i Tm}{1-z^{-1}}$
Derivative	$k_d s$	$\frac{k_d(z-1)}{z}$	$\frac{k_d(1-z^{-1})}{Tm}$

Source: The authors. Adapted from [34,36].

Using the PID controller equation from Figure 25 and the Ziegler–Nichols rules from Table 3, the resulting implemented PI equation is

$$Gc(s) = u(s)/e(s) = k_p + k_i/s + k_d\, s \tag{17}$$

$$Gc(z) = k_p + (k_i\, Tm)/(1 - z\hat{}(-1)) + (k_d\, (1 - z\hat{}(-1)))/Tm, \quad \text{with } Tm = 0.1\ s \tag{18}$$

$$Gc(z) = P(z) + I(z) + D(z) \tag{19}$$

with

$$p(z) = P(z)e(z) = k_pe(z) \tag{20}$$

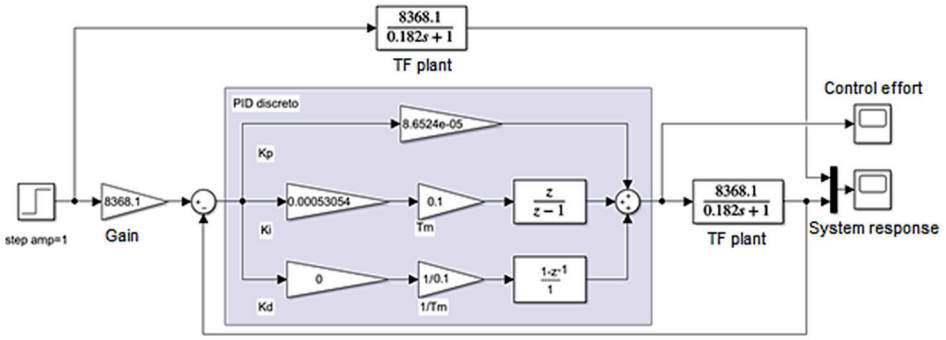


Figure 25. PI controller structure in discrete time for the fly-cutting plant. Source: The authors.

$$i(z) = I(z)e(z) = (k_i\, Tm)/(1 - z^{-1})e(z) \tag{21}$$

$$d(z) = D(z)e(z) = ((k_d\, (1 - z^{-1}))/Tm)e(z) \tag{22}$$

Finally, the PID controller expressed in difference equations is

$$p[k] = k_p\, e[k] \tag{23}$$

$$d[k] = (kd/Tm)(e[k] - e[k - 1]) \tag{24}$$

$$i[k] = [e[k]]k_iTm + i[k - 1] \tag{25}$$

Using Equations (23)–(25), we implemented the closed-loop control as shown in Figure 25.

The fly-cutting plant was subsequently subjected to simulation and operational tests following the modeling and simulation of all circuits. The system responses are presented over a period of six seconds, during which the conveyor belt remains inactive, representing a duty cycle of 0%, and the cutting system is in its initial position. At 6.8 s, a duty cycle of 50% is applied to the conveyor belt, and 1.8 s elapses for the belt speed to stabilize at approximately 6330 rpm. At this point, a set point for the conveyor belt movement is established, causing the geared motor to transition from a resting state (red signal).

Between 8 and 9.7 s, the cutting trace is generated as both speeds equalize. At 9.8 s, a downward peak is observed in the red signal, accompanied by an increase in revolutions to approximately 8344. This is attributed to a directional change, during which the X-axis geared motor pauses momentarily before returning to its initial position at maximum speed.

The X-axis geared motor, now at its initial position, ceases movement after 11.9 s, reverses direction, and generates another peak in revolutions. This prevents the motor's

inertia from causing additional movement or contact with the left stop. Following the application of the brake to the X-axis motor, the system responds after 12.1 s, synchronizing the X-axis speed with the belt to produce a second stroke. The operation summary is shown in Figure 26.

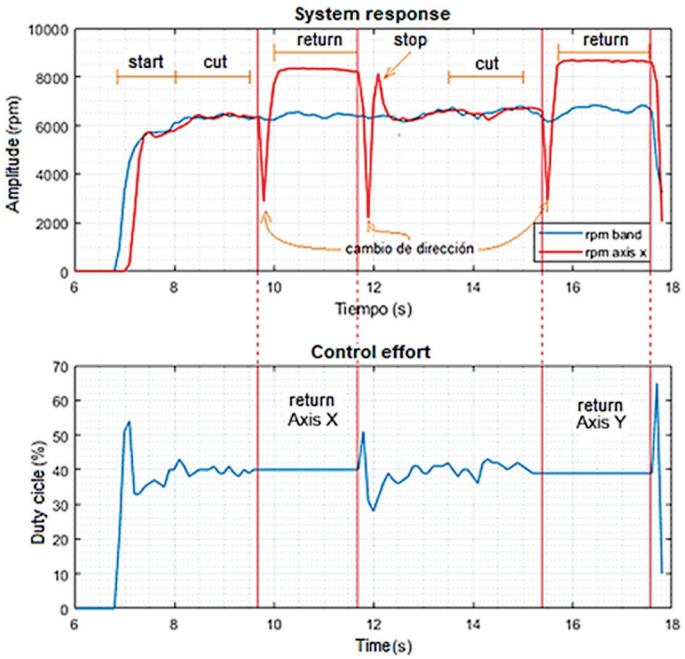


Figure 26. The programming structure developed for the simulation of the fly-cutting plant. Source: The authors.

From the operation of the cutting plant in a dynamic state, it was determined that the system exhibited a response time of 3.28 milliseconds following the application of the signal. Regarding the displacement of pieces from one end to the other, the system showed an error of 4.83%. This discrepancy arose because the theoretical time for this displacement was 4.55 s, while the system required 4.77 s to complete the task.

Additionally, certain displacement and cutting system parameters had to be recalculated due to differences between the theoretical maximum speed of the X-axis motor (100 rpm) and the actual achieved speed (88.2 rpm), resulting in an error of 11.8%. Due to these discrepancies in the response times of specific cutting plant elements, a compensation factor was incorporated into the controller to ensure precision and uniformity in the cuts. Without this compensation factor, the cuts exhibited a deviation of approximately 35°.

4. Conclusions

This research represents a significant advancement in the development of industrial plant prototypes within an academic framework, providing an invaluable resource for engineers training in control and automation. The primary methodological innovation lies in the design and simulation of a cost-effective, small-scale cutting plant that replicates continuous cutting processes found in industry. This approach enables students to understand and manage extensive processes within a laboratory environment, offering practical experience comparable to industrial settings without the high costs associated with commercial equipment.

From an engineering perspective, this study integrates key concepts from control theory, mechanical engineering, and physical principles to construct a prototype that facilitates education and hands-on experimentation with real-time control systems. The validation of the design through MATLAB simulations, along with the use of tools such

as Code Composer Studio for developing real-time controllers, underscores the plant's effectiveness as both an educational and research model, adaptable to various contexts.

Furthermore, this research is pivotal in bridging the gap between theoretical knowledge and practical application in engineering education. The developed prototype allows students to engage with authentic industrial processes, enriching their understanding of complex system dynamics and fostering the development of innovative and efficient industrial solutions using emerging technologies like 3D printing and augmented reality. These attributes make this prototype a valuable asset in training future engineers, equipping them to meet the challenges of industrial automation and control.

Author Contributions: Conceptualization, D.F.R.-J. and C.A.T.V.; methodology, D.F.R.-J. and C.A.T.V.; software, D.F.R.-J. and C.A.T.V.; validation, D.F.R.-J. and C.A.T.V.; formal analysis, D.F.R.-J. and C.A.T.V.; investigation, D.F.R.-J. and C.A.T.V.; resources, D.F.R.-J. and C.A.T.V.; data curation, D.F.R.-J. and C.A.T.V.; writing—original draft preparation, D.F.R.-J. and C.A.T.V.; writing—review and editing, D.F.R.-J. and C.A.T.V.; visualization, D.F.R.-J. and C.A.T.V.; supervision D.F.R.-J. and C.A.T.V.; project administration, D.F.R.-J. and C.A.T.V.; funding acquisition, D.F.R.-J. and C.A.T.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

The files required to perform the simulations can be downloaded from https://drive.google.com/file/d/1WNoosUDp7Nm2SyEC7THFFODBir-vfcfP/view?usp=drive_link (accessed on 8 December 2024).

References

1. Facultad de Humanidades y Ciencias Sociales. Evolución de las Tecnologías Educativas Desde Su Origen. Available online: https://www.ui1.es/blog-ui1/evolucion-tecnologias-educativas (accessed on 9 July 2024).
2. El Impacto de las TIC en la Educación Universitaria. Available online: https://www.ui1.es/blog-ui1/el-impacto-de-las-tic-en-la-educacion-universitaria (accessed on 9 July 2024).
3. Nkomo, L.M.; Daniel, B.K.; Butson, R.J. Synthesis of student engagement with digital technologies: A systematic review of the literature. *Int. J. Educ. Technol. High. Educ.* **2021**, *18*, 34. [PubMed]
4. Kerimbayev, N.; Umirzakova, Z.; Shadiev, R.; Jotsov, V. A student-centered approach using modern technologies in distance learning: A systematic review of the literature. *Smart Learn. Environ.* **2023**, *10*, 61. [CrossRef]
5. Crompton, H.; Burke, D. Artificial intelligence in higher education: The state of the field. *Int. J. Educ. Technol. High. Educ.* **2023**, *20*, 22. [CrossRef]
6. Lugo, J.G.C.; Padilla, J.J.; Ybarra, E.; Romero, A. Metodología para realizar una automatización utilizando PLC. *Rev. De Ing. Eléctrica Electrónica Comput.* **2005**, *1*, 18–21.
7. Medina, A.P.L.; Saba, G.H.; Silva, J.H.; De Guevara Durán, E.L. Los Laboratorios Virtuales y Laboratorios Remotos en la Enseñanza de la Ingeniería. *Rev. Int. De Educ. En Ing.* **2011**, *4*, 24–30.
8. Losada-Gutierrez, C.; Espinosa, F.; Santos-Perez, C.; Marron-Romera, M.; Rodriguez-Ascariz, J.M. Remote Control of a Robotic Unit: A Case Study for Control Engineering Formation. *IEEE Trans. Educ.* **2020**, *63*, 246–254. [CrossRef]
9. Madhivanan, N.; Narayanan, K. FPGA implementation of general purpose real time controller-applied to flying shears in hot rolling mills. In Proceedings of the International Conference on Computing, Electronics and Electrical Technologies (ICCEET), Nagercoil, India, 21–22 March 2012; pp. 521–525. [CrossRef]
10. Zhang, X.; Mei, R.; Chen, S.; Sun, J.; Zhang, D. Optimizing Design of Control System for Flying Shear in Tandem Cold Rolling. *Eng. Technol. Res.* **2017**. [CrossRef] [PubMed]
11. Wang, Y.Q.; Li, L.; Luo, Y.X.; Yan, X.C. Study on the Dynamics Performance of the Drive Train of Flying Shear Machine. *Appl. Mech. Mater.* **2011**, *121*, 1397–1401. [CrossRef]

12. Hirano, H.; Hikichi, K.; Tanaka, S. The wafer-level vacuum sealing and electrical interconnection using electroplated gold bumps planarized by single-point diamond fly cutting. In Proceedings of the 2015 Transducers-2015 18th International Conference on Solid-State Sensors, Actuators and Microsystems (TRANSDUCERS), Anchorage, AK, USA, 21–25 June 2015; pp. 1283–1286. [CrossRef]
13. Farisi, M.S.A.; Tanaka, K.; Hirano, H.; Tanaka, S. 250 °C wafer-level vacuum sealing using electroplated copper bonding frame planarized by fly-cutting. In Proceedings of the 19th International Conference on Solid-State Sensors, Actuators and Microsystems (TRANSDUCERS), Kaohsiung, Taiwan, 18–22 June 2017; pp. 1191–1194. [CrossRef]
14. Qi, L.; Shuo, L.; Ying, W.; Kai, L.; Baohua, L.; Huili, M.; Jinran, L. Design of Flying Shear Device and Algorithm Based On Motion Control System. In *Topics in Chemical & Material Engineering*; Volkson Press: Kuala Lumpur, Malaysia, 2018; pp. 257–259. [CrossRef]
15. Solomon, M.-G.; Gaiceanu, M. The Optimization of the Dynamic Regime of the Asynchronous Vectorial AC Motor in Flying Shear. In Proceedings of the 7th International Symposium on Electrical and Electronics Engineering (ISEEE), Galati, Romania, 28–30 October 2021; pp. 1–6. [CrossRef]
16. Wang, C.; Wang, X.; Zhang, B. Design of the flying shear servo control system. In Proceedings of the Chinese Automation Congress (CAC), Jinan, China, 20–22 October 2017; pp. 1659–1664. [CrossRef]
17. Peric, N.; Petrovic, I. Flying shear control system. *IEEE Trans. Ind. Applicat.* **1990**, *26*, 1049–1056. [CrossRef]
18. Ng, G.; Deleanu, S.; Prevost, J.-P.; Carpenter, D. Improving the operation of a flying dividing shear by using Direct Torque Control. In Proceedings of the International Conference on Modern Power Systems (MPS), Cluj-Napoca, Romania, 6–9 June 2017; pp. 1–8. [CrossRef]
19. Das, S.; Kumar, G.; Lal, D.K.; Sharma, G.; Kumar, A.; Adhikary, M.; Mukhopadhyay, G. Metallurgical Investigation of Flying Shear Blades at Hot Rolling Rebar Mills. *J. Fail. Anal. Preven.* **2023**, *23*, 1851–1859. [CrossRef]
20. Bi, J.; Fan, W.; Huang, H.; Liu, B. Reliability Design of an Electronic Cam Curve for Flying Shear Machine in Short Materials Cutting. *J. Shanghai Jiaotong Univ.* **2020**, *25*, 246–252. [CrossRef]
21. Song, J.C.; Wang, C.Z.; Xu, D. Dynamic Simulation and Control Strategy of Centrifugal Flying Shear. *Appl. Mech. Mater.* **2009**, *16*, 278–282. [CrossRef]
22. Huertas, A.S.; Madrid, O.A.L. Diseño de una Cizalla Industrial para el Corte de Láminas de Acero al Silicio Utilizadas en Transformadores Secos de Energía Eléctrica. Ph.D. Thesis, Universidad Tecnológica de Pereira, Facultad de Tecnologías, Pereira, Colombia, 2015.
23. Bravo, E.C.; Bacca, E.B.; Calvache, B.A.; Aristizabal, J.E.C.; Buitrago, J.A. Laboratorio distribuido con acceso remoto para la enseñanza de la robótica. *Rev. Educ. En Ing.* **2009**, *7*, 51–61.
24. Aravena, J.; Carrasco, D.; Uriarte, M.; Contreras, E.; Díaz, M.; Donoso, F. Design and Implementation of an Interface Board for a Real-Time Control System based on the LAUNCH-XL F28379D board. In Proceedings of the IEEE Chilean Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), Valparaíso, Chile, 6–9 December 2021; pp. 1–5. [CrossRef]
25. Rosinová, D.; Schwarz, R. Teaching State-Space Control Methods Using a Floatshield Laboratory Plant. In Proceedings of the 24th International Conference on Process Control (PC), Strbske Pleso, Slovakia, 6–9 June 2023; pp. 261–266. [CrossRef]
26. Gutiérrez, J.G.H.; Cardona, J.E.; Muñoz-Gutiérrez, P.; Ramírez-Jiménez, D.F. Rapid control prototyping using free software of the position control of a two degree of freedom robot. In Proceedings of the XIV Congreso Internacional de Electrónica y Tecnologìas de Avanzada, Pamplona, Spain, 20–22 October 2020; pp. 1–6.
27. Productos, Rishbin Wuxi Co., Ltd. Available online: http://es.rishbin.com (accessed on 5 July 2024).
28. Ghosh, K. A Study on Kinematic Analysis of Crank-Type Flying Shear Mechanism for Production-Oriented Design and Motion Control. *J. Inst. Eng. India Ser. C* **2019**, *100*, 851–858. [CrossRef]
29. Schneider Electric España, Schneider Electric: Motion control. Corte al Vuelo T2-T3, Application Know-How, (8 April 2019). Available online: https://www.youtube.com/watch?v=NKDQ9n3aAbw (accessed on 5 July 2024).
30. MATLAB Documentation. Available online: https://www.mathworks.com/help/matlab/index.html?s_tid=CRUX_lftnav (accessed on 15 July 2024).
31. Tune PID Controllers—MATLAB. Available online: https://www.mathworks.com/help/control/ref/pidtuner-app.html?s_tid=srchtitle_site_search_1_pid%20tuner (accessed on 15 July 2024.).
32. Navigating the SketchUp Interface | SketchUp Help. Available online: https://help.sketchup.com/en/sketchup/user-interface (accessed on 15 July 2024).
33. Ramirez-Jimenez, D.F.; Parrado, A.L.; Medina, J.V. Overview of a framework for Implementation of digital controllers in Energia IDE using Texas Instruments microcontrollers. In Proceedings of the IEEE 5th Colombian Conference on Automatic Control (CCAC), Ibague, Colombia, 19–22 October 2021; pp. 13–18. [CrossRef]
34. Ramírez-Jiménez, D.F.; Bedoya-Benítez, P.A.; Munoz-Gutierrez, P.A. Visualization and Control System for a Wastewater Laboratory Plant with Biological Treatment. *Cienc. E Ing. Neogranadina* **2024**, *34*, 1. [CrossRef]

35. Ramírez-Jiménez, D.F. Sistema de medición y control de temperatura para un prototipo de planta de tratamiento de aguas residuals. *Investig. Innovación Ing.* **2021**, *9*, 100–113. [CrossRef]
36. Smith, C.A.; Corripio, A.B. *Principles and Practice of Automatic Process Control*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 1997.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

machines

MDPI

Article

TOPSIS-Based Methodology for Selecting Fused Filament Fabrication Machines

Vignesh Venkat Raman, Rakshith Badarinath and Vittaldas V. Prabhu *

Marcus Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA 16802, USA; vvr5204@psu.edu (V.V.R.); rakshithb174@gmail.com (R.B.)
* Correspondence: vittal.prabhu@psu.edu

Abstract: Additive manufacturing (AM) has been gaining increased traction in the manufacturing industry due to its ability to fabricate prototypes and end use parts in low volumes at a much lower cost compared to conventional manufacturing processes. There has been research to select an AM process appropriate for fabricating particular parts. However, there is little extant research to select appropriate AM machines even though there is a growing number of AM machines with interesting topologies, structures, and systems. This paper proposes a methodology that aims to assist Technical Experts in selecting a machine for Fused Filament Fabrication (FFF). The methodology is built around a weighted Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), which uses the concept of relative closeness and attribute weights to rank the machines. The paper uses Monte Carlo simulations for sensitivity analysis to evaluate the impact of randomizing attribute scoring, perturbing weights assigned, and probability distributions used to model human decision variability. The methodology and the sensitivity analysis were applied to three case studies, with five FFF machines and seven attributes, and top machines ranked for a specific part were found to be largely robust.

Keywords: Fused Filament Fabrication; TOPSIS; material extrusion; machine selection; Monte Carlo sensitivity analysis

1. Introduction

Additive manufacturing (AM) is a manufacturing process where a part is built layer by layer using one of the following processes: material extrusion, stereolithography, powder bed fusion, sheet lamination, material jetting, binder jetting, and directed energy deposition. There has been a lot of research and many algorithms developed to help the user select a suitable process for a part and its application [1]. In metals, the mechanical properties obtained usually influence both the selected process and the parameters that are selected for that process [2]. However, in polymer extrusion, some properties, such as material behavior and mechanical performance of the finished part, are dependent primarily on the extruder and features such as temperature, flow rate, nozzle diameter, etc. [3]. A similar methodology could be conducted to compare the different extruders, and both the extruder and the machine ranking must be taken into consideration when a final decision is made. Conceptually, in subtractive or conventional manufacturing, this is akin to selecting the machining center and the tooling for parts. Although the machining center constrains the compatible tool, it is the final combination that influences the overall performance. Incorrectly selecting machines can have detrimental effects on the desired part. This could occur in the form of early part failure and inconsistent tolerance compliance. A higher-level

issue that can be caused is the inability to meet throughput requirements, which could be caused due to the longer loading times, larger support structure requirements, and longer post-processing times. Within material extrusion AM (Fused Filament Fabrication) alone, there exist several unique systems that build upon the advantages and disadvantages of the process by utilizing different nozzle/extruder mounting techniques.

Machine selection is a key gap in the literature space of FFF, as most of the selection-based literature seeks to help identify the type of AM process used, but not the specific machine type. S. Raja et al. [4] compared different AM methods by applying the Analytical Hierarchy Process (AHP). Y. Wang et al. [5] compare the different AM processes based on their suitability for Design for Additive Manufacturing (DfAM). Their approach is very suitable as a first step for additive manufacturing decision-making. However, once a suitable AM process has been determined, it becomes very challenging to identify a suitable machine within that AM process. For example, within FFF alone, there are several different machine types.

M. Palanisamy et.al. [6] compare commercially available machines within the selective laser sintering scope of AM. The authors have effectively captured the different elements of a decision-making process and have applied the best–worst method of MCDM. This approach is very suitable for selective laser sintering applications due to the increased complexity of the process. However, for a process like FFF, this approach would not work, as the quality of the manufactured part is more heavily dependent on the machine used. Conventionally and most frequently available machines have an extruder nozzle mounted on a gantry. However, due to the limitation of a gantry system in terms of print volume and design complexity capabilities, a lot of new novel approaches have been in development, and some have already been developed. For example, to counter the print volume limitation of the conventional gantry system, companies have introduced a belt-based system that extends the print volume unidirectionally [7]. These novel machines also have some limitations, usually cost or, in some cases, accessibility.

Kaill et al. [8] provides a comparative study between conventional gantry (3-axis) systems to robotic arm (5-axis) systems by comparing the performance of the manufactured samples. This is a good direction that takes advantage of the ability of researchers comparing their novel techniques to conventional ones. However, industry cannot apply these techniques, as it would require access to both the systems, making the comparison expensive and non-feasible. For a selection purpose, the problem is comparing the available techniques on an overall basis based on the requirements of the industry/company.

The aim of this paper is to present Technical Experts, decision-makers, and key stakeholders with a proposed solution to this problem. Using FFF as an example, the paper provides a step-by-step methodology that will help compare the different types of machines within an AM process scope. The proposed methodology is based on the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) method, which has been modified to account for different attribute weights. Three parts of different complexities are used for evaluating the proposed methodology. However, as TOPSIS relies on subjective decisions, a sensitivity analysis is performed to evaluate the robustness of the results obtained from the case studies.

2. Methodology

This methodology is designed to support Technical Experts (TEs) and decision-makers in identifying and ranking suitable Fused Filament Fabrication (FFF) systems for part production. Considering that decision-makers may not possess deep analytical or technical expertise, the input requirements have been deliberately simplified, and the objectives are linearly structured. This design choice aligns with findings in decision science, which indicate that individuals perform better when presented with simplified comparisons, especially in pairwise formats [9].

To develop the decision framework, three multi-criteria decision-making (MCDM) methods were evaluated and compared: the Analytic Hierarchy Process (AHP) [10], VIšekriterijumsko KOmpromisno Rangiranje (VIKOR) [11], and the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS). The AHP was excluded due to its complexity and susceptibility to inconsistency and bias in group decision-making scenarios. VIKOR offered robust trade-off modeling but was ultimately not selected. TOPSIS was chosen as the preferred method for its capability to effectively manage both benefit and cost-type criteria, scalability to large decision spaces, and interpretability by non-expert stakeholders.

The proposed methodology employs a weighted TOPSIS approach and proceeds through seven sequential steps, summarized below. The first four steps require expert input, while the final three involve computational procedures.

2.1. Compilation of Machine-Attribute Matrix

TEs compile a list of candidate FFF machines and associated performance attributes. Attributes may include capital cost, build volume, precision, and other relevant technical specifications. Let:

- M_j represent the jth machine, where $j = 1, 2, 3, \ldots, J$;
- A_k represent the kth attribute, where $k = 1, 2, 3, \ldots, K$.

The resulting matrix R of dimension $J \times K$ is structured as:

$$R = \begin{bmatrix} A_{11} & \ldots & A_{1J} \\ \vdots & \ddots & \vdots \\ A_{K1} & \ldots & A_{KJ} \end{bmatrix}$$

Here, A_{kj} denotes the value of attribute k for machine j. All entries in a given column are expressed in consistent units.

2.2. Definition of Ideal and Non-Ideal Attribute Values

TEs define the ideal (most desirable) and non-ideal (least desirable) scores for each attribute. These form the basis for evaluating relative performance:

- I = Ideal score vectors
- NI = Non-ideal score vectors

2.3. Scoring and Normalizing

Each machine is evaluated against the defined attributes based on specification data and expert judgment. A normalized score matrix S is constructed:

$$S = \begin{bmatrix} S_{11} & \ldots & S_{1J} \\ \vdots & \ddots & \vdots \\ S_{K1} & \ldots & S_{KJ} \end{bmatrix}$$

Normalization ensures that all attributes are scaled appropriately to allow valid comparison across different units and magnitudes.

2.4. Assignment of Attribute Weights

TEs assign relative importance weights $W = [W_1, W_2, \ldots, W_K]$ to each attribute. These weights may be derived from literature, stakeholder priorities, or requirements specific to the intended part applications. In cases involving multiple experts, differences in assigned weights can be reconciled using methods such as the rating or Borda count techniques.

2.5. Calculation of Distances from Ideal and Non-Ideal Solutions

Euclidean distances are computed between each machine's performance and the ideal/non-ideal vectors:

$$D^+ = I - S$$
$$D^- = S - NI$$

These distances quantify how close each alternative is to the ideal and how far it is from the non-ideal solution.

2.6. Computation of Relative Closeness

The relative closeness of each machine to the ideal solution is then calculated:

$$C = \frac{D^-}{(D^+ + D^-)}$$

Higher values of C_{kj} indicate greater desirability.

2.7. Aggregation and Ranking

The final score T_j for each machine is obtained by weighing the closeness values:

$$T_i = \sum_{n=1}^{k} \left(W_n * C_{kj} \right) \forall j$$

Machines are then ranked in descending order of T_j, with the highest score indicating the most suitable machine.

2.8. Sensitivity Analysis

To assess the robustness of the rankings and account for potential bias in the weights or scores, a sensitivity analysis is recommended. Monte Carlo simulation can be used to perturb the attribute weights within a predefined range and observe variations in the resulting rankings. This provides insight into the stability and confidence of the selection outcome.

Figure 1 illustrates the full methodology workflow. Solid lines denote mandatory steps, while dashed lines indicate optional procedures, such as group-based weight reconciliation.

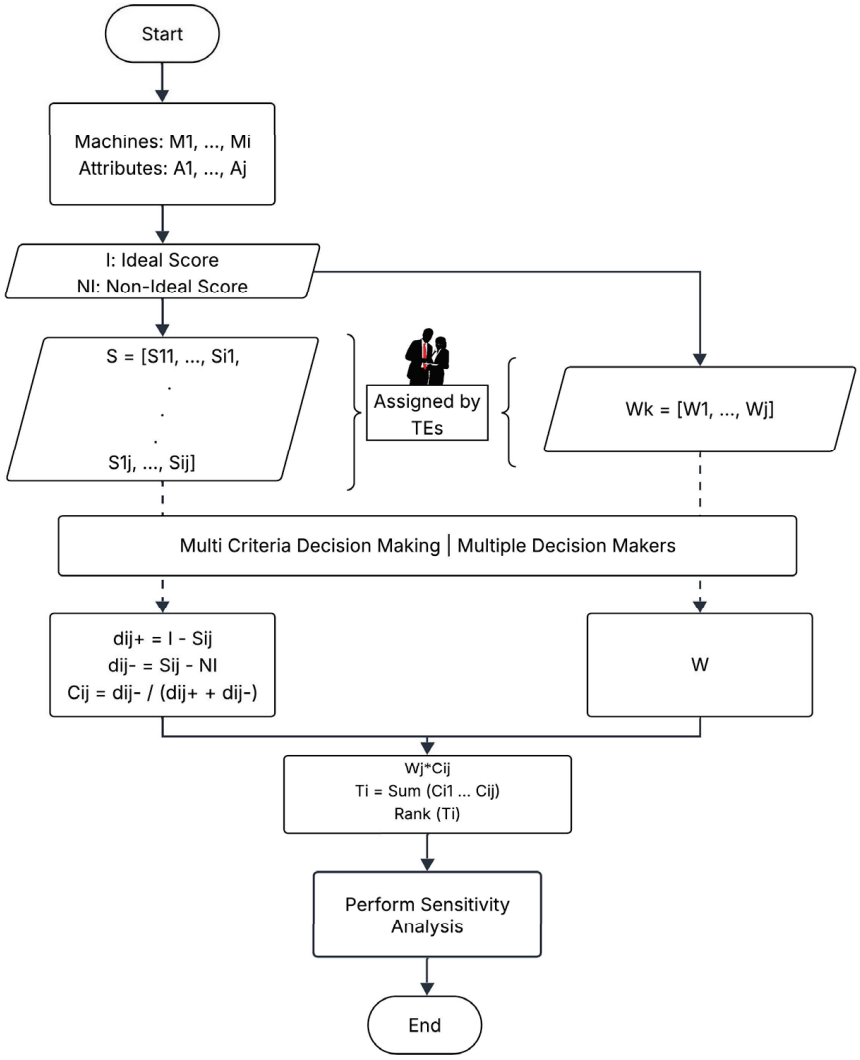


Figure 1. Methodology flow chart.

When an organization evaluates FFF adoption, it typically has specific part candidates in mind, with known dimensions, tolerances, production volumes, and cost constraints. The proposed methodology enables ranking of machines for each part. If a single machine consistently ranks highest across all parts, it becomes the clear choice. Otherwise, further analysis may be required to resolve conflicting rankings and optimize machine selection.

3. Evaluation and Discussion of Proposed Methodology

The proposed methodology is evaluated by performing three case studies using a set of 5 machines, 7 attributes, and 3 parts that have different weights. The machines and attributes selected are provided in Tables 1 and 2, respectively.

Table 1. List of machines selected for evaluation of the methodology.

	Machines:
M1	Retrofit
M2	Gantry
M3	Assembled robot
M4	Belt Gantry
M5	Fixed nozzle

Table 2. List of attributes selected for evaluation of the methodology.

	Attributes:
A1	Build volume
A2	Design freedom
A3	Support independence
A4	Max print speeds
A5	Ease of use
A6	Cost effectiveness
A7	G-code generation effort

Machine 1 is a retrofit FFF system [12], which is a custom-built 6-degree-of-freedom robotic arm with an extruder attached to the end. This machine is most suitable for single-use purposes, as it can be repurposed later. Machine 2 is a conventional gantry system. The extruder can be moved using either the Cartesian or the polar coordinate systems. This is the most commonly occurring FFF system, as most desktop systems are of this kind. Machine 3 is an assembled robot [13] and involves buying a whole pre-built robotic arm setup. The benefit of this is cost versus convenience. It is more suitable for large companies that want to use the system for mass production. Machine 4 is a belt gantry system [7]. It is a customization on the conventional gantry system, where instead of a traditional build plate, a moving belt is used to create an infinite axis. This is most suitable for bulk production or for unidirectionally large parts. This system is more expensive due to the customization of the conventional gantry system. Finally, machine 5 is a fixed nozzle system [14]. This is a novel take on the retrofit system where, instead of the extruder being attached to the robotic arm, the build plate is attached to the robotic arm. The extruder instead is fixed. This machine is most suitable for cases where multiple manufacturing processes are to be conducted on the same part.

The weights and scores are from Technical Experts (TEs); these are subjective and can vary from one expert to another even though they may be largely in agreement [15]. Specifically, these subjective values may vary based on the experiences of each expert as well as their frame of mind when evaluating the case. To test the sensitivity of the rankings obtained to changes in these subjective values, a Monte Carlo simulation [16] of 10,000 iterations was performed on the model. This was performed by modeling the scores as a triangular distribution with the assigned score as the mode and the lower and upper limits as 1 and 10, respectively. The rationale for the model is that the score assigned by the TEs may differ, but the probability of these differences decreases with increasing difference. This ensures each iteration of the simulation has randomized scores, with the assigned score being the most frequently appearing score.

To further evaluate the robustness of the model, the weights for Case 1 (Nema-23 adapter plate) were perturbed. The perturbation was performed by reducing the highest weighed attribute and proportionally distributing the reduced amount to the remaining weights (−10%, −20%). The perturbation was also performed by increasing the highest weighed attribute and proportionally removing the increased amount from the remaining weights (+10%, +20%). Figure 2 is the proposed evaluation methodology that has been used. The methodology for the Monte Carlo sensitivity analysis is outlined below:

1. The first step of the sensitivity analysis is to repeat the initial steps of the TOPSIS methodology, which include compiling the R matrix, defining the I and NI values, assigning scores to the S matrix, and giving each attribute a weight (W matrix).
2. Once the preparatory steps are completed, the next step of the sensitivity analysis is to use the values in the score matrix S as the mode to generate a matrix with a triangular distribution for each corresponding element. This triangular distribution will be used

in a loop (based on the number of iterations) to generate a random number for the element score in each iteration.

3. The third step of the sensitivity analysis is to perform the TOPSIS methodology at every iteration and store the ranks obtained. Once the ranks are determined and stored, the frequency of each rank appearing is observed and plotted.
4. The next step is to perturb the weights and restart the TOPSIS methodology. The weight of the highest weighed parameter is perturbed, and the score is either distributed to or taken from the other weights. This also ensures that the weights remain normalized.

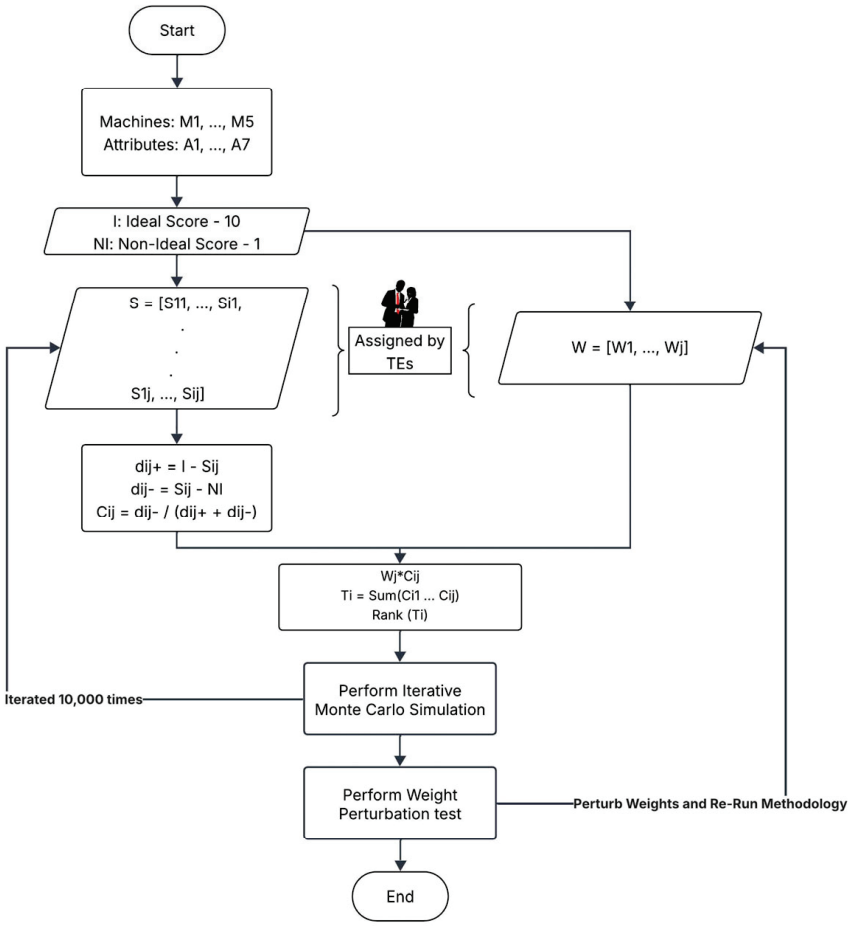


Figure 2. Evaluation flow chart.

3.1. Nema-23 Motor Adapter Plate [17]

This part, displayed in Figure 3, has a relatively small build volume and simple geometry without overhangs, so it does not require significant support material. Therefore, attributes that are important for this part are print speed, cost of the printer, setup simplicity, and G-code generation simplicity. Accordingly, weights are assigned and presented in Table 3, and Table 4 presents the weighted scores for each of the machines under consideration. Based on this, the best machine would be the gantry (score of 0.78), followed by the belt gantry (score 0.71). The ranked order is M2 gantry, M4 belt gantry, M5 fixed nozzle, M1 retrofit, and M3 assembled robot.

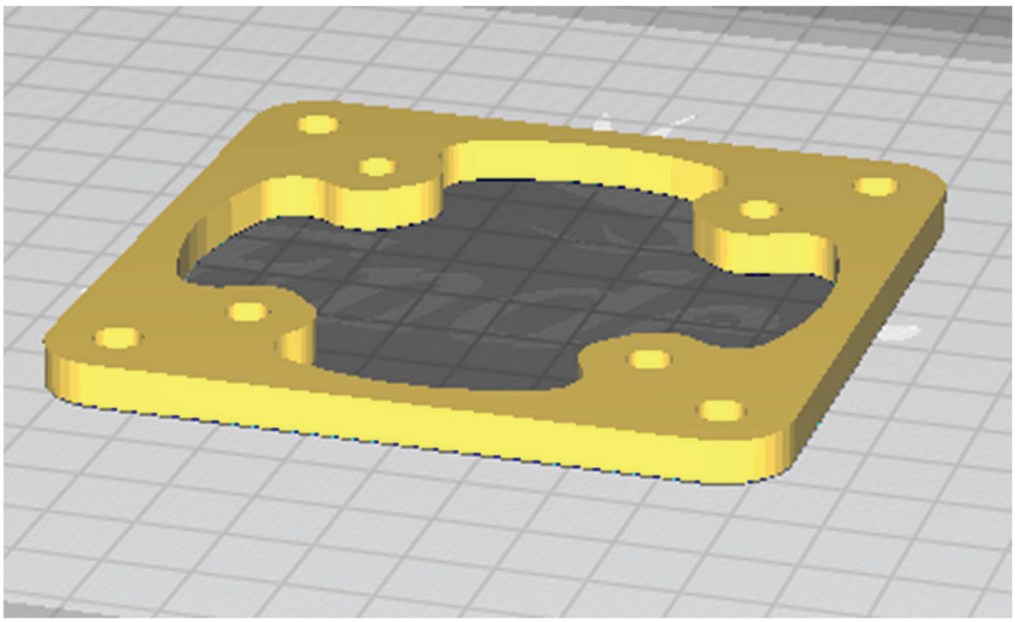


Figure 3. Nema-23 motor adapter plate.

Table 3. Assigned weights for the case of the Nema-23 motor adapter plate.

Criteria/Parameter	Weights
Build volume	0.05
Design freedom	0.05
Support independence	0.05
Max print speeds	0.3
Ease of use	0.25
Cost effectiveness	0.2
G-code generation effort	0.1

Table 4. Weighted score calculation for the case of the Nema-23 motor adapter plate.

Criteria/Parameter	Weighted Scores (Weight × Score)				
	Retrofit	Gantry	Assembled Robot	Belt Gantry	Fixed Nozzle
Build volume	0.02	0.03	0.02	0.04	0.03
Design freedom	0.04	0.02	0.04	0.03	0.04
Support independence	0.04	0.01	0.04	0.02	0.04
Max print speeds	0.13	0.27	0.13	0.27	0.13
Ease of use	0.08	0.22	0.14	0.17	0.08
Cost effectiveness	0.16	0.13	0.04	0.09	0.16
G-code generation effort	0.02	0.10	0.08	0.10	0.02
Net	**0.49**	**0.78**	**0.49**	**0.71**	**0.51**

3.2. Five-Blade Impeller [18]

This part, displayed in Figure 4, has a relatively small build volume but has a complex geometry with significant overhanging features, so support independence becomes important. Because of these part requirements, the cost, G-code generation, and maximum print speeds are given higher weights. Accordingly, weights were assigned and are presented in Table 5, and Table 6 presents the weighted scores for each of the machines under consideration. Based on this, the best machine would be the fixed nozzle (score of 0.69), followed closely by the assembled robot (score 0.66) and retrofit (score 0.65).

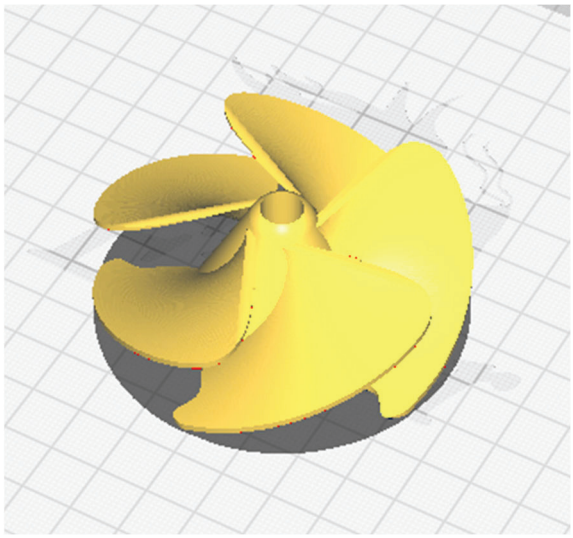


Figure 4. Five-blade impeller.

Table 5. Assigned weights for the case of the five-blade impeller.

Criteria/Parameter	Weights
Build volume	0.05
Design freedom	0.3
Support independence	0.3
Max print speeds	0.1
Ease of use	0.05
Cost effectiveness	0.1
G-code generation effort	0.1

Table 6. Weighted score calculation for the case of the five-blade impeller.

Criteria/Parameter	Weighted Scores (Weight × Score)				
	Retrofit	Gantry	Assembled Robot	Belt Gantry	Fixed Nozzle
Build volume	0.02	0.03	0.02	0.04	0.03
Design freedom	0.23	0.13	0.23	0.17	0.23
Support independence	0.23	0.03	0.23	0.10	0.27
Max print speeds	0.04	0.09	0.04	0.09	0.04
Ease of use	0.02	0.04	0.03	0.03	0.02
Cost effectiveness	0.08	0.07	0.02	0.04	0.08
G-code generation effort	0.02	0.10	0.08	0.10	0.02
Net	**0.65**	**0.49**	**0.66**	**0.58**	**0.69**

3.3. Vertical-Axis Wind Turbine [19]

This part, displayed in Figure 5, has a relatively large build volume with a complex geometry, so support independence becomes important. Because of this part's requirements, the build volume, design freedom, and support independence are given higher weights. Accordingly, weights are assigned and presented in Table 7, and Table 8 presents the weighted scores for each of the machines under consideration. Based on this, the best machine would be the belt gantry (score of 0.7), and other machines are clustered together but have significantly lower weighted scores.

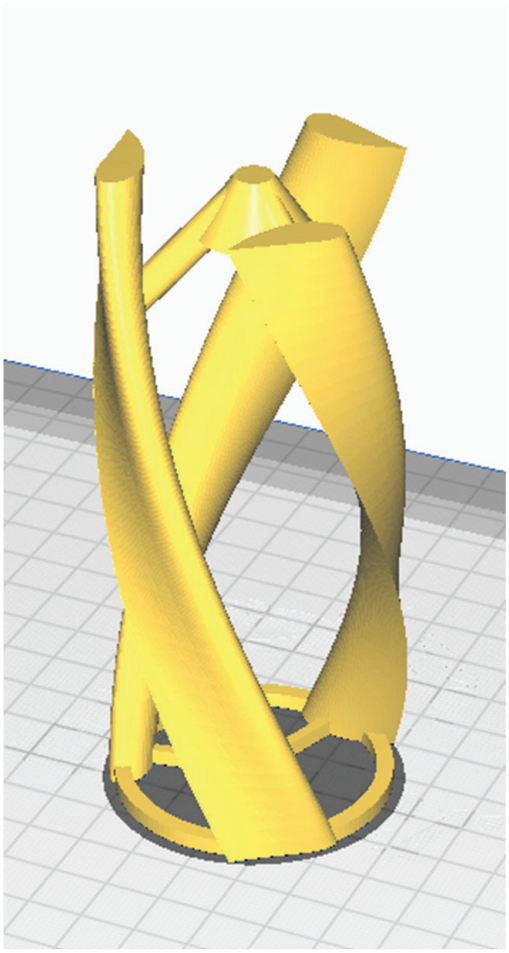


Figure 5. Vertical-axis wind turbine.

Table 7. Assigned weights for the case of the vertical-axis wind turbine.

Criteria/Parameter	Weights
Build volume	0.3
Design freedom	0.15
Support independence	0.15
Max print speeds	0.1
Ease of use	0.1
Cost effectiveness	0.1
G-code generation effort	0.1

Table 8. Weighted score calculation for the case of the vertical-axis wind turbine.

Criteria/Parameter	Weighted Scores (Weight × Score)				
	Retrofit	Gantry	Assembled Robot	Belt Gantry	Fixed Nozzle
Build volume	0.13	0.17	0.13	0.27	0.17
Design freedom	0.12	0.07	0.12	0.08	0.12
Support independence	0.12	0.02	0.12	0.05	0.13
Max print speeds	0.04	0.09	0.04	0.09	0.04
Ease of use	0.03	0.09	0.06	0.07	0.03
Cost effectiveness	0.08	0.07	0.02	0.04	0.08
G-code generation effort	0.02	0.10	0.08	0.10	0.02
Net	**0.54**	**0.59**	**0.57**	**0.70**	**0.59**

3.4. Sensitivity Analysis

Given that the scores and weights are subjective inputs, it becomes important to assess how sensitive the rankings provided by the methodology are, particularly when the top ranked ones are close to each other. For example, an input from a decision-maker may not precisely represent their opinion and may be somewhat approximate. The challenge therefore is to assess how the rankings for the machines change if the scores and weights are changed. Therefore, sensitivity analysis was performed using Monte Carlo simulations for all three cases.

Figure 6 shows the number of times a specific ranking of the five machines occurred for Case 1 in 10,000 runs of Monte Carlo simulation. It can be observed from the figure that the ranking [4,6,8,20,21] occurred 445 times, i.e., M2 was the top, followed by M4, M5, M1, and M3 at the bottom. This is consistent with the baseline ranked order of M2 gantry, M4 belt gantry, M5 fixed nozzle, M1 retrofit, and M3 assembled robot. Furthermore, in Figure 6, it is observed that there was a significant drop in rank frequency when there was a change in the top two machines. If we compare the data in Figure 6 with those in Table 4, it is observed that the scores of the gantry and the belt gantry system were significantly higher than the other machines, which is further confirmed by the data in Figure 6. Therefore, in this case, the rank order can be considered robust to small variations in human input.

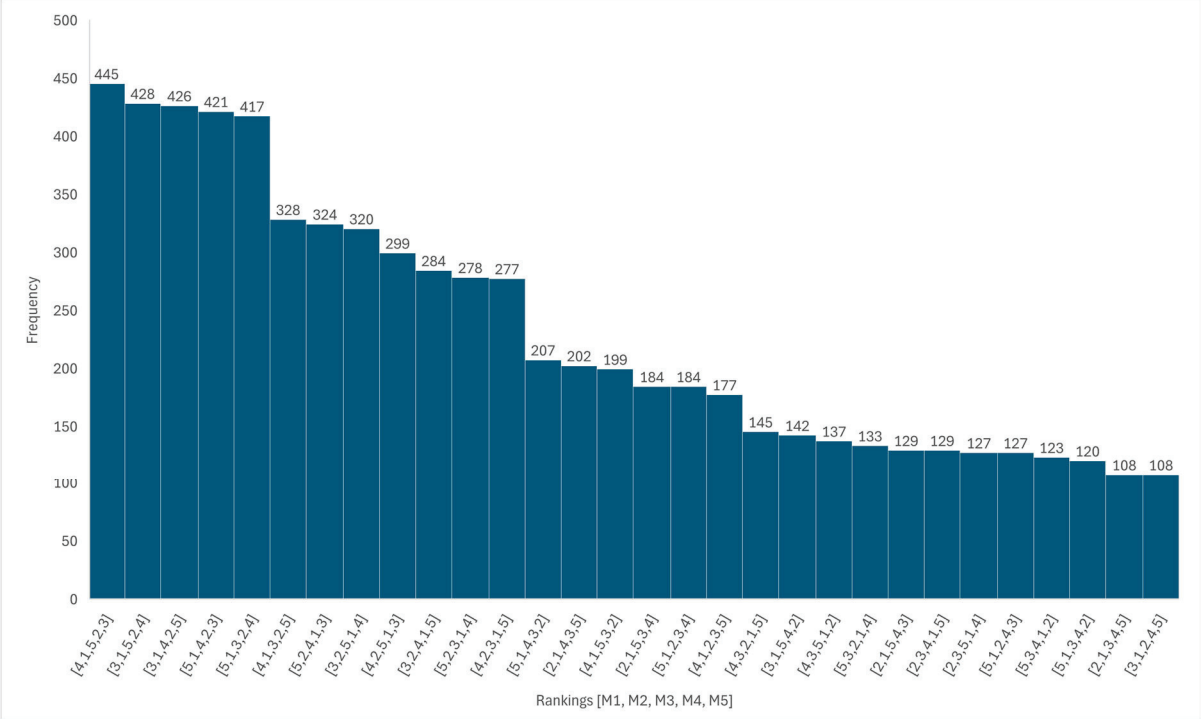


Figure 6. Monte Carlo simulation frequency of top 30 machine ranking orders for the Nema-23 motor adapter plate.

Similarly, Figures 7 and 8 show Cases 2 and 3, respectively, with the number of times a particular rank order is obtained after 10,000 runs of Monte Carlo simulation. It is observed that the top two machines that have been recommended based on baseline human input match the ones from the simulation about 40% of the time. In Figure 7, there are three machines that are rotating in the top two recommended machines. This can be attributed to scores that are relatively close to each other in Table 6, further confirming that the baseline human decision is robust in this case.

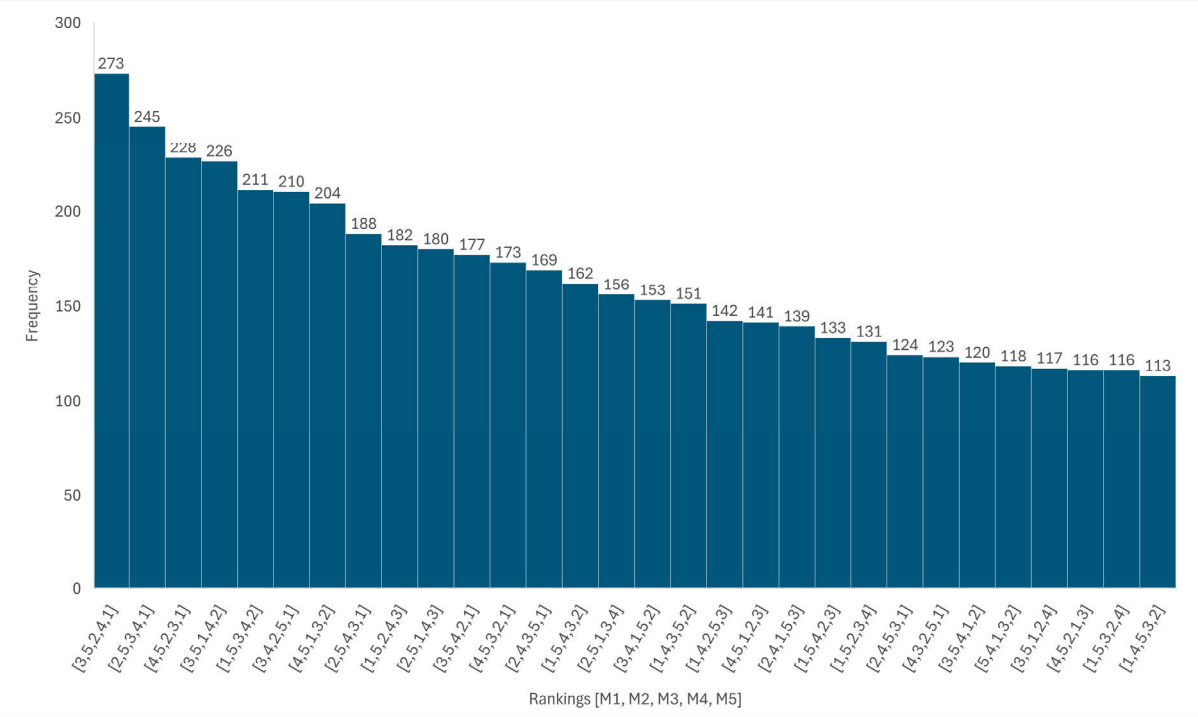


Figure 7. Monte Carlo simulation frequency of top 30 machine ranking orders for the 5-blade impeller.

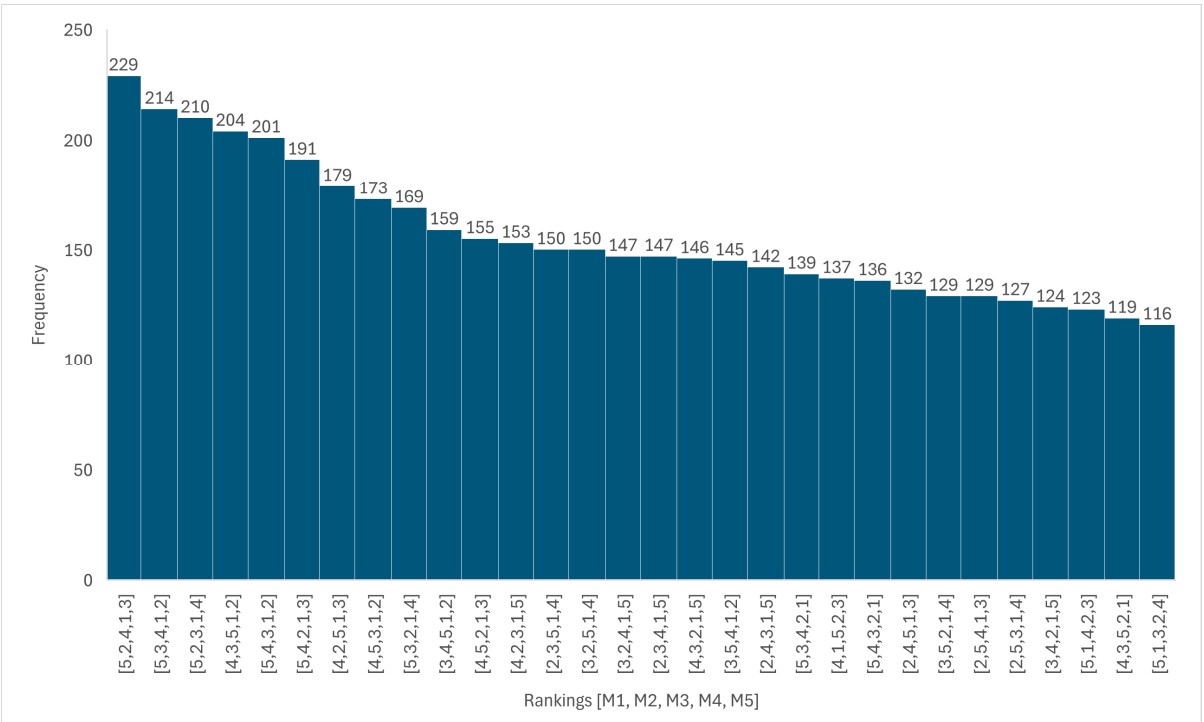


Figure 8. Monte Carlo simulation frequency of top 30 machine ranking orders for the vertical-axis wind turbine.

In Table 8, the belt gantry system has the best score by a significant margin from the other machines, whose scores are all close to each other. From Figure 8, it is observed that the belt gantry system is top ranked most often. Since the other systems have scores close to each other, there is no significant difference between them, and the baseline human decision can be considered robust in this case also.

From the Monte Carlo simulation of the scores, the recommended printer matched the ones from the case studies. From this, it is evident that the TEs' assigned scores display some level of flexibility and the overall methodology can be considered robust. The third case study is the only one that shows a deviation in the ranking of printers, albeit the rankings of the worst were interchanged. This is acceptable, as their scores are relatively close to each other.

From the Monte Carlo simulations of the varied weights, we can observe that the ranking of the printers did not change with perturbations. In particular, the machine that was top ranked remained the same regardless of the amount of perturbation or direction of perturbation, as shown in Table 9. This is because the methodology depends on the overall ratio of all parameters rather than on the weight of a particular parameter, which makes it robust.

Table 9. Frequency of each machine ranking first in the Monte Carlo simulation after perturbing weights.

Machine	Perturbation%						
	−20%	−10%	0%	10%	20%		
Retrofit	7.10%	6.74%	6.58%	6.50%	6.40%		High
Gantry	48.07%	48.21%	47.88%	47.72%	47.44%		
Assembled robot	7.05%	6.58%	6.52%	6.36%	6.38%		
Belt gantry	29.85%	31.35%	32.14%	32.70%	33.24%		
Fixed nozzle	7.93%	7.12%	6.88%	6.72%	6.54%		Low

Another aspect of robustness is the choice of probability distribution used to characterize variation in human input and the number of simulation runs: e.g., whether the top ranked machine remains the same for 1000 simulations, 5000 simulations, and 10,000 simulations. All the above Monte Carlo simulations model human input variation as a triangular distribution. Alternatively, instead of triangular distribution, normal or beta distributions can be considered.

Figure 9 illustrates a histogram of 10,000 computed samples for triangular distribution for machine 1, attribute 1, with a specified mode of 5, an upper limit of 10, and a lower limit of 1. Similarly, Figures 10 and 11 illustrate histograms for normal and beta distributions. The specified (and computed) parameters of the triangular distribution are mean 5 (5.34), mode 5 (5.13), and standard deviation 1.5 (1.82). The specified (and computed) parameters of the normal distribution are mean 5 (4.99), mode 5 (4.93), and standard deviation 1.5 (1.53). The specified (and computed) parameters of the beta distribution are mean 5 (5.0), mode 5 (4.95), and standard deviation 1.5 (1.34). Given the close agreement between specified and computed values of the parameters, the Monte Carlo simulation program can be considered an accurate computational model.

The impact of triangular, normal, and beta distributions were evaluated by running 10,000 simulations for the case of Nema 23 plate. Specifically, for the triangular distribution, the mode was set to the score assigned by the TEs along with lower and upper bounds of 1 and 10, respectively. For the normal distribution, the mean is the TE-assigned score and the standard deviation is 1.5. For the beta distribution, to have a mean of 5, mode of 5, and standard deviation of 1.5, the shape parameters specified are $\alpha = 4.44$, $\beta = 5.56$, with a concentration of 10. Minor adjustments are made in the Monte Carlo simulation program to ensure that scores computed are bounded between 1 and 10 by rounding up/down to the nearest bound. As shown in Table 10, the top two machines were consistent regardless of the distribution used. Therefore, in this case, the methodology can be considered robust to the probability distribution used to model the variation in human input.

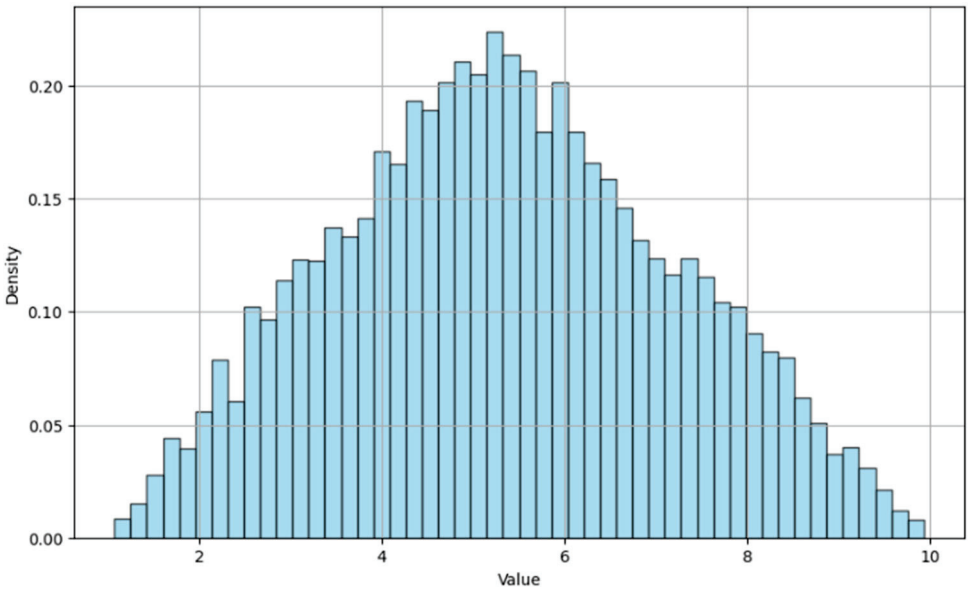


Figure 9. Triangular distribution for the build volume for the retrofit robot for the case of the Nema-23 motor adapter plate.

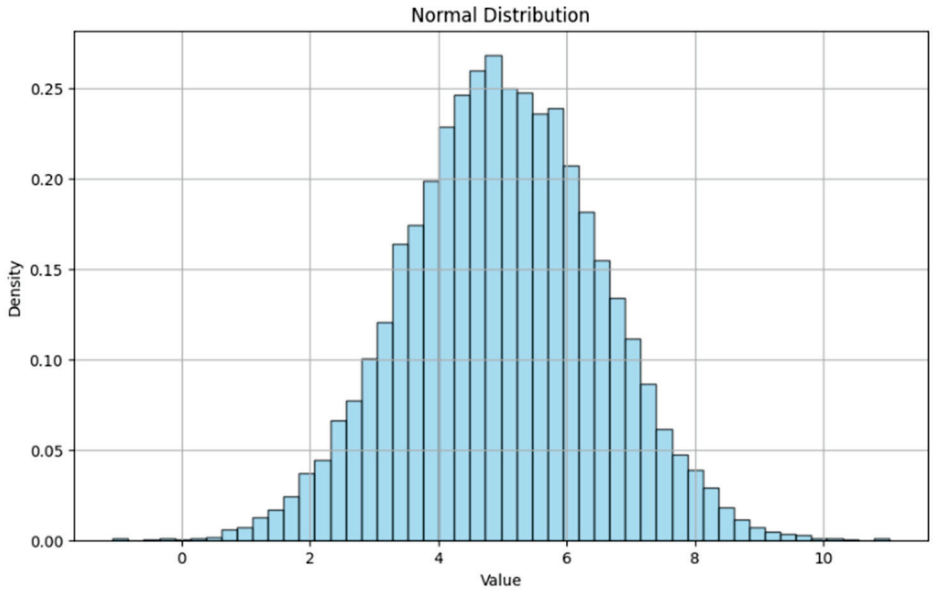


Figure 10. Normal distribution for the build volume for the retrofit robot for the Nema-23 motor adapter plate.

Table 10. Results of the Monte Carlo simulations with different distributions.

Type of Distribution	Triangular	Normal	Beta
Recommended machines	2, 4	2, 4	2, 4
Frequency	4141	5469	7184
Frequency percentage	41.41%	54.69%	71.84%

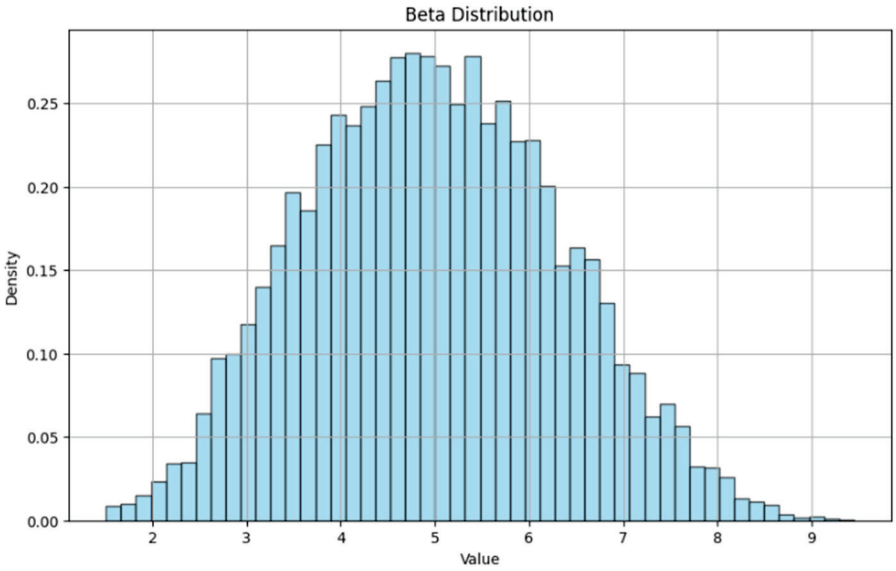


Figure 11. Beta distribution for the build volume for the retrofit robot for the Nema-23 motor adapter plate.

The triangular distribution had the lowest frequency of occurrence of the top two recommended machines. From the frequency plots, all three distributions followed the same trend. The difference in the frequencies between different recommendations was more severe in the normal and beta distributions. There was very little difference between the orders within the top two machines, which was observed in the triangular distribution. Figure 12 provides the frequency of the top eight recommended systems for each distribution.

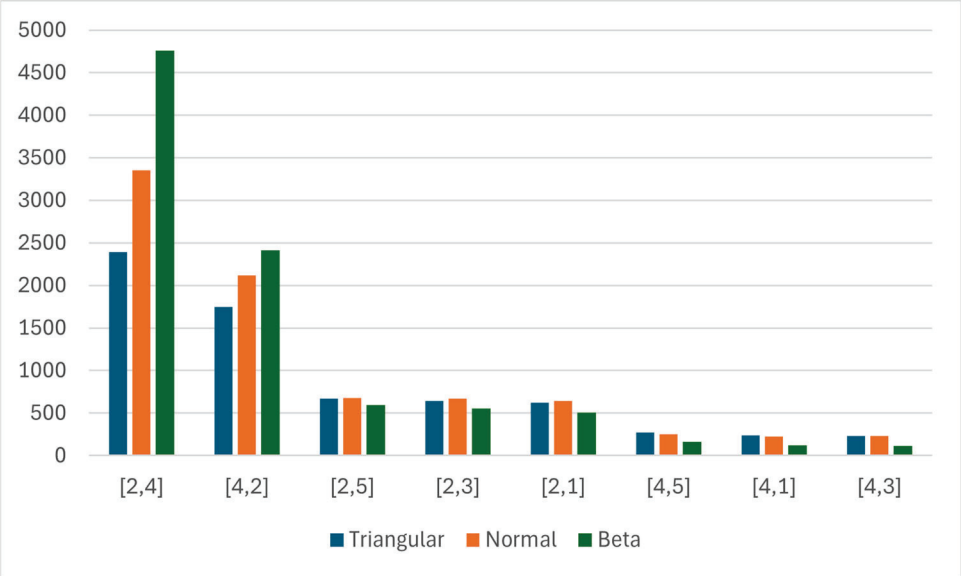


Figure 12. Monte Carlo simulation frequency of top 8 machine recommendations for different distributions in the Nema-23 motor adapter plate.

To evaluate the sensitivity of the ranked order to the number of simulation runs, Monte Carlo simulations were run for 1000, 5000, and 10,000 runs. As shown in Table 11, the top two machines were consistently M2 and M4 for all runs. Therefore, in this case, the methodology can be robust to the number of simulation runs beyond 1000.

Table 11. Results of the Monte Carlo simulations with different runs for the case of the Nema-23 motor adapter plate.

Number of Iterations	1000	5000	10,000
Recommended machines	2, 4	2, 4	2, 4
Frequency	430	2078	4117
Frequency percentage	43.00%	41.56%	41.17%

4. Conclusions

This paper proposed a TOPSIS-based methodology for selecting FFF machines and applied the methodology to rank machines for fabricating three different parts. The parts considered were a NEMA 23 adapter plate, an impeller blade, and a vertical turbine, which varied in complexity and the need for support structures. Five FFF machines with different architectures were considered. This methodology has several advantages:

1. Computationally simple. No complex algorithm that requires high computational power is used.
2. Easily expandable/modifiable. As the scores do not mutually influence each other, additional machines can be added in the future without having to repeat the human input elicitation for earlier machines.
3. Part specific performance: This methodology ranks machines based on the requirements for the specific part under consideration.

One limitation of the methodology is that it relies on human inputs for scoring machines for various criteria and weights for these criteria, all of which are subjective and can vary based on the individual's expertise and experience. Monte Carlo simulation was used to perform several sensitivity analyses to evaluate the impact of changes in human input to the ranked order of machines for the three parts. It was found that the top two machines consistently remained on top even when scores were changed randomly in the simulation. Moreover, +/− 10% or +/− 20% perturbations in the most important weight did not change the top ranked machine for any of the three parts. Furthermore, the proposed methodology was also found to be robust to the probability distribution used to characterize variability in human input across triangular, normal, and beta distributions. Lastly, top-ranked machines did not change for 1000, or 5000, or 10,000 Monte Carlo simulations. Therefore, it can be reasonably concluded that the proposed methodology is robust in terms of variability in human inputs of scores, weights, and the probability distribution used to model human variability.

5. Future Work

The focus of this paper was selecting machines for FFF for a given part. Future work should include ranking extruders for various materials and mechanical properties. Much of the proposed methodology can be reused for other additive manufacturing machines and processes. Another worthy topic would be to include assistance in the selection of suitable weights for each criterion. The use of explainable MCDMs such as VIKOR could also be implemented alongside the methodology to provide additional details on the tradeoffs that help the decision-makers understand the reasoning behind the TOPSIS ranking.

Author Contributions: Conceptualization, V.V.R., R.B., and V.V.P.; software, V.V.R.; validation, V.V.R. and R.B.; formal analysis, V.V.R.; investigation, V.V.R. and R.B.; resources, V.V.P.; data curation, V.V.R. and R.B.; writing—original draft preparation, V.V.R.; writing—review and editing, V.V.R., R.B., and V.V.P.; visualization, V.V.R. and R.B.; supervision, V.V.P.; project administration, V.V.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data relevant to the research are included in the body of the manuscript and in Appendix A.

Acknowledgments: The author would like to acknowledge Anshul Balamwar 1 for his assistance in developing the Monte Carlo Code for the sensitivity analysis. The authors acknowledge the use of AI-based tools and OpenAI's ChatGPT (version: GPT-4) for language support.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AM	Additive Manufacturing
FFF	Fused Filament Fabrication
TOPSIS	Technique for Order of Preference by Similarity to Ideal Solution
MCDA	Multi-Criteria Decision Analysis
DfAM	Design for Additive Manufacturing
AHP	Analytical Hierarchy Process
VIKOR	VIšekriterijumsko KOmpromisno Rangiranje
SME	Subject Matter Expert
TE	Technical Expert

Appendix A

Table A1. Euclidean distances from the ideal vector.

Criteria/Parameter	Retrofit	Gantry	di+ Assembled Robot	Belt Gantry	Fixed Nozzle
Build volume	5	4	5	1	4
Design freedom	2	5	2	4	2
Support independence	2	8	2	6	1
Max print speeds	5	1	5	1	5
Ease of use	6	1	4	3	6
Cost effectiveness	2	3	7	5	2
G-code generation effort	7	0	2	0	7

Table A2. Euclidean distances from the non-ideal vector.

Criteria/Parameter	Retrofit	Gantry	di− Assembled-Robot	Belt-Gantry	Fixed Nozzle
Build volume	4	5	4	8	5
Design freedom	7	4	7	5	7
Support independence	7	1	7	3	8
Max print speeds	4	8	4	8	4
Ease of use	3	8	5	6	3
Cost effectiveness	7	6	2	4	7
G-code generation effort	2	9	7	9	2

Table A3. Relative closeness values.

Criteria/Parameter	Retrofit	Gantry	C Assembled Robot	Belt Gantry	Fixed Nozzle
Build volume	0.44	0.56	0.44	0.89	0.56
Design freedom	0.78	0.44	0.78	0.56	0.78
Support independence	0.78	0.11	0.78	0.33	0.89
Max print speeds	0.44	0.89	0.44	0.89	0.44
Ease of use	0.33	0.89	0.56	0.67	0.33
Cost effectiveness	0.78	0.67	0.22	0.44	0.78
G-code generation effort	0.22	1.00	0.78	1.00	0.22

Table A4. Perturbed weights.

Parameter	Perturbation%				
	−10%	−20%	0%	+10%	+20%
Build volume	0.0521	0.0543	0.05	0.0479	0.0457
Design freedom	0.0521	0.0543	0.05	0.0479	0.0457
Support independence	0.0521	0.0543	0.05	0.0479	0.0457
Max print speeds	0.27	0.21	0.3	0.33	0.36
Ease of use	0.2607	0.2714	0.25	0.2393	0.2286
Cost effectiveness	0.2086	0.2171	0.2	0.1914	0.1829
G-code generation effort	0.1043	0.1086	0.1	0.0957	0.0914

References

1. Tavcar, J.; Nordin, A. Multi-criteria assessment and process selection model for additive manufacturing in the conceptual phase of design. *Proc. Des. Soc.* **2021**, *1*, 2197–2206. [CrossRef]
2. Khedr, M.; Li, W.; Min, N.; Abd-Elaziem, W.; Jin, X. Strengthening contributions of mechanical twinning and dislocations to the flow stress of hadfield high-manganese steel: Quantitative analysis. *J. Mater. Eng. Perform.* **2023**, *32*, 501–511. [CrossRef]
3. Kechagias, J.D.; Vidakis, N.; Petousis, M. Parameter effects and process modeling of FFF-TPU mechanical response. *Mater. Manuf. Process.* **2023**, *38*, 341–351. [CrossRef]
4. Raja, S.; Rajan, A.J.; Kumar, V.P.; Rajeswari, N.; Girija, M.; Modak, S.; Kumar, R.V.; Mammo, W.D.; Gupta, P. Selection of additive manufacturing machine using analytical hierarchy process. *Sci. Program.* **2022**, *2022*, 1596590. [CrossRef]
5. Wang, Y.; Blache, R.; Xu, X. Selection of additive manufacturing processes. *Rapid Prototyp. J.* **2017**, *23*, 434–447. [CrossRef]
6. Palanisamy, M.; Pugalendhi, A.; Ranganathan, R. Selection of suitable additive manufacturing machine and materials through best–worst method (BWM). *Int. J. Adv. Manuf. Technol.* **2020**, *107*, 2345–2362. [CrossRef]
7. Villi, Ö.; Villi, O.; Yavuz, H. Concept Design for Optimizing Mass Production Processes with 3D Printer in the Industry. *Int. J. 3D Print. Technol. Digit. Ind.* **2024**, *8*, 8–19. [CrossRef]
8. Kaill, N.; Campbell, R.I.; Pradel, P.; Bingham, G.A. A comparative study between 3-axis and 5-axis additively manufactured samples and their ability to resist compressive loading. In Proceedings of the International Solid Freeform Fabrication Symposium, Austin, TX, USA, 12–14 August 2019.
9. Kim, C.N.; Yang, K.H.; Kim, J. Human decision-making behavior and modeling effects. *Decis. Support Syst.* **2008**, *45*, 517–527. [CrossRef]
10. Lin, Z.-C.; Yang, C.-B. Evaluation of machine selection by the AHP method. *J. Mater. Process. Technol.* **1996**, *57*, 253–258. [CrossRef]
11. Chatterjee, P.; Chakraborty, S. A comparative analysis of VIKOR method and its variants. *Decis. Sci. Lett.* **2016**, *5*, 469–486. [CrossRef]
12. Badarinath, R.; Prabhu, V. Integration and evaluation of robotic fused filament fabrication system. *Addit. Manuf.* **2021**, *41*, 101951. [CrossRef]
13. What Are the Robotic Arm 3D Printing Solutions on the Market? Available online: https://www.3dnatives.com/en/robotic-arms-3d-printing-141020226/# (accessed on 13 June 2025).
14. Keating, S.; Oxman, N. Compound fabrication: A multi-functional robotic platform for digital design and fabrication. *Robot. Comput.-Integr. Manuf.* **2013**, *29*, 439–448. [CrossRef]
15. Gwet, K. *Handbook of Inter-Rater Reliability*; STATAXIS Publishing Company: Gaithersburg, MD, USA, 2001; pp. 223–246.
16. Rubinstein, R.Y.; Kroese, D.P. *Simulation and the Monte Carlo Method*; John Wiley & Sons: Hoboken, NJ, USA, 2016.

17. NEMA-34 Mount to NEMA-23 Motor Adapter Plate. Available online: https://www.thingiverse.com/thing:526424 (accessed on 3 February 2025).
18. Impeller Design 70mm—3/4/5-Blades. Available online: https://www.thingiverse.com/thing:3834122 (accessed on 3 February 2025).
19. Penciltop Wind Turbine. Available online: https://www.thingiverse.com/thing:2136314 (accessed on 3 February 2025).
20. Gardner, J.A.; Nethercott-Garabet, T.; Kaill, N.; Campbell, R.I.; Bingham, G.A.; Engstrøm, D.S.; Balc, N.O. Aligning material extrusion direction with mechanical stress via 5-axis tool paths. In Proceedings of the International Solid Freeform Fabrication Symposium, Austin, TX, USA, 13–15 August 2018.
21. Ceballos, B.; Lamata, M.T.; Pelta, D.A. A comparative analysis of multi-criteria decision-making methods. *Prog. Artif. Intell.* **2016**, *5*, 315–322. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

Machines Editorial Office
E-mail: machines@mdpi.com
www.mdpi.com/journal/machines



Disclaimer/Publisher's Note: The title and front matter of this reprint are at the discretion of the Guest Editor. The publisher is not responsible for their content or any associated concerns. The statements, opinions and data contained in all individual articles are solely those of the individual Editor and contributors and not of MDPI. MDPI disclaims responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.





Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-6086-9