computers

Special Issue Reprint

# AI in Its Ecosystem

Edited by
Behrouz Zolfaghari, Amin Ramezani, Mohsen Hadian and Firooz Saghezchi

mdpi.com/journal/computers

MDPI

# AI in Its Ecosystem

# AI in Its Ecosystem

Guest Editors

**Behrouz Zolfaghari**
**Amin Ramezani**
**Mohsen Hadian**
**Firooz Saghezchi**

MDPI

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

*Guest Editors*
Behrouz Zolfaghari
The Department of
Cybersecurity and
Information Technology
University of West Florida
Pensacola
USA

Firooz Saghezchi
Distributed Signal Processing
(DSP)
RWTH Aachen University
Aachen
Germany

Amin Ramezani
Division of Cardiac Surgery
Harvard Medical School
Boston
USA

Mohsen Hadian
Health Data Science
University Health Network
Toronto
Canada

*Editorial Office*
MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Special Issue, published open access by the journal *Computers* (ISSN 2073-431X), freely accessible at: https://www.mdpi.com/journal/computers/special_issues/1C327Q7Z14.

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 978-3-7258-8741-5 (Hbk)**
**ISBN 978-3-7258-8742-2 (PDF)**
**https://doi.org/10.3390/books978-3-7258-8742-2**

© 2026 by the authors. Articles in this reprint are Open Access and distributed under the Creative Commons Attribution (CC BY) license. The reprint 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 Editors** . . . . . **vii**

**Apostolos Glenis and George Vouros**
Scalable Univariate and Multivariate Time-Series Classifiers with Deep Learning Methods Exploiting Symbolic Representations
Reprinted from: *Computers* **2025**, *14*, 563, https://doi.org/10.3390/computers14120563 . . . . . **1**

**Indira Karymsakova, Dinara Kozhakhmetova, Dariga Bekenova, Danila Ostroukh, Roza Bekbayeva, Lazat Kydyralina, et al.**
Development of an Early Lung Cancer Diagnosis Method Based on a Neural Network
Reprinted from: *Computers* **2025**, *14*, 397, https://doi.org/10.3390/computers14090397 . . . . . **38**

**Aleksandar Milenkovic, Andjelija Djordjevic, Dragan Jankovic, Petar Rajkovic, Kofi Edee and Tatjana Gric**
Healthcare AI for Physician-Centered Decision-Making: Case Study of Applying Deep Learning to Aid Medical Professionals
Reprinted from: *Computers* **2025**, *14*, 320, https://doi.org/10.3390/computers14080320 . . . . . **59**

**Andrea La Fauci De Leo, Pooneh Bagheri Zadeh, Kiran Voderhobli and Akbar Sheikh Akbari**
A New AI Framework to Support Social-Emotional Skills and Emotion Awareness in Children with Autism Spectrum Disorder
Reprinted from: *Computers* **2025**, *14*, 292, https://doi.org/10.3390/computers14070292 . . . . . **82**

**Hassan Ayaz, Kashif Sultan, Muhammad Sheraz and Teong Chee Chuah**
From Patterns to Predictions: Spatiotemporal Mobile Traffic Forecasting Using AutoML, TimeGPT and Traditional Models
Reprinted from: *Computers* **2025**, *14*, 268, https://doi.org/10.3390/computers14070268 . . . . . **100**

**Oumeima Thâalbi and Moulay A. Akhloufi**
EMGP-Net: A Hybrid Deep Learning Architecture for Breast Cancer Gene Expression Prediction
Reprinted from: *Computers* **2025**, *14*, 253, https://doi.org/10.3390/computers14070253 . . . . . **122**

**Rui Ramos, Joaquim Casaca and Rui Patrício**
Adoption Drivers of Intelligent Virtual Assistants in Banking: Rethinking the Artificial Intelligence Banker
Reprinted from: *Computers* **2025**, *14*, 209, https://doi.org/10.3390/computers14060209 . . . . . **146**

**Duo Peng and Weiqi Yan**
Test-Time Training with Adaptive Memory for Traffic Accident Severity Prediction
Reprinted from: *Computers* **2025**, *14*, 186, https://doi.org/10.3390/computers14050186 . . . . . **166**

**Su Jin Mun and Won Ho Choi**
Artificial Intelligence in Neoplasticism: Aesthetic Evaluation and Creative Potential
Reprinted from: *Computers* **2025**, *14*, 130, https://doi.org/10.3390/computers14040130 . . . . . **186**

**Agathoklis A. Krimpenis and Loukas Athanasakos**
"Optimizing the Optimization": A Hybrid Evolutionary-Based AI Scheme for Optimal Performance
Reprinted from: *Computers* **2025**, *14*, 97, https://doi.org/10.3390/computers14030097 . . . . . **209**

**Vladimir Sonkin and Cătălin Tudose**
Beyond Snippet Assistance: A Workflow-Centric Framework for End-to-End AI-Driven Code Generation
Reprinted from: *Computers* **2025**, *14*, 94, https://doi.org/10.3390/computers14030094 . . . . . **230**

**Davide Merolla, Vittorio Latorre, Antonio Salis and Gianluca Boanelli**
Improving Road Safety with AI: Automated Detection of Signs and Surface Damage
Reprinted from: *Computers* **2025**, *14*, 91, https://doi.org/10.3390/computers14030091 . . . . . **255**

**Leon Wiese, Lennart Hinz, Eduard Reithmeier, Philippe Korn and Michael Neuhaus**
Detection of Surgical Instruments Based on Synthetic Training Data
Reprinted from: *Computers* **2025**, *14*, 69, https://doi.org/10.3390/computers14020069 . . . . . **274**

**Avisek Kundu, Seeboli Ghosh Kundu, Santosh Kumar Sahu and Nitesh Dhar Badgayan**
Leveraging Azure Automated Machine Learning and CatBoost Gradient Boosting Algorithm for Service Quality Prediction in Hospitality
Reprinted from: *Computers* **2025**, *14*, 32, https://doi.org/10.3390/computers14020032 . . . . . **290**

**Yiannis Kiouvrekis and Theodor Panagiotakopoulos**
Electromagnetic Field Distribution Mapping: A Taxonomy and Comprehensive Review of Computational and Machine Learning Methods
Reprinted from: *Computers* **2025**, *14*, 373, https://doi.org/10.3390/computers14090373 . . . . . **304**

**Dorothea S. Adamantiadou and Loukas Tsironis**
Leveraging Artificial Intelligence in Project Management: A Systematic Review of Applications, Challenges, and Future Directions
Reprinted from: *Computers* **2025**, *14*, 66, https://doi.org/10.3390/computers14020066 . . . . . **336**

**Aggeliki Fanarioti and Kostas Karpouzis**
Artificial Intelligence and the Future of Mental Health in a Digitally Transformed World
Reprinted from: *Computers* **2025**, *14*, 259, https://doi.org/10.3390/computers14070259 . . . . . **360**

# About the Editors

**Behrouz Zolfaghari**

Behrouz Zolfaghari received his Ph.D. in Computer Engineering from AmirKabir University of Technology, Tehran, Iran. He has completed three postdocs in the Indian University of Technology Guwahati, the University of Guelph (Ontario, Canada), and Waseda University (Tokyo, Japan). He has worked as an assistant professor at Osaka University (Japan) and as a visiting assistant professor at Miami University (Ohio) before joining the University of West Florida. His research interests include high-performance computing and cybersecurity. Behrouz Zolfaghari has published over 70 papers and books. BookAuthority has ranked his book on Physical Uncloneable Functions (PUFs) as the best book of all time in the area of cryptographic hardware. He has won the ScienceFather Best Researcher Award for his work in cryptographic hardware.

**Amin Ramezani**

Amin Ramezani received his B.Sc. degree in Electronic Engineering from Shahid Beheshti University, Tehran, Iran, in 2001, his M.Sc. degree in Electrical Engineering from Sharif University of Technology in 2003, and his Ph.D. degree in Control Engineering from the University of Tehran in 2011. He subsequently served as an Assistant Professor at Tarbiat Modares University, where his research focused on process control and automation, advanced instrumentation, fault-tolerant control systems, discrete event systems, tele-operation, and intelligent transportation systems. Throughout his academic career, he has authored and co-authored numerous scientific publications and has been a member of the IEEE since 2006. Dr. Ramezani currently serves as a Research Specialist at Harvard Medical School and Brigham and Women's Hospital, where he develops artificial intelligence and digital health technologies for cardiovascular medicine and critical care. His current research integrates machine learning, large language models, physiological digital twins, biomedical signal processing, and predictive analytics to create next-generation clinical decision-support systems. His work bridges engineering, artificial intelligence, and medicine, with a particular focus on translating advanced computational methods into practical tools that improve patient outcomes and healthcare delivery.

**Mohsen Hadian**

Mohsen Hadian is a leading researcher in artificial intelligence and machine learning, with a deep commitment to advancing these technologies in both biomedical and ecological contexts. Currently a Postdoctoral Fellow at the University of Toronto and University Health Network, Mohsen Hadian specializes in developing AI-driven solutions for complex systems, including neuroscience and cognitive systems. He earned dual Ph.D. degrees in Biomedical Engineering and Electrical Engineering, with research integrating AI techniques such as deep learning, reinforcement learning, and neural network modeling. Mohsen Hadian's work spans a broad range of AI applications, from designing predictive models for healthcare to implementing machine learning algorithms for environmental sustainability. His expertise includes natural language processing, computer vision, and AI-based optimization, enabling innovative approaches to ecosystem management and medical diagnostics. He has contributed significantly to robust control systems, AI in healthcare, and process automation, with his research cited extensively in international journals. As a sessional faculty member at the University Canada West, Mohsen Hadian teaches courses in business analytics and machine learning, contributing to the development of the next generation of AI professionals.

**Firooz Saghezchi**

Firooz Saghezchi is a Lead Scientist at the Chair for Distributed Signal Processing (DSP), RWTH Aachen University, Germany, where he contributes to the 6GEM+ Research Transfer Hub and conducts research on value-oriented 6G communication systems for the connected digital industry. His research interests lie at the intersection of AI and wireless communications, with a particular focus on AI-native wireless systems, MIMO-OFDM signal processing, radio resource management, and energy-efficient wireless transceiver design. His current work focuses on the development of programmable and AI-enabled 6G communication platforms, including FPGA-accelerated wireless PHY processing and intelligent radio access technologies for industrial automation, mobility, and immersive communication applications. Prior to joining RWTH Aachen University in 2023, he worked at Instituto de Telecomunicações in Aveiro, Portugal, as a Senior Researcher (2016–2023) and Research Assistant (2010–2016), contributing to several national and European research projects, including H2020-ECSEL DAIS (Distributed Artificial Intelligence Systems), H2020-ICT 5GENESIS (5th Generation End-to-End Network, Experimentation, System Integration, and Showcasing), H2020-ECSEL SEMI40 (Power Semiconductor and Electronics Manufacturing 4.0), EUREKA-CATRENE MOBITRUST (Secure Communications for Next-Generation Public Protection and Disaster Relief Systems), ENIAC-JU E2SG (Energy to Smart Grid), and FP7-ICT C2POWER (Cognitive Radio and Cooperative Strategies for Power Saving in Multi-Standard Wireless Devices). He received his Ph.D. in Electrical Engineering (Telecommunications) jointly from the Universities of Aveiro, Porto, and Minho, Portugal, in 2016. He is an IEEE Senior Member and has authored more than 60 peer-reviewed journal and conference publications, one book, and several book chapters. He received the Best Paper Award from the MDPI *Electronics* journal in 2022. He has served as an Associate Editor of *IEEE Access* since 2020 and has contributed to the technical organization of numerous international IEEE conferences, including IEEE GLOBECOM, ICC, VTC, WCNC, and CAMAD, as a Technical Program Committee member and Special Session Chair.

computers

MDPI

Article

# Scalable Univariate and Multivariate Time-Series Classifiers with Deep Learning Methods Exploiting Symbolic Representations

**Apostolos Glenis * and George Vouros ***

Department of Digital Systems, University of Piraeus, 185 34 Piraeus, Greece
* Correspondence: apostglen46@gmail.com (A.G.); georgev@unipi.gr (G.V.)

**Abstract:** Time-series classification (TSC) is an important task across sciences. Symbolic representations (especially SFA) are very effective at combating noise. In this paper, we employ symbolic representations to create state-of-the-art time-series classifiers, with the aim to advance scalability without sacrificing accuracy. First, we create a graph representation of the time series based on SFA words. We use this representation together with graph kernels and an SVM classifier to create a scalable time-series classifier. Next, we use the graph representation together with a Graph Convolutional Neural Network to test how it fares against state-of-the-art time-series classifiers. Additionally, we devised deep neural networks exploiting the SFA representation, inspired by the text classification domain, to study how they fare against state-of-the-art classifiers. The proposed deep learning classifiers have been adapted and evaluated for the multivariate time-series case and also against state-of-the-art time-series classification algorithms based on symbolic representations.

**Keywords:** time-series classification; symbolic representations; scalability

## 1. Introduction

Time-series classification (TSC) is an important task across many branches of science. Indicative applications of TSC are seizure detection [1], earthquake monitoring [2], insect classification [3], applications in power systems [4], and other applications [5–13]. Symbolic representations (especially SFA) [14] have been proven to be very effective in combating noise. To this end, this paper aims to incorporate symbolic representations to create state-of-the-art time-series classification algorithms, aiming to advance their scalability without sacrificing accuracy. More specifically, this paper aims to answer the following research questions (RQs):

- RQ1. How would a classifier built on top of graph representations of SFA words perform in terms of accuracy and execution time? Our first attempt to answer this question is through the use of graph kernels, which are more scalable than Graph Neural Networks, with an SVM classifier on top.
- RQ2. Closely linked with RQ1, we take the graph representation with SFA words and combine it with Graph Convolutional Neural Networks to see how the resulting classifier would fare against the state of the art.
- RQ3. For this research question, we aim to answer whether SFA, together with state-of-the-art deep learning methods adapted from the text classification domain,

can also provide state-of-the-art accuracy and execution times for multivariate time-series classification.

- RQ4. Finally, we aim to answer whether SCALE-BOSS-MR [15], a state-of-the-art symbolic time-series classifier, can be adapted to the multivariate use case and whether it provides state-of-the-art accuracy and execution time.

To answer the above research questions, this paper makes the following contributions:

1. We use a graph representation together with a symbolic representation to represent time-series as a graph. First, we use this representation together with graph kernels and an SVM classifier to create SCALE-BOSS-GRAPH.
2. We use the graph representation of time-series in conjunction with a Graph Convolutional Neural Network to see whether it can attain state-of-the-art accuracy and execution time.
3. We adapt state-of-the-art neural network architectures from the text classification domain, and we use them in conjunction with symbolic representation to create state-of-the-art deep learning symbolic time-series classifiers. We also adapt the proposed deep learning methods to the multivariate use case and compare them against state-of-the-art deep learning methods.
4. We adapt SCALE-BOSS-MR to the multivariate use case, and we compared the adapted version SCALE-BOSS-MR-MV to state-of-the-art time-series classifiers.

## 2. Related Work

### *2.1. Methods Using Symbolic Representations*

Symbolic representations have been used widely in the time-series literature for dimensionality reduction, smoothing, and combating noise. Most of the recent TSC literature has used SFA [14] or SFA in conjunction with SAX [16]. SFA is a state-of-the-art symbolic representation that uses Fourier coefficients to represent the time-series. SFA has been proven to be very efficient in combating noise.

In [17], the authors present SAX-VSM. SAX-VSM uses SAX as its symbolic representation. SAX-VSM pairs SAX representations with Term Frequency–Inverse Document Frequency (TF-IDF) weighting. After the training set is transformed into SAX words, a tf-idf vector is created for each class as a whole. In the next step, the set of targets time-series is transformed into SAX words, and the Term-Frequency (TF) vector is created for each time-series in the target set. To classify a time-series, the cosine similarity between the TF vectors of that series and of the classes' centers is computed.

BOSS [18] is the first time-series classification algorithm to use SFA. However, the high computational load of BOSS made it impractical for many applications. To this end, in [19], the authors present BOSS-VS, a time-series classification algorithm that uses SFA but is geared towards scalability.

In [20], the authors present WEASEL, a middle ground between BOSS and BOSS-VS in terms of computational complexity and accuracy. WEASEL uses multiple window sizes and bigrams with a chi-squared feature extraction pipeline and a linear classifier on top.

In [21], the authors present K-BOSS-VS, a time-series classifier that is a middle ground between BOSS and BOSS-VS in terms of computational complexity and accuracy. The main idea of K-BOSS-VS is that, instead of having a single representative for each class as in BOSS-VS, the K-means clustering algorithm is applied to pick K representatives for each class.

In [22,23], the authors present MRSQM, a framework that uses multiple symbolic representations for scalable and accurate TSC. MRSQM uses both SFA and SAX representations and efficient sequence mining to extract important time-series features.

In [24], the authors present BOSS-SCALE, which generalizes the process of K-BOSS-VS to use a pluggable classifier on top of the SFA Term-Frequency vectors.

In [25], the authors present WEASEL 2.0, an updated version of WEASEL that incorporates first-order differences and dilation together with randomization to provide a state-of-the-art scalable classification algorithm.

In [15], the authors extend SCALE-BOSS into SCALE-BOSS-MR. SCALE-BOSS-MR uses multiple window sizes, first-order differences, and dilation. On top of these representations, SCALE-BOSS-MR uses a pluggable classifier like SCALE-BOSS. Instead of using randomization to achieve scalability, SCALE-BOSS-MR uses carefully crafted optimization parameters that aim to balance between accuracy and scalability.

SCALE-BOSS-MR proved to be significantly faster than both MRSQM and WEASEL 2.0 with minimal loss in accuracy.

In summary, SAX-VSM and BOSS-VS are aimed towards scalability. BOSS aims only at high accuracy and it is not scalable. WEASEL is proposed as a middle ground between BOSS and BOSS-VS, achieving high accuracy and relative scalability. SCALE-BOSS achieves great accuracy with execution time only marginally worse than BOSS-VS. WEASEL 2.0 provides very high accuracy with execution time better than WEASEL thanks to randomization. Finally, SCALE-BOSS-MR achieves state-of-the-art accuracy while retaining state-of-the-art execution time.

There are not many works proposing symbolic classifiers for multivariate time-series. In [26] the authors present WEASEL+MUSE, which extends WEASEL to the multivariate use case while incorporating first-order differences.

### *2.2. Methods Using Convolutional Kernels*

In [27] the authors present ROCKET, a scalable and highly accurate TSC classifier. ROCKET uses a large number (10000) of Convolutional kernels paired with a linear classifier.

In [28] the authors present a significantly faster version of ROCKET that retains state-of-the-art accuracy called MiniRocket. It manages to do so by using a significantly lower number of Convolutional kernels.

### *2.3. Methods Using Deep Learning*

In [29] the authors present the MLP, FCN and ResNet architectures. The MLP architecture uses three fully connected layers, each with 500 neurons. The architecture uses dropout [30] to prevent overfitting and the ReLU activation function [31] to prevent saturation of the gradient. Regarding FCN, the basic blocks of the architecture are a convolutional layer followed by Batch Normalization [32]. The architecture stacks three such blocks, followed by a global average pooling layer and a softmax layer. The authors adapt the ResNet architecture [33] for time-series classification. The proposed architecture uses three residual blocks with filters of size $k = \{64, 128, 128\}$. The three residual blocks are stacked and followed by a global average pooling layer and a softmax layer.

In [34] the authors present a Graph Convolutional Neural Network (GCN) for graph classification. Their architecture stacks two Graph Convolution layers followed by a Mean Pooling operation, two fully connected layers, and finally a SoftMax layer.

In [35] the authors present CNNClassifier. The architecture comprises two 1-D convolutional layers followed by a fully connected layer.

The authors in [36] aim to build a model that is conceptually straightforward and computationally lightweight. They propose the EncoderClassifier. The model consists of convolutional layers, an attention layer to summarize the time axis, and finally a fully connected layer to obtain the desired dimensionality. The convolutional network consists

of three convolutional blocks with two factor max pooling layers between them. Each convolutional block comprises 1-D convolutions followed by instance normalization, PRelu, and a dropout layer. After the convolutional layer the results are passed to an attention layer. The results of the attention layer are passed through a fully connected layer and an instance normalization layer.

The InceptionTime classifier proposed in [37] is an ensemble of deep Convolutional Neural Network (CNN) models, inspired by the Inception-v4 architecture. InceptionTime proposes the use of BottleNeck and Multiplex Convolutions. The BottleNeck operation consists of 1-D Convolutions with unit kernel size. The main idea of multiplex Convolutions is to learn in parallel different Convolution layers of different kernel sizes. The composition of an Inception network classifier contains two different residual blocks, compared to ResNet, which comprises three. For the Inception network, each block is composed of three Inception modules rather than traditional fully convolutional layers. Each residual block's input is transferred via a shortcut linear connection to be added to the next block's input, thus mitigating the vanishing gradient problem by allowing a direct flow of the gradient.

In [38] the authors create hand-crafted filters in order to detect certain patterns in time-series. Those filters are fixed and do not change during training. The authors modify existing architectures (FCN and Inception time) and through their evaluation show that incorporating the fixed filters improves accuracy. More specifically, the authors propose three types of filters: (a) increasing trend, (b) decreasing trend, and (c) peak detection.

In [39] the authors present the LITE classifier. LITE uses 9814 training parameters while achieving state-of-the-art performance. LITE uses DepthWiseSeperableConvolutions, which are boosted by three techniques:(a) multiplexing convolutions [37,40], (b) dilated convolutions [41], and (c) custom filters [38].

### *2.4. Multivariate Time Series Classification*

In [42] the authors present two methods of Multivariate Time Series classification (MTS) representation using SAX. In the first one, multivariate words are generated by combining the symbols of each attribute at a particular time, and a bag-of-patterns approach is used to classify MTS. This way, the method considers the correlation between the time-series by combining individual representations of attributes. However, this approach significantly increases the size of the individual word for a large number of attributes. This potentially limits the effectiveness in terms of accuracy of the method since longer words could potentially carry less information because of the curse of dimensionality. The authors also propose a stacked bag-of-patterns approach. This approach concatenates the representations of multiple univariate time-series into one by stacking the term frequency vectors for each attribute column-wise.

In [43] the authors propose a method that uses a symbolic representation to perform multivariate time-series (MTS) classification. The authors propose a new symbolic representation for MTS denoted as SMTS. SMTS considers all attributes of MTS simultaneously, rather than separately, to extract information contained in relationships. An elementary representation is used that consists of the time index and the values (as well as first-order differences) of the individual time-series as columns. Each MTS is concatenated vertically and each instance is labeled with the class of the corresponding time-series. A tree learner is applied to the simple representation and each instance is assigned to a terminal node of the tree. This means that a symbol is associated for each time index (instance), instead of intervals of the time-series. With R terminal nodes, the dictionary contains R symbols. After each instance has been associated with a symbol from each tree of the ensemble, the

term frequency vector for each symbol is computed and the term frequency vectors are stacked column-wise as in the Stacked-Bag-of-Patterns in [42].

In [26] the authors extend WEASEL to work with multivariate datasets. The authors use the WEASEL pipeline for each feature separately and then they create the final term frequency vector by stacking the separate term frequency vectors column-wise. The WEASEL-MUSE also uses first-order differences to enhance accuracy.

In order to adapt ROCKET [27] to multivariate datasets, kernels are randomly assigned dimensions and then weights are generated for each channel. This way, ROCKET retains scalability in the multivariate use case.

The interested reader can read a detailed survey of multivariate time-series classification in [44].

## 3. Graph Kernel Preliminaries

The idea behind graph kernels is to create a kernel matrix that can be used with off-the-shelf machine learning models to make predictions. The kernel matrix measures the similarity between all pairs of graphs. In order for the kernel matrix to be valid, the matrix has to be symmetric with positive eigenvalues.

The key idea of kernel methods is to design a feature representation of a given graph as a bag-of-words. In this case, we need a feature representation $\phi(.)$ such that $K(G1, G2) = (\phi(G1))^{\intercal} \cdot \phi(G2)$, where $K$ is the kernel matrix, $G1$ is the first graph and $G2$ is the second graph. The naive representation is to use words as nodes. The problem with this approach is that we can have the same representation for very different graphs. This happens because this approach does not consider connectivity between the nodes of each graph.

To address this issue, we are using the Weisfeiler–Lehman (WL) graph kernel. The main idea of the WL graph kernel [45] is to have an efficient feature descriptor $f(G)$ by exploiting the neighborhood structure of the graph to iteratively enrich the node vocabulary, i.e., the set of possible labels for the graph nodes. The WL is a generalized version of node degrees because node degrees provide 1-hop information, whereas WL generalizes this to the k-hop. The algorithm to achieve this is called WL isomorphism test or color refinement.

*Color Refinement*

This section describes the color refinement procedure behind the WL kernel.

Given a graph $G$ with nodes $V$, the process assigns an initial color $c^{(0)}(v)$ to each node $v$. Then, it iteratively refines colors as follows: $c^{(k+1)}(v) = HASH(c^{(k)}(v), c^{(k)}(u))$, with $u \in neib(v)$ , and $HASH$ maps different inputs (node labels) to different colors. After $k$ steps of color refinement, it summarizes the structure of the K-Hop neighborhood.

Figures 1–4 provide an example of the application of the WL graph kernel for two graphs.

Initially, all nodes of the two graphs get a color (label) of 1. Next, for each node it computes the next label. The next label for each node is created by merging the colors of its neighbors together with its previous color. For example, the top left node of $G1$ gets a label of (1,111), whereas the top left node of $G2$ gets a label of (1,11).

After it has created the new color for each node, the processes pass it through the hash function. For example, according to the hash table, (1,111) gets a new color 4, while (1,11) gets a new color 3.

Figure 3 shows the second iteration of the color refinement procedure.

Figure 4 provides the feature representation of the two graphs, by counting the total number of color occurrences in each graph.

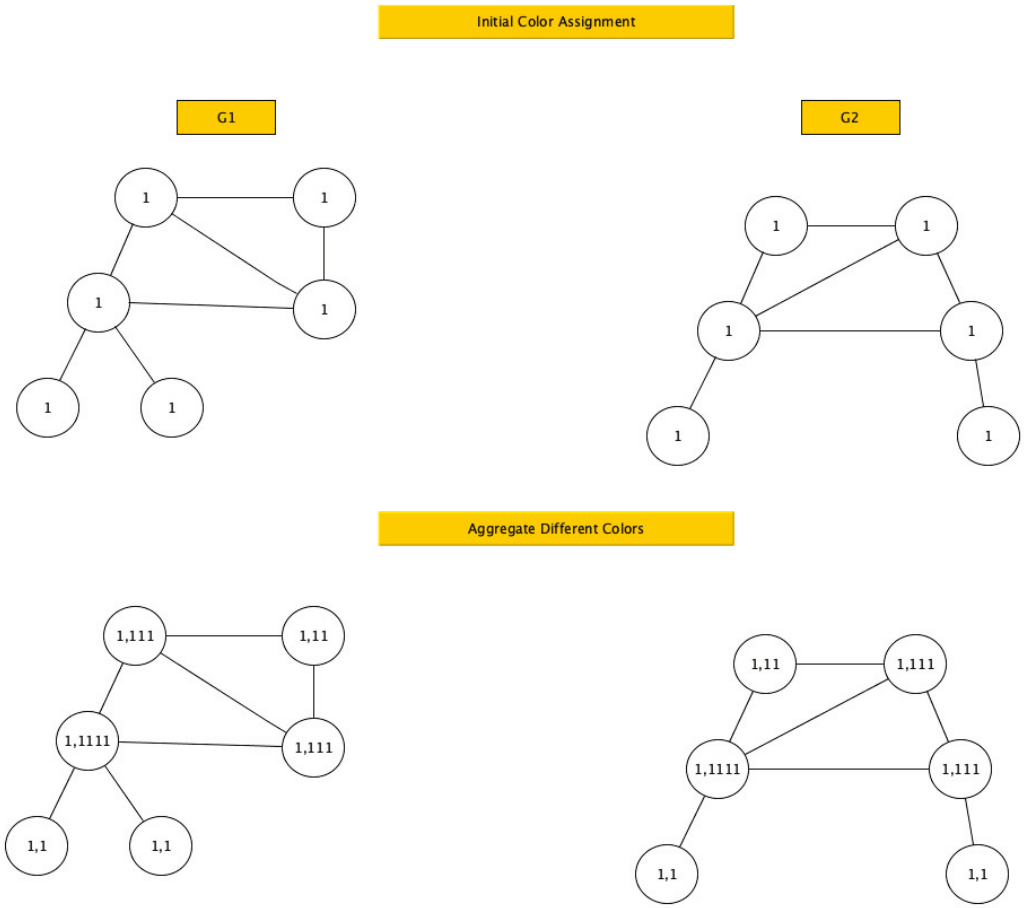



**Figure 1.** Iteration 0 of color refinement.

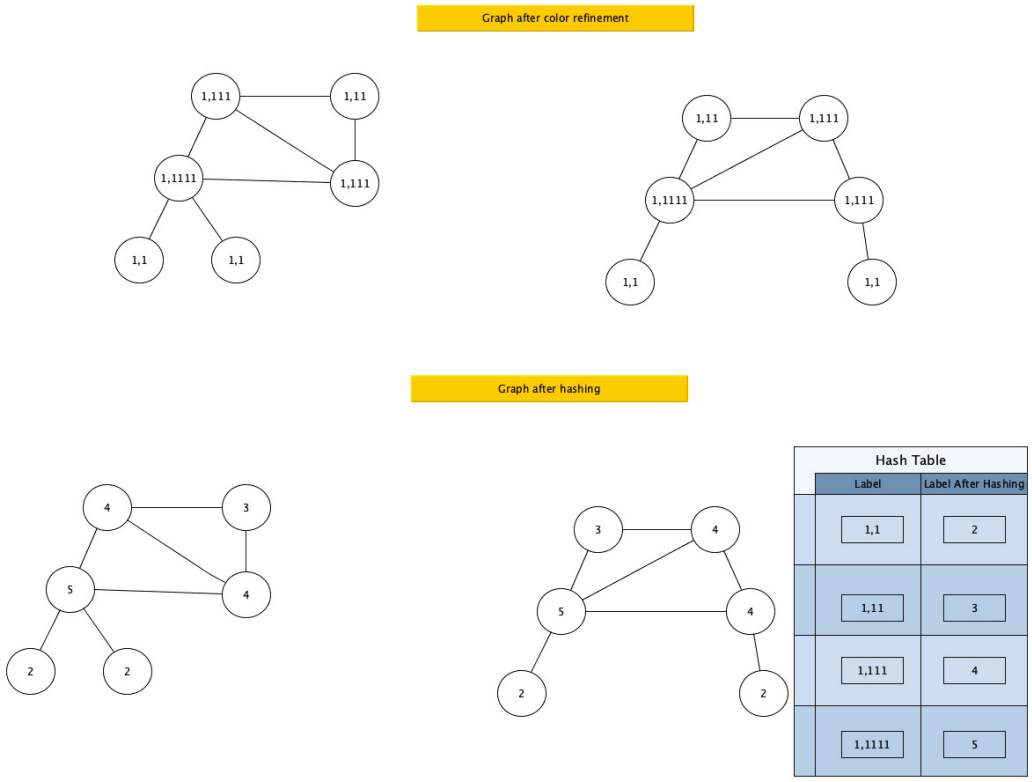



**Figure 2.** Hashing after the first iteration of color refinement.

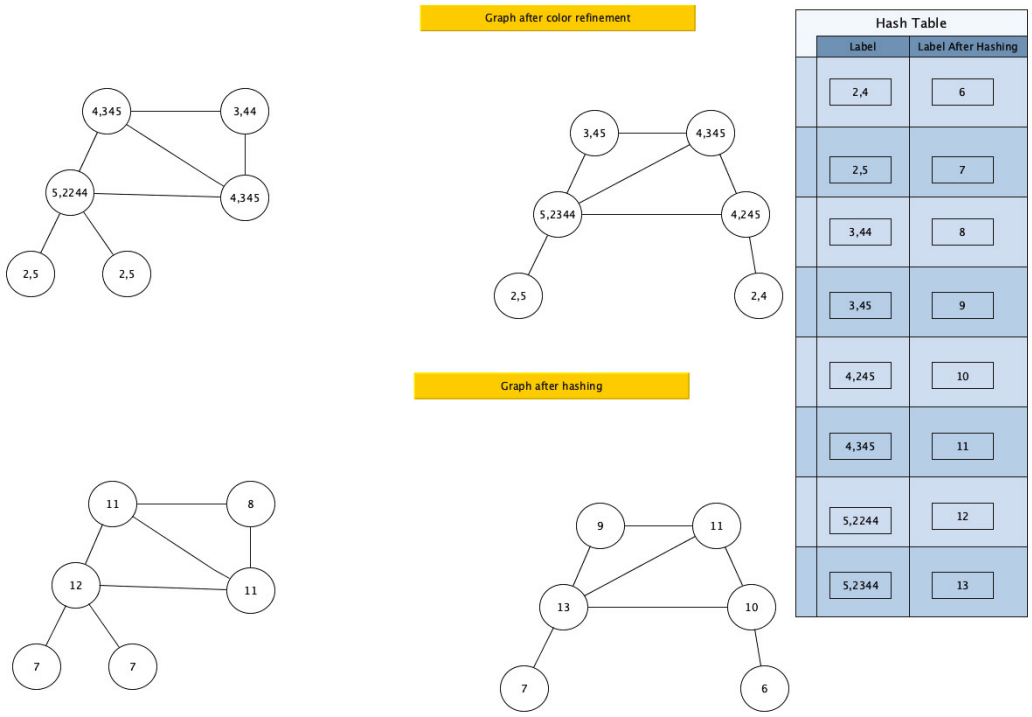



**Figure 3.** Hashing after the second iteration of color refinement.

Color Counts Table

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G1 | 6 | 2 | 1 | 2 | 1 | 0 | 2 | 1 | 0 | 0 | 0 | 2 | 1 |
| G2 | 6 | 2 | 1 | 2 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 |

$$K(G1,G2)=\mathrm{dot}(\mathrm{transpose}(\varphi(G1)),\varphi(G2)) = 49$$

**Figure 4.** Vector representation of the two graphs.

## 4. Deep Learning Preliminaries

### *4.1. Convolutional Neural Networks*

The original purpose of Convolutional Neural Networks (CNNs) [46,47] is to detect certain features in an image. The early layers of a CNN detect local visual features, and later layers of a CNN take those features and combine them to produce higher-level features.

In order for a CNN layer to detect features, the main idea is to pass a filter across the image (or more generally, a tensor) with a process called convolution.

The convolution operation retains spatial information and essentially performs a point-wise multiplication of the patch of an image with the filter followed by an addition of the point-wise multiplication results to extract the final value.

Usually, convolutional layers are followed by a pooling operation for dimensionality reduction. Pooling is an operation of applying a small filter that returns either the maximum value (max pooling) or the average (average pooling) in the specific window of the tensor.

### *4.2. Recurrent Neural Networks*

RNNs [48–50] are created to handle sequence information. The RNN has a single hidden layer, but it also has a memory buffer that stores the output of the hidden layer and feeds it back into the layer along with the next element from the input sequence. This means that the RNN processes each input of the sequence together with context information from the previous inputs. One problem with RNNs is that, while they can process a sequence of inputs, they have the problem of vanishing gradients. This is because training an RNN to process a sequence of inputs requires the error to be backpropagated through the entire length of the RNN.

### *4.3. Long Short Term Memory*

In [51] the authors present the Long Short Term Memory (LSTM) cells that aim to solve the vanishing gradient problem.

In LSTMs, the propagation of the activations within the cell through time is controlled by three components called gates: the forget gate, the input gate, and the output gate. The forget gate determines which activations in the cell should be forgotten at each time step; the input gate decides how the activations in the cell should be updated in response to the new input. Finally, the output gate controls what activations should be used to generate the output in response to the current input.

The forget gate takes the concatenation of the input and the hidden state and passes this vector through a layer of neurons that uses a sigmoid activation function. The output of this forget layer is a vector of values in the range 0 to 1. The cell state is then multiplied by this forget vector. The result of this multiplication is that activations in the cell state that are multiplied by components in the forget vector with values near 0 are forgotten, and activations that are multiplied by components with values near to 1 are remembered. Next, the input gate decides what information should be added to the cell state.

The procedure works as follows. The gate decides which elements in the cell state should be updated and what information should be included in the update. The concatenated input $x_t$ plus hidden state $h_{t-1}$ is passed through a layer of sigmoid units to generate a vector of elements, the same width as the cell, where each element in the vector is in the range 0 (no update) to 1 (to be updated). At the same time that the filter vector is generated, the concatenated input and hidden state are also passed through a layer of *tanh* units. The final stage of processing in an LSTM is to decide which elements of the cell should be output in response to the current input.

A candidate output vector is generated by passing the cell through a *tanh* layer. At the same time, the concatenated input and propagated hidden state vector are passed through a layer of sigmoid units to create another filter vector. The actual output vector is then calculated by multiplying the candidate output vector by this filter vector. An RNN with LSTM cells is called an LSTM network.

### *4.4. Attention*

In [52] the authors present the attention mechanism to address the bottleneck problem that arises with the use of a fixed-length encoding vector, where the decoder would have limited access to the information provided by the input. This is thought to become especially problematic for long sequences.

The main idea of the attention mechanism is to have a context vector per time step. The intuition behind the context vector is that now the context vector attends to the relevant part of the input sequence.

The attention mechanism is composed of three parts:

1. Alignment scores: The alignment scores $e$ indicate how well the elements of the input sequence align with the current output. The alignment scores are computed by $e = tanh(dot(x, W) + b)$, where $W$ is the weight matrix of the attention layer, $b$ is the bias of the attention layer and $x$ is the input of the attention layer.
2. Weights: The weights $a$ are computed by passing a softmax to $e$. That is, $a = softmax(e)$
3. Context Vector: The context vector $c$ is computed by the dot product of $x$ and $a$. That is, $c = dot(x, a)$

## 5. Proposed Methods

### *5.1. SCALE-BOSS-GRAPH*

The SCALE-BOSS-GRAPH framework fuses graph and symbolic representations.

The idea of the graph-based symbolic representation comes from the input of BOSS, BOSS-VS and K-BOSS-VS. More specifically, after each time-series is encoded in SFA words, each successive word is connected to the one after it to create a graph. Figure 5 shows examples of graph-based symbolic representations. For example, for the first time-series in the figure, the first word "acbb" is connected with the second word "acbb", which in turn is connected with the third word "acbc" and so on. Each time-series in the dataset is encoded as a graph and the nodes of the graph are labeled using the SFA representation.

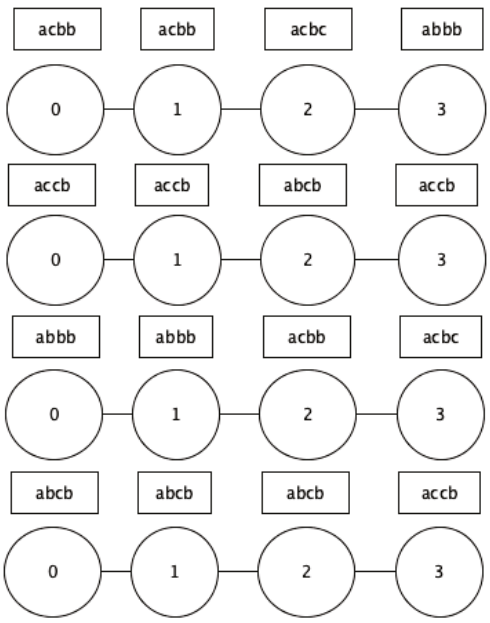



**Figure 5.** Graph-based symbolic representation.

The workflow of SCALE-BOSS-GRAPH is as follows:

1. In the first step the training set is converted to its symbolic representation. In our instantiation we have chosen SFA but the choice of the symbolic representation is orthogonal to the framework.
2. Next, we create the graph-based symbolic representation for the training set.

3. In the third step, we create the Kernel for the training set. For this instantiation, we have chosen the WL graph kernels [45] with Vertex Histogram kernel [53] as the base kernel. The choice of the kernel is orthogonal to the method. We have chosen WL with Vertex Histogram as a base kernel for scalability reasons. More concretely WL proved to be amongst the fastest graph kernels with very good accuracy results. We experimented with other graph kernels, but most were orders of magnitude slower than WL with no gains in terms of accuracy. In addition to this, as already pointed out, the WL kernel is a generalization of nodes' degrees since these provide one-hop neighborhood information, whereas WL kernel generalized this to K-hop information.
4. Then, we convert the test set to its symbolic representation, and the Graph-based symbolic representation.
5. Then, we create the kernel representation for the test set.
6. The SVC classifier is trained on the precomputed kernel for the training set.

We also used the graph-based symbolic representation as an input to a CNN [34].

### *5.2. Proposed Networks for the Univariate Case*

In this section we describe the neural network architectures that have been studied to exploit the symbolic representation of univariate time-series. In all of the proposed neural network architectures, the input is constructed as follows:

1. First, the raw time-series is converted to the SFA representation.
2. Then each of the resulting SFA words is converted to integers according to the following scheme.
   (a) The most common SFA word is assigned the integer 2.
   (b) The second most common SFA word is assigned the integer 3 and so on for all the next in frequency SFA words.
   (c) The integer 0 is used for padding and the integer 1 is reserved for the Out-of-Vocabulary token.

Figure 6 depicts the architecture of the LSTM-CNN neural network. The first layer is an embedding layer that takes the sequence of integers and creates embeddings. Embedding vectors have a length of 32. Then, the embedding layer is fed into a CNN followed by a MaxPooling layer. The CNN has a filter size 32 and kernel size of 3. After that, the results of the MaxPooling layer are fed into an LSTM layer. The LSTM is followed by a batch normalization layer that aims to help the network train faster and achieve better accuracy. Next, we have a Dropout layer to prevent overfitting. Finally, a fully connected layer does the classification.

Different CNN architectures could be used if we would like to make the proposed architecture deeper. However, keeping the architecture relatively shallow has positive effects on execution time. Also, according to our tests, making the architecture deeper did not necessarily help in terms of accuracy.

In our evaluation section, we compare our proposed architectures with ResNet and it appears to be significantly faster and more accurate.

Figure 7 shows the architecture of a network that combines Bidirectional LSTM with a self-attention layer. A Bidirectional LSTM adds a second LSTM layer that processes the data in reverse (right-to-left). Thus, Bidirectional LSTM provides the model with a complete understanding of the context surrounding a given point in the sequence. The first layer is an embedding layer that takes the sequence of integers and creates embeddings. The embeddings have a vector length of 32. The embeddings are fed into an LSTM layer that is followed by an attention layer, which is followed by a batch normalization layer and a dropout layer. Finally, a Fully Connected layer does the classification.

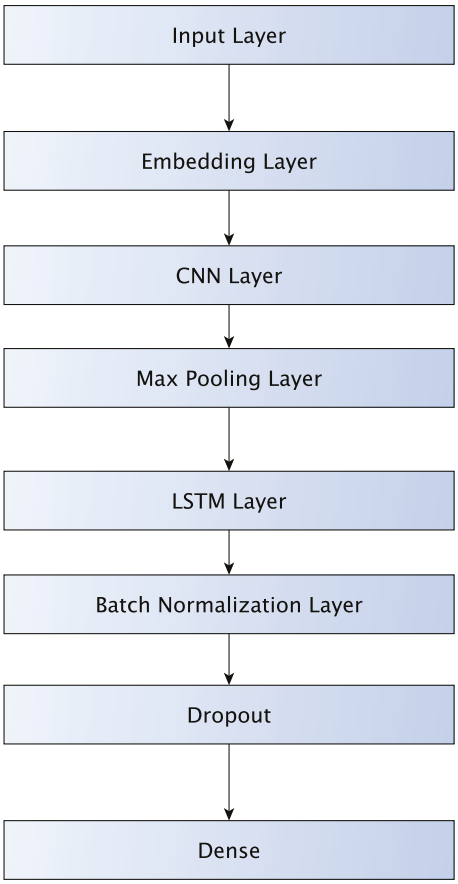



**Figure 6.** LSTM-CNN network.

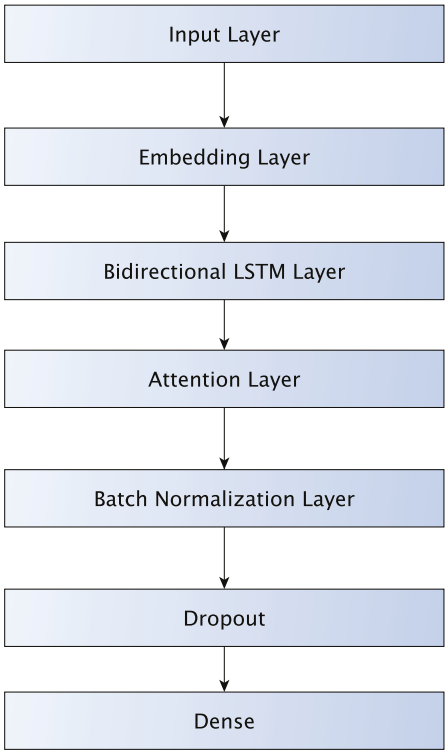



**Figure 7.** LSTM-Attention network.

Figure 8 shows a neural network architecture that augments the LSTM-Attention network with a CNN layer. The CNN layer is placed before the Bidirectional LSTM layer and is followed by a MaxPooling layer. The CNN has a filter size of 32 and a kernel size of 3. The purpose of the CNN layer is to improve accuracy, but also to significantly reduce execution time. The CNN layer performs dimensionality reduction (via Max Pooling) and thus reduces the complexity of the step after it.

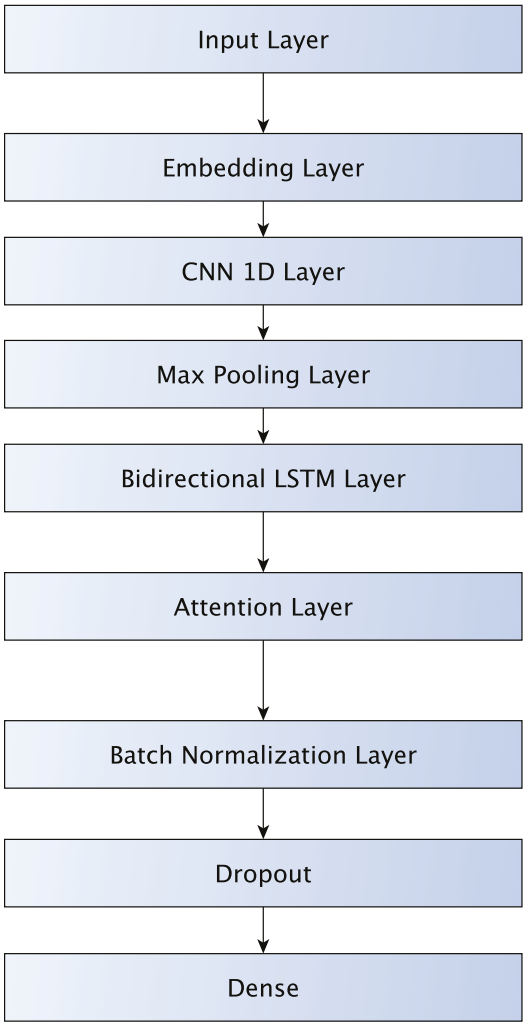



**Figure 8.** LSTM-Attention-CNN network.

Finally, in Figure 9 we see the LSTM-Attention-CNN2 network that augments the LSTM-Attention-CNN network with an additional CNN layer followed by a MaxPooling layer. The first CNN has a filter size of 32 while the second CNN has a filter size of 16, and both have a kernel size of 3. The extra CNN Layer aims to reduce execution time.

The algorithms proposed were implemented on top of pyts [54], GraKeL [55], scikit-learn [56], Keras [57] and StellarGraph [58].

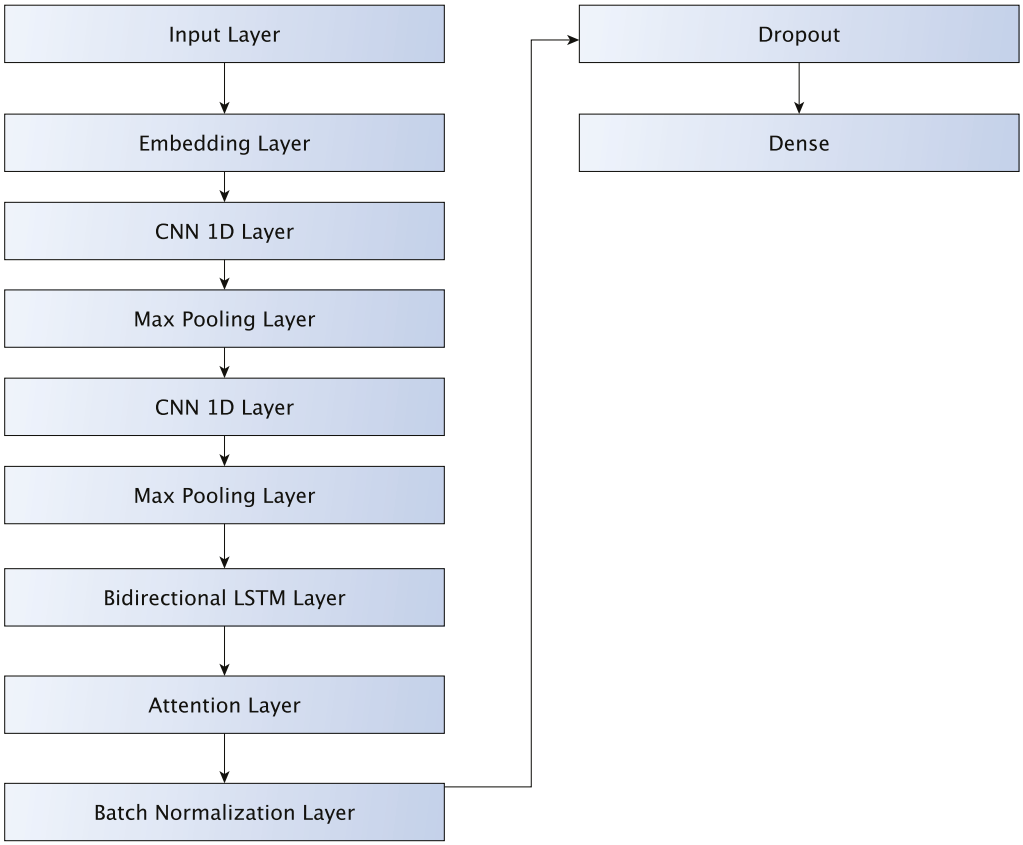



**Figure 9.** LSTM-Attention-CNN2 network.

### *5.3. Merging Strategies for the Multivariate Use Case*

In the case of multivariate time-series, there are two strategies to handle the multiple attributes: The BAG and the STACK strategies that are used in both the proposed Deep Learning methods and the proposed extension of SCALE-BOSS-MR (SCALE-BOSS-MR-MV or SBMR-MV). More concretely, suppose we have a time-series with two attributes Attr0 and Attr1. Attr0 contains the sequence "aaaa bbbb" and Attr1 contains the sequence "cccc dddd". In the BAG case we create the sequence "aaaa bbbb cccc ddd" and then we create the Term-Frequency vector (in the SBMR-MV case) or we pass it into the embedding layer (in the Deep Learning algorithm case.) In the STACK case we create a term frequency vector (or a vectorized representation by an Embedding Layer) for Attr0 and then a term frequency vector (or an Embedding Layer output) for Attr1. Then we concatenate the two vectorized representations horizontally (column-wise) to obtain the final vectorized representation.

### *5.4. Proposed Multivariate Neural Networks*

In the BAG case, each attribute of a multivariate time-series is transformed into its symbolic representation, and the symbolic representations are then fused. Then these are passed on to an embedding layer. In the BAG case, the relationships between attributes are not retained.

From then on the process is the same as for the univariate time-series.

Figure 10 shows the workflow for LSTM-Attention-CNN2 with the BAG workflow.

In the STACK case, each attribute is transformed into a symbolic representation, which is fed into an embedding layer. Then, the embeddings are concatenated to form the final embedding of the input sequence.

In Figure 11 we can see the workflow for LSTM-Attention-CNN2 with the STACK workflow.

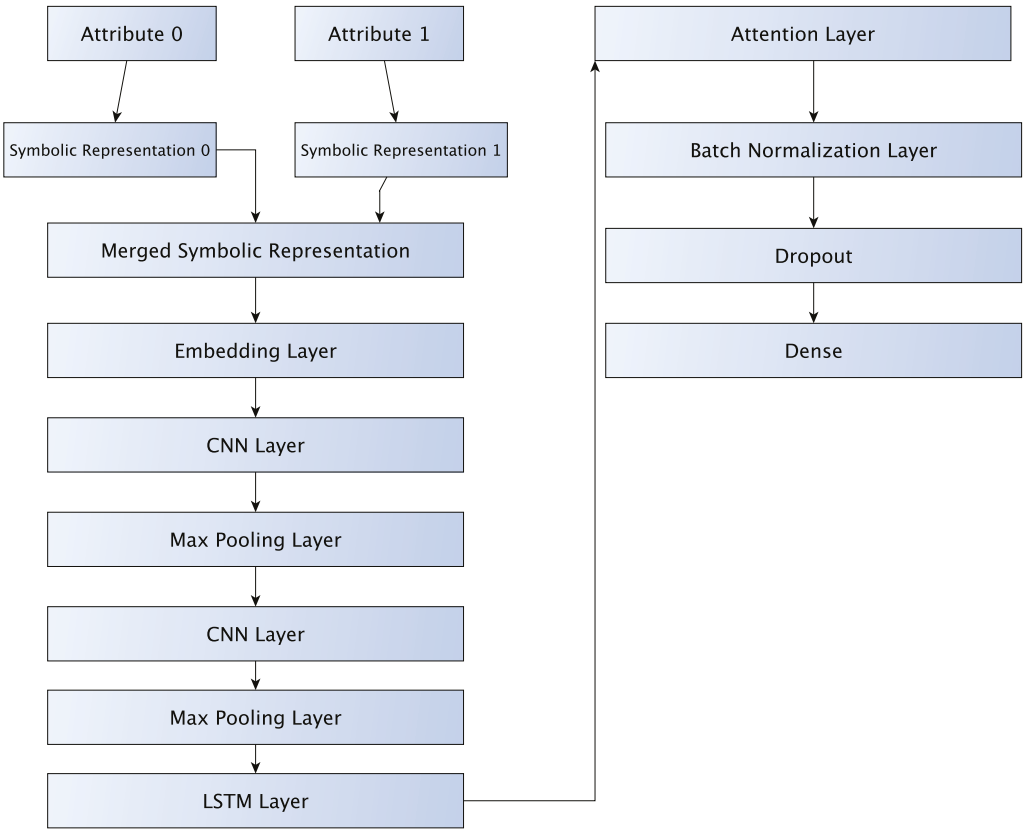



**Figure 10.** Depiction of the BAG version of LSTM-Attention-CNN2 neural network.

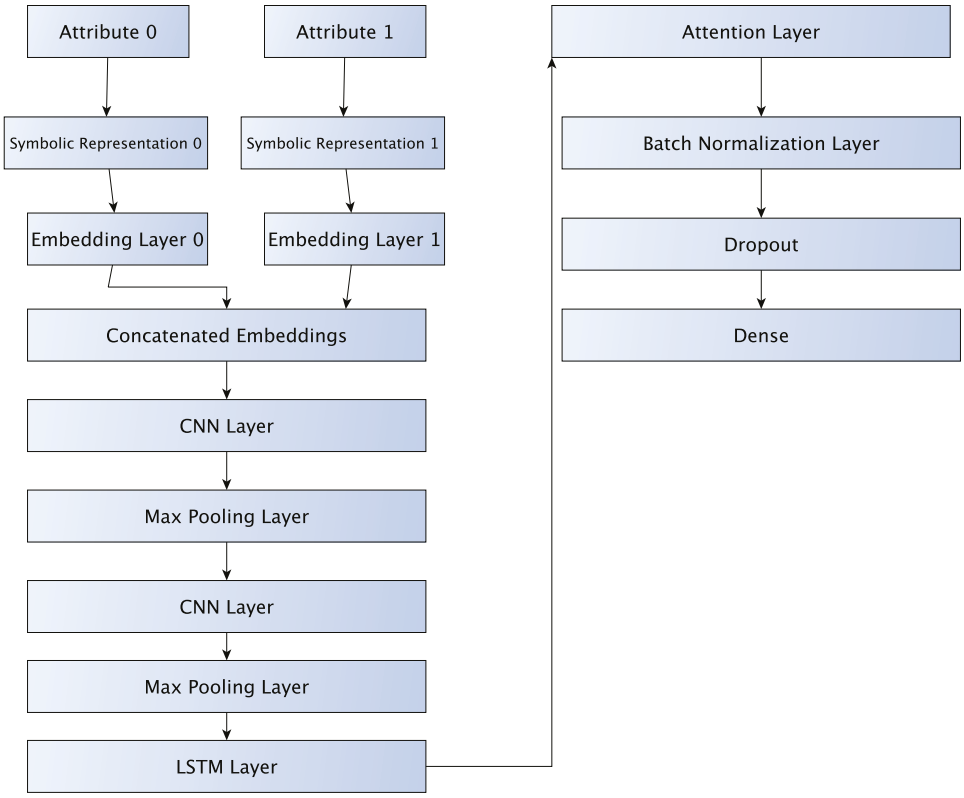



**Figure 11.** Depiction of the STACK version of LSTM-Attention-CNN2 neural network.

### *5.5. The SCALE-BOSS-MR-MV (SBMR-MV) Algorithm*

Before diving deeper into the specifics of SBMR-MV, we explain in more detail the Merging used in SBMR-MV. More concretely in the BAG case, the vocabularies for all the attributes of the time-series get merged to create a single vocabulary and then a term frequency vector is created. For example, the word aaaa for attribute 0 and the word aaaa for attribute 1 will be in the same bag, which now has a term frequency of 2 for time-series 0. This is because the word aaaa for attribute 0 has a term frequency value of 1 for time-series 0 and the word aaaa for attribute 1 also has a term frequency of 1. When the two get merged to create the final Term-Frequency vector for the BAG strategy, we get a Term-Frequency value of 2. The BAG merging strategy is shown in Figure 12.

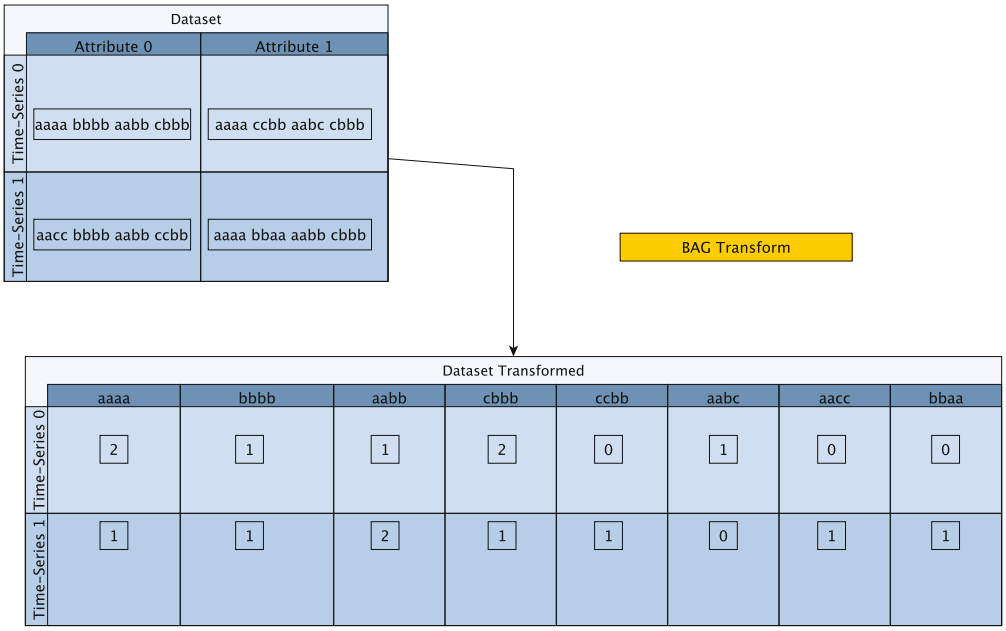



**Figure 12.** BAG merging strategy example for SBMR-MV.

In the STACK case, each attribute is transformed into its symbolic representation and then a Term Frequency vector is created for each attribute. Then the term frequency vectors are stacked column-wise to produce the final term frequency vector. In contrast to the BAG strategy, when using the STACK merging strategy, the vocabularies for each attribute remain distinct and the term frequency vectors are in effect stacked horizontally (column-wise). For example the word aaaa for attribute 0 is denoted as "aaaa Atr0" and is different than the word aaaa for attribute 1 that is denoted as "aaaa Atr1". In the STACK merging strategy, both "aaaa Atr0" and "aaaa Atr1" get a Term-Frequency value of 1 for time-series 0.

The STACK merging strategy is shown in Figure 13.

The final Term Frequency vector is then passed to the classifier and the procedure from then on is the same as in SCALE-BOSS-MR.

Algorithm 1 describes the SCALE-BOSS-MR-MV (SBMR-MV) process in detail. Lines 1 through 16 describe the inner method of SBMR-MV.

In lines 4 to 7, we apply dilation to each attribute of the input time-series.

Then, in lines 8 through 12, we create the term frequency vector and transform it either using bag or stack merging strategy. Then, in line 13 we merge the current term frequency vector with the global term frequency vector to create the final term frequency vector that will be passed to the classifier. In lines 18 and 19, we create the global term frequency vectors for the train and test sets, respectively. In line 20 we run SBMR-MV-inner. This is a helper function for creating the term frequency vector from the input time-series for the training set. In line 23 we run the SBMR-MV-inner for the first-order differences of the training set. In line 24 we fit the classifier with the term frequency vector of the training set. In line 25 we run SBMR-MV inner for the test set. In line 28 we run the SBMR-MV-inner for the first-order differences of the test set. Finally, in line 29 we get the predicted values for the test set.

The algorithm has two sets of hyperparameters, the window size and the dilation size. The window size allows SBMR-MV to see the input at different granularities similar to WEASEL. Dilation allows the algorithm to see parts of the input that are further ahead.

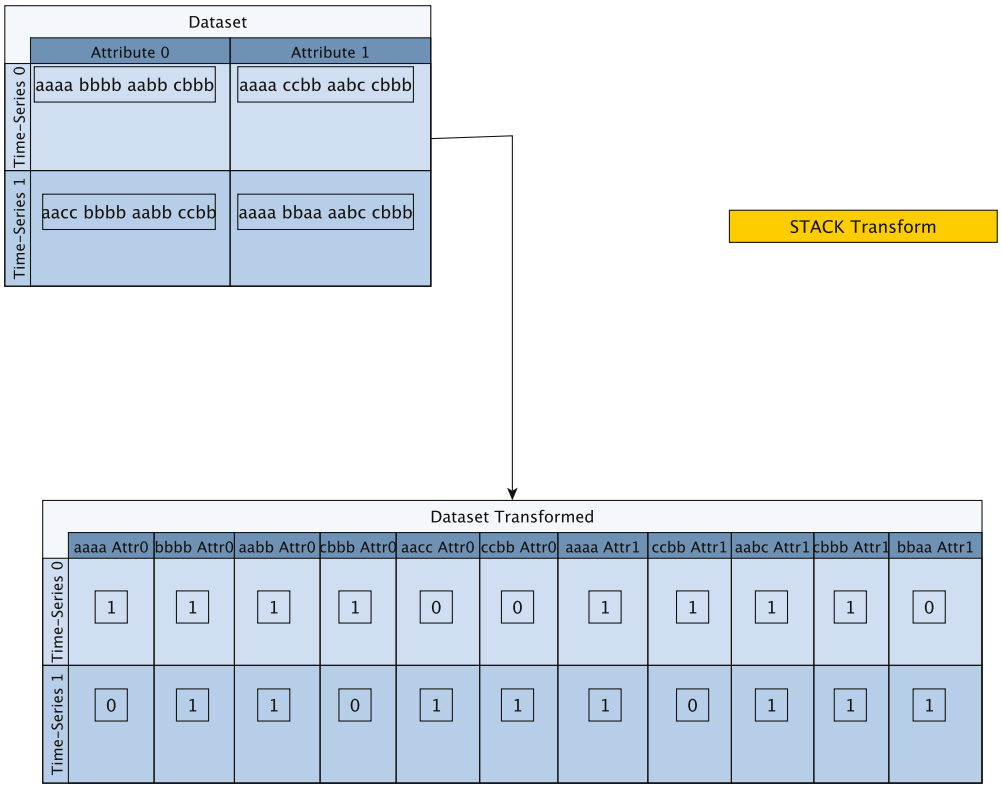



**Figure 13.** STACK merging strategy example.

**Algorithm 1:** SCALE-BOSS-MR-MV (SBMR-MV) algorithm.

```
Input  : X: the training set
Input  : y_train: the labels for the training set
Input  : X_test: the test set
Input  : window_configs: the window sizes and steps
Input  : dilation_filter_configs: the dilation filter sizes
Input  : clf: Classifier that can work on term-frequency vectors
Output : predicted: the predictions for the test set
Function SBMR-MV-inner(X, window_configs, dilation_filter_configs,
  global_term_frequency_vector):
    for curr_window_config in window_configs do
        for curr_dilation_filter in dilation_filter_configs do
            for curr_attribute in attributes do
                X_dilated ← dilation (X[curr_attribute],curr_dilation_filter)
                curr_symbolic_representation← process_time_series
                  (X_dilated,curr_window_config)
                  symbolic_representation_list.append(curr_symbolic_representation)
            end for
            if BAG then
                curr_term_frequency_vector ← bag_merging_strategy
                  (symbolic_representation_list)
            else
                curr_term_frequency_vector ← stack_merging_strategy
                  (symbolic_representation_list)
            end if
            global_term_frequency_vector ← hstack
              (global_term_frequency_vector,curr_term_frequency_vector)
        end for
    end for
    return

global_term_frequency_vector_train ← array ()
global_term_frequency_vector_test ← array ()
SBMR-MV-inner (X,window_configs, dilation_filter_configs,
  global_term_frequency_vector_train)
if doTrend then
    X_trend_train ← diff (X)
    SBMR-MV-inner (X_trend_train,window_configs, dilation_filter_configs,
      global_term_frequency_vector_train)
clf.fit (global_term_frequency_vector_train,y_train)
SBMR-MV-inner (X_test,window_configs, dilation_filter_configs,
  global_term_frequency_vector_test)
if doTrend then
    X_trend_test ← diff (X_test)
    SBMR-MV-inner (X_trend_test,window_configs, dilation_filter_configs,
      global_term_frequency_vector_test)
predicted ← clf.predict (global_term_frequency_vector_test)
return predicted
```

## 6. Evaluation

### *6.1. Experimental Setting*

We have chosen mean accuracy and mean execution time as the measures of choice.

We define execution time (Total_time) as the total execution time (training time plus test time) of each method for each dataset. Total_Time_Mean is the mean execution time across all datasets. Accuracy_Mean is the mean accuracy across all datasets.

We did that because all of the methods in the literature [15,18–20,24,25] use accuracy as a measure of qualitative performance. In addition to that, we used mean execution time as a measure of computational efficiency.

Our tests were performed on a mid 2015 Macbook with a 2.2 Ghz Intel Core i7 and 16 GB of RAM.

Table 1 shows the characteristics of the datasets used for the evaluation of the univariate case. We have chosen the eight datasets from the UCR archive [59] with the largest training set.

**Table 1.** Characteristics of the datasets.

| Name | Train_Size | Test_Size | n_Classes | n_Timestamps |
|---|---|---|---|---|
| Crop | 7200 | 16,800 | 24 | 46 |
| FordB | 3636 | 810 | 2 | 500 |
| FordA | 3601 | 1320 | 2 | 500 |
| NonInvasiveFetalECGThorax2 | 1800 | 1965 | 42 | 750 |
| NonInvasiveFetalECGThorax1 | 1800 | 1965 | 42 | 750 |
| PhalangesOutlinesCorrect | 1800 | 858 | 2 | 80 |
| HandOutlines | 1000 | 370 | 2 | 2709 |
| TwoPatterns | 1000 | 4000 | 4 | 128 |

In Table 2 we see the characteristics of the multivariate datasets.

We have chosen five UEA multivariate datasets. We have chosen datasets that have large enough trainSize, contain no null values, and have at least 40 timestamps.

In our evaluation, we measured the following:

1. The performance, in terms of accuracy and execution time for SCALE-BOSS-GRAPH for the univariate use case.
2. The performance, in terms of accuracy and execution time of the proposed SFA-enhanced neural network classifiers in the univariate use case.
3. The performance, in terms of accuracy and execution time of the proposed SFA-enhanced neural network classifiers in the multivariate use case.
4. The performance, in terms of accuracy and execution time of the adaptation of SCALE-BOSS-MR (SCALE-BOSS-MR-MV) to the multivariate use case.

We provide the detailed Accuracy and Total Execution Time for each algorithm and dataset in the Appendix A of the paper.

**Table 2.** Description of the multivariate datasets.

| Name | Train_Size | Test_Size | NumDimensions | n_Timestamps | n_Classes |
|---|---|---|---|---|---|
| ArticularyWordRecognition | 275 | 300 | 9 | 144 | 25 |
| EthanolConcentration | 261 | 263 | 3 | 1751 | 4 |
| FingerMovements | 316 | 100 | 28 | 50 | 2 |
| SelfRegulationSCP1 | 268 | 293 | 6 | 896 | 2 |
| UWaveGestureLibrary | 120 | 320 | 3 | 315 | 8 |

### *6.2. SCALE-BOSS Results*

Table 3 shows the mean accuracy and mean execution time for a selection of SCALE-BOSS instantiations used as a baseline to judge the performance of SCALE-BOSS-GRAPH. The selected instantiations were chosen because they are the best in terms of accuracy and execution time.

**Table 3.** Accuracy and total execution time of the SCALE-BOSS classifiers.

| Algorithm | Accuracy_Mean | Accuracy_std | Total_Time_Mean | Total_Time_std |
|---|---|---|---|---|
| BOSS-RF | 0.796 | 0.088 | 9.7 | 5.5 |
| MB-K-BOSS-VS | 0.727 | 0.101 | 11.5 | 6.9 |

### *6.3. State-of-the-Art Algorithm Results*

Table 4 shows the mean accuracy and mean execution time for the state-of-the-art algorithms.

We have chosen WEASEL 2.0 and MRSQM as state-of-the-art algorithms that use symbolic representations, and Rocket and miniRocket as state-of-the-art algorithms in terms of accuracy, which are also quite efficient in terms of execution time.

The mean accuracy and mean execution time for the state-of-the-art algorithms are used as a yardstick of success to compare the algorithms proposed. In bold we can see the algorithm with the best accuracy and the algorithm with the best execution time.

**Table 4.** Accuracy and execution time of state-of-the-art algorithms.

| Algorithm | Accuracy_Mean | Accuracy_Pop_std | Total_Time_Mean | Total_Time_Pop_std |
|---|---|---|---|---|
| WEASEL_V2 | **0.904** | 0.078 | 339.6 | 189.9 |
| Rocket | 0.901 | 0.084 | 309.3 | 167.7 |
| MiniRocket | 0.900 | 0.082 | **18.7** | 20.1 |
| MRSQM_nsax0_nsfa5 | 0.889 | 0.093 | 418.8 | 239.4 |
| MRSQM_nsax1_nsfa1 | 0.878 | 0.101 | 176.2 | 101.1 |
| MRSQM_nsax0_nsfa1 | 0.861 | 0.109 | 117.2 | 76.1 |
| MRSQM_nsax1_nsfa0 | 0.839 | 0.113 | 104.2 | 59.0 |

### *6.4. SCALE-BOSS-GRAPH Results*

Table 5 shows the mean accuracy and mean execution time for several instantiations of the SCALE-BOSS-GRAPH framework.

The SVC classifier has two main parameters: (a) the number of iterations and (b) the regularization parameter C.

In the first row, we can see the result of a SVC classifier with 2 iterations and using cross validation for three C values 2.0, 3.0 and 3.5. In the second row, we can see the results for a SVC classifier with 2 iterations and a C value of 3.5. In the third row, we can see the result of a SVC classifier for 2 iterations and a C value of 1.0.

Table 5 answers RQ1. This question aims to answer whether graph kernel-based algorithms fused with SFA representation are accurate and efficient. As we can see in Table 5, the WL kernel together with the SFA representation and a kernel SVM is moderately accurate (it is less accurate than MB-K-BOSS-VS) with good execution time (twice the execution time of BOSS-RF and K-BOSS-VS, but significantly less than that of other state-of-the-art algorithms).

**Table 5.** Mean accuracy and execution time of the proposed algorithms.

| Algorithm | Accuracy_Mean | Accuracy_Pop_std | Total_Time_Mean | Total_Time_Pop_std |
|---|---|---|---|---|
| SVCcv-2_2.0_3.0_3.5 | **0.699** | 0.079 | 24.699 | 12.783 |
| SVC-2_3.5 | 0.697 | 0.079 | **20.734** | 9.494 |
| SVC-2_1.0 | 0.680 | 0.089 | 21.094 | 9.576 |

### *6.5. Evaluation of the Deep Neural Network Architectures*

As part of our evaluation, we used the deep learning architectures provided by aeon [60] as a baseline to compare against our own.

Table 6 shows the average accuracy and execution time of the state-of-the-art neural classifiers. The ep20 notation means that the algorithm ran for 20 epochs, while the no-bn notation means that there is no batch normalization layer. The IndividualLite classifier trained for 100 epochs achieves the best accuracy with a reasonable execution time. The CNNClassifier trained for 200 epochs achieves very high accuracy with the best trade-off between accuracy and execution time.

**Table 6.** Accuracy and total execution time of the neural classifiers.

| Algorithm | Accuracy_Mean | Accuracy_Pop_std | Total_Time_Mean | Total_Time_Pop_std |
|---|---|---|---|---|
| IndividualLITEClassifier-ep100 | **0.819** | 0.121 | 599.293 | 408.343 |
| EncoderClassifier-ep20 | 0.811 | 0.096 | 1045.826 | 665.514 |
| CNNClassifier-ep200 | 0.794 | 0.110 | 59.852 | 25.028 |
| IndividualLITEClassifier-ep50 | 0.770 | 0.139 | 328.929 | 223.978 |
| ResNet-ep20 | 0.705 | 0.181 | 881.376 | 589.755 |
| IInception-ep20 | 0.642 | 0.242 | 1117.784 | 784.749 |
| IndividualLITEClassifier-ep20 | 0.634 | 0.222 | 135.210 | 90.210 |
| FCNClassifier-ep20 | 0.546 | 0.315 | 385.911 | 282.955 |
| MLPClassifier-ep200 | 0.493 | 0.225 | 66.662 | 33.646 |
| CNNClassifier-ep20 | 0.365 | 0.293 | 6.757 | 2.414 |
| MLPClassifier-ep20 | 0.353 | 0.264 | **7.605** | 3.389 |

In Table 7 we can see the average accuracy and average execution time for the SFA-enhanced neural networks. The LSTM-attention-cnn2 trained for 50 epochs achieves the best accuracy with a very competitive execution time. The LSTM-attention-cnn trained for 50 epochs follows in terms of accuracy, but with double the execution time. The LSTM-attention-cnn2 trained for 20 epochs achieves very high accuracy with the best execution time. The addition of batch normalization gives an improvement in accuracy of 4% for the 50 epoch case. The Graph Convolutional Neural network (GCN) achieves the worst performance with relatively high execution time compared to the other networks.

Table 7 answers RQ2, whether using GCN instead of the WL kernel is more efficient. As we can see, GCN is only slightly worse in accuracy compared to the WL kernel (as shown in Table 5) but is significantly worse in terms of execution time.

The rest of Table 7 answers RQ3 on how the proposed neural network architectures compare against the state-of-the-art neural network classifiers. As we can see, the proposed neural network architectures score the same or better in terms of accuracy compared to the state-of-the-art methods, but are significantly more efficient in terms of execution time.

**Table 7.** Accuracy and total execution time of the SFA-enhanced neural classifiers.

| Algorithm | Accuracy_Mean | Accuracy_Pop_std | Total_Time_Mean | Total_Time_Pop_std |
|---|---|---|---|---|
| LSTM-attention-cnn2-ep50 | **0.805** | 0.106 | 157.001 | 147.923 |
| LSTM-attention-cnn-ep50 | 0.798 | 0.103 | 302.599 | 360.341 |
| LSTM-attention-ep50 | 0.798 | 0.102 | 610.468 | 556.662 |
| LSTM-attention-cnn2-ep20 | 0.795 | 0.107 | **94.234** | 73.159 |
| LSTM-attention-cnn-ep20 | 0.792 | 0.100 | 175.713 | 143.131 |
| LSTM-attention-ep20 | 0.783 | 0.107 | 391.999 | 334.938 |
| LSTM-CNN-ep50 | 0.718 | 0.130 | 138.903 | 102.731 |
| LSTM-CNN-ep20 | 0.709 | 0.123 | 118.926 | 95.803 |
| GCN-ep20 | 0.681 | 0.115 | 576.484 | 825.842 |
| LSTM-CNN-no-bn-ep50 | 0.680 | 0.171 | 181.424 | 149.712 |
| LSTM-CNN-no-bn-ep20 | 0.665 | 0.195 | 109.958 | 79.093 |

In Figure 14 we can see the average accuracy and the average execution time for the SFA-enhanced neural classifiers. The SFA-enhanced neural classifiers are the classifiers we created by fusing the SFA symbolic representation and neural networks. LSTM-attention-cnn2-ep50 is the best in terms of balancing execution time and accuracy. It achieves the best accuracy with a modest 157 seconds of mean execution time. The same model trained for 20 epochs follows closely with 1% lower mean accuracy but with the best execution time.

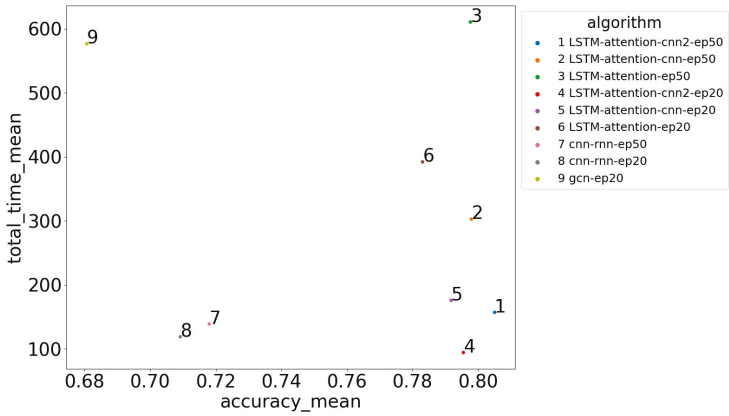



**Figure 14.** Execution time and accuracy for the neural classifiers.

Figure 15 shows the average accuracy and average execution time for the state-of-the-art neural classifiers.

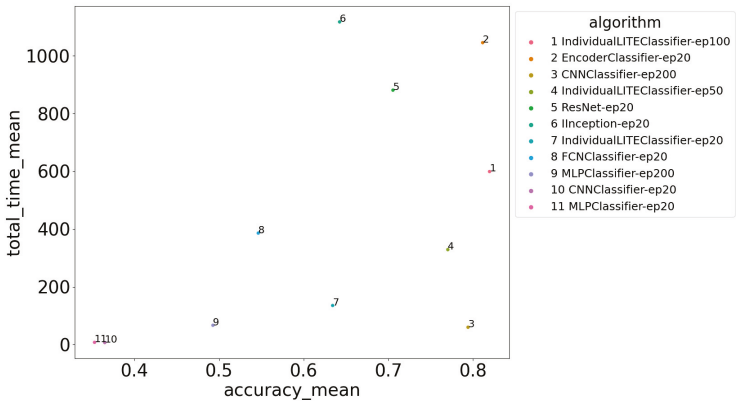



**Figure 15.** Execution time and accuracy for the state-of-the-art neural classifiers.

### *6.6. Multivariate Evaluation*

We have chosen to reuse the windowing and dilation configurations that proved be the best in SCALE-BOSS-MR [15]. These are shown in Tables 8 and 9.

**Table 8.** Window configurations.

| Name | Window_Configs (Window Sizes) | Window Step |
|---|---|---|
| W0 | 24 | 1 |
| W8 | 4, 8, 12, 16, 20, 24, 28, 32, 36, 40 | 1 |
| W11 | 12, 16, 20, 24, 28, 32, 36, 40 | 4 |
| W14 | 12, 16, 20, 24, 28, 32 | 2 |
| W15 | 12, 16, 20, 24, 28, 32, 36, 40 | 8 |

**Table 9.** Dilation configurations.

| Name | Dilation_Filter_Configs |
|---|---|
| D0 | 1 (no dilation) |
| D2 | 1, 7, 9, 11 |
| D4 | 1, 11 |
| D6 | 1, 9 |

The W0 configuration has a single window size with step 1, which is the baseline configuration. The W8 configuration has many window sizes with step 1; this configuration is the most compute-intensive but gave the best results in terms of accuracy in the SCALE-BOSS-MR case. The W11, W14 and W15 are very similar in terms of window size and the main difference is in the window step parameter.

Regarding dilation, larger dilation sizes mean that the algorithm can look further in time. Having multiple dilation sizes in theory gives the algorithm multiple resolution of the time-series. The D0 parameter is to have dilation 1 which means no dilation. This is the configuration that works as a baseline, to see the effects of dilation in the other configurations. D2 has the most dilation sizes (4 dilation sizes in total) aiming to provide the best accuracy. D4 and D6 have a single dilation size parameter beside 1.

In Table 10 we see the mean execution time and mean accuracy for the novel neural network architectures. As we can see, adding self-attention gives a 7 % boost in accuracy and that the STACK strategy is significantly more accurate compared to the BAG strategy.

**Table 10.** Proposed neural network methods' mean accuracy and execution time.

| Algorithm | Accuracy_Mean | Accuracy_Pop_std | Total_Time_Mean | Total_Time_Pop_std |
|---|---|---|---|---|
| LSTM-attention-cnn2-ep50-stack | **0.678** | 0.210 | 31.987 | 19.480 |
| LSTM-CNN-ep50-stack | 0.601 | 0.189 | 40.720 | 41.213 |
| LSTM-attention-cnn2-ep50-bag | 0.581 | 0.192 | 47.251 | 22.522 |
| LSTM-CNN-ep20-stack | 0.574 | 0.215 | **16.052** | 9.202 |
| LSTM-attention-cnn2-ep20-stack | 0.569 | 0.217 | 23.072 | 15.059 |
| LSTM-attention-cnn2-ep20-bag | 0.500 | 0.199 | 49.631 | 34.406 |

Table 11 provides the mean execution time and mean accuracy for the state-of-the-art neural networks. The CNNClassifier is really scalable and achieves good accuracy compared to other neural network methods. We can see that our proposed neural networks are significantly more accurate than the state-of-the-art classifiers.

**Table 11.** State-of-the-art neural network methods' mean accuracy and execution time.

| Algorithm | Accuracy_Mean | Accuracy_Pop_std | Total_Time_Mean | Total_Time_Pop_std |
|---|---|---|---|---|
| EncoderClassifier-ep20 | **0.633** | 0.227 | 128.007 | 121.132 |
| CNNClassifier-ep200 | 0.593 | 0.218 | 11.475 | 5.231 |
| IInception-ep20 | 0.578 | 0.239 | 153.699 | 147.896 |
| IndividualLITEClassifier-ep50 | 0.575 | 0.285 | 53.479 | 47.540 |
| IndividualLITEClassifier-ep100 | 0.548 | 0.279 | 100.233 | 90.002 |
| CNNClassifier-ep20 | 0.367 | 0.269 | **2.303** | 0.658 |
| ResNet-ep20 | 0.299 | 0.154 | 103.053 | 101.543 |

In Table 12 we see the mean execution time and mean accuracy for the single window SCALE-BOSS-MR Multivariate (SBMR-MV).

SBMR-MV-ridge-cv-W0-trend-D2-BG-STACK achieves the best accuracy but at the expense of execution time.

On the other hand, SBMR-MV-RF-W0-trend-D4-BG-STACK and SBMR-MV-ridge-cv-W0-trend-D4-BG-STACK achieve similar accuracy with significantly lower execution time. SBMR-MV-RF-W0-trend-D0-UG-STACK achieves an accuracy of 0.657 with a very low execution time. SBMR-MV-RF-W0-noTrend-D0-UG-STACK does not contain any trend information and gives an accuracy of 0.65. The above results show that adding trend information and dilation helps improve accuracy.

In Table 13 we see the mean execution time and mean accuracy for the multiple windows SBMR-MV. This table shows the effect of using multiple windows and chi squared test for feature selection and on top of dilation and trend information.

SBMR-MV-ridge-cv-W15-trend-D2-BG-BAG-chi gives the best accuracy of 0.733 while being relatively scalable. SBMR-MV-ridge-cv-W15-trend-D2-BG-STACK-chi is not far behind but with higher execution time. SBMR-MV-ridge-cv-W15-trend-D6-BG-STACK-chi is also very close with accuracy 0.729 while being significantly more scalable compared to other window configurations. SBMR-MV-ridge-cv-W15-trend-D6-BG-BAG-chi is not far behind with accuracy of 0.725 while being even more scalable compared to SBMR-MV-ridge-cv-W15-trend-D6-BG-STACK-chi. SBMR-MV-ridge-cv-W15-trend-D6-BG-BAG achieves accuracy of 0.713, which shows that the chi squared test gives a slight bump in accuracy when using multiple windows.

We can also observe that the W15 configuration is generally better than W11 while also being significantly more scalable. W14 is not better than W11 or W15, neither in terms of performance nor in accuracy. W8 has by far the slowest execution time while not producing better results.

Tables 12 and 13 answer RQ4 on how SCALE-BOSS-MR can be adapted to the multivariate use case and how it performs in terms of accuracy and execution time.

Table 14 shows the mean execution time and the mean accuracy for the Rocket family of classifiers. Both ROCKET and miniRocket proved to be very accurate, but ROCKET is significantly less scalable (in terms of execution time) than SBMR-MV.

**Table 12.** SCALE-BOSS-MR Multivariate (SBMR-MV) mean accuracy and execution time for a single window.

| Algorithm | Accuracy_Mean | Accuracy_Pop_std | Total_Time_Mean | Total_Time_Pop_std |
|---|---|---|---|---|
| SBMR-MV-ridge-cv-W0-trend-D2-BG-STACK | 0.700 | 0.269 | 59.023 | 43.971 |
| SBMR-MV-RF-W0-trend-D4-BG-STACK | 0.691 | 0.216 | 29.784 | 22.312 |
| SBMR-MV-ridge-cv-W0-trend-D4-BG-STACK | 0.684 | 0.246 | 29.422 | 22.086 |
| SBMR-MV-RF-W0-trend-D2-BG-STACK | 0.684 | 0.237 | 59.453 | 44.198 |
| SBMR-MV-ridge-cv-W0-trend-D2-BG-BAG | 0.682 | 0.238 | 57.667 | 44.877 |
| SBMR-MV-RF-W0-trend-D6-UG-STACK | 0.667 | 0.239 | 26.334 | 20.425 |
| SBMR-MV-RF-W0-trend-D0-UG-STACK | 0.657 | 0.225 | 13.416 | 10.537 |
| SBMR-MV-RF-W0-noTrend-D0-UG-STACK | 0.650 | 0.192 | 6.987 | 5.445 |
| SBMR-MV-RF-W0-trend-D2-BG-BAG | 0.647 | 0.183 | 60.185 | 47.299 |
| SBMR-MV-ridge-cv-W0-trend-D6-UG-BAG | 0.646 | 0.230 | 25.831 | 20.542 |
| SBMR-MV-ridge-cv-W0-trend-D4-BG-BAG | 0.646 | 0.224 | 28.723 | 22.514 |
| SBMR-MV-ridge-cv-W0-trend-D0-UG-STACK | 0.643 | 0.226 | 13.293 | 10.460 |
| SBMR-MV-RF-W0-trend-D4-BG-BAG | 0.641 | 0.187 | 29.460 | 23.218 |
| SBMR-MV-RF-W0-trend-D6-UG-BAG | 0.637 | 0.194 | 27.080 | 21.417 |
| SBMR-MV-RF-W0-trend-D0-UG-BAG | 0.631 | 0.173 | 13.208 | 10.507 |
| SBMR-MV-ridge-cv-W0-noTrend-D0-UG-STACK | 0.629 | 0.227 | 6.847 | 5.475 |
| SBMR-MV-ridge-cv-W0-trend-D0-UG-BAG | 0.576 | 0.196 | 12.982 | 10.392 |
| SBMR-MV-RF-W0-noTrend-D0-UG-BAG | 0.571 | 0.170 | 6.870 | 5.493 |
| SBMR-MV-ridge-cv-W0-noTrend-D0-UG-BAG | 0.549 | 0.196 | 6.770 | 5.463 |

**Table 13.** SCALE-BOSS-MR Multivariate (SBMR-MV) mean accuracy and execution time for multiple windows.

| Algorithm | Accuracy_Mean | Accuracy_Pop_std | Total_Time_Mean | Total_Time_Pop_std |
|---|---|---|---|---|
| SBMR-MV-ridge-cv-W15-trend-D2-BG-BAG-chi | 0.733 | 0.217 | 62.364 | 41.879 |
| SBMR-MV-ridge-cv-W15-trend-D2-BG-STACK-chi | 0.730 | 0.242 | 71.386 | 39.949 |
| SBMR-MV-ridge-cv-W15-trend-D6-BG-STACK-chi | 0.729 | 0.253 | 38.239 | 22.024 |
| SBMR-MV-ridge-cv-W15-trend-D6-BG-BAG-chi | 0.725 | 0.245 | 31.643 | 21.584 |
| SBMR-MV-ridge-cv-W11-trend-D6-BG-STACK-chi | 0.722 | 0.262 | 64.418 | 41.776 |
| SBMR-MV-ridge-cv-W11-trend-D6-BG-BAG-chi | 0.718 | 0.233 | 59.652 | 43.653 |
| SBMR-MV-ridge-cv-W11-trend-D4-BG-STACK-chi | 0.717 | 0.250 | 64.460 | 42.175 |
| SBMR-MV-ridge-cv-W11-trend-D6-BG-STACK | 0.717 | 0.268 | 67.959 | 41.675 |
| SBMR-MV-ridge-cv-W15-trend-D4-BG-STACK-chi | 0.715 | 0.251 | 35.160 | 20.303 |
| SBMR-MV-ridge-cv-W15-trend-D6-BG-BAG | 0.713 | 0.254 | 31.486 | 21.149 |
| SBMR-MV-ridge-cv-W8-trend-D6-BG-STACK | 0.713 | 0.263 | 298.052 | 210.282 |
| SBMR-MV-ridge-cv-W15-trend-D6-BG-STACK | 0.710 | 0.268 | 36.318 | 19.974 |
| SBMR-MV-ridge-cv-W11-trend-D4-BG-BAG-chi | 0.709 | 0.231 | 59.747 | 43.973 |
| SBMR-MV-ridge-cv-W11-trend-D6-BG-BAG | 0.708 | 0.253 | 75.683 | 56.347 |
| SBMR-MV-ridge-cv-W14-trend-D6-BG-STACK | 0.703 | 0.265 | 89.956 | 61.231 |
| SBMR-MV-ridge-cv-W15-trend-D4-BG-BAG-chi | 0.699 | 0.248 | 31.711 | 21.682 |
| SBMR-MV-RF-W15-trend-D6-BG-STACK | 0.698 | 0.213 | 37.453 | 20.296 |
| SBMR-MV-ridge-cv-W8-trend-D6-BG-BAG | 0.695 | 0.257 | 287.476 | 219.437 |
| SBMR-MV-RF-W11-trend-D6-BG-BAG | 0.687 | 0.201 | 59.510 | 42.949 |
| SBMR-MV-RF-W11-trend-D6-BG-STACK | 0.670 | 0.253 | 67.362 | 41.766 |
| SBMR-MV-RF-W15-trend-D6-BG-BAG | 0.651 | 0.188 | 32.324 | 21.439 |

**Table 14.** Rocket method mean accuracy and execution time.

| Algorithm | Accuracy_Mean | Accuracy_Pop_std | Total_Time_Mean | Total_Time_Pop_std |
|---|---|---|---|---|
| miniRocket | 0.765 | 0.228 | 7.439 | 5.958 |
| ROCKET | 0.747 | 0.227 | 118.868 | 109.902 |

## 7. Conclusions

During our evaluation, we arrived at the following conclusions:

1. The SCALE-BOSS-GRAPH algorithm proved very scalable, but its accuracy is lower than the state of the art.
2. The GCNs were on par in terms of accuracy with the state-of-the-art deep learning methods, but were not that scalable compared to other methods.
3. The proposed deep learning methods inspired by the text classification domain proved to be very effective both in terms of accuracy as well as in terms of execution time when compared with state-of-the-art deep learning methods both in the univariate and multivariate use cases. The optimizations we employed were instrumental for this result; for example, batch normalization gave a 5 % improvement in accuracy over the baseline, whereas adding an attention layer to the model gave a 10 % increase in accuracy but with a non-negligible cost in execution time. Adding convolutional layers to the architecture did not contribute much in accuracy but proved very beneficial in terms of execution time.
4. The adaptation of SCALE-BOSS-MR to the multivariate use case namely SCALE-BOSS-MR-MV proved very scalable and accurate compared to the state of the art. More specifically, SCALE-BOSS-MR-MV proved to be more scalable than ROCKET while being marginally less accurate.

For future work, we plan to try transformer architectures fused with the SFA symbolic representation. Another line of research would be to try other Graph Neural Networks with more complex graph representations fused with the SFA symbolic representation.

**Author Contributions:** Conceptualization, A.G. and G.V.; methodology, A.G. and G.V.; software, A.G.; validation, A.G. and G.V.; formal analysis, A.G. and G.V.; investigation, A.G.; resources, not applicable; data curation, A.G.; writing—original draft preparation, A.G. and G.V.; writing—review and editing, A.G. and G.V.; visualization, A.G. and G.V.; supervision, G.V.; project administration, G.V. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Publicly available datasets were analyzed in this study. The UCR and UEA datasets used in this study can be found at: http://www.timeseriesclassification.com/dataset.php (accessed on 24 November 2025) and https://www.cs.ucr.edu/%7Eeamonn/time_series_data_2018/ (accessed on 24 November 2025).

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

## Appendix A

The appendix shows detailed accuracy and total execution time results for all the methods and datasets. The detailed tables help compare the proposed methods with other methods proposed in the literature, as well as compare the proposed methods against each other on a per-dataset basis.

In Table A1 we can see the accuracy of the instantiations of SCALE-BOSS-GRAPH for all the datasets.

In Table A2 we can see the execution time of the instantiations of SCALE-BOSS-GRAPH for all the datasets.

In Table A3 we can see the accuracy of the state-of-the-art neural classifiers for all the datasets.

In Table A4 we can see the execution time of the state-of-the-art neural classifiers for all the datasets.

In Table A5 we can see the accuracy of the SFA-enhanced neural implementation for all the datasets.

Crop, NonInvasiveFetalECGThorax2 and NonInvasiveFetalECGThorax1 have class count higher than or equal to 8. Both LSTM-attention-cnn2-ep50 and LSTM-attention-cnn2-ep50 have accuracy higher than or equal to the best state-of-the-art neural classifiers. This shows that our proposed method works well with datasets with high class count.

In Table A6 we can see the execution time of the SFA-enhanced neural implementation for all the datasets.

In Table A7 we see the accuracy for the single-window SBMR across all datasets.

ArticularyWordRecognition and UWaveGestureLibrary have class count higher than or equal to 8. We can see that the proposed algorithm even with a single window can achieve high accuracy. It is worth noting that the accuracy on those datasets is sometimes lower compared to datasets with lower class count. This suggests that (a) the proposed algorithm works well with datasets with high class count and (b) that the class count is not a deciding factor for classification accuracy.

In Table A8 we see the execution time for single-window SBMR across all datasets.

In Table A9 we see the accuracy for the multiple-window SBMR across all datasets.

We can see that SBMR-MV with multiple window sizes achieves near perfect accuracy for the datasets with high class count. This also reinforces the argument that high class count is not a very good indicator of classification difficulty of a dataset.

In Table A10 we see the execution time for multiple-window SBMR across all datasets.

In Table A11 we see the accuracy for the novel neural network architectures across all datasets.

We can see that the proposed algorithms and especially LSTM-attention-cnn2-ep50-stack perform really well on datasets with high class count especially compared with state-of-the-art neural network architectures. We can also observe that the datasets with high class count are not the most difficult to classify.

In Table A12 we see the execution time for novel neural networks across all datasets.

In Table A13 we see the accuracy for the state-of-the-art neural network architectures across all datasets.

In Table A14 we see the execution time for state-of-the-art neural networks across all datasets.

In Table A15 we see the accuracy for the Rocket family of algorithms across all datasets.

In Table A16 we see the execution time for the Rocket family of algorithms across all datasets.

**Table A1.** Accuracy of the instantiations of SCALE-BOSS-GRAPH.

| Dataset<br>Algorithm | Crop | FordA | FordB | HandOutlines | NonInvasiveFetalECGThorax1 | NonInvasiveFetalECGThorax2 | PhalangesOutlinesCorrect | TwoPatterns |
|---|---|---|---|---|---|---|---|---|
| SVC-2_1.0 | 0.605 | 0.813 | 0.598 | 0.700 | 0.534 | 0.699 | 0.704 | 0.786 |
| SVC-2_3.5 | 0.612 | 0.811 | 0.620 | 0.695 | 0.588 | 0.732 | 0.726 | 0.794 |
| SVCcv-2_2.0_3.0_3.5 | 0.612 | 0.811 | 0.617 | 0.711 | 0.588 | 0.735 | 0.723 | 0.793 |

**Table A2.** Execution time of the instantiations of SCALE-BOSS-GRAPH.

| Dataset<br>Algorithm | Crop | FordA | FordB | HandOutlines | NonInvasiveFetalECGThorax1 | NonInvasiveFetalECGThorax2 | PhalangesOutlinesCorrect | TwoPatterns |
|---|---|---|---|---|---|---|---|---|
| SVC-2_1.0 | 29.307 | 26.350 | 22.957 | 29.297 | 26.089 | 24.783 | 2.726 | 7.243 |
| SVC-2_3.5 | 29.857 | 26.383 | 23.948 | 27.087 | 24.552 | 24.667 | 2.413 | 6.966 |
| SVCcv-2_2.0_3.0_3.5 | 46.939 | 29.536 | 27.208 | 28.834 | 27.345 | 26.861 | 2.990 | 7.879 |

**Table A3.** Accuracy of the state-of-the-art neural network classifiers.

| Dataset<br>Algorithm | Crop | FordA | FordB | HandOutlines | NonInvasiveFatalECGThorax1 | NonInvasiveFatalECGThorax2 | PhalangesOutlinesCorrect | TwoPatterns |
|---|---|---|---|---|---|---|---|---|
| CNNClassifier-ep20 | 0.049 | 0.510 | 0.509 | 0.859 | 0.018 | 0.018 | 0.613 | 0.344 |
| CNNClassifier-ep200 | 0.599 | 0.905 | 0.759 | 0.884 | 0.790 | 0.856 | 0.650 | 0.909 |
| EncoderClassifier-ep20 | 0.664 | 0.942 | 0.801 | 0.900 | 0.831 | 0.848 | 0.657 | 0.847 |
| FCNClassifier-ep20 | 0.505 | 0.896 | 0.752 | 0.641 | 0.030 | 0.042 | 0.667 | 0.838 |
| IInception-ep20 | 0.696 | 0.927 | 0.704 | 0.359 | 0.257 | 0.500 | 0.696 | 1.000 |
| IndividualLITEClassifier-ep100 | 0.720 | 0.948 | 0.809 | 0.781 | 0.589 | 0.883 | 0.828 | 1.000 |
| IndividualLITEClassifier-ep20 | 0.649 | 0.906 | 0.833 | 0.665 | 0.304 | 0.250 | 0.775 | 0.690 |
| IndividualLITEClassifier-ep50 | 0.706 | 0.937 | 0.775 | 0.822 | 0.537 | 0.683 | 0.700 | 1.000 |
| MLPClassifier-ep20 | 0.067 | 0.568 | 0.506 | 0.722 | 0.032 | 0.046 | 0.613 | 0.270 |
| MLPClassifier-ep200 | 0.270 | 0.698 | 0.598 | 0.854 | 0.205 | 0.219 | 0.614 | 0.484 |
| ResNet-ep20 | 0.652 | 0.928 | 0.716 | 0.668 | 0.371 | 0.700 | 0.611 | 0.999 |

**Table A4.** Execution time of the state-of-the-art neural network classifiers.

| Dataset<br>Algorithm | Crop | FordA | FordB | HandOutlines | NonInvasiveFatalECGThorax1 | NonInvasiveFatalECGThorax2 | PhalangesOutlinesCorrect | TwoPatterns |
|---|---|---|---|---|---|---|---|---|
| CNNClassifier-ep20 | 9.292 | 8.740 | 8.856 | 8.097 | 6.642 | 6.664 | 2.929 | 2.835 |
| CNNClassifier-ep200 | 62.032 | 81.318 | 80.168 | 88.313 | 66.284 | 61.911 | 20.709 | 18.079 |
| EncoderClassifier-ep20 | 543.388 | 1464.732 | 1477.982 | 2208.398 | 1173.432 | 1165.267 | 186.202 | 147.210 |
| FCNClassifier-ep20 | 107.983 | 547.794 | 546.054 | 908.470 | 473.204 | 410.372 | 48.388 | 45.022 |
| IInception-ep20 | 405.621 | 1622.802 | 1674.065 | 2536.508 | 1170.444 | 1210.527 | 177.632 | 144.673 |
| IndividualLITEClassifier-ep100 | 277.578 | 836.207 | 888.090 | 1347.996 | 641.307 | 631.699 | 91.441 | 80.029 |
| IndividualLITEClassifier-ep20 | 64.009 | 186.180 | 187.645 | 305.871 | 145.600 | 146.959 | 23.528 | 21.885 |
| IndividualLITEClassifier-ep50 | 148.385 | 482.603 | 495.227 | 725.426 | 338.380 | 348.790 | 49.652 | 42.970 |
| MLPClassifier-ep20 | 15.172 | 9.065 | 8.830 | 7.562 | 6.117 | 6.087 | 4.518 | 3.489 |
| MLPClassifier-ep200 | 139.296 | 83.352 | 83.041 | 64.771 | 51.888 | 50.578 | 36.980 | 23.391 |
| ResNet-ep20 | 316.447 | 1341.227 | 1398.291 | 1800.079 | 979.298 | 953.220 | 146.271 | 116.175 |

**Table A5.** Accuracy of the SFA-enhanced neural network classifiers.

| Dataset<br>Algorithm | Crop | FordA | FordB | HandOutlines | NonInvasiveFetalECGThorax1 | NonInvasiveFetalECGThorax2 | PhalangesOutlinesCorrect | TwoPatterns |
|---|---|---|---|---|---|---|---|---|
| LSTM-CNN-ep20 | 0.602 | 0.836 | 0.720 | 0.603 | 0.544 | 0.688 | 0.742 | 0.940 |
| LSTM-CNN-ep50 | 0.614 | 0.880 | 0.709 | 0.643 | 0.523 | 0.700 | 0.726 | 0.950 |
| LSTM-CNN-no-bn-ep20 | 0.612 | 0.900 | 0.683 | 0.646 | 0.343 | 0.450 | 0.721 | 0.969 |
| LSTM-CNN-no-bn-ep50 | 0.625 | 0.855 | 0.662 | 0.659 | 0.377 | 0.538 | 0.755 | 0.967 |
| LSTM-attention-cnn-ep20 | 0.614 | 0.908 | 0.731 | 0.854 | 0.720 | 0.819 | 0.753 | 0.935 |
| LSTM-attention-cnn-ep50 | 0.615 | 0.902 | 0.735 | 0.881 | 0.751 | 0.845 | 0.719 | 0.935 |
| LSTM-attention-cnn2-ep20 | 0.594 | 0.911 | 0.742 | 0.889 | 0.722 | 0.820 | 0.758 | 0.929 |
| LSTM-attention-cnn2-ep50 | 0.599 | 0.907 | 0.744 | 0.903 | 0.785 | 0.838 | 0.732 | 0.933 |
| LSTM-attention-ep20 | 0.614 | 0.903 | 0.740 | 0.881 | 0.692 | 0.777 | 0.718 | 0.941 |
| LSTM-attention-ep50 | 0.614 | 0.909 | 0.737 | 0.878 | 0.768 | 0.832 | 0.712 | 0.930 |
| gcn-ep20 | 0.525 | 0.912 | 0.730 | 0.676 | 0.582 | 0.606 | 0.642 | 0.772 |

**Table A6.** Execution time of the SFA-enhanced neural network classifiers.

| Dataset<br>Algorithm | Crop | FordA | FordB | HandOutlines | NonInvasiveFetalECGThorax1 | NonInvasiveFetalECGThorax2 | PhalangesOutlinesCorrect | TwoPatterns |
|---|---|---|---|---|---|---|---|---|
| LSTM-CNN-ep20 | 29.245 | 106.697 | 145.504 | 321.038 | 158.140 | 156.379 | 12.654 | 21.750 |
| LSTM-CNN-ep50 | 33.414 | 130.198 | 139.316 | 277.145 | 236.521 | 260.698 | 14.464 | 19.466 |
| LSTM-CNN-no-bn-ep20 | 25.209 | 148.240 | 157.016 | 254.593 | 130.238 | 131.012 | 14.881 | 18.473 |
| LSTM-CNN-no-bn-ep50 | 45.386 | 200.621 | 108.940 | 463.126 | 292.452 | 295.635 | 20.117 | 25.116 |
| LSTM-attention-cnn-ep20 | 39.236 | 177.046 | 204.562 | 479.402 | 231.187 | 227.560 | 20.807 | 25.905 |
| LSTM-attention-cnn-ep50 | 46.482 | 204.554 | 164.150 | 1176.196 | 339.760 | 447.863 | 18.086 | 23.698 |
| LSTM-attention-cnn2-ep20 | 26.532 | 89.626 | 101.880 | 246.840 | 127.765 | 130.393 | 13.175 | 17.661 |
| LSTM-attention-cnn2-ep50 | 28.055 | 104.636 | 102.962 | 456.134 | 250.488 | 280.236 | 14.411 | 19.087 |
| LSTM-attention-ep20 | 67.502 | 554.118 | 364.036 | 1097.373 | 490.547 | 484.485 | 35.522 | 42.412 |
| LSTM-attention-ep50 | 100.004 | 409.937 | 497.081 | 1666.853 | 1145.013 | 970.118 | 38.963 | 55.774 |
| gcn-ep20 | 126.940 | 434.270 | 410.670 | 2714.180 | 437.547 | 425.622 | 24.379 | 38.267 |

**Table A7.** SBMR-MV accuracy for a single window across all datasets.

| Dataset<br>Algorithm | ArticularyWordRecognition | EthanolConcentration | FingerMovements | SelfRegulationSCP1 | UWaveGestureLibrary |
|---|---|---|---|---|---|
| SBMR-MV-RF-W0-noTrend-D0-UG-BAG | 0.807 | 0.350 | 0.500 | 0.727 | 0.472 |
| SBMR-MV-RF-W0-noTrend-D0-UG-STACK | 0.953 | 0.399 | 0.500 | 0.734 | 0.666 |
| SBMR-MV-RF-W0-trend-D0-UG-BAG | 0.873 | 0.357 | 0.620 | 0.737 | 0.569 |
| SBMR-MV-RF-W0-trend-D0-UG-STACK | 0.963 | 0.285 | 0.570 | 0.765 | 0.700 |
| SBMR-MV-RF-W0-trend-D2-BG-BAG | 0.833 | 0.357 | 0.510 | 0.792 | 0.744 |
| SBMR-MV-RF-W0-trend-D2-BG-STACK | 0.957 | 0.331 | 0.480 | 0.819 | 0.831 |
| SBMR-MV-RF-W0-trend-D4-BG-BAG | 0.870 | 0.327 | 0.550 | 0.751 | 0.706 |
| SBMR-MV-RF-W0-trend-D4-BG-STACK | 0.963 | 0.354 | 0.550 | 0.751 | 0.838 |
| SBMR-MV-RF-W0-trend-D6-UG-BAG | 0.870 | 0.331 | 0.500 | 0.758 | 0.725 |
| SBMR-MV-RF-W0-trend-D6-UG-STACK | 0.950 | 0.331 | 0.440 | 0.788 | 0.828 |
| SBMR-MV-ridge-cv-W0-noTrend-D0-UG-BAG | 0.823 | 0.274 | 0.470 | 0.713 | 0.463 |
| SBMR-MV-ridge-cv-W0-noTrend-D0-UG-STACK | 0.990 | 0.316 | 0.480 | 0.706 | 0.653 |
| SBMR-MV-ridge-cv-W0-trend-D0-UG-BAG | 0.873 | 0.308 | 0.490 | 0.713 | 0.497 |
| SBMR-MV-ridge-cv-W0-trend-D0-UG-STACK | 0.983 | 0.323 | 0.480 | 0.734 | 0.697 |
| SBMR-MV-ridge-cv-W0-trend-D2-BG-BAG | 0.983 | 0.312 | 0.520 | 0.826 | 0.769 |
| SBMR-MV-ridge-cv-W0-trend-D2-BG-STACK | 0.987 | 0.274 | 0.500 | 0.850 | 0.891 |
| SBMR-MV-ridge-cv-W0-trend-D4-BG-BAG | 0.963 | 0.308 | 0.500 | 0.737 | 0.722 |
| SBMR-MV-ridge-cv-W0-trend-D4-BG-STACK | 0.993 | 0.312 | 0.500 | 0.761 | 0.856 |
| SBMR-MV-ridge-cv-W0-trend-D6-UG-BAG | 0.937 | 0.281 | 0.500 | 0.761 | 0.753 |

**Table A8.** SBMR-MV execution time for a single window across all datasets.

| Dataset<br>Algorithm | ArticularyWordRecognition | EthanolConcentration | FingerMovements | SelfRegulationSCP1 | UWaveGestureLibrary |
|---|---|---|---|---|---|
| SBMR-MV-RF-W0-noTrend-D0-UG-BAG | 3.105 | 14.333 | 2.333 | 12.768 | 1.810 |
| SBMR-MV-RF-W0-noTrend-D0-UG-STACK | 3.242 | 14.439 | 2.543 | 12.773 | 1.939 |
| SBMR-MV-RF-W0-trend-D0-UG-BAG | 5.992 | 26.609 | 4.489 | 25.454 | 3.495 |
| SBMR-MV-RF-W0-trend-D0-UG-STACK | 6.176 | 27.464 | 4.869 | 25.030 | 3.543 |
| SBMR-MV-RF-W0-trend-D2-BG-BAG | 27.984 | 116.769 | 21.084 | 119.072 | 16.017 |
| SBMR-MV-RF-W0-trend-D2-BG-STACK | 29.462 | 113.139 | 25.120 | 113.507 | 16.038 |
| SBMR-MV-RF-W0-trend-D4-BG-BAG | 13.639 | 58.592 | 10.274 | 57.004 | 7.791 |
| SBMR-MV-RF-W0-trend-D4-BG-STACK | 14.583 | 57.821 | 12.483 | 56.129 | 7.905 |
| SBMR-MV-RF-W0-trend-D6-UG-BAG | 12.494 | 54.386 | 9.202 | 52.044 | 7.275 |
| SBMR-MV-RF-W0-trend-D6-UG-STACK | 12.421 | 51.986 | 9.732 | 50.525 | 7.008 |
| SBMR-MV-ridge-cv-W0-noTrend-D0-UG-BAG | 3.029 | 14.389 | 2.269 | 12.406 | 1.756 |
| SBMR-MV-ridge-cv-W0-noTrend-D0-UG-STACK | 3.084 | 14.438 | 2.410 | 12.546 | 1.756 |
| SBMR-MV-ridge-cv-W0-trend-D0-UG-BAG | 5.856 | 26.680 | 4.375 | 24.617 | 3.384 |
| SBMR-MV-ridge-cv-W0-trend-D0-UG-STACK | 6.158 | 26.524 | 4.742 | 25.588 | 3.454 |
| SBMR-MV-ridge-cv-W0-trend-D2-BG-BAG | 27.256 | 111.711 | 20.511 | 113.173 | 15.684 |
| SBMR-MV-ridge-cv-W0-trend-D2-BG-STACK | 29.062 | 112.905 | 24.957 | 112.336 | 15.853 |
| SBMR-MV-ridge-cv-W0-trend-D4-BG-BAG | 13.406 | 57.022 | 10.161 | 55.374 | 7.650 |
| SBMR-MV-ridge-cv-W0-trend-D4-BG-STACK | 14.341 | 56.462 | 12.269 | 56.234 | 7.807 |
| SBMR-MV-ridge-cv-W0-trend-D6-UG-BAG | 11.783 | 51.678 | 8.750 | 50.138 | 6.808 |

**Table A9.** SBMR-MV accuracy for multiple windows across all datasets.

| Dataset<br>Algorithm | ArticularyWordRecognition | EthanolConcentration | FingerMovements | SelfRegulationSCP1 | UWaveGestureLibrary |
|---|---|---|---|---|---|
| SBMR-MV-RF-W11-trend-D6-BG-BAG | 0.897 | 0.350 | 0.570 | 0.788 | 0.831 |
| SBMR-MV-RF-W11-trend-D6-BG-STACK | 0.973 | 0.338 | 0.410 | 0.751 | 0.878 |
| SBMR-MV-RF-W15-trend-D6-BG-BAG | 0.793 | 0.327 | 0.550 | 0.771 | 0.812 |
| SBMR-MV-RF-W15-trend-D6-BG-STACK | 0.953 | 0.361 | 0.560 | 0.765 | 0.853 |
| SBMR-MV-ridge-cv-W11-trend-D4-BG-BAG-chi | 0.997 | 0.350 | 0.550 | 0.788 | 0.859 |
| SBMR-MV-ridge-cv-W11-trend-D4-BG-STACK-chi | 0.997 | 0.335 | 0.520 | 0.819 | 0.916 |
| SBMR-MV-ridge-cv-W11-trend-D6-BG-BAG | 0.993 | 0.316 | 0.510 | 0.857 | 0.863 |
| SBMR-MV-ridge-cv-W11-trend-D6-BG-BAG-chi | 0.993 | 0.342 | 0.570 | 0.823 | 0.863 |
| SBMR-MV-ridge-cv-W11-trend-D6-BG-STACK | 0.997 | 0.285 | 0.530 | 0.857 | 0.916 |
| SBMR-MV-ridge-cv-W11-trend-D6-BG-STACK-chi | 0.997 | 0.323 | 0.500 | 0.874 | 0.916 |

**Table A9.** *Cont.*

| Dataset<br>Algorithm | ArticularyWordRecognition | EthanolConcentration | FingerMovements | SelfRegulationSCP1 | UWaveGestureLibrary |
|---|---|---|---|---|---|
| SBMR-MV-ridge-cv-W14-trend-D6-BG-STACK | 0.993 | 0.297 | 0.490 | 0.833 | 0.903 |
| SBMR-MV-ridge-cv-W15-trend-D2-BG-BAG-chi | 0.987 | 0.373 | 0.610 | 0.846 | 0.850 |
| SBMR-MV-ridge-cv-W15-trend-D2-BG-STACK-chi | 0.990 | 0.342 | 0.560 | 0.850 | 0.909 |
| SBMR-MV-ridge-cv-W15-trend-D4-BG-BAG-chi | 0.993 | 0.304 | 0.540 | 0.775 | 0.884 |
| SBMR-MV-ridge-cv-W15-trend-D4-BG-STACK-chi | 0.993 | 0.327 | 0.520 | 0.826 | 0.906 |
| SBMR-MV-ridge-cv-W15-trend-D6-BG-BAG | 0.990 | 0.308 | 0.530 | 0.860 | 0.878 |
| SBMR-MV-ridge-cv-W15-trend-D6-BG-BAG-chi | 0.990 | 0.312 | 0.590 | 0.857 | 0.878 |
| SBMR-MV-ridge-cv-W15-trend-D6-BG-STACK | 0.990 | 0.281 | 0.520 | 0.843 | 0.916 |
| SBMR-MV-ridge-cv-W15-trend-D6-BG-STACK-chi | 0.990 | 0.300 | 0.590 | 0.850 | 0.916 |
| SBMR-MV-ridge-cv-W8-trend-D6-BG-BAG | 0.997 | 0.285 | 0.520 | 0.846 | 0.828 |
| SBMR-MV-ridge-cv-W8-trend-D6-BG-STACK | 0.997 | 0.289 | 0.530 | 0.853 | 0.894 |

**Table A10.** SBMR-MV execution time for multiple windows across all datasets.

| Dataset<br>Algorithm | ArticularyWordRecognition | EthanolConcentration | FingerMovements | SelfRegulationSCP1 | UWaveGestureLibrary |
|---|---|---|---|---|---|
| SBMR-MV-RF-W11-trend-D6-BG-BAG | 31.708 | 110.903 | 25.473 | 112.669 | 16.795 |
| SBMR-MV-RF-W11-trend-D6-BG-STACK | 41.358 | 115.328 | 42.269 | 119.476 | 18.377 |
| SBMR-MV-RF-W15-trend-D6-BG-BAG | 18.695 | 58.289 | 16.674 | 58.350 | 9.610 |
| SBMR-MV-RF-W15-trend-D6-BG-STACK | 25.418 | 61.010 | 28.893 | 61.293 | 10.652 |
| SBMR-MV-ridge-cv-W11-trend-D4-BG-BAG-chi | 31.945 | 111.858 | 23.221 | 114.698 | 17.014 |
| SBMR-MV-ridge-cv-W11-trend-D4-BG-STACK-chi | 40.846 | 112.995 | 32.580 | 117.638 | 18.243 |
| SBMR-MV-ridge-cv-W11-trend-D6-BG-BAG | 38.256 | 145.062 | 31.322 | 143.597 | 20.178 |
| SBMR-MV-ridge-cv-W11-trend-D6-BG-BAG-chi | 32.462 | 112.871 | 23.014 | 112.695 | 17.219 |
| SBMR-MV-ridge-cv-W11-trend-D6-BG-STACK | 40.630 | 115.386 | 44.945 | 120.181 | 18.654 |
| SBMR-MV-ridge-cv-W11-trend-D6-BG-STACK-chi | 41.561 | 113.599 | 32.276 | 115.983 | 18.670 |
| SBMR-MV-ridge-cv-W14-trend-D6-BG-STACK | 49.472 | 160.643 | 47.373 | 167.632 | 24.661 |
| SBMR-MV-ridge-cv-W15-trend-D2-BG-BAG-chi | 38.014 | 112.522 | 27.867 | 113.787 | 19.632 |
| SBMR-MV-ridge-cv-W15-trend-D2-BG-STACK-chi | 55.559 | 116.257 | 42.339 | 120.743 | 22.030 |
| SBMR-MV-ridge-cv-W15-trend-D4-BG-BAG-chi | 18.775 | 58.028 | 14.462 | 57.993 | 9.299 |
| SBMR-MV-ridge-cv-W15-trend-D4-BG-STACK-chi | 25.860 | 58.994 | 21.088 | 59.509 | 10.347 |
| SBMR-MV-ridge-cv-W15-trend-D6-BG-BAG | 17.784 | 57.440 | 16.034 | 56.868 | 9.303 |
| SBMR-MV-ridge-cv-W15-trend-D6-BG-BAG-chi | 18.816 | 58.599 | 13.885 | 57.084 | 9.833 |
| SBMR-MV-ridge-cv-W15-trend-D6-BG-STACK | 24.287 | 58.777 | 27.667 | 60.622 | 10.238 |
| SBMR-MV-ridge-cv-W15-trend-D6-BG-STACK-chi | 28.181 | 63.833 | 22.487 | 65.002 | 11.692 |
| SBMR-MV-ridge-cv-W8-trend-D6-BG-BAG | 141.573 | 543.507 | 105.930 | 566.511 | 79.861 |
| SBMR-MV-ridge-cv-W8-trend-D6-BG-STACK | 157.743 | 546.896 | 145.858 | 560.055 | 79.709 |

**Table A11.** Neural network architecture accuracy for multiple windows across all datasets.

| Dataset<br>Algorithm | ArticularyWordRecognition | EthanolConcentration | FingerMovements | SelfRegulationSCP1 | UWaveGestureLibrary |
|---|---|---|---|---|---|
| LSTM-attention-cnn2-ep20-bag | 0.470 | 0.255 | 0.530 | 0.853 | 0.394 |
| LSTM-attention-cnn2-ep20-stack | 0.927 | 0.266 | 0.500 | 0.498 | 0.653 |
| LSTM-attention-cnn2-ep50-bag | 0.647 | 0.266 | 0.600 | 0.860 | 0.531 |
| LSTM-attention-cnn2-ep50-stack | 0.907 | 0.346 | 0.530 | 0.853 | 0.756 |
| LSTM-CNN-ep20-stack | 0.913 | 0.247 | 0.510 | 0.638 | 0.562 |
| LSTM-CNN-ep50-stack | 0.887 | 0.327 | 0.480 | 0.686 | 0.625 |

**Table A12.** Neural network architecture execution time across all datasets.

| Dataset<br>Algorithm | ArticularyWordRecognition | EthanolConcentration | FingerMovements | SelfRegulationSCP1 | UWaveGestureLibrary |
|---|---|---|---|---|---|
| LSTM-attention-cnn2-ep20-bag | 30.993 | 85.282 | 17.878 | 96.897 | 17.104 |
| LSTM-attention-cnn2-ep20-stack | 12.719 | 48.457 | 11.653 | 32.342 | 10.190 |
| LSTM-attention-cnn2-ep50-bag | 45.070 | 71.921 | 19.318 | 73.720 | 26.229 |
| LSTM-attention-cnn2-ep50-stack | 22.647 | 67.122 | 14.955 | 38.722 | 16.491 |
| LSTM-CNN-ep20-stack | 10.642 | 32.644 | 9.252 | 19.476 | 8.247 |
| LSTM-CNN-ep50-stack | 13.897 | 119.332 | 12.221 | 45.039 | 13.111 |

**Table A13.** State-of-the-art neural network classifiers' accuracy in all datasets.

| Dataset<br>Algorithm | ArticularyWordRecognition | EthanolConcentration | FingerMovements | SelfRegulationSCP1 | UWaveGestureLibrary |
|---|---|---|---|---|---|
| CNNClassifier-ep20 | 0.083 | 0.255 | 0.460 | 0.846 | 0.191 |
| CNNClassifier-ep200 | 0.457 | 0.297 | 0.530 | 0.877 | 0.806 |
| EncoderClassifier-ep20 | 0.847 | 0.232 | 0.540 | 0.812 | 0.734 |
| IInception-ep20 | 0.920 | 0.255 | 0.480 | 0.778 | 0.456 |
| IndividualLITEClassifier-ep100 | 0.987 | 0.281 | 0.550 | 0.696 | 0.228 |
| IndividualLITEClassifier-ep50 | 0.930 | 0.308 | 0.510 | 0.884 | 0.244 |
| ResNet-ep20 | 0.163 | 0.247 | 0.460 | 0.502 | 0.125 |

**Table A14.** State-of-the-art neural network classifiers' execution time in all datasets.

| Dataset<br>Algorithm | ArticularyWordRecognition | EthanolConcentration | FingerMovements | SelfRegulationSCP1 | UWaveGestureLibrary |
|---|---|---|---|---|---|
| CNNClassifier-ep20 | 1.969 | 3.426 | 1.835 | 2.652 | 1.631 |
| CNNClassifier-ep200 | 9.026 | 20.174 | 6.734 | 14.693 | 6.750 |
| EncoderClassifier-ep20 | 41.563 | 334.121 | 27.218 | 199.412 | 37.719 |
| IInception-ep20 | 52.853 | 414.763 | 32.317 | 222.469 | 46.094 |
| IndividualLITEClassifier-ep100 | 38.759 | 257.808 | 28.145 | 144.833 | 31.622 |
| IndividualLITEClassifier-ep50 | 20.604 | 138.162 | 16.901 | 74.098 | 17.632 |
| ResNet-ep20 | 33.791 | 283.649 | 19.412 | 147.360 | 31.053 |

**Table A15.** Rocket accuracy across all datasets.

| Dataset<br>Algorithm | ArticularyWordRecognition | EthanolConcentration | FingerMovements | SelfRegulationSCP1 | UWaveGestureLibrary |
|---|---|---|---|---|---|
| ROCKET | 0.993 | 0.426 | 0.530 | 0.850 | 0.934 |
| miniRocket | 0.990 | 0.475 | 0.500 | 0.918 | 0.941 |

**Table A16.** Rocket execution time across all datasets.

| Dataset<br>Algorithm | ArticularyWordRecognition | EthanolConcentration | FingerMovements | SelfRegulationSCP1 | UWaveGestureLibrary |
|---|---|---|---|---|---|
| ROCKET | 40.319 | 284.413 | 10.389 | 215.686 | 43.534 |
| miniRocket | 3.452 | 16.734 | 1.318 | 12.168 | 3.523 |

## References


1. Chaovalitwongse, W.A.; Prokopyev, O.A.; Pardalos, P.M. Electroencephalogram (EEG) time series classification: Applications in epilepsy. *Ann. Oper. Res.* **2006**, *148*, 227–250. [CrossRef]
2. Arul, M.; Kareem, A. Applications of shapelet transform to time series classification of earthquake, wind and wave data. *Eng. Struct.* **2021**, *228*, 111564. [CrossRef]
3. Potamitis, I. Classifying insects on the fly. *Ecol. Inform.* **2014**, *21*, 40–49. [CrossRef]
4. Susto, G.A.; Cenedese, A.; Terzi, M. Chapter 9—Time-Series Classification Methods: Review and Applications to Power Systems Data. In *Big Data Application in Power Systems*; Arghandeh, R., Zhou, Y., Eds.; Elsevier: Amsterdam, The Netherlands, 2018; pp. 179–220. [CrossRef]
5. Ismail Fawaz, H.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.A. Evaluating surgical skills from kinematic data using convolutional neural networks. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, 16–20 September 2018; Proceedings, Part IV 11; Springer: Berlin/Heidelberg, Germany, 2018; pp. 214–221.
6. Tao, L.; Elhamifar, E.; Khudanpur, S.; Hager, G.D.; Vidal, R. Sparse hidden markov models for surgical gesture classification and skill evaluation. In Proceedings of the Information Processing in Computer-Assisted Interventions: Third International Conference, IPCAI 2012, Pisa, Italy, 27 June 2012; Proceedings 3; Springer: Berlin/Heidelberg, Germany, 2012; pp. 167–177.
7. Forestier, G.; Lalys, F.; Riffaud, L.; Trelhu, B.; Jannin, P. Classification of surgical processes using dynamic time warping. *J. Biomed. Inform.* **2012**, *45*, 255–264. [CrossRef]
8. Devanne, M.; Wannous, H.; Berretti, S.; Pala, P.; Daoudi, M.; Del Bimbo, A. 3-D human action recognition by shape analysis of motion trajectories on riemannian manifold. *IEEE Trans. Cybern.* **2014**, *45*, 1340–1352. [CrossRef]
9. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231. [CrossRef] [PubMed]
10. Pinto, J.P.; Pimenta, A.; Novais, P. Deep learning and multivariate time series for cheat detection in video games. *Mach. Learn.* **2021**, *110*, 3037–3057. [CrossRef]
11. Younis, R.; Zerr, S.; Ahmadi, Z. Multivariate time series analysis: An interpretable cnn-based model. In Proceedings of the 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), Shenzhen, China, 13–16 October 2022; IEEE: Piscataway Township, NJ, USA, 2022; pp. 1–10.
12. Madrid, F.; Singh, S.; Chesnais, Q.; Mauck, K.; Keogh, E. Matrix profile xvi: Efficient and effective labeling of massive time series archives. In Proceedings of the 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Washington, DC, USA, 5–8 October 2019; IEEE: Piscataway Township, NJ, USA, 2019; pp. 463–472.
13. Devanne, M.; Rémy-Néris, O.; Le Gals-Garnett, B.; Kermarrec, G.; Thepaut, A. A co-design approach for a rehabilitation robot coach for physical rehabilitation based on the error classification of motion errors. In Proceedings of the 2018 Second IEEE International Conference on Robotic Computing (IRC), Laguna Hills, CA, USA, 31 January–2 February 2018; IEEE: Piscataway Township, NJ, USA, 2018; pp. 352–357.
14. Schäfer, P.; Högqvist, M. SFA: A symbolic fourier approximation and index for similarity search in high dimensional datasets. In Proceedings of the 15th International Conference on Extending Database Technology, Berlin, Germany, 27–30 March 2012; ACM: New York, NY, USA, 2012; pp. 516–527.
15. Glenis, A.; Vouros, G.A. SCALE-BOSS-MR: Scalable Time Series Classification Using Multiple Symbolic Representations. *Appl. Sci.* **2024**, *14*, 689. [CrossRef]
16. Lin, J.; Keogh, E.; Wei, L.; Lonardi, S. Experiencing SAX: A novel symbolic representation of time series. *Data Min. Knowl. Discov.* **2007**, *15*, 107–144. [CrossRef]
17. Senin, P.; Malinchik, S. Sax-vsm: Interpretable time series classification using sax and vector space model. In Proceedings of the 2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, 7–10 December 2013; IEEE: Piscataway Township, NJ, USA, 2013; pp. 1175–1180.
18. Schäfer, P. The BOSS is concerned with time series classification in the presence of noise. *Data Min. Knowl. Discov.* **2015**, *29*, 1505–1530. [CrossRef]
19. Schäfer, P. Scalable time series classification. *Data Min. Knowl. Discov.* **2016**, *30*, 1273–1298.
20. Schäfer, P.; Leser, U. Fast and accurate time series classification with weasel. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; ACM: New York, NY, USA, 2017; pp. 637–646.
21. Glenis, A.; Vouros, G.A. Balancing between scalability and accuracy in time-series classification for stream and batch settings. In Proceedings of the Discovery Science: 23rd International Conference, DS 2020, Thessaloniki, Greece, 19–21 October 2020; Proceedings 23; Springer: Berlin/Heidelberg, Germany, 2020; pp. 265–279.
22. Nguyen, T.L.; Ifrim, G. MrSQM: Fast time series classification with symbolic representations. *arXiv* **2021**, arXiv:2109.01036.

23. Nguyen, T.L.; Ifrim, G. Fast time series classification with random symbolic subsequences. In Proceedings of the International Workshop on Advanced Analytics and Learning on Temporal Data, Grenoble, France, 19–23 September 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 50–65.
24. Glenis, A.; Vouros, G.A. SCALE-BOSS: A framework for scalable time-series classification using symbolic representations. In Proceedings of the 12th Hellenic Conference on Artificial Intelligence, Corfu, Greece, 7–9 September 2022; pp. 1–9.
25. Schäfer, P.; Leser, U. WEASEL 2.0–A Random Dilated Dictionary Transform for Fast, Accurate and Memory Constrained Time Series Classification. *arXiv* **2023**, arXiv:2301.10194.
26. Schäfer, P.; Leser, U. Multivariate time series classification with WEASEL+ MUSE. *arXiv* **2017**, arXiv:1711.11343.
27. Dempster, A.; Petitjean, F.; Webb, G.I. ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Min. Knowl. Discov.* **2020**, *34*, 1454–1495. [CrossRef]
28. Dempster, A.; Schmidt, D.F.; Webb, G.I. Minirocket: A very fast (almost) deterministic transform for time series classification. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Singapore, 14–18 August 2021; pp. 248–257.
29. Wang, Z.; Yan, W.; Oates, T. Time series classification from scratch with deep neural networks: A strong baseline. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; IEEE: Piscataway Township, NJ, USA, 2017; pp. 1578–1585.
30. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
31. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 807–814.
32. Ioffe, S. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167. [CrossRef]
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
34. Monti, F.; Frasca, F.; Eynard, D.; Mannion, D.; Bronstein, M.M. Fake news detection on social media using geometric deep learning. *arXiv* **2019**, arXiv:1902.06673. [CrossRef]
35. Zhao, B.; Lu, H.; Chen, S.; Liu, J.; Wu, D. Convolutional neural networks for time series classification. *J. Syst. Eng. Electron.* **2017**, *28*, 162–169. [CrossRef]
36. Serra, J.; Pascual, S.; Karatzoglou, A. Towards a universal neural network encoder for time series. In *Artificial Intelligence Research and Development*; IOS Press: Amsterdam, The Netherlands, 2018; pp. 120–129.
37. Ismail Fawaz, H.; Lucas, B.; Forestier, G.; Pelletier, C.; Schmidt, D.F.; Weber, J.; Webb, G.I.; Idoumghar, L.; Muller, P.A.; Petitjean, F. Inceptiontime: Finding alexnet for time series classification. *Data Min. Knowl. Discov.* **2020**, *34*, 1936–1962. [CrossRef]
38. Ismail-Fawaz, A.; Devanne, M.; Weber, J.; Forestier, G. Deep learning for time series classification using new hand-crafted convolution filters. In Proceedings of the 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 17–20 December 2022; IEEE: Piscataway Township, NJ, USA, 2022; pp. 972–981.
39. Ismail-Fawaz, A.; Devanne, M.; Berretti, S.; Weber, J.; Forestier, G. Lite: Light inception with boosting techniques for time series classification. In Proceedings of the 2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA), Thessaloniki, Greece, 9–13 October 2023; IEEE: Piscataway Township, NJ, USA, 2023; pp. 1–10.
40. Cui, Z.; Chen, W.; Chen, Y. Multi-scale convolutional neural networks for time series classification. *arXiv* **2016**, arXiv:1603.06995. [CrossRef]
41. Wickstrøm, K.; Kampffmeyer, M.; Mikalsen, K.Ø.; Jenssen, R. Mixing up contrastive learning: Self-supervised representation learning for time series. *Pattern Recognit. Lett.* **2022**, *155*, 54–61. [CrossRef]
42. Ordonez, P.; Armstrong, T.; Oates, T.; Fackler, J. Using modified multivariate bag-of-words models to classify physiological data. In Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, Canada, 11 December 2011; IEEE: Piscataway Township, NJ, USA, 2011; pp. 534–539.
43. Baydogan, M.G.; Runger, G. Learning a symbolic representation for multivariate time series classification. *Data Min. Knowl. Discov.* **2015**, *29*, 400–422. [CrossRef]
44. Ruiz, A.P.; Flynn, M.; Large, J.; Middlehurst, M.; Bagnall, A. The great multivariate time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Min. Knowl. Discov.* **2021**, *35*, 401–449. [CrossRef] [PubMed]
45. Shervashidze, N.; Schweitzer, P.; Van Leeuwen, E.J.; Mehlhorn, K.; Borgwardt, K.M. Weisfeiler-Lehman graph kernels. *J. Mach. Learn. Res.* **2011**, *12*, 2539–2561.
46. LeCun, Y. Generalization and network design strategies. *Connect. Perspect.* **1989**, *19*, 18.

47. LeCun, Y.; Bengio, Y. Convolutional networks for images, speech, and time series. *Handb. Brain Theory Neural Netw.* **1995**, *3361*, 1995.
48. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*; MIT Press: Cambridge, MA, USA, 1986; pp. 318–362.
49. Elman, J.L. Finding structure in time. *Cogn. Sci.* **1990**, *14*, 179–211. [CrossRef]
50. Jordan, M.I. Serial order: A parallel distributed processing approach. In *Advances in Psychology*; Elsevier: Amsterdam, The Netherlands, 1997; Volume 121; pp. 471–495.
51. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
52. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
53. Sugiyama, M.; Borgwardt, K. Halting in random walk kernels. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1639–1647.
54. Faouzi, J.; Janati, H. pyts: A Python Package for Time Series Classification. *J. Mach. Learn. Res.* **2020**, *21*, 1–6.
55. Siglidis, G.; Nikolentzos, G.; Limnios, S.; Giatsidis, C.; Skianis, K.; Vazirgiannis, M. Grakel: A graph kernel library in python. *J. Mach. Learn. Res.* **2020**, *21*, 1–5.
56. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
57. Chollet, F. *Deep Learning with Python*; Simon and Schuster: New York, NY, USA, 2021.
58. Data61, C. StellarGraph Machine Learning Library. 2018. Available online: https://github.com/stellargraph/stellargraph (accessed on 9 September 2025).
59. Dau, H.A.; Bagnall, A.; Kamgar, K.; Yeh, C.C.M.; Zhu, Y.; Gharghabi, S.; Ratanamahatana, C.A.; Keogh, E. The UCR time series archive. *IEEE/CAA J. Autom. Sin.* **2019**, *6*, 1293–1305. [CrossRef]
60. Middlehurst, M.; Ismail-Fawaz, A.; Guillaume, A.; Holder, C.; Guijo-Rubio, D.; Bulatova, G.; Tsaprounis, L.; Mentel, L.; Walter, M.; Schäfer, P.; et al. aeon: A Python toolkit for learning from time series. *J. Mach. Learn. Res.* **2024**, *25*, 1–10.


**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.

computers

MDPI

Article

# Development of an Early Lung Cancer Diagnosis Method Based on a Neural Network

**Indira Karymsakova [1], Dinara Kozhakhmetova [1,*], Dariga Bekenova [2], Danila Ostroukh [1], Roza Bekbayeva [1], Lazat Kydyralina [1], Alina Bugubayeva [3] and Dinara Kurushbayeva [1]**

[1] Graduate School Digital Technologies and Construction, Department of Automation and Information Technologies, Shakarim University, St. Glinka, 20A, Semey 071412, Kazakhstan; karymsakovaib@gmail.com (I.K.); dao.ost@yandex.kz (D.O.); roza.toghan3194@gmail.com (R.B.); lazat_75@mail.ru (L.K.); sh_din097@mail.ru (D.K.)

[2] Department of Information Technologies, Higher School of Business and Digital Technologies, University "Turan-Astana", Y Dukenuly, 29a Street, Astana 010000, Kazakhstan; darigs111@gmail.com

[3] Department of Digital Engineering and IT-Analytics, Faculty of Finance, Logistics and Digital Technologies, Karaganda University of Kazpotrebsouz, 9 Academic St., Karaganda 100009, Kazakhstan; alina.bugubayeva88@gmail.com

* Correspondence: d.kojahmetova@shakarim.kz

**Abstract:** Cancer is one of the most lethal diseases in the modern world. Early diagnosis significantly contributes to prolonging the life expectancy of patients. The application of intelligent systems and AI methods is crucial for diagnosing oncological diseases. Primarily, expert systems or decision support systems are utilized in such cases. This research explores early lung cancer diagnosis through protocol-based questioning, considering the impact of nuclear testing factors. Nuclear tests conducted historically continue to affect citizens' health. A classification of regions into five groups was proposed based on their proximity to nuclear test sites. The weighting coefficient was assigned accordingly, in proportion to the distance from the test zones. In this study, existing expert systems were analyzed and classified. Approaches used to build diagnostic expert systems for oncological diseases were grouped by how well they apply to different tumor localizations. An online questionnaire based on the lung cancer diagnostic protocol was created to gather input data for the neural network. To support this diagnostic method, a functional block diagram of the intelligent system "Oncology" was developed. The following methods were used to create the mathematical model: gradient boosting, multilayer perceptron, and Hamming network. Finally, a web application architecture for early lung cancer detection was proposed.

**Keywords:** expert system; information systems; artificial intelligence; neural network; Hamming network; analysis; intellectual systems; databases

## 1. Introduction

The high incidence of oncological diseases across various age groups is a critical issue. Hence, early detection and prevention of cancer remain crucial tasks. In regions that were exposed to nuclear testing, there is a high incidence of cancer, and hereditary factors also play a role. The hereditary factor plays a significant role in cancer development. Lung cancer does have a genetic origin in a subset of cases, primarily among young patients with adenocarcinoma [1].

In this regard, one of the most important tasks is the early detection of oncological diseases and their prevention. Early diagnosis and accurate diagnostic assessment enhance the

likelihood of effective treatment measures and extend patient survival. When diagnosing lung cancer, heredity and smoking play an important role.

The integration and utilization of information technologies in medicine are rapidly advancing. Currently, numerous decision support systems (DSSs) exist within the medical field, widely applied in preliminary diagnostics for various diseases, including oncology, where expert systems have been significantly developed. Depending on the type of disease and localization, different mathematical apparatus is used in these expert systems.

Development of a method for early diagnosis of lung cancer will allow us to identify the disease at an early stage and to take the necessary measures to preserve the full life of patients. The proposed expert system-based method for early detection of lung cancer facilitates the assessment of individual risk levels for lung cancer, as well as for other pulmonary conditions. This approach has been validated in clinical settings and has demonstrated promising results. Its implementation could enable timely intervention in high-risk cases, thereby improving early diagnostic outcomes.

## 2. Materials and Methods

The present study uses methods of scientific analysis, synthesis, and classification of approaches in the development of expert systems, along with the application of Hamming neural networks.

Numerous expert systems have been developed for diagnosing oncological diseases, employing diverse approaches and methodologies. For instance, Karabatak et al. described constructing an expert system for breast cancer detection using association rules (ARs) and neural networks (NNs). This method involves developing an automated diagnostic system based on AR and NN, employing a three-fold cross-validation method for performance evaluation during testing. This approach facilitates quick automatic diagnostic systems for other diseases; however, it has a high dependency on feature selection [2].

Oyelade et al. proposed an approach utilizing semantic networks and natural language processing (NLP). They introduced a formalized input generation model addressing the challenge of raw data inputs through inference processes, a breast cancer lexicon, rule sets, and NLP. An input-generation algorithm using Python's 3.13.1 NLP capabilities initially filters and generates a preliminary input collection. This collection serves as input for the inference mechanism, which employs ontology-based rules and lexicon to generate valid tokens, subsequently processed by medical expert systems or diagnostic decision support systems (DDSSs). This model improved performance by 64% compared to systems processing raw patient data [3].

Physical research methods for expert systems in oncology diagnostics were proposed in [4], focusing on multimedia training systems for physicians. Disease localizations studied include esophagus, stomach, intestines, pancreas, and liver. The system's practical application enhances the efficiency of oncologists' postgraduate training and professional development [4].

Shereen A. Taie et al. developed a prototype expert system evaluating breast cancer risk based on criterion-based parameters. The system assesses lifestyle factors against specified criteria to determine breast cancer development risk, providing preventive recommendations that encourage lifestyle modifications. Although knowledge within this system derives from expert opinions and published research, further expert validation is needed for practical application [5].

Rahul Boadh et al. discussed fuzzy expert systems for diagnosing various cancers. Fuzzy reasoning has recently emerged as a dynamic strategy in disease recognition, particularly for malignant neoplasms. While identifying malignancies involves uncertainties and errors, fuzzy-based models have significantly advanced diagnostics. Integrating fuzzy

logic and expert recommendations allows early analysis of human infections, reducing mortality. The Ant Colony Optimization (ACO) algorithm optimized fuzzy IF/THEN rule sets for enhanced diagnostic performance [6].

Intelligent screening methods for diagnosing and treating oncological diseases were introduced in [7]. This approach predicts patient conditions beyond medical history by simulating qualified physician prognostic processes and therapeutic decision-making. Unlike traditional statistical models, this approach builds knowledge databases based on patterns from similar patient situations. These knowledge-based models provide superior accuracy compared to classical statistical models [7]. Ritna Wahyuni et al. proposed a hybrid expert system construction method employing forward chaining for symptom tracking and certainty factor methods for accuracy in symptom condition assessment. The system facilitates early education in diagnosing and treating lymph node diseases, achieving an 80% user accuracy rating during initial system testing [8].

Tanyukevich et al. introduced an expert system classifying cancer risk groups based on a production model for formalizing medical knowledge. This study presented a Markov chain-based model for predicting cancer patient numbers, incorporating standard oncological classifications and latent disease periods. This expert system enables personalized risk-reduction programs, outperforming traditional systems [9].

Ali Keleş et al. investigated a neuro-fuzzy rule-based expert system for breast cancer diagnostics. Neuro-fuzzy methods identified rules employed in the Ex-DBC system's inference engine, providing reliable diagnostic performance with 97% specificity, 76% sensitivity, 96% positive predictive value, and 81% negative predictive value. High positive predictive value facilitates unnecessary biopsy avoidance and can also serve educational purposes for medical students [10].

Shinsuke Morio et al. developed an expert system for early breast cancer detection aimed at nonspecialist users. The system facilitates interactions between users and a microcomputer, collecting symptom descriptions and offering preventive advice. The system, implemented in Prolog, outlines early detection methods and suggests subsequent actions [11].

In [12], new approaches using fuzzy expert systems for breast cancer prediction and diagnosis were examined. This study investigated developing an enhanced mobile fuzzy expert system (EMFES) employing type-2 fuzzy logic for accuracy and uncertainty management. Key aspects included comparative effectiveness, uncertainty handling, mobile feature integration, programming language selection, and dynamic fuzzy rule creation. Challenges and deployment opportunities for mobile EMFES were explored, highlighting the necessity for comprehensive, adaptable mobile solutions [13].

A Mamdani inference system with 67 fuzzy rules derived from expert opinions was implemented in MATLAB 2017a [14] and evaluated through 50 clinical case scenarios. Sensitivity was 92.1%, and specificity was 83.1%, demonstrating effectiveness [14–19].

The authors in [20] present a fuzzy expert system for diagnosing heart diseases. The system receives precise input data and performs a fuzzification step. After that, defuzzification is carried out to convert the fuzzy output of the system into a crisp value, which is easier for humans to interpret. The system measures the risk level of heart disease, providing one of the following outputs: low, high, or risky. The results demonstrate that the system is effective in identifying the appropriate risk levels [20].

In [21], the authors propose a hybrid approach wherein probabilistic rules generated from a decision tree are integrated with expert-provided rules. The aim of this integration is to augment the data derived from the decision tree with expert knowledge. The system is applied to the diagnosis of diabetes mellitus. Experimental results demonstrate that this combined approach outperforms the use of each rule set independently [21].

In [22], the Hamming method is applied. Subset accuracy evaluates the proportion of correctly classified labels (identical to the ground truth), which is generally considered too strict a metric for multi-label classification models. Hamming loss assesses performance based on errors in instance-label pair predictions; an error here indicates either a relevant label is missed or an irrelevant label is predicted [22].

In [23], it was established that the choice of similarity function has a significant impact on the system's performance. The random forest similarity function was found to be the most accurate for the given datasets. This study highlights the importance of careful selection of the similarity function when developing a case-based reasoning system for medical diagnosis [23].

Approaches for expert systems in cancer diagnosis were analyzed, identifying strengths and weaknesses, and classified according to disease localization applicability.

## 3. Results

Analyzing these approaches enables classification according to applicability, as discussed in Table 1:

- Class 1: Methods applicable across various localizations.
- Class 2: Methods applicable across fewer localizations.
- Class 3: Methods applicable exclusively to specific localizations.

**Table 1.** Classification of approaches in expert systems.

| Class | Approaches |
|---|---|
| Class 1 | Neural network approach |
| Class 2 | Approach using factor analysis and physical research methods |
| Class 3 | Full formalization, associative rules, semantic networks, natural language analysis, and fuzzy logic approaches |

The approaches used in the development of expert systems for cancer diagnosis were classified based on their applicability depending on the localization of the disease.

Class 1—most frequently used methods.

Class 2—less frequently used methods.

Class 3—least frequently used methods.

The conducted classification of approaches used in expert systems allows identifying the most optimal methods applicable across a broad range of localizations for subsequent use in developing a prototype of an information system for cancer diagnostics [24–30]. Based on this analysis, the decision was made to utilize neural networks in constructing an expert system for preliminary diagnosis.

The following stages are proposed for building an intelligent network based on neural networks for analyzing and providing recommendations for cancer treatment:

- Knowledge acquisition and analysis, identification, and conceptualization.
- Knowledge modeling and formalization.
- Development of a script-based question database and neural network training.
- Creation of an intelligent system prototype.

1. Knowledge acquisition and analysis, identification, and conceptualization.

At this stage, an analysis of the subject domain was performed, tasks were comprehended, requirements were formed, and an informal description was compiled highlighting

the main characteristics, key concepts/objects, their input/output values, the anticipated type of solution, and relevant knowledge.

2. Knowledge modeling and formalization.

Data collection regarding medical indicators of patients diagnosed with lung cancer over the past five years within the region was conducted, and primary risk assessment indicators for lung cancer susceptibility were identified. Medical indicators and responses were determined for inclusion in a survey questionnaire, forming the knowledge base of the intelligent diagnostic system. This questionnaire was developed as an online survey following the lung cancer diagnostic protocol.

Two additional questions regarding patients' place of birth and residence were included to consider the radioactive contamination factor resulting from nuclear testing. The weight of each response is allocated proportionally to its overall significance to facilitate subsequent development of the Hamming neural network.

To simulate oncological experts' reasoning, a total of 21 rules were derived based on medical expert opinions. The relative weight of each response was assigned by oncologists based on the results of studies involving data from cancer patients. Each rule includes 2 to 5 responses, each with a numerical value corresponding to the degree of impact in Table 1 [31].

The online survey was completed by 1283 patients. The results of the online survey showed the following percentages in Tables 2 and 3.

**Table 2.** Rules and answers of the expert system.

| **Question** | **Answers** | | | | |
|---|---|---|---|---|---|
| Entry 1 | 1st answer | 2nd answer | 3rd answer | 4th answer | 5th answer |
| Age | 30–40 | 40–50 | 50–60 | 60 and above | 20–30 |
| | 0.2 | 0.5 | 0.5 | 0.5 | |
| Gender | Male | Female | | | |
| | 1 | 0.6 | | | |
| Do you smoke? | Yes | No | | | |
| | 1 | 0.5 | | | |
| Smoking experience | Up to 10 years | From 10 to 20 years | From 20 to 30 years | From 30 and above | I do not smoke |
| | 0.7 | 0.9 | 1 | 1 | |
| How many packs/pieces of cigarettes do you consume per day? | Up to 10 pieces | Up to 1 pack | 1 pack | From 1 pack and above | |
| | 0.8 | 1 | 1 | 1 | |
| How many times a year do you get sick with ARVI? | 1 time | 2 times | periodically | Was not sick | |
| | 0.4 | 0.5 | 0.8 | 0.4 | |
| Oncologic history of parents and close relatives | One of the parents | Both parents | Among close relatives on the maternal side | Among close relatives on the paternal side | No one has |
| | 0.9 | 0.9 | 0.5 | 0.5 | |
| Have you ever had hemoptysis? | Yes | No | | | |
| | 0.9 | 0.4 | | | |
| Do you ever have shortness of breath, a feeling of lack of air? | Periodically | Often | During physical exertion | | |
| | 0.5 | 1 | 1 | | |
| Was there a change in the voice? | Yes | No | | | |
| | 0.8 | 0.2 | | | |

**Table 2.** *Cont.*

| Question | Answers | | | | |
|---|---|---|---|---|---|
| Do you ever have weakness that is not related to anything? | Periodically | Often | No | | |
| | 0.3 | 0.5 | 0.2 | | |
| Have you been coughing? | Not at all | A little | Not so little | Very much | |
| | 0.2 | 0.2 | 0.3 | 0.8 | |
| Have you had problems swallowing? | Not at all | A little | Not so little | Very much | |
| | 0.3 | 0.3 | 0.4 | 1 | |
| Have you had chest pain? | Not at all | A little | Not so little | Very much | |
| | 0.5 | 0.6 | 1 | 1 | |
| Have you had pain in your arm or shoulders? | Not at all | A little | Not so little | Very much | |
| | 0.5 | 0.5 | 0.7 | 1 | |
| Have you ever had a dry cough? | Not at all | A little | Not so little | Very much | |
| | 0.4 | 0.6 | 0.8 | 0.8 | |
| Have you experienced unrelated weight loss? | Not at all | A little | Not so little | Very much | |
| | 0.5 | 0.5 | 1 | | |
| Have you experienced a loss of appetite? | Not at all | A little | Not so little | Very much | |
| | 1 | 1 | 1 | 1 | |
| Are you registered with a pulmonologist for other lung diseases? | Yes | No | | | |
| | 1 | 0.5 | | | |
| Birthplace | Region 1 | Region 2 | Region 3 | Region 4 | Other |
| | 0.3 | 0.5 | 0.8 | 0.9 | 0.3 |
| Place of residence | Region 1 | Region 2 | Region 3 | Region 4 | Other |
| | 0.3 | 0.5 | 0.8 | 0.9 | 0.3 |

**Table 3.** Percentage of responses received.

| Question | Answers | | | | |
|---|---|---|---|---|---|
| Entry 1 | 1st answer | 2nd answer | 3rd answer | 4th answer | 5th answer |
| Age | 30–40 | 40–50 | 50–60 | 60 and above | 20–30 |
| | 12.3 | 18.3 | 18.7 | 11 | 19.6 |
| Gender | Male | Female | | | |
| | 46.6 | 53.4 | | | |
| Do you smoke? | Yes | No | | | |
| | 24.7 | 75.3 | | | |
| Smoking experience | Up to 10 years | From 10 to 20 years | From 20 to 30 years | From 30 and above | I do not smoke |
| | 9.4 | 5.7 | 6.6 | 1.9 | 76.4 |
| How many packs/pieces of cigarettes do you consume per day? | Up to 10 pieces | Up to 1 pack | 1 pack | From 1 pack and above | |
| | 25.4 | 41.3 | 9.5 | 23.8 | |
| How many times a year do you get sick with ARVI? | 1 time | 2 times | periodically | Was not sick | |
| | 26.5 | 22.8 | 19.2 | 31.5 | |

**Table 3.** *Cont.*

| Question | Answers | | | | |
|---|---|---|---|---|---|
| Oncologic history of parents and close relatives | One of the parents | Both parents | Among close relatives on the maternal side | Among close relatives on the paternal side | No one has |
| | 15.1 | 4.1 | 5.5 | 7.8 | 67.6 |
| Have you ever had hemoptysis? | Yes | No | | | |
| | 93.6 | 6.4 | | | |
| Do you ever have shortness of breath, a feeling of lack of air? | Periodically | Often | During physical exertion | | |
| | 5.9 | 2.3 | 31.3 | 60.7 | |
| Was there a change in the voice? | Yes | No | | | |
| | 18.7 | 81.3 | | | |
| Do you ever have weakness that is not related to anything? | Periodically | Often | No | | |
| | 26.9 | 10.5 | 62.6 | | |
| Have you been coughing? | Not at all | A little | Not so little | Very much | |
| | 31.1 | 53 | 8.2 | 7.8 | |
| Have you had problems swallowing? | Not at all | A little | Not so little | Very much | |
| | 63 | 26.9 | 7.3 | 2.7 | |
| Have you had chest pain? | Not at all | A little | Not so little | Very much | |
| | 60.3 | 31.5 | 5.5 | 2.7 | |
| Have you had pain in your arm or shoulders? | Not at all | A little | Not so little | Very much | |
| | 49.8 | 38.8 | 7.3 | 4.1 | |
| Have you ever had a dry cough? | Not at all | A little | Not so little | Very much | |
| | 40.2 | 48.9 | 6.4 | 4.6 | |
| Have you experienced unrelated weight loss? | Not at all | A little | Not so little | Very much | |
| | 72.1 | 19.6 | 8.2 | | |
| Have you experienced a loss of appetite? | Not at all | A little | Not so little | Very much | |
| | 62.1 | 26 | 6.4 | 5.5 | |
| Are you registered with a pulmonologist for other lung diseases? | Yes | No | | | |
| | 7.8 | 92.2 | | | |
| Birthplace | Region 1 | Region 2 | Region 3 | Region 4 | Other |
| | 0.3 | 0.5 | 0.8 | 0.9 | 0.3 |
| Place of residence | Region 1 | Region 2 | Region 3 | Region 4 | Other |
| | 0.3 | 0.5 | 0.8 | 0.9 | 0.3 |

3. Development of a script-based question database and neural network training.

At this stage, a database of questions was created, and the neural network was trained on data obtained from patients' responses to the online questionnaire.

4. Creation of an intelligent system prototype.

At this stage, the architecture and structure of the intelligent system were developed, including data file exchange, an information-logical model, and data, physical, and logical

design of the system. A web application for medical organization staff was also developed. The functional structure of the expert system is shown in Figure 1.

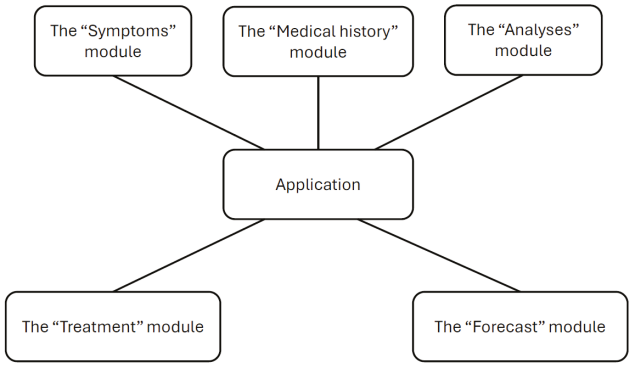



**Figure 1.** Functional diagram of the intelligent system "Oncology".

The intelligent system consists of six modules: Symptoms Module, Medical History Module, Tests Module, Application Module, Treatment Module, and Prognosis Module.

The Symptoms Module stores information about symptoms associated with respiratory system diseases, diagnoses, and questions based on the protocol used during patient interviews. The Medical History Module stores information about the medical histories of patients who have visited the center with respiratory system complaints. The Tests Module stores information about the results of patient tests recorded in the database.

Data from the Symptoms, Medical History, and Tests modules are transferred to the Application Module. These data are processed in Power BI 2.145.1105.0, after which a mathematical model is created in Python, and patient response data are stored in an SQL database.

The Treatment Module stores information about possible diagnoses for respiratory system diseases and treatment protocols.

The Prognosis Module allows forecasting of respiratory diseases, including lung cancer specifically.

Knowledge modeling and formalization.

At this stage, questions and answers from patients' oncology history cards are formalized for subsequent processing by the expert system.

In Equation (1), we define the model of the intelligent system (Figure 2):

$$Y = \{A, B, C, E\}, \tag{1}$$

where Y is the set of resulting data.

A is the set of questions on the oncologic history.

B is the set of answers on the oncologic history.

C is the set of weights of answers on the oncologic history.

E is the set of diagnoses.

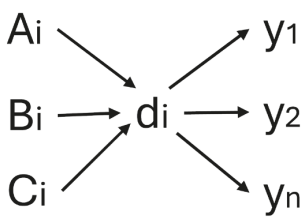



**Figure 2.** Neural network's model. Here, $A_i$, $B_i$, and $C_i$ are input sets; $d_i$ is the activation function; and $y_n$ are the resulting values.

The maximization function is used as the activation function $d_i$; the answer with the maximum weight is selected from the answer options. In Equation (2), we define the activation functions:

$$d_i = f\{\max(C_i)\} \tag{2}$$

Creating a database of scripted questions, training the neural network.

At this stage, a database of questions is created, and the neural network is trained using the data obtained from the online questionnaire of the respondents.

The neural network training process is shown in Figure 3.

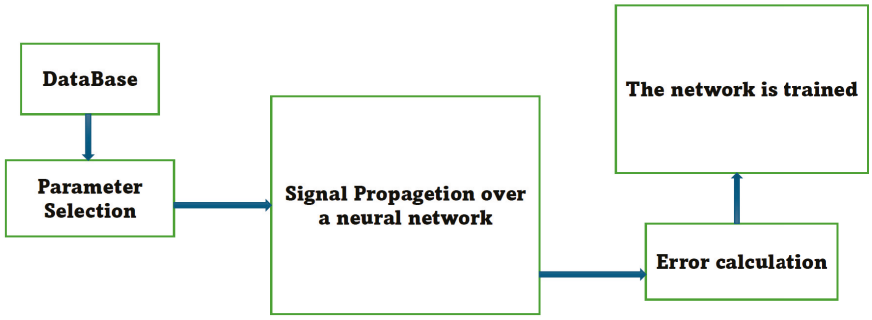



**Figure 3.** NN training process.

The process of signal propagation according to one criterion is shown in Figure 4.

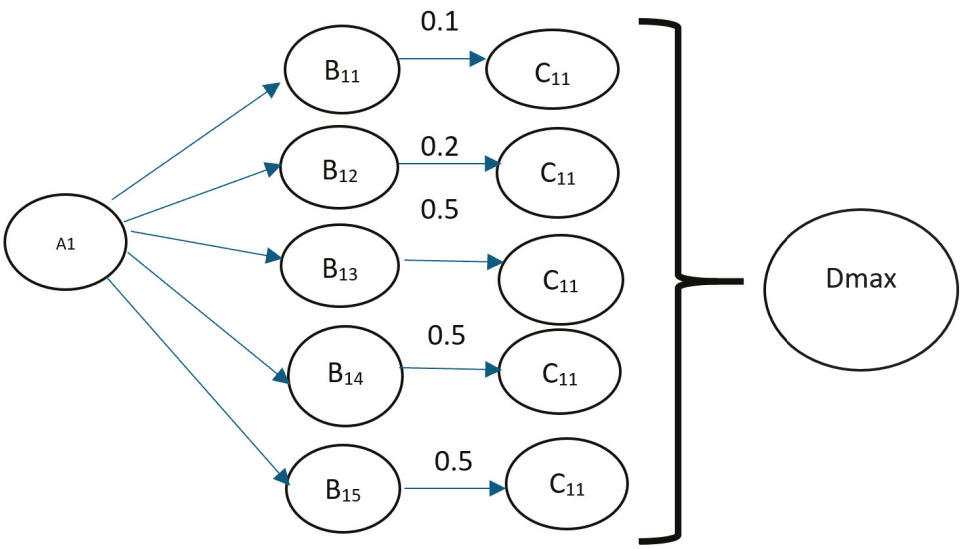



**Figure 4.** The process of signal propagation according to one criterion.

The functioning of this network can be described as follows:

The input layer receives data consisting of respondents' questions and answers for processing and passes them along the network.

Hidden layer(s) perform computations by selecting the maximum value among the weighted responses and processing the data and then transmit the results to the next layer.

The output layer predicts the likelihood of lung cancer susceptibility and classifies the results based on the processed data.

Each neuron in the neural network is connected to the adjacent layers through weights, which determine the importance of each input signal. The input values and the corresponding weights are multiplied, and the results are summed. If a certain threshold is reached,

the neuron's activation function is triggered, allowing the neuron to transmit its output to the next layer.

A mathematical model was created in Python, utilizing a neural network. The input parameters for this model are the weighted answers, with the endogenous factor being the response to the question "Have you ever had cancer?" Each answer is assigned a specific weight, and based on these weights, the system calculates a resulting value. Consequently, the script outputs the probability of lung cancer susceptibility.

The following methods were used to create the mathematical model: gradient boosting, multilayer perceptron, and Hamming network.

Alongside widely used models such as multilayer perceptrons and gradient boosting methods, we selected the Hamming network for the following reasons:

1. Robustness to Input Noise.

The Hamming network is known for its robustness to input distortions, maintaining correct classification even in the presence of noise—provided the input remains sufficiently similar to the reference patterns. In our study, the responses are selected from five predefined answer options, resulting in inputs that are generally close to the reference vectors. Given these conditions, we consider the Hamming network to be a well-suited choice for achieving accurate classification.

2. Simplicity of Implementation.

The architecture of the Hamming network consists of only two layers:

- The first layer computes the correlation between the input vector and the reference patterns.
- The second layer implements a winner-takes-all mechanism, selecting the closest matching pattern.

This structure makes the model both transparent and easy to implement, whether in hardware or software.

3. High Recognition Speed.

Thanks to its simple architecture, the Hamming network can quickly identify the closest reference pattern without requiring extensive training or complex computations. This leads to high recognition speed, which is especially advantageous in applications that demand rapid decision-making.

4. Efficiency with Small Datasets.

Unlike deep neural networks that require large volumes of training data, the Hamming network can operate effectively with a limited number of reference patterns. This is particularly important for tasks where data are scarce or expensive to obtain. In our study, the dataset is based on medical indicators. Conducting surveys within medical institutions involves significant time and numerous administrative approvals from institutional leadership. As a result, the survey was conducted online, which limited the number of respondents. Given these constraints, the Hamming network was selected as an appropriate model for initial experimentation.

5. Applicability to Classification and Recognition Tasks.

The Hamming network is widely used for the following tasks:

- Pattern and character recognition.
- Signal and time-series analysis.
- Data integrity verification.

- Development of systems with limited computational resources.

Since this experiment was conducted online and tested on a single physician's workstation, computational resources were limited. Therefore, the Hamming network was chosen for model validation due to its low computational requirements.

In summary, the choice of the Hamming network is justified by its simplicity, robustness to noise, fast processing speed, and suitability for tasks with constrained resources. In future research, with a larger volume of input data, deep neural network models will be explored and evaluated.

Creating a model using the Hamming network.

The code that calculates the risk probabilities is shown in Figure 5.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import joblib
import optuna

from sklearn.model_selection import train_test_split, cross_val_score, StratifiedKFold
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.impute import SimpleImputer
from sklearn.compose import ColumnTransformer
from sklearn.metrics import classification_report, roc_auc_score, log_loss
from sklearn.ensemble import RandomForestClassifier
from sklearn.calibration import CalibratedClassifierCV
from sklearn.base import clone
from sklearn.pipeline import Pipeline


1 usage  new *
def convert_range_to_numeric(x):
    if isinstance(x, str) and "-" in x:
        try:
            a, b = x.split("-")
            return (float(a.strip()) + float(b.strip())) / 2.0
        except Exception:
            return np.nan
    return x


1 usage  new *
def load_dataframe():
    try:
        return pd.read_excel( io: "data_train_1000_balanced_cancer_40_60.xlsx", sheet_name="data")
    except Exception:
        try:
            return pd.read_excel( io: "data_train.xlsx", sheet_name="TrainingData")
        except Exception:
            return pd.read_excel( io: "data_train.xlsx", sheet_name="data")


df = load_dataframe()
```

**Figure 5.** The code that calculates the risk probabilities.

The most significant factors identified were age, smoking duration, daily cigarette consumption, and the annual frequency of ARVI infections.

To build a forecast model, the following numerical features were taken:

*numeric_features = [*
*" Your age ",*
*" Smoking experience ",*
*" How many packs/pieces do you consume per day? ",*
*" How many times a year do you get SARS?"*

The test and training samples were generated as follows:

*X_train, X_test, y_train, y_test = train_test_split(*
*X_train_full,*
*y_train_full,*
*test_size = 0.2,*
*random_state = 42*
*)*

The following parameters were defined as categorical features:

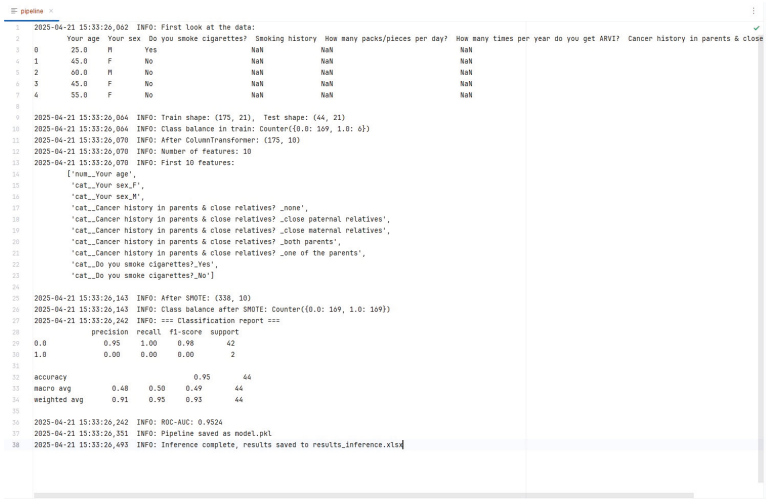
*categorical_features = [*
*" Your gender ",*
*"Oncoanamnesis of parents and close relatives? ",*
*" Do you use cigarettes? "*
*]*

The output of the intermediate values of the neural network is shown in Figure 6.

Intermediate results were obtained in the first layer.

*// Train shape: (175, 21), Test shape: (44, 21)*

After splitting the dataset into training and testing sets, we obtained 175 rows with 21 features in the training set and 44 rows with the same number of features in the test set. This confirms the correct 80/20 split and that all features were retained.



```
pipeline
2025-04-21 15:33:26,062  INFO: First look at the data:
      Your age  Your sex  Do you smoke cigarettes?  Smoking history  How many packs/pieces per day?  How many times per year do you get ARVI?  Cancer history in parents & close
0        25.0         M                       Yes              NaN                             NaN                                       NaN
1        45.0         F                        No              NaN                             NaN                                       NaN
2        60.0         M                        No              NaN                             NaN                                       NaN
3        45.0         F                        No              NaN                             NaN                                       NaN
4        55.0         F                        No              NaN                             NaN                                       NaN

2025-04-21 15:33:26,064  INFO: Train shape: (175, 21),  Test shape: (44, 21)
2025-04-21 15:33:26,064  INFO: Class balance in train: Counter({0.0: 169, 1.0: 6})
2025-04-21 15:33:26,070  INFO: After ColumnTransformer: (175, 10)
2025-04-21 15:33:26,070  INFO: Number of features: 10
2025-04-21 15:33:26,070  INFO: First 10 features:
         ['num__Your age',
         'cat__Your sex_F',
         'cat__Your sex_M',
         'cat__Cancer history in parents & close relatives? _none',
         'cat__Cancer history in parents & close relatives? _close paternal relatives',
         'cat__Cancer history in parents & close relatives? _close maternal relatives',
         'cat__Cancer history in parents & close relatives? _both parents',
         'cat__Cancer history in parents & close relatives? _one of the parents',
         'cat__Do you smoke cigarettes?_Yes',
         'cat__Do you smoke cigarettes?_No']

2025-04-21 15:33:26,143  INFO: After SMOTE: (338, 10)
2025-04-21 15:33:26,143  INFO: Class balance after SMOTE: Counter({0.0: 169, 1.0: 169})
2025-04-21 15:33:26,242  INFO: === Classification report ===
              precision  recall  f1-score  support
0.0            0.95     1.00     0.98       42
1.0            0.00     0.00     0.00        2

accuracy                          0.95       44
macro avg       0.48     0.50     0.49       44
weighted avg    0.91     0.95     0.93       44

2025-04-21 15:33:26,242  INFO: ROC-AUC: 0.9524
2025-04-21 15:33:26,351  INFO: Pipeline saved as model.pkl
2025-04-21 15:33:26,493  INFO: Inference complete, results saved to results_inference.xlsx
```

**Figure 6.** Output of intermediate values of a neural network.

Class balance in the training set: *Counter({0.0: 169, 1.0: 6})*

The training set exhibited significant imbalance: 169 negative instances (0) and only 6 positive instances (1). This imbalance necessitated balancing; hence, SMOTE was subsequently employed.

After ColumnTransformer: (175, 10).

Following numerical imputation, standardization, and one-hot encoding, the original 21 columns transformed into 10 features. The feature matrix size became 175 $\times$ 10, indicating that categories were correctly encoded and numerical features underwent scaling.

Number of features: 10.

Verification confirms that the final feature vector indeed consists of 10 columns after all transformations.

First 10 features:

Feature names output (get_feature_names_out) revealed one numerical feature (num_age) and nine categorical one-hot features. The names are meaningful, with no duplicates.

*After SMOTE: (338, 10)*

The SMOTE method generated synthetic samples for the minority class, increasing the number of rows to 338 while retaining 10 features. Each row continues to be described by the same feature set.

Class balance after SMOTE: *Counter({0.0: 169, 1.0: 169})*

The classes are now fully balanced, with 169 examples each. The model will be trained on equal numbers of positive and negative cases.

Classification report.

On the test dataset, the model achieves approximately 0.95 accuracy for the zero class but fails to recognize the positive class (precision and recall for class 1 are 0). This indicates that under current settings, RandomForest has not learned to identify the rare class in independent data.

*ROC-AUC: 0.9524*

Despite poor metrics for class 1, the overall ROC-AUC metric is high (~0.95). This occurs because the test dataset contains only two positive samples; even a single error significantly affects precision/recall but has minimal impact on the ROC curve.

The precision and recall values for the positive class are equal to 0. This is due to the fact that there are only two positive instances in the test set. While the model is capable of identifying positive cases, the likelihood of doing so is low. This is primarily because the survey was conducted among 1283 respondents, of whom only a small fraction tested positive for cancer.

In future study, the questionnaire is expected to be administered across multiple medical institutions, including a specialized oncology center.

In particular, a model was built using the original input parameters based on gradient boosting with decision trees as the base algorithm. The model was implemented using the LightGBM framework. Class balancing was performed using the SMOTE method, while hyperparameter optimization was carried out with Optuna. The obtained results are presented in Figure 7.

```
========== SAMPLE PREDICTIONS (first 10 of test) ==========
   actual     proba  pred@0.50  pred@0.29
0     0.0  0.003819          0          0
1     0.0  0.015741          0          0
2     0.0  0.004650          0          0
3     0.0  0.001777          0          0
4     0.0  0.003787          0          0
5     0.0  0.060287          0          0
6     0.0  0.001938          0          0
7     0.0  0.116701          0          0
8     0.0  0.001938          0          0
9     0.0  0.001777          0          0
```

**Figure 7.** Results of the model obtained using the gradient boosting method.

Based on the operation of the neural network, the following data were obtained, as presented in the corresponding Figure 8.

```
=== MODEL SUMMARY ===
Model: "sequential"
```

| Layer (type) | Output Shape | Param # |
|---|---|---|
| dense (Dense) | (None, 64) | 704 |
| dropout (Dropout) | (None, 64) | 0 |
| dense_1 (Dense) | (None, 32) | 2,080 |
| dropout_1 (Dropout) | (None, 32) | 0 |
| dense_2 (Dense) | (None, 1) | 33 |

```
 Total params: 8,453 (33.02 KB)
 Trainable params: 2,817 (11.00 KB)
 Non-trainable params: 0 (0.00 B)
 Optimizer params: 5,636 (22.02 KB)

=== CLASSIFICATION REPORT ===
              precision    recall  f1-score   support

         0.0       1.00      0.91      0.95        43
         1.0       0.33      1.00      0.50         2

    accuracy                           0.91        45
   macro avg       0.67      0.95      0.73        45
weighted avg       0.97      0.91      0.93        45

Accuracy: 0.9111111111111111
Precision: 0.3333333333333333
Recall: 1.0
F1-score: 0.5
ROC-AUC: 0.9767441860465116
```

**Figure 8.** Parameters of the neural network.

Subsequently, a model using a multilayer perceptron was implemented in Keras. The final performance metrics of this model are presented in Figure 9.

The problem of zero accuracy for the positive class arose due to the following constraints:

- Insufficient amount of input data, as the survey was conducted among residents of regions affected by nuclear testing.
- A low number of positive responses indicating cancer incidence. Since the majority of respondents were healthy, the percentage of positive cases was consequently low.

```
=== CONFUSION MATRIX ===
[[39  4]
 [ 0  2]]

=== CLASS BALANCE (Test set) ===
had_cancer
0.0    0.955556
1.0    0.044444
Name: proportion, dtype: float64

=== SAMPLE PREDICTIONS ===
   Actual  Predicted  Probability
0     0.0          0     0.017884
1     0.0          0     0.000673
2     0.0          0     0.080154
3     0.0          0     0.020091
4     0.0          0     0.017884
5     0.0          0     0.023210
6     0.0          0     0.006847
7     0.0          1     0.700024
8     0.0          0     0.006847
9     0.0          0     0.020091
```

**Figure 9.** Results of the model obtained using a multilayer perceptron.

We have taken the following steps to address the issue of zero precision for the positive class:

1. Class Imbalance Adjustment:

Using an oversampling technique, the size of the positive class was increased. The proportion of positive instances was raised to 40%. Additionally, class weighting was applied in the loss function to ensure that the model pays greater attention to the positive class.

2. Use of Additional Metrics:

To further evaluate the models, the following metrics were also employed: precision, recall, F1-score, and AUC-ROC. All three models demonstrated strong performance according to these metrics:

- Gradient boosting using LightGBM: Cross-Validation ROC-AUC (best from Optuna) = 0.9059; Test ROC-AUC = 1.0000.
- Multilayer perceptron model implemented in Keras: ROC-AUC (sklearn) = 0.9459.
- Hamming network model: ROC-AUC (sklearn) = 0.95.

3. Classification Threshold Adjustment:

To improve the model's sensitivity to the positive class, the classification threshold was adjusted from 0.3 to 0.5.

4. Model Architecture Improvement:

To prevent overfitting, dropout regularization was applied to the models.

After adjusting the models according to the above-described procedure, the following results were obtained, as shown in Figures 10–12:

After model adjustment, all three models demonstrated adequate probability estimates, indicating the development of models suitable for predicting cancer risk.

To date, validation has been conducted using three types of neural networks: multilayer perceptron, gradient boosting method, and Hamming network. It should be noted that this study is not conclusive. As a larger cohort of respondents is planned, the amount of input data will increase accordingly. Therefore, future studies will focus on developing models using different types of neural networks depending on the data volume and computational resource constraints. A more detailed comparison of the results obtained from multiple model types will then be performed.

```
Test: loss=0.0678, acc=0.9850, auc(keras)=0.9969
=== Classification report (test, tuned threshold) ===
              precision    recall  f1-score   support

           0       1.00      0.97      0.99       120
           1       0.96      1.00      0.98        80

    accuracy                           0.98       200
   macro avg       0.98      0.99      0.98       200
weighted avg       0.99      0.98      0.99       200

ROC-AUC (sklearn, test) = 0.9969270833333332
```

**Figure 10.** Implementation of the model using a multilayer perceptron.

```
=== Classification report (test) ===
              precision    recall  f1-score   support

           0       1.00      0.97      0.99       120
           1       0.96      1.00      0.98        80

    accuracy                           0.98       200
   macro avg       0.98      0.99      0.98       200
weighted avg       0.99      0.98      0.99       200

Optuna best CV ROC-AUC: 0.9965
Test ROC-AUC: 0.9921
Test LogLoss: 0.0586
```

**Figure 11.** Implementation of the model using the gradient boosting method.

```
=== Classification report (test) ===
              precision    recall  f1-score   support

           0       1.00      0.97      0.99       120
           1       0.96      1.00      0.98        80

    accuracy                           0.98       200
   macro avg       0.98      0.99      0.98       200
weighted avg       0.99      0.98      0.99       200


Optuna best CV ROC-AUC: 0.9968
Test ROC-AUC: 0.9936
Test LogLoss: 0.0596
```

**Figure 12.** Implementation of the model using the Hamming network.

The model calculates a predicted_proba value indicating the numerical probability of susceptibility to lung cancer. This resulting value is classified into four risk categories as follows:

Class 1: 0–0.01, low risk.

Class 2: 0.01–0.1, moderate risk.

Class 3: 0.1–0.54, high risk.

Class 4: 0.55–1, very high risk.

Based on responses from 1283 respondents, a diagram illustrating the percentage distribution of risk classes was obtained, as shown in Figure 13.

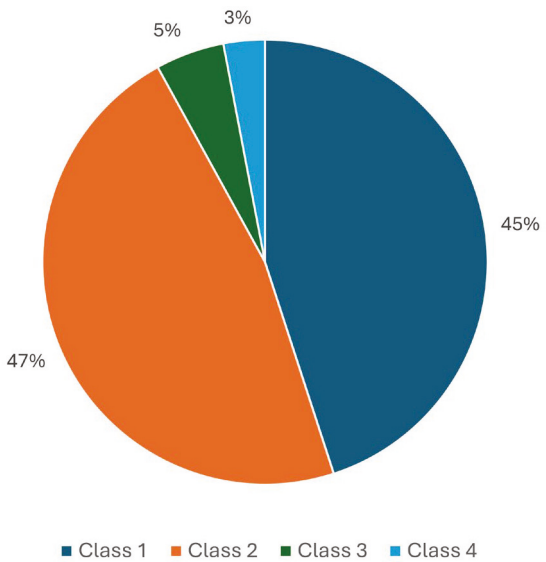



**Figure 13.** Lung cancer risk status.

Creation of an intelligent system prototype.

At this stage, the architecture and structure of the intelligent system were developed, including data file exchange, an information-logical model, and data, physical, and logical system design. Additionally, a web application layout was proposed for medical organization staff.

The system architecture comprises a survey questionnaire, server, mathematical module, application, expert decision-maker (oncologist), medical indicators verification module, and diagnostic module (Figure 14).

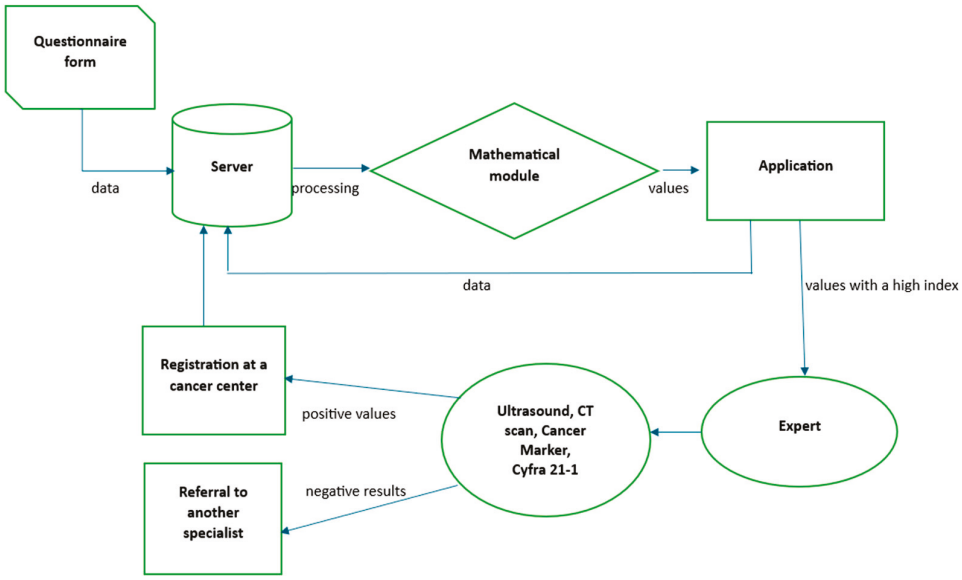



**Figure 14.** System architecture.

The structure of the system is shown in Figure 15.

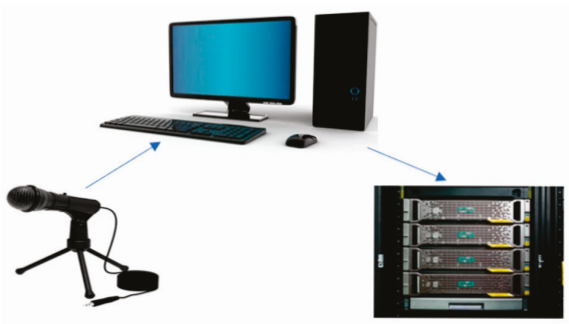


**Figure 15.** System structure.

The Questionnaire Module contains questions focused on the early diagnosis of lung cancer. The collected data are uploaded to the server and stored in a database. The web system is presently undergoing refinement.

## 4. Discussion

Three methods were tested to create a mathematical model: gradient boosting, multilayer perceptron, and Hamming network. The Hamming network was chosen to create the mathematical model, as it combines sufficient accuracy with ease of implementation and interpretation, which is especially important for expert systems. In the mathematical module, the obtained data are processed using Hamming's neural network in Python. As a result, the system generates values representing the risk of lung cancer susceptibility.

The input layer of the network receives data comprising respondents' questions and answers for processing and subsequently forwards it through the network. The hidden layer performs computations by selecting the maximum value among the weighted responses and processes the data before transmitting the results to the next layer. The output layer predicts the probability of susceptibility to lung cancer and classifies the outcomes based on the processed information.

The Application Module includes questions and corresponding responses related to the early detection of lung cancer. The module includes questions based on the lung cancer diagnostic protocol, covering disease symptoms, medical history, heredity, and other relevant factors. Within this application, the patient answers questions and receives a risk assessment for potential lung cancer.

High-risk values are then forwarded to the Expert Decision-Maker (an oncologist), who evaluates the results and directs the patient to appropriate medical tests. If these tests confirm the presence of malignant tumors, the patient is registered at an oncology center. Otherwise, the patient is referred to another specialist.

A patient completed the lung cancer risk assessment questionnaire. The resulting score was 0.936779, which was classified as Class 4—very high risk. The patient was advised to undergo a chest X-ray, which revealed a formation in the upper right lung. A subsequent contrast-enhanced CT scan of the chest was performed. The conclusion: CT signs of peripheral cancer in S1 of the upper lobe of the right lung, lymphadenopathy (mts) of the paratracheal intrathoracic lymph nodes, fibrotic-focal changes in S2 and S6 of the right lung and S1–2 and S6 of the left lung, chronic bronchitis, bullous emphysema, and atherosclerosis of the thoracic aorta.

Surgical treatment was carried out as follows: single-port video-assisted thoracoscopic biopsy of the upper lobe of the right lung, right-sided anterolateral thoracotomy, right upper lobectomy with lymph node dissection, and pleural cavity drainage on the right side.

The proposed method for early detection of lung cancer was initially evaluated on a single patient, demonstrating positive outcomes. Nevertheless, extensive validation involving a larger patient population is required to establish its clinical utility. We anticipate that, upon successful validation, the implementation of this method will facilitate prompt identification and management of high-risk cases. Future research plans include the integration of an ultrasound image recognition module into the system, which is expected to further enhance diagnostic accuracy. This method may also assist in the detection of other pulmonary diseases.

Future research plans include integrating an ultrasound image recognition module into the system, which will enhance diagnostic accuracy and facilitate timely intervention for subsequent patient treatment.

## 5. Conclusions

The detection and treatment of oncological diseases remain among the most pressing issues today. The mortality rate associated with cancer is extremely high. Therefore, early diagnosis of cancer is one of the most critical tasks in both medicine and science. Early detection allows for prolonged patient life expectancy and reduced mortality.

An algorithm for constructing a prototype intelligent system for the early diagnosis of cancer has been developed. Population-based surveys were employed for knowledge acquisition and analysis in lung cancer diagnostics. A conceptual system model was created, detailing the principal processes and their interrelationships. Domain knowledge was subsequently formalized for neural-network design: the questions and answers of

the lung-diagnosis protocol were systematized, and response weights were assigned in accordance with their relative importance. The following methods were used to create the mathematical model: gradient boosting, multilayer perceptron, and Hamming network.

A neural network was then built to estimate the risk of lung cancer susceptibility. Finally, an architecture for a web-based intelligent system enabling early lung cancer diagnosis was proposed, with eventual deployment in clinical practice envisaged.

The proposed method for the early diagnosis of lung cancer was tested on one patient and gave positive results. Implementing such a method will enable a rapid response to high-risk cases. Future research plans include integrating an ultrasound image recognition module into the system, which will further enhance diagnostic effectiveness. This method can also help identify other diseases of the lungs: pulmonary tuberculosis (disseminated), pulmonary sarcoidosis, pneumonia, and pulmonary fibrosis.

The integration of an ultrasound image recognition module into the system facilitates early detection of a range of pulmonary diseases and allows for the determination of disease progression stages.

The following challenges can be identified in the research: if this application were implemented in local clinics, the coverage of the surveyed population would be greater. The integration of ultrasound imaging into the system requires selecting an image recognition method that is optimally suited for this expert system.

**Author Contributions:** Conceptualization, writing—review and editing, I.K. and D.B.; writing—original draft preparation, D.K. (Dinara Kozhakhmetova); investigation, methodology, R.B. and L.K.; software, D.O. and A.B.; supervision, D.K. (Dinara Kurushbayeva). All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP23485656).

**Data Availability Statement:** The original contributions presented in this study are included in the article. The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

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

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| DSS | decision support systems |
| AR | association rules |
| NN | neural networks |
| NLP | natural language processing |
| DDSS | diagnostic decision support systems |
| ACO | Ant Colony Optimization |

## References

1. Benusiglio, P.R.; Fallet, V.; Sanchis-Borja, M.; Coulet, F.; Cadranel, J. Lungcancer isalsoahereditary disease. *Eur. Respir. Rev.* **2021**, *30*, 210045. [CrossRef]
2. Karabatak, M.; Ince, M.C. An expert system for detection of breast cancer based on association rules and neural network. *Expert Syst. Appl.* **2009**, *36*, 3465–3469. [CrossRef]
3. Oyelade, O.N.; Obiniyi, A.A.; Junaidu, S.B.; Adewuyi, S.A. Patient symptoms elicitation process for breast cancer medical expert systems: A semantic web and natural language parsing approach. *Future Comput. Inform. J.* **2018**, *3*, 72–81. [CrossRef]
4. Goldshtein, M.; Amin, G.; Rod, D. An NLP-Based Exploration of Variance in Student Writing and Syntax: Implications for Automated Writing Evaluation. *Computers* **2024**, *13*, 160. [CrossRef]

5. Hassan, L.; Saleh, A.; Singh, V.K.; Puig, D.; Abdel-Nasser, M. Detecting Breast Tumors in Tomosynthesis Images Utilizing Deep Learning-Based Dynamic Ensemble Approach. *Computers* **2023**, *12*, 220. [CrossRef]
6. Gavade, A.B.; Nerli, R.; Kanwal, N.; Gavade, P.A.; Pol, S.S.; Rizvi, S.T.H. Automated Diagnosis of Prostate Cancer Using mpMRI Images: A Deep Learning Approach for Clinical Decision Support. *Computers* **2023**, *12*, 152. [CrossRef]
7. Kalra, G.; Rajoria, Y.K.; Boadh, R.; Rajendra, P.; Pandey, P.; Khatak, N.; Kumar, A. Study of fuzzy expert systems towards prediction and detection of fraud case in health care insurance. *Mater. Today Proc.* **2022**, *56*, 477–480. [CrossRef]
8. Bakhtadze, N.N.; Belenkiy, V.M.; Pyatetsky, V.E.; Sakrutina, E.A.; Nikulina, I.V. Intelligent Screening for Diagnostic and Treatment of Cancer Diseases. *Procedia Comput. Sci.* **2017**, *112*, 1238–1245. [CrossRef]
9. Keleş, A.; Keleş, A.; Yavuz, U. Expert system based on neuro-fuzzy rules for diagnosis breast cancer. *Expert Syst. Appl.* **2011**, *38*, 5719–5726. [CrossRef]
10. Apasiya, E.; Salifu, A.; Agbedemnab, P. New Approaches to the Prognosis and Diagnosis of Breast Cancer Using Fuzzy Expert Systems. *J. Comput. Commun.* **2024**, *12*, 151–169. [CrossRef]
11. Safdari, R.; Arpanahi, H.K.; Langarizadeh, M.; Ghazisaiedi, M.; Dargahi, H.; Zendehdel, K. Design a Fuzzy Rule-based Expert System to Aid Earlier Diagnosis of Gastric Cancer. *Acta Inform. Medica* **2018**, *26*, 19–23. [CrossRef] [PubMed]
12. Übeyli, E.D. Implementing automated diagnostic systems for breast cancer detection. *Expert Syst. Appl.* **2007**, *33*, 1054–1062. [CrossRef]
13. Houssami, N.; Lee, C.I.; Buist, D.S.M.; Tao, D. Artificial intelligence for breast cancer screening: Opportunity or hype? *Breast* **2017**, *36*, 31–33. [CrossRef]
14. Teramoto, A.; Kiriyama, Y.; Michiba, A.; Yazawa, N.; Tsukamoto, T.; Imaizumi, K.; Fujita, H. Automated Generation of Lung Cytological Images from Image Findings Using Text-to-Image Technology. *Computers* **2024**, *13*, 303. [CrossRef]
15. Melyani, M.; Frasetyo, T.; Rahadjeng, I.; Mufid, Z.; Rafik, A.; Shaura, R.; Daniel, D.; Emita, I. Design Framework of Expert System Program in Otolaryngology Disease Diagnosis use Extreme Programming (XP) Method (Case Study in THB Bekasi Hospital). *J. Technol. Inform. Eng.* **2024**, *3*, 397. [CrossRef]
16. Kunhimangalam, R.; Ovallath, S.; Joseph, P.K. A Novel Fuzzy Expert System for the Identification of Severity of Carpal Tunnel Syndrome. *BioMed Res. Int.* **2013**, *2013*, 846780. [CrossRef]
17. Gashteroodkhani, O.; Majidi, M.; Etezadi-Amoli, M. A Fuzzy-based Control Scheme for Recapturing Waste Energy in Water Pressure Reducing Valves. In Proceedings of the IEEE Power and Energy Society General Meeting (PESGM), Portland, OR, USA, 5–10 August 2018; pp. 1–5. [CrossRef]
18. Wójcik, W.; Mezhiievska, I.; Pavlov, S.V.; Lewandowski, T.; Vlasenko, O.V.; Maslovskyi, V.; Volosovych, O.; Kobylianska, I.; Moskovchuk, O.; Ovcharuk, V.; et al. Medical Fuzzy-Expert System for Assessment of the Degree of Anatomical Lesion of Coronary Arteries. *Int. J. Environ. Res. Public Health* **2023**, *20*, 979. [CrossRef]
19. Mazhar, T.; Nasir, Q.; Haq, I.; Kamal, M.M.; Ullah, I.; Kim, T.; Mohamed, H.G.; Alwadai, N. A Novel Expert System for the Diagnosis and Treatment of Heart Disease. *Electronics* **2022**, *11*, 3989. [CrossRef]
20. Aguilera-Venegasa, G.; Roanes-Lozanob, E.; Rojo-Martínezc, G.; Galán-Garcíaa, J. A proposal of a mixed diagnostic system based on decision trees andprobabilistic experts rules. *J. Comput. Appl. Math.* **2023**, *427*, 115130. [CrossRef]
21. Sarabi, S.; Han, Q.; Vries, B.; Georges, L. Methodology for development of an expert system to derive knowledge from existing nature-based solutions experiences. *MethodsX* **2023**, *10*, 101978. [CrossRef]
22. Mustafa, E.; Saad, M.; Rizkallah, L. Building an enhanced case-based reasoning and rule-based systems for medical diagnosis. *J. Eng. Appl. Sci.* **2023**, *70*, 139. [CrossRef]
23. Tariq, R. *Make Your Own Neural Network*; CreateSpace Independent Publishing Platform: North Charleston, SC, USA, 2016; p. 222.
24. Chen, M.; Challita, U.; Saad, W.; Yin, C.; Debbah, M. Artificial Neural Networks-Based Machine Learning for Wireless Networks: A Tutorial. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 3039–3071. [CrossRef]
25. Molina, E.; Parraga-Alava, J. Redes Neuronales Artificiales en la Resolución de Problemas de Clasificación: Una Revisión Sistemática de la Literatura. *ENFOQUE UTE* **2024**, *15*, 1–10. [CrossRef]
26. Vásquez-Coronel, J.A.; Mora, M.; Vilches, K. A Review of multilayer extreme learning machine neural networks. *Artif. Intell. Rev.* **2023**, *56*, 13691. [CrossRef]
27. Syahidin, Y.; Ismail, A.; Siraj, F. Application of Artificial Neural Network Algorithms to Heart Disease Prediction Models with Python Programming. *J. E-Komtek* **2022**, *6*, 292–302. [CrossRef]
28. Wang, Y.H.; Lin, G.Y. Design of Medical Diagnostic System Based on Artificial Intelligence. *Int. J. Appl. Inf. Manag.* **2023**, *3*, 170–176. [CrossRef]
29. Tiwari, A.; Mishra, S.; Kuo, T.R. Current AI technologies in cancer diagnostics and treatment. *Mol. Cancer* **2025**, *24*, 159. [CrossRef]

30. Liao, J.; Li, X.; Gan, Y.; Han, S.; Rong, P.; Wang, W.; Li, W.; Zhou, L. Artificial intelligence assists precision medicine in cancer treatment. *Front. Oncol.* **2023**, *12*, 998222. [CrossRef]
31. Adilgazyuly, S.; Bulegenov, T.; Mussakhanova, A.; Kenesh Dzhusupov, K.; Altybayeva Zh Medeulova, A.; Tokenova Zh Karymsakova, I.; Uruzbayeva, G.; Kussainova, A. Assessment of physicians' perspectives on the quality and accessibility of rehabilitation services for lung cancer patients across all levels of healthcare delivery in the Republic of Kazakhstan. *Cent. Asian J. Med. Hypotheses Ethics* **2025**, *6*, 142–153. [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.

computers

MDPI

Article

# Healthcare AI for Physician-Centered Decision-Making: Case Study of Applying Deep Learning to Aid Medical Professionals

**Aleksandar Milenkovic [1,*], Andjelija Djordjevic [1], Dragan Jankovic [1], Petar Rajkovic [1], Kofi Edee [2] and Tatjana Gric [3]**

[1] Department of Computer Science, Faculty of Electronic Engineering in Nis, University of Nis, 18000 Nis, Serbia; andjelija.djordjevic@elfak.ni.ac.rs (A.D.); dragan.jankovic@elfak.ni.ac.rs (D.J.); petar.rajkovic@elfak.ni.ac.rs (P.R.)

[2] Institut Pascal, CNRS, Clermont Auvergne INP, Université Clermont Auvergne, F-63000 Clermont-Ferrand, France; kofi.edee@uca.fr

[3] Department of Electronic Systems, Vilnius Gediminas Technical University, LT-10223 Vilnius, Lithuania; tatjana.gric@vilniustech.lt

* Correspondence: aleksandar.milenkovic@elfak.ni.ac.rs

**Abstract:** This paper aims to leverage artificial intelligence (AI) to assist physicians in utilizing advanced deep learning techniques integrated into developed models within electronic health records (EHRs) in medical information systems (MISes), which have been in use for over 15 years in health centers across the Republic of Serbia. This paper presents a human-centered AI approach that emphasizes physician decision-making supported by AI models. This study presents two developed and implemented deep neural network (DNN) models in the EHR. Both models were based on data that were collected during the COVID-19 outbreak. The models were evaluated using five-fold cross-validation. The convolutional neural network (CNN), based on the pre-trained VGG19 architecture for classifying chest X-ray images, was trained on a publicly available smaller dataset containing 196 entries, and achieved an average classification accuracy of $91.83 \pm 2.82\%$. The DNN model for optimizing patient appointment scheduling was trained on a large dataset (341,569 entries) and a rich feature design extracted from the MIS, which is daily used in Serbia, achieving an average classification accuracy of $77.51 \pm 0.70\%$. Both models have consistent performance and good generalization. The architecture of a realized MIS, incorporating the positioning of developed AI tools that encompass both developed models, is also presented in this study.

**Keywords:** AI; deep learning; medical information systems; electronic health records; healthcare; physician-centered decision-making

## 1. Introduction

AI in medical diagnosis represents a powerful tool that assists physicians in their daily work. With the help of AI systems, physicians receive valuable support that enables them to quickly identify potential problems and notice details that the "human eye" might miss. AI can analyze large amounts of data in record time, saving valuable time while also providing additional accuracy in analyzing images or other medical data.

However, the final decision still needs to be made by the physicians. Based on the suggestions provided by the AI system, the physician considers all relevant information from the patient's medical history (from the EHR) and decides on treatment. AI can clearly show the physician which aspects of the data it has highlighted, such as parts of an image that are considered suspicious, ensuring transparency in the decision-making process [1].

The greatest advantage of this approach is that responsibility for the patient's treatment remains entrusted to the physician. AI systems are not autonomous; they simply assist the physician in making more informed decisions by providing additional information and recommendations, while the physician retains a pivotal role in the diagnostic process.

The benefits of this integrated approach are manifold. First, it enables faster and more accurate diagnostics, thereby reducing patient waiting times and enhancing the quality of healthcare. AI provides greater access to medical expertise, particularly in rural areas where specialists may not be readily available. It also reduces the number of errors that can occur due to fatigue or stress. Meanwhile, the AI system, through continuous learning from new cases, gradually becomes more accurate, always under the supervision of experts.

Today, modern MIS and EHR systems should incorporate AI-based tools to enhance the quality of healthcare services, improve overall performance, and streamline daily operations. Deep learning [2] is a branch of AI and machine learning (ML). Deep learning algorithms are based on artificial neural network (ANN) models. Deep learning techniques include DNNs, recurrent neural networks, CNNs, and deep belief networks. These techniques are used to solve complex problems such as natural language processing (NLP), computer vision, bioinformatics, audio recognition, medical image analysis, and speech recognition. Deep learning provides valuable solutions for medical image analysis problems and is a key method for future applications [3]. A DNN is an ANN with more hidden layers between the input and output layers. A CNN [4] is a type of DNN model that enables the extraction of higher representations of image content. CNNs are a state-of-the-art image classification technique [5], which uses some features of the visual cortex. In this paper, a state-of-the-art technique for the classification of chest X-ray images is presented. The DNN model is used to optimize scheduling time slots. Both models are created in order to combat the COVID-19 pandemic.

The most efficient and reliable way to diagnose patients infected by COVID-19 during the pandemic and monitor their recovery is the chest X-ray and CT image [6]. By using the CNN, the model presented in this paper was developed, which rapidly classifies patients according to the chest X-ray image into two categories: one category is patients with COVID-19, while the other is those without COVID-19. The automatic classification assists the physicians in establishing the COVID-19 diagnosis.

Due to social distancing and other measures that limited the movement of patients during the COVID-19 pandemic, a considerable percentage of patients did not appear for their scheduled appointments. The appointments for an examination with a physician and/or diagnostic device remained reserved; however, the patients did not appear for them. Those are also often scheduled examinations on expensive diagnostic devices, for which patients must wait for an extended period of time, sometimes even 6 months. In this regard, there was a need to develop a tool to optimize patient scheduling in available slots. During the COVID-19 pandemic, primary healthcare institutions operated at reduced capacity (with fewer available slots for scheduling appointments), making it essential to optimize and increase the occupancy rate of these slots. The tool presented in this paper utilizes the developed DNN to classify patients with scheduled appointments as either arriving or not arriving at their scheduled time slots. Patients classified as those who were "potentially not coming" to their scheduled appointments should have been contacted by phone to confirm their arrival. An option was also supported for scheduling a higher number of patients classified as "potentially not coming" in the same time slots, which was modeled according to airline companies' overbooking of seat procedures [7].

The objective of this paper is to assist physicians in making final decisions by utilizing deep learning techniques incorporated into AI tools in MIS and EHRs. This study showcases

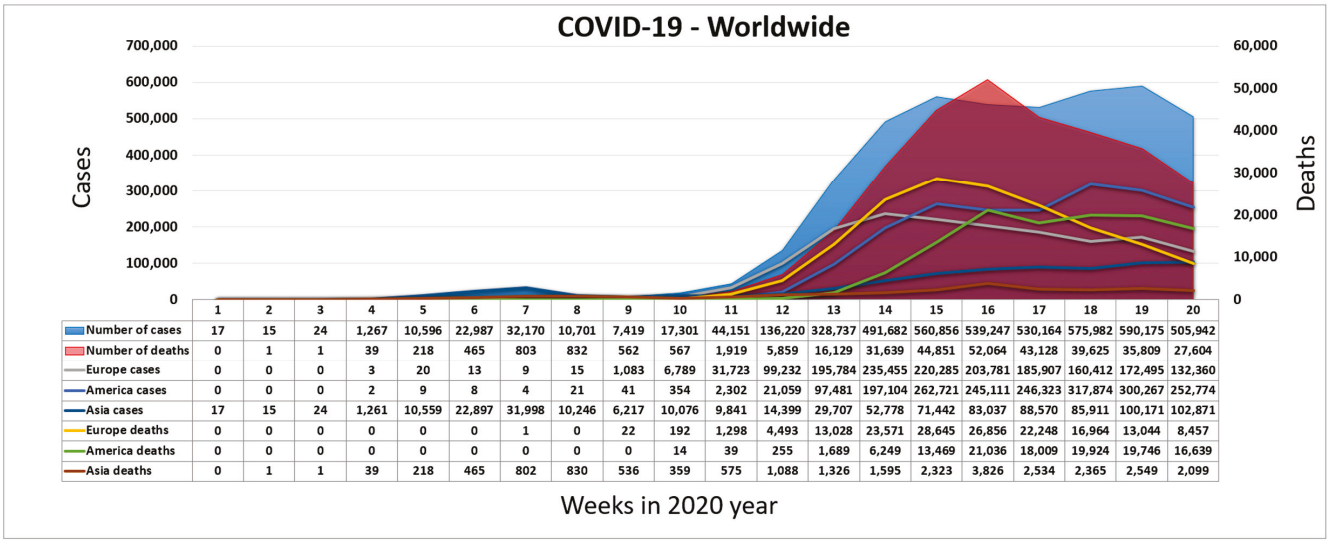
the effective application of AI in addressing critical situations, such as the COVID-19 pandemic, by prioritizing AI-assisted Physician-Centered Decision-Making. The model for rapidly classifying chest X-ray images into two classes, non-COVID-19 and COVID-19, is designed using a CNN. The model for the classification of appointments of patients in two classes during the pandemic: 0—patient is predicted to arrive at the scheduled time slot, 1—patient is predicted not to arrive, is created using the DNN.

This study presents the need for physician-centered decision-making in fast-reacting situations, utilizing AI tools that encompass both developed DNN models. The primary goal of this research is to integrate AI into existing MISes and enhance EHR with AI, rather than competing with sophisticated models in the field of study. Overall, the developed models presented in this paper have achieved remarkable accuracy and performance. Outbreaks, such as COVID-19, are exact situations where rapid physician-centered decision-making, assisted by AI tools, is crucial.

After the Introduction section, the Background and the Materials and Methods sections follow, respectively. The authors then present the results in the Results section and discuss them in the Discussion section. After the Discussion, the Conclusion section and the Reference list follow.

## 2. Background

The coronavirus SARS-CoV-2 appeared by the end of 2019 [8]. It was discovered in China in the city of Wuhan, the province of Hubei [9]. The disease it causes is called COVID-19. For COVID-19, the World Health Organization (WHO) assigned the urgent ICD-10 diagnosis U07.1 [10]. The newly emerged virus was connected to the virus SARS-CoV since 2013 [11]. The basic symptoms caused by COVID-19 are described in [9]. Figure 1 shows the exponential rise in the number of infected and deceased people worldwide during the COVID-19 pandemic [12]. The main crisis centers became Europe and North America, after the emergence of the virus in China. Due to the characteristics of COVID-19, the WHO decided to mark the epidemic of the virus SARS-CoV-2 as a pandemic on 11 March 2020.




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of cases | 17 | 15 | 24 | 1,267 | 10,596 | 22,987 | 32,170 | 10,701 | 7,419 | 17,301 | 44,151 | 136,220 | 328,737 | 491,682 | 560,856 | 539,247 | 530,164 | 575,982 | 590,175 | 505,942 |
| Number of deaths | 0 | 1 | 1 | 39 | 218 | 465 | 803 | 832 | 562 | 567 | 1,919 | 5,859 | 16,129 | 31,639 | 44,851 | 52,064 | 43,128 | 39,625 | 35,809 | 27,604 |
| Europe cases | 0 | 0 | 0 | 3 | 20 | 13 | 9 | 15 | 1,083 | 6,789 | 31,723 | 99,232 | 195,784 | 235,455 | 220,285 | 203,781 | 185,907 | 160,412 | 172,495 | 132,360 |
| America cases | 0 | 0 | 0 | 2 | 9 | 8 | 4 | 21 | 41 | 354 | 2,302 | 21,059 | 97,481 | 197,104 | 262,721 | 245,111 | 246,323 | 317,874 | 300,267 | 252,774 |
| Asia cases | 17 | 15 | 24 | 1,261 | 10,559 | 22,897 | 31,998 | 10,246 | 6,217 | 10,076 | 9,841 | 14,399 | 29,707 | 52,778 | 71,442 | 83,037 | 88,570 | 85,911 | 100,171 | 102,871 |
| Europe deaths | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 22 | 192 | 1,298 | 4,493 | 13,028 | 23,571 | 28,645 | 26,856 | 22,248 | 16,964 | 13,044 | 8,457 |
| America deaths | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 14 | 39 | 255 | 1,689 | 6,249 | 13,469 | 21,036 | 18,009 | 19,924 | 19,746 | 16,639 |
| Asia deaths | 0 | 1 | 1 | 39 | 218 | 465 | 802 | 830 | 536 | 359 | 575 | 1,088 | 1,326 | 1,595 | 2,323 | 3,826 | 2,534 | 2,365 | 2,549 | 2,099 |

**Figure 1.** Worldwide spread of COVID-19 during the first weeks of 2020.

Deep learning techniques for biomedical and health informatics are described in depth in [13]. Automated diagnosis of tuberculosis from chest X-rays has been tackled with either hand-crafted algorithms or ML approaches such as support vector machines and CNNs [14].

The detailed descriptions of deep learning for medical image analysis and the principles and methods of neural networks and deep learning concepts are given in [3]. The authors show COVID-NET [15], a tailored deep convolutional neural network design for detecting COVID-19 cases from chest X-ray images with an accuracy of 92.6%. Markus Schmitt shows nine ways of machine learning that help to combat the viral pandemic [16]. The approach that enables the automatic detection of the coronavirus disease using X-ray images and deep convolutional networks is shown in [17]. The authors used the pre-trained models ResNET50, Inception V3, and Inception-ResNetV2. They achieved the best performance with the pre-trained model ResNET50. Models with a location-attention mechanism could more accurately diagnose COVID-19 at the chest radiography with an overall accuracy rate of 86.7% and could be a promising supplementary diagnostic method for frontline clinical physicians [18]. The symptoms and establishing the COVID-19 diagnosis based on chest X-ray images are shown in the article [19]. Developing an intelligent healthcare system [20] is a crucial factor for further improvements from the perspective of patients, healthcare employees, and medical science researchers. To support early detection efforts, researchers proposed a Genetic Deep Learning Convolutional Neural Network model enhanced with Huddle Particle Swarm Optimization instead of traditional gradient descent. Using public chest X-ray datasets, the model effectively identifies COVID-19 and other pneumonia types, achieving 97.23% accuracy, 98.62% sensitivity, 97.0% specificity, and 93.0% precision. Authors emphasize that future work should involve expanding to larger datasets and broader lung disease prediction [21]. Many institutions improved their IT infrastructure to prepare for the exponential rise of patients suffering from COVID-19 [22]. The prediction about the visits of patients can be helpful for stakeholders in hospitals or other health centers to make decisions in managing human resources and materials [23].

The survey paper [24] comprehensively examines various interpretation and visualization techniques applied to deep learning models in medical imaging. This paper reviews methodologies, discusses their applications, and evaluates their effectiveness in enhancing the interpretability, reliability, and clinical relevance of deep learning models in medical image analysis. While significant progress has been made, challenges related to scalability, clinical integration, robustness, standardization, and ethical considerations persist.

This paper [25] presents a comprehensive review of ChatGPT's applications in healthcare, focusing on its role in enhancing patient engagement, symptom assessment, and decision support to improve diagnostic accuracy. It evaluates ChatGPT's usefulness across medical specialties and highlights its benefits in clinical, educational, and administrative settings. While the model seems promising, the review also addresses key challenges, including potential inaccuracies, ethical issues, and regulatory concerns. The goal is to provide a balanced perspective for healthcare professionals, researchers, and policymakers, emphasizing the need for continual updates as medical knowledge evolves.

## 3. Materials and Methods

This research uses two datasets and two different deep learning techniques for binary classifications: one for X-ray image classification using CNN and another for optimizing time slot scheduling using DNN. The Graphics Card NVIDIA GTX 1080 Ti 11 GB GDDR5X, Intel Core i7-9700K, 32 GB RAM, and 512 GB SSD are used to train proposed models. As software tools during research, we used TensorFlow 2.18.0 [26], Keras 3.7.0 [27], NumPy 2.0.2 [28], pandas 2.2.3 [29], matplotlib 3.10.0 [30], OpenCV 4.12.0 [31], scikit-learn 1.6.0 [32], and PyCharm 2024.3.1.1 for Python 3.12.8 [33].

### *3.1. Model for the Chest X-Ray Image Classification*

This research uses a dataset, shared by Dr. Joseph Cohen [34], to classify chest X-ray images. The dataset contains chest X-ray/CT images with COVID-19, acute respiratory distress syndrome (ARDS) [35], MERS [36], SARS [37], and pneumonia. The authors used 98 X-ray images that were labeled with COVID-19. Also, 98 non-COVID-19 chest X-ray images were selected from the dataset Chest X-Ray Images (Pneumonia) [38]. Our research is based on 98 COVID-19 and 98 patients with Non-COVID-19 disease (a total of 196 images). All images are preprocessed: adjusted contrast, adjusted saturation, removed black bars, and saved in 224 × 224 image format.

Deep CNN [39] models may take longer to train on large datasets. Transfer learning [40] involves using models trained on one problem but used on a related similar problem. Transfer learning decreases the time necessary for training deep neural network models and provides a better reduction of the generalization error. A way to shorten this process is to reuse the model weights from pre-trained models developed for standard computer vision benchmark datasets, such as the ImageNet [41] image recognition tasks [42]. We created the CNN model based on the VGG19 [43] pre-trained model to classify chest X-ray images into two classes: Non-COVID-19 and COVID-19 (Figure 2).

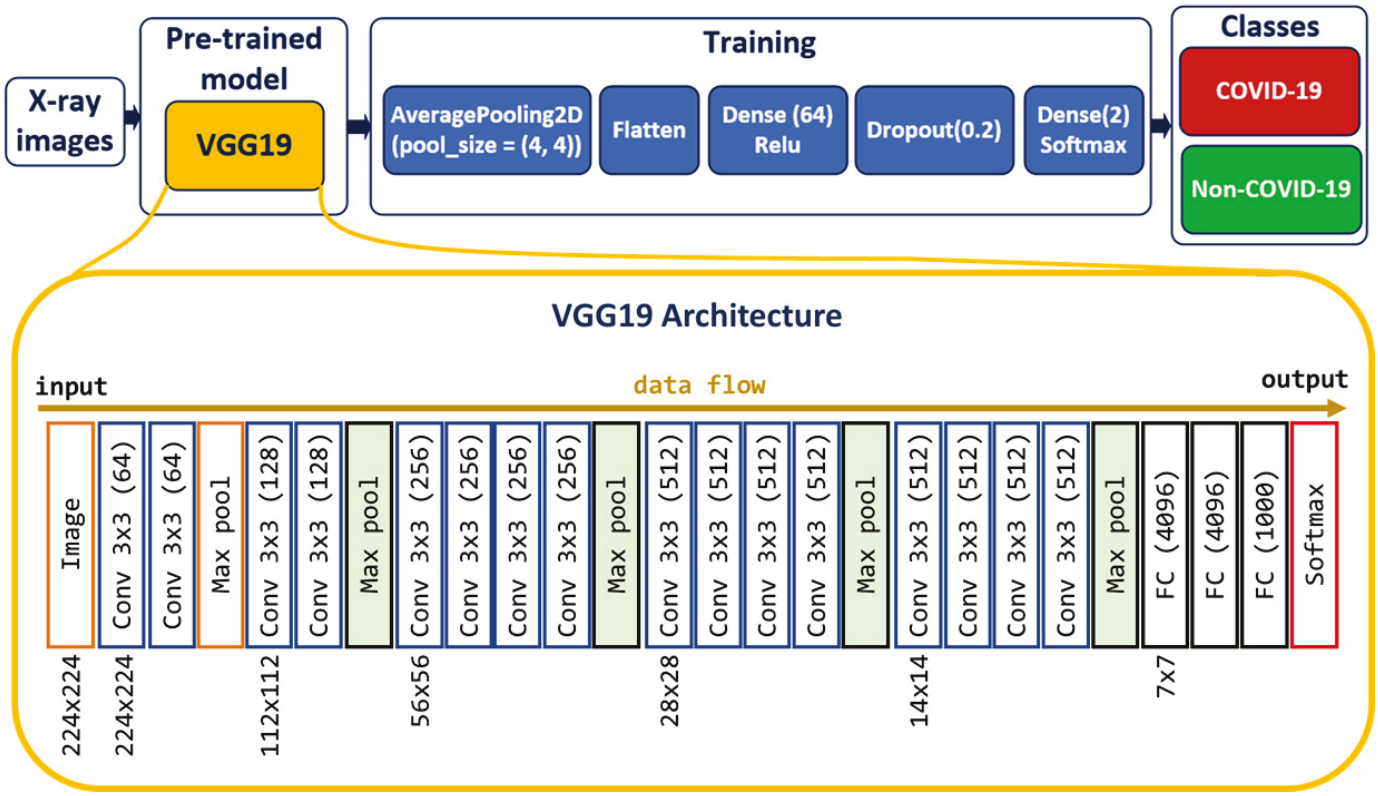



**Figure 2.** Model for predicting COVID-19 and non-COVID-19 cases.

All images are randomly split into two separate datasets with a ratio of 80% (for training) and 20% (for testing). The k-fold cross-validation [44] is used to evaluate the CNN model. The value for k is set to 5 (Figure 3). The batch size is set to 32, the number of epochs to 300, and the learning rate to 0.001 (hyperparameters were determined experimentally).

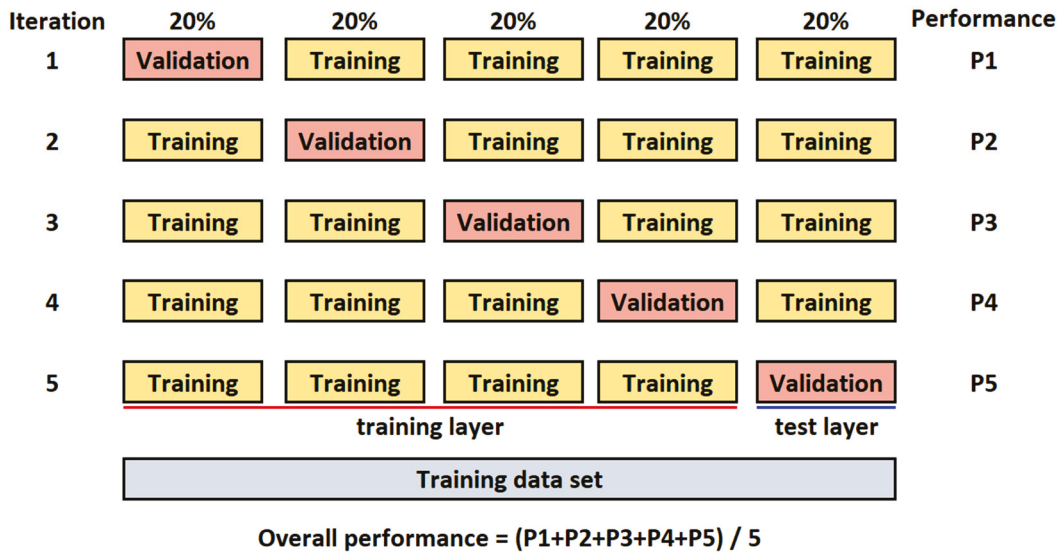



**Figure 3.** K-fold cross-validation for k = 5.

### *3.2. Model for the Optimization of Time Slots Scheduling*

The City of Nis is the third-largest city in the Republic of Serbia (RS) by area and population and became the second-largest COVID-19 hotspot in the country. It is served by a single health center with 63 branches, making the Health Center Nis (HCN) the largest primary healthcare provider in Serbia and the Balkans, serving around half a million patients. In RS, several nationwide licensed MISes are used, one of which is MEDIS.NET [45]. Developed at the Faculty of Electronic Engineering in Nis, it is used in 25 primary healthcare facilities across southeastern RS and has been in operation at HCN for over 15 years.

The dataset used for training the DNN model is extracted from the database of HCN. It contains features presented in Table 1, which are related to the time slot scheduling of patients. The data in the dataset are extracted from the database for a period of 5 years (from 2015 to 2019).

**Table 1.** List of features in the dataset.

| Features | Features Tag | Descriptions |
|---|---|---|
| Doctor | D | Physician's ID |
| Gender | G | 1—Male, 0—Female |
| Chronic diagnosis | Cd | 0—No, 1—Yes |
| Day of week | dow | [1–7], 1—the first day of the week (Monday) |
| Quartal | Q | [1–4], annual quarter |
| Week number | Wn | Week number of the year |
| Family feast | Ff | The scheduled examination is on the date of the top 10 family feasts in Serbia |
| New Year's Day | nyd | 30th, 31st December and 1st, 2nd,3rd January |
| Christmas | Ch | 6th, 7th, 8th January |

**Table 1.** *Cont.*

| Features | Features Tag | Descriptions |
|---|---|---|
| Type of health insurance | Thi | 1—employed (codes: 10 **)<br>2—unemployed (codes: 9999, 11 **)<br>3—pensioners (codes: 12 **)<br>4—other types of insurance (codes: 15 **)<br>5—sensitive groups (codes: 1820, 421)<br>6—foreign citizens (codes: 17 **)<br>7—state health care users (codes: 18 **)<br>8—special regulations (codes: 19 **)<br>9—soldiers (codes: 90 **, 434, 1012, 1913)<br>10—conventions (codes: 20 **)<br>11—pregnant women (codes: 412, 1811) |
| Patient age | Pa | 1—[less than 1] years<br>2—[1–6] years<br>3—[7–18] years<br>4—[19–34] years<br>5—[35–49] years<br>6—[50–64] years<br>7—[65–74] years<br>8—[75+] years |
| Seasons | S | 1—Spring<br>2—Summer<br>3—Autumn<br>4—Winter |
| Specialty for scheduling | Ss | 1—neuropsychiatry<br>2—dentist<br>3—psychiatry<br>4—medical doctor<br>5—dermatovenerology<br>6—otorhinolaryngology<br>7—physical medicine and rehabilitation<br>8—internal medicine<br>9—pediatrics<br>10—ophthalmology<br>11—gynecology and obstetrics<br>12—general medicine |
| Number of days until the scheduled date | ndusd | Number of days from the moment of scheduling an examination slot until the actual examination |
| Percent of non-attendance | pna | Percentage of missed scheduled appointments by patients |
| Status | stat | 0—patient is predicted to arrive at the scheduled time slot<br>1—patient is predicted not to arrive at the scheduled time slot |

**—Represent combinations of other available types of health insurance codes in the Republic of Serbia.

The relationship between different input features in a dataset is presented in Figure 4 and indicates how much they tend to change together. The distribution of data within a dataset is shown in Figure 5.

**Figure 4.** Features correlation between different input features in a dataset. A darker color represents a stronger correlation between the corresponding features.

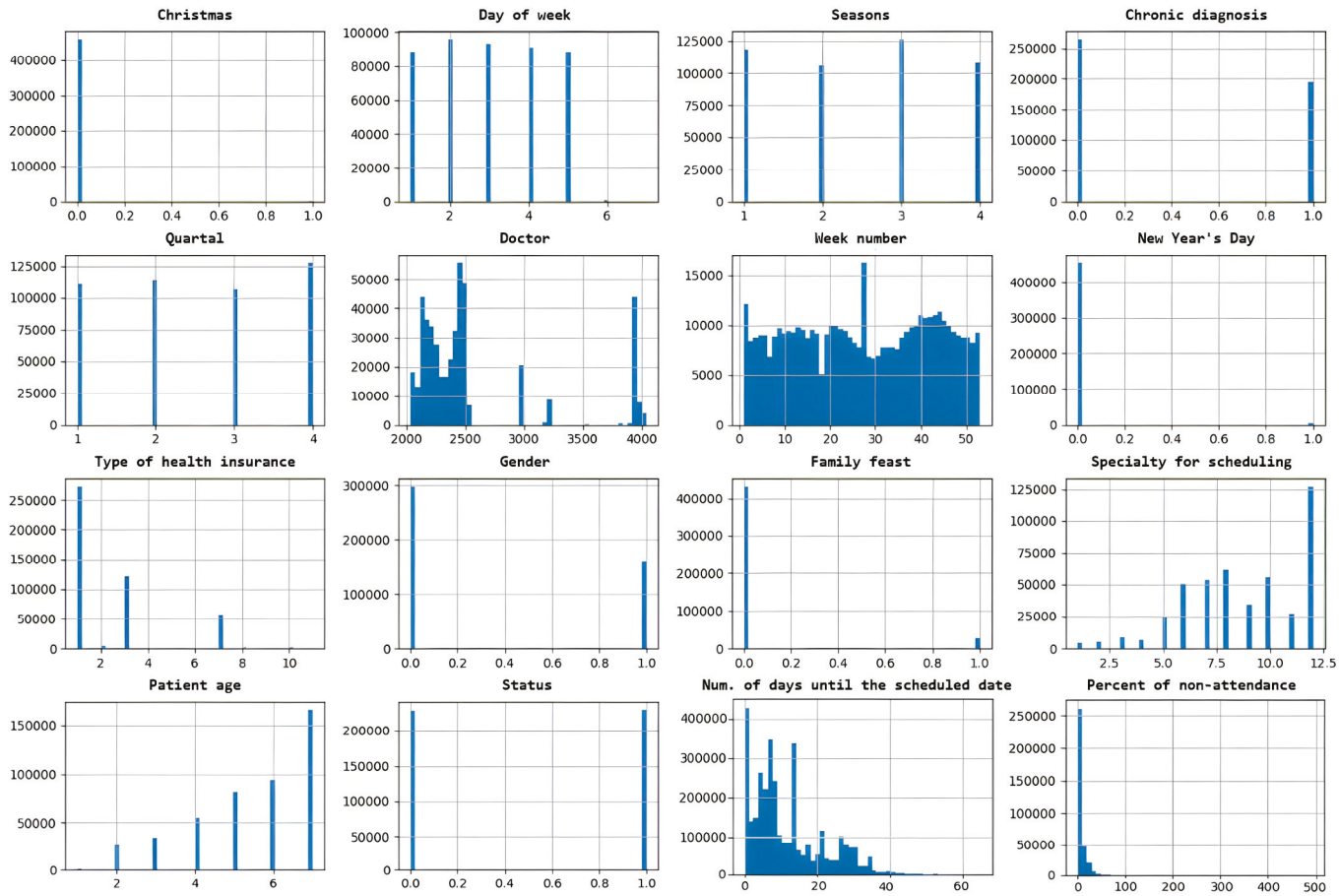



**Figure 5.** Distribution of data within a dataset.

The DNN model is created to binary classify scheduled time slots by feature status (0—patient is predicted to arrive at the scheduled time slot, 1—patient is predicted not to arrive). The dataset initially contained 388,194 rows. After processing (removing rows with nullable values, all values are normalized), the data dataset contains 341,569 (88%) rows. The dataset is very well balanced by feature status (Figure 5), with almost 50% for classes 0 and 1. It is randomly split into two divided datasets with a ratio of 80% (for training) and 20% (for testing). The k-fold cross-validation is used to evaluate the DNN model. The value for k is set to 5. The batch size is set to 512, the number of epochs to 100, and the learning rate to 0.01 (hyperparameters were determined experimentally). The model for binary classification is shown in Figure 6.

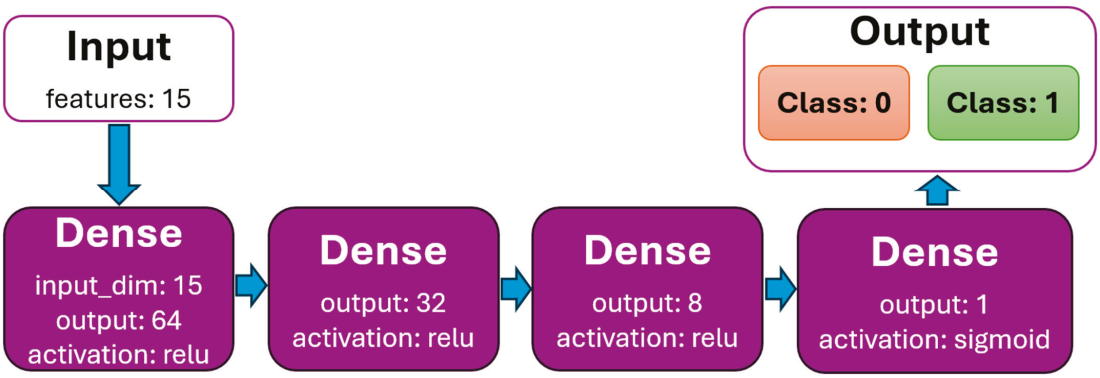



**Figure 6.** DNN model graph with dense layers.

### *3.3. Performance Metrics*

In order to measure the performance of proposed models, a Classification report [46] is used. Variables TP, FP, TN, and FN are True Positive, False Positive, True Negative, and False Negative, respectively. The accuracy, sensitivity, specificity, precision, and F1-score [47] were used to calculate the model's performance.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{F1-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

The accuracy determines how correctly the values are predicted. The precision determines how many of the predictions are correct. The recall determines how many of the correct results are discovered. The F1-score uses precision and recall calculations. It is a measure of balanced average results.

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

Sensitivity is called the true positive rate (TPR), recall, or probability of detection. This is a measure of actual positives that are correctly identified (e.g., the ability of a test to identify patients with a disease correctly).

$$\text{specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{6}$$

Specificity is called the true negative rate (TNR). This is a measure of actual negatives that are correctly identified (e.g., the ability of a test to identify a patient without the disease correctly).

## 4. Results

This section presents the achieved results and performance of the two developed DNN models, as well as their application in MIS.

### *4.1. Model for the Chest X-Ray Image Classification*

Figure 7 presents the performance of the model for chest X-ray image classification. The confusion matrix and classification report are shown in Figure 8. The maximum accuracy, precision, sensitivity, specificity, and F1-score are 94.87%, 94.74%, 94.74%, 95.00% and 94.74%, respectively, for fold number 3. The average accuracy, precision, sensitivity, specificity, and F1-score across all folds are 91.83 ± 2.82%, 92.78 ± 4.36%, 90.83 ± 2.22%, 92.90 ± 4.42% and 91.76 ± 2.78%. Table 2 shows metrics for a model for chest X-ray image classification.

**Table 2.** Metrics for a model for chest X-ray image classification.

| Fold Number | Accuracy | Precision | Sensitivity | Specificity | F1-score |
|---|---|---|---|---|---|
| 1 | 0.8718 | 0.8500 | 0.8947 | 0.8500 | 0.8718 |
| 2 | 0.9231 | 0.9474 | 0.9000 | 0.9474 | 0.9231 |
| 3 | 0.9487 | 0.9474 | 0.9474 | 0.9500 | 0.9474 |
| 4 | 0.9231 | 0.9444 | 0.8947 | 0.9500 | 0.9189 |
| 5 | 0.9250 | 0.9500 | 0.9048 | 0.9474 | 0.9268 |
| Mean: | 0.9183 | 0.9278 | 0.9083 | 0.9290 | 0.9176 |
| 95% CI: | [0.8834, 0.9533] | [0.8738, 0.9819] | [0.8807, 0.9359] | [0.8741, 0.9838] | [0.8830, 0.9522] |
| Standard Deviation: | 0.9183 ± 0.0282 | 0.9278 ± 0.0436 | 0.9083 ± 0.0222 | 0.9290 ± 0.0442 | 0.9176 ± 0.0278 |

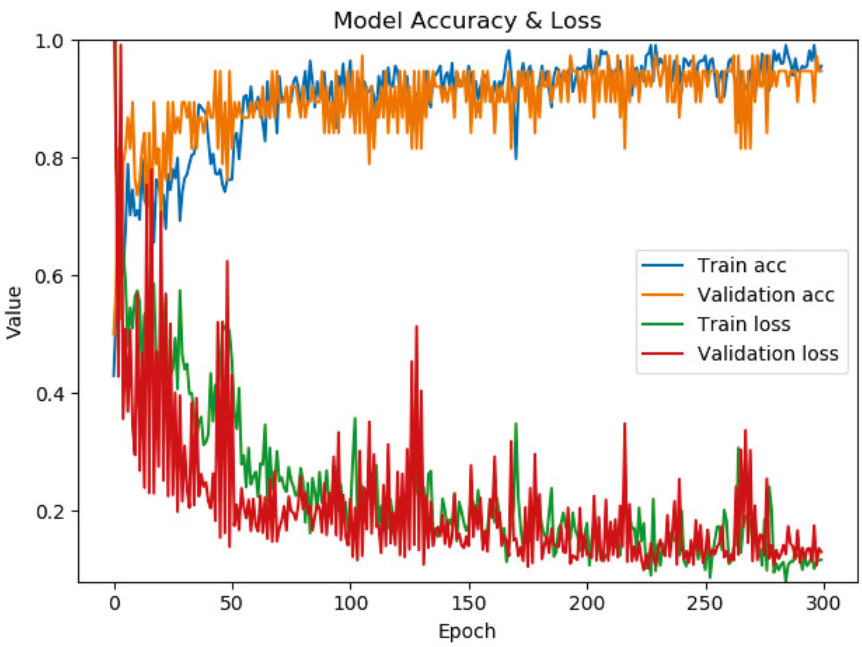



**Figure 7.** The performance of the model for chest X-ray image classification.

| Predicted \ Actual | COVID-19 | Non-COVID-19 |
|---|---|---|
| COVID-19 | TP<br>94.7%<br>18 | FP<br>5%<br>1 |
| Non-COVID-19 | FN<br>5.3%<br>1 | TN<br>95%<br>19 |

**A** - Confusion matrix results of the model for chest X-ray images classification for binary classification.

| Classes | Precision | Recall | F1-score |
|---|---|---|---|
| Non-COVID-19 | 0.95 | 0.95 | 0.95 |
| COVID-19 | 0.95 | 0.95 | 0.95 |
| Macro avg | 0.95 | 0.95 | 0.95 |
| Weighted avg | 0.95 | 0.95 | 0.95 |

**B** - Classification report of model for chest X-ray images classification.

**Figure 8.** Confusion matrix results and classification report of the model for chest X-ray images classification for binary classification.

TP is the number of patients (94.7%) that the model classified as having COVID-19, and they are patients with that diagnosis. TN is the number of patients (95%) predicted by the model not to have the disease COVID-19, and who actually do not have the disease. Patients without and with the disease can be identified with an accuracy of 95% using the proposed model. FP is the number of patients (5%) predicted by the model to be positive for COVID-19, but actually they are not. This is where the problem lies because such patients would most likely be sent to a temporary COVID-19 hospital and thus stay in a highly infectious environment. There is a high chance that such patients will become infected with COVID-19 after additional exposure to the viral environment. FN is the number of patients (5.3%) predicted by the model to not be positive for COVID-19 but are actually positive. Such patients would be discharged without an established diagnosis of COVID-19 and could infect a larger population of people with whom they have social contact and would not receive adequate healthcare. Nonetheless, chest X-ray imaging is performed for patients who have a distinct picture of the disease of COVID-19, so they would certainly not be immediately discharged home without additional testing (RT-PCR testing). In order to increase the performance of the model, it is necessary to increase the percentage of TP and TN and decrease FP and FN.

### *4.2. Model for the Optimization of Time Slots Scheduling*

Figure 9 presents the model's performance for optimizing time slot scheduling. The confusion matrix and classification report are shown in Figure 10. The maximum accuracy, precision, sensitivity, specificity, and F1-score calculated on the confusion matrix are 78.23%, 82.82%, 72.59%, 84.16%, and 77.37%, respectively, for fold number 4. The average accuracy, precision, sensitivity, specificity, and F1-score across all folds are 77.51 $\pm$ 0.7%, 82.53 $\pm$ 0.46%, 71.57 $\pm$ 1.08%, 83.86 $\pm$ 0.38% and 76.66 $\pm$ 0.70%. Table 3 shows metrics for a model for the optimization of time slot scheduling.

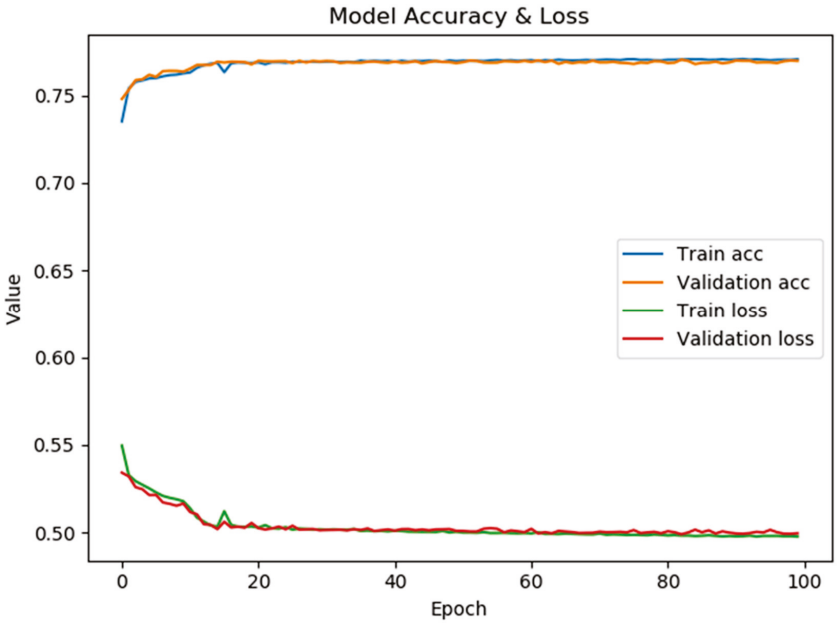
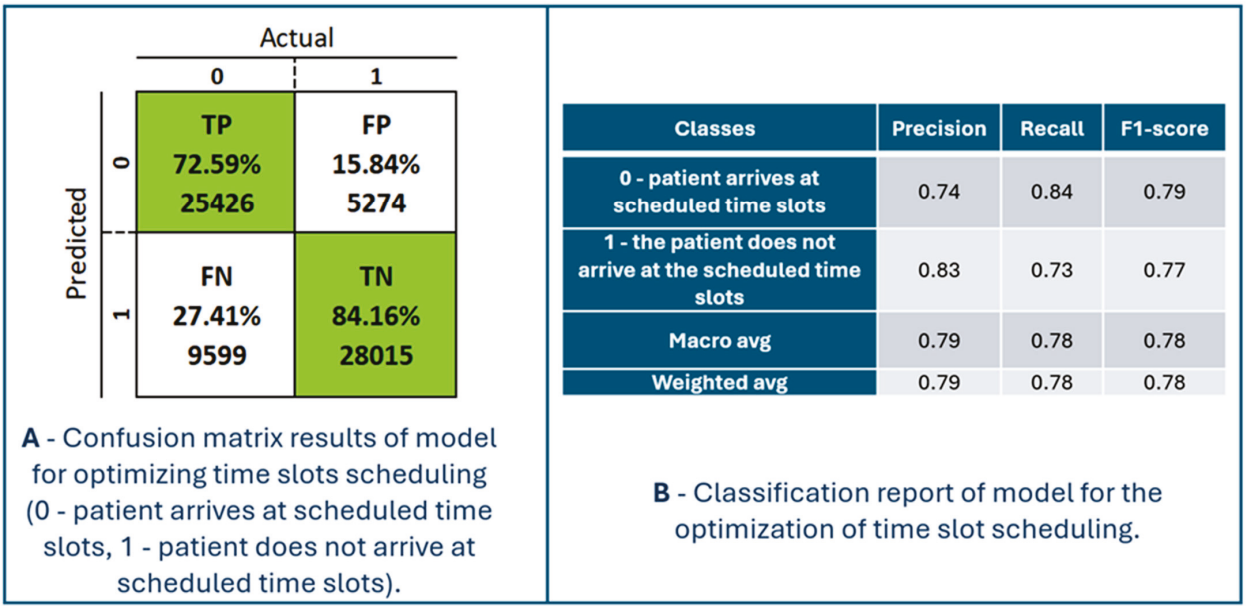



**Figure 9.** The performance of the model for the optimization of time slot scheduling.




| Classes | Precision | Recall | F1-score |
|---|---|---|---|
| 0 - patient arrives at scheduled time slots | 0.74 | 0.84 | 0.79 |
| 1 - the patient does not arrive at the scheduled time slots | 0.83 | 0.73 | 0.77 |
| Macro avg | 0.79 | 0.78 | 0.78 |
| Weighted avg | 0.79 | 0.78 | 0.78 |

**Figure 10.** Confusion matrix results and classification report of the model for optimizing time slots scheduling (0—patient arrives at scheduled time slots, 1—patient does not arrive at scheduled time slots).

TP is the number of patients (72.59%) the model predicted would arrive at the scheduled time. Those patients would definitely come at their scheduled time. This has no positive or negative effects on the model. TN is the number of patients (84.16%) that the model classified as not coming to the scheduled examination, and they really did not come. It is possible to double appointments in those appointments to ensure a higher degree of occupancy of the available appointments. FP is the number of patients (15.84%) for whom the system predicted they would come for an examination but who did not come for the scheduled examination. Therefore, those appointments remain unfilled. To increase the occupancy of this time slot, two patients can be scheduled at the same time. FN is the number of patients (27.41%) for whom the model predicted that they would not come for examination, but they did come. Such patients are generally a potential problem for

crowding and overbooking appointment slots. In real situations, a little space is always left between two adjacent slots to schedule an examination to provide medical services to such patients. The duration of those breaks between two appointment slots is defined depending on the specialization and the type of service in the health center. For the system to have a better performance, it is necessary to increase the percentages of TP and TN and decrease the percentages for FP and FN, in order to increase accuracy, sensitivity (recall), specificity, precision, and F1-score.

**Table 3.** Metrics for a model for the optimization of time slot scheduling.

| Fold Number | Accuracy | Precision | Sensitivity | Specificity | F1-score |
|---|---|---|---|---|---|
| 1 | 0.7697 | 0.8231 | 0.7074 | 0.8367 | 0.7609 |
| 2 | 0.7819 | 0.8234 | 0.7284 | 0.8376 | 0.7730 |
| 3 | 0.7748 | 0.8318 | 0.7119 | 0.8433 | 0.7672 |
| 4 | 0.7823 | 0.8282 | 0.7259 | 0.8416 | 0.7737 |
| 5 | 0.7669 | 0.8202 | 0.7047 | 0.8338 | 0.7581 |
| Mean: | 0.7751 | 0.8253 | 0.7157 | 0.8386 | 0.7666 |
| 95% CI: | [0.7665, 0.7838] | [0.8196, 0.8311] | [0.7022, 0.7291] | [0.8338, 0.8434] | [0.7579, 0.7753] |
| Standard Deviation: | 0.7751 ± 0.0070 | 0.8253 ± 0.0046 | 0.7157 ± 0.0108 | 0.8386 ± 0.0038 | 0.7666 ± 0.0070 |

### *4.3. AI Tools in MIS That Use Developed Models*

Figure 11 shows two modules in MEDIS.NET that use described and developed models to classify chest X-ray images and optimize time slot scheduling. The physician can activate the tool for the chest X-ray to diagnose COVID-19 automatically from the patient's health record. The tool relies on the developed CNN model for classifying chest X-ray images.

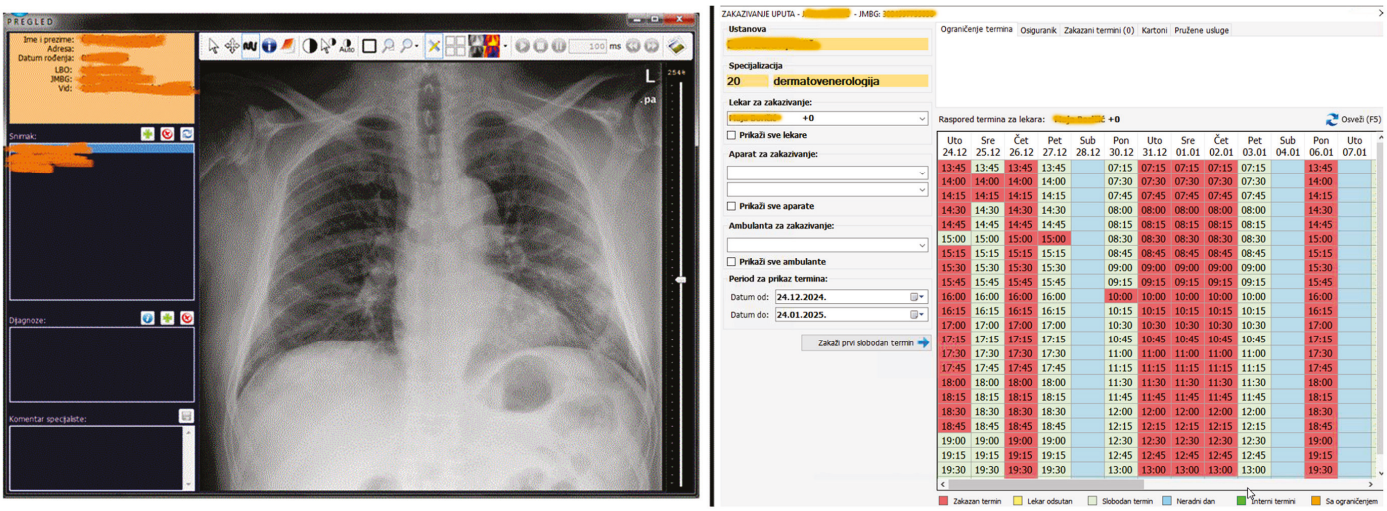



**Figure 11.** The MIS MEDIS.NET modules use previously described, developed, and trained models (**Left**—Module for the Department of Radiology, **Right**—Module for scheduling patient appointments).

During the examination appointment for each selected patient, an estimate can be obtained as to whether the patient will appear at the scheduled time. That functionality relies on the developed DNN model.

Figures 12 and 13 show how both developed models are integrated into MIS MEDIS.NET and EHR.

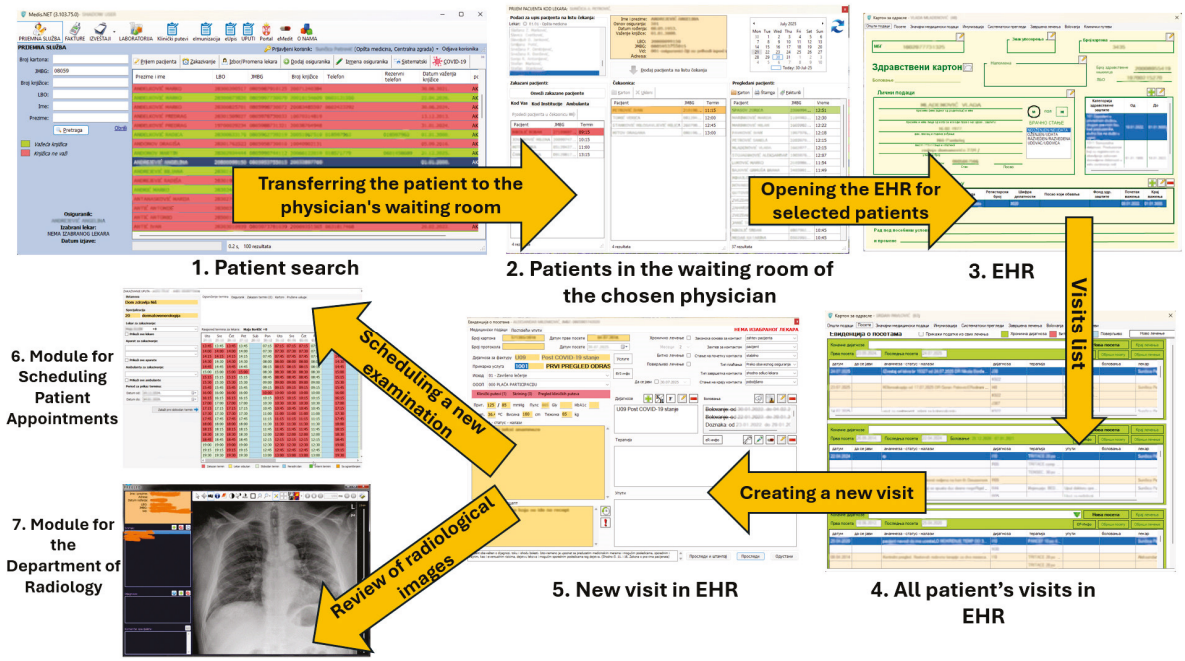



**Figure 12.** Scheduling an appointment for a new examination and using the Module for the Department of Radiology via the EHR.

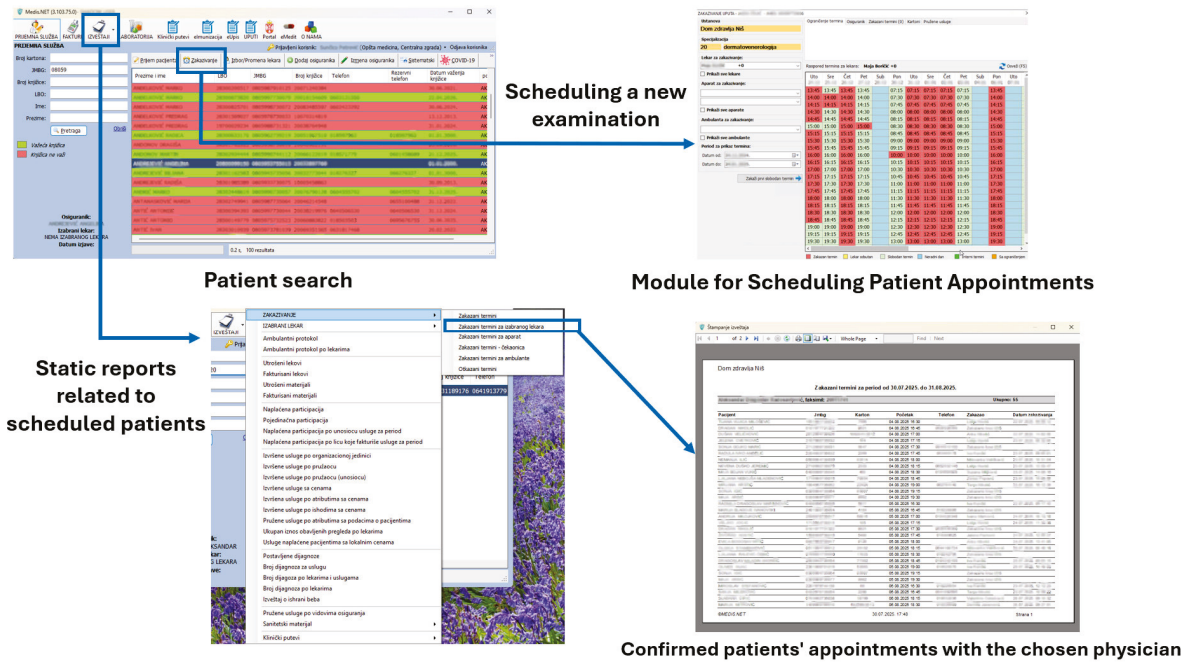



**Figure 13.** Scheduling an appointment for a new examination through the Call Center of a health center that uses the MIS MEDIS.NET.

## 5. Discussion

In this research, chest X-ray images are used to predict COVID-19, and an average accuracy of 91.83 ± 2.82% is accomplished by using the pre-trained VGG19 model for the classification of non-COVID-19 and COVID-19 cases. Accuracy values are increased through the training process of 300 epochs, while loss values are decreased (Figure 7). The performance metrics of the model for X-ray image classification are presented in Figure 7. In [48], the authors proposed a detection of COVID-19 by using X-ray images based on

deep features and SVM. They obtained an accuracy of 95.38% for ResNet50 and SVM. The authors in [18] proposed a prediction model for distinguishing COVID-19 from influenza-A viral pneumonia by using CT images. They used deep learning techniques and achieved an accuracy of 86.7%. The Inception transfer learning model is used by the authors in [49] to predict COVID-19 cases by CT images. They accomplished an accuracy of 89.5%. In [17], the author compares three pre-trained models for X-ray image classification: InceptionV3, ResNet50, and Inception ResNetV2, and obtained the accuracy of 97%, 98%, and 97%, respectively. In their research, the best result was obtained by using a pre-trained ResNet50 model with an accuracy of 98%.

Our trained model for the classification of chest X-ray images can be used at the Department of Radiology at the HCN as a supporting tool for physicians during and after the COVID-19 pandemic (Figure 11, left side). Radiologists in the HCN can use this new tool within the already existing radiology module. This tool can be well-accepted and more appealing to young physicians and physicians at the Department of Radiology in HCN, during the COVID-19 pandemic. Despite the satisfactory classification average accuracy of 91.83 $\pm$ 2.82%, the radiologists are those who make the final decision to establish the diagnosis of COVID-19 and to decide whether to hospitalize the patient.

The developed DNN model for the optimization of time slot scheduling is used for the implementation of the subsystem of MEDIS.NET for the intelligent identification and assessment of patients who will not come to the appointment with the chosen physician or/and to expensive diagnostic examinations for which a patient needs to wait, sometimes even for several months. The subsystem is part of the Scheduling Module of MEDIS.NET, shown in Figure 11 on the right side. The developed DNN model is used for the classification of scheduled patient examinations into two classes: 0—patient is predicted to arrive at the scheduled time slot, 1—patient is predicted not to arrive at the scheduled time slot. For each created patient appointment, the physician receives a message that informs him/her about the possibility of the patient's non-arrival in the scheduled time slot. During the COVID-19 pandemic, patients were able to schedule an examination appointment directly with the chosen physician by calling the Call Centre of the HCN (Figure 13). At the moment of scheduling, using the developed DNN model, a Call Center operator can receive an estimate, the classification of the scheduled appointment is determined into the two previously mentioned classes, and information is provided about the potential non-arrival of a patient to the scheduled appointment. The operator can mark the slot in that way, allowing the other patient to be scheduled simultaneously. Usually, in those overlapping slots, those patients are scheduled for whom the system issues a warning during the scheduling process that they will not arrive at the scheduled appointment. Airline companies worldwide have used similar principles for overbooking seats for many years. The paper [50] shows optimal overbooking strategies in airlines by using a dynamic programming approach in continuous time. By using the developed DNN model, upon the request of the health center employees, each patient who had a scheduled examination appointment could check whether he/she would arrive at the time of the appointment, and a corresponding report is generated. After the system check, daily reports were created about the patients who were not going to arrive at the scheduled appointments as per the assessment of the system. After the check, the patients were contacted by phone to confirm their arrival at the scheduled appointments. It was required to contact patients with scheduled appointments on expensive diagnostic devices, like CT and MRI, by phone and who were classified as patients who were potentially not arriving at the examination (class 1). During the COVID-19 pandemic, all physicians at the health center worked according to the changed, reduced schedule only a few days a week. The number of

possible time slots for scheduling appointments was significantly reduced (72%), making it more difficult for patients to book their desired appointments. Due to the reduction of possible slots for the scheduling of appointments, a check about the possible non-arrival of patients at their appointments was necessary. Even beyond the work reduction, the number of scheduled appointments on a daily basis was immense. The average number of scheduled slots daily was 1738. Because of such a huge number of appointments, Call Center operators were not able to call each patient separately to confirm their arrival at the scheduled appointment. The average percent of non-arrivals of patients at the scheduled slots during the COVID-19 pandemic was 62%. If more patients arrived for one scheduled slot, the patients would then be examined by the first available physicians. Such a case is unlikely to happen for scheduled slots on expensive diagnostic devices (CT, MRI) since the patients were contacted about their arrival at the appointment in advance.

The satisfaction of end users, who directly or indirectly use the software, is always a key goal during the development of large and complex information systems, especially those such as medical information systems. During the COVID-19 pandemic, we were unable to interview doctors and patients who were both direct and indirect users of the system. In regular circumstances, overbooking patients for one appointment slot is unfeasible. This does not mean that today this model is unusable, but that it can be used for resource planning (appointments are made with the chosen physicians, but also with diagnostic devices). According to the regulations of the Republic of Serbia, it is possible to schedule only one patient per time slot. In Serbia, a central registry for scheduling examinations was developed, and all MISes were integrated with it.

The model's average accuracy for optimizing time slot scheduling is 77.51 $\pm$ 0.7%. Accuracy values are increased through a training process of 100 epochs, while loss values are decreased (Figure 9). In order to increase the efficacy of the classification of time slot scheduling, the authors of this paper have developed two more models by using Random Forest [51] and Logistic Regression [52] for the same dataset created for the training of the neural network model. After training the models, the accuracy for the developed models that used Random Forest (Figure 14) and logistic regression is 74% and 73%, respectively. With these three models, the voting system is created. The voting system is used to more precisely determine whether the patient is arriving at the scheduled appointment. The soft voting method is applied, treating all classifiers equally. The final output is generated by averaging the outputs of each individual classifier [53] as shown in Figure 15. Using the voting system, the number of non-arrivals is reduced to below 10%.

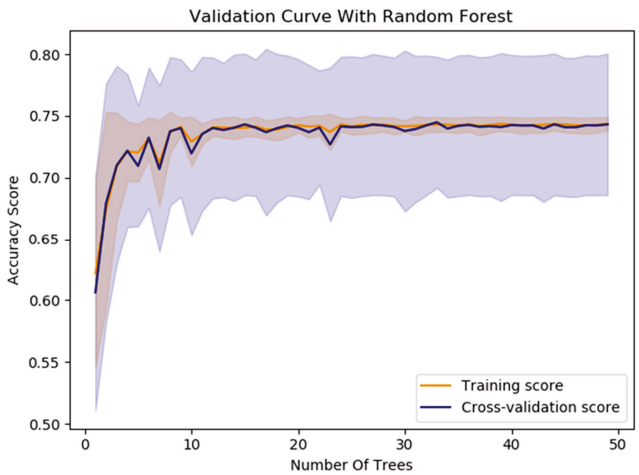



**Figure 14.** The Random Forest model's performance for optimizing time slot scheduling.

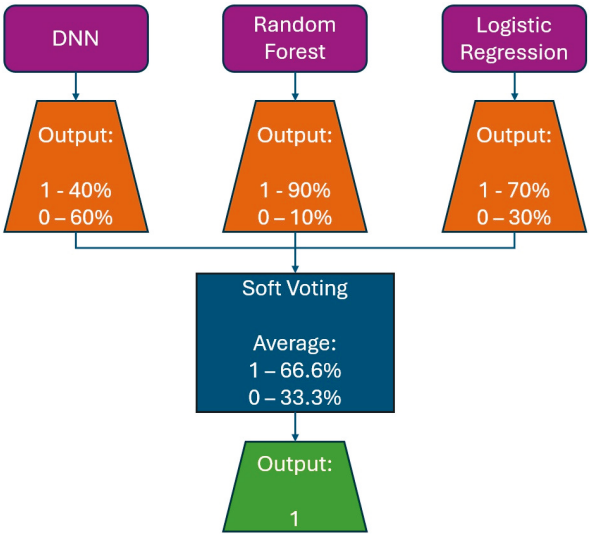



**Figure 15.** Example for soft voting ensemble of DNN, Random Forest, and logistic regression classifiers.

### *5.1. Architecture of MIS MEDIS.NET and Position of AI Tools*

The architecture of MIS MEDIS.NET is shown in Figure 16. The MIS MEDIS.NET is composed of several interconnected modules designed to support healthcare professionals in performing their tasks more efficiently, while also ensuring comprehensive storage and analysis of medical data. These modules significantly contribute to the efficient and coordinated operation of healthcare institutions. At the heart of the MIS is the EHR, which serves as its foundational component. Beyond the EHR, the system also incorporates modules for financial management, administration, and user privileges, static and dynamic reporting, as well as mHealth and TeleHealth functionalities. In addition, it includes Laboratory and Radiological Information Systems, along with a set of artificial intelligence tools, which contain the two previously described modules. Both realized modules (Module for the Department of Radiology, Module for Scheduling Patient Appointments) are directly dependent on artificial intelligence tools, electronic health records, and the Radiological Information System. The models presented in this paper have been developed and integrated into artificial intelligence tools.

Figure 16 shows the collaboration between MIS and other available services in the Republic of Serbia. The MIS MEDIS.NET is equipped with services that enable communication and data exchange with external systems. In the RS, a Vertical Integration has been established, enabling the secure exchange of key patient information between different healthcare providers and MISes. This system allows physicians to access medical data from other institutions upon obtaining patient consent, which is particularly important for providing more personalized and continuous care. Moreover, centralized data exchange on a national level supports monitoring of rare and chronic diseases, electronic prescription management, school enrollment procedures, appointment scheduling, preventive care, public health surveillance, along with other functionalities aimed at improving the accessibility and quality of care.

This paper [54] identifies key challenges and solutions for the implementation of modern MIS MEDIS.NET in a Serbian healthcare organization, including issues such as poor infrastructure, complex user interfaces, and low IT literacy. It outlines strategies for software architecture, development, and deployment to improve system usability and adoption. In the paper [55], the use and acceptance of the MIS MEDIS.NET in Serbian primary healthcare are analyzed. The study evaluates the success rate of entering various medical records and assesses user acceptance through the Technology Acceptance Model,

focusing on perceived ease of use and usefulness. Results show high acceptance and accuracy for frequently used features, such as service registration. In contrast, specialized functionalities (e.g., for adult physical exams) are used less, indicating areas for further system improvement.

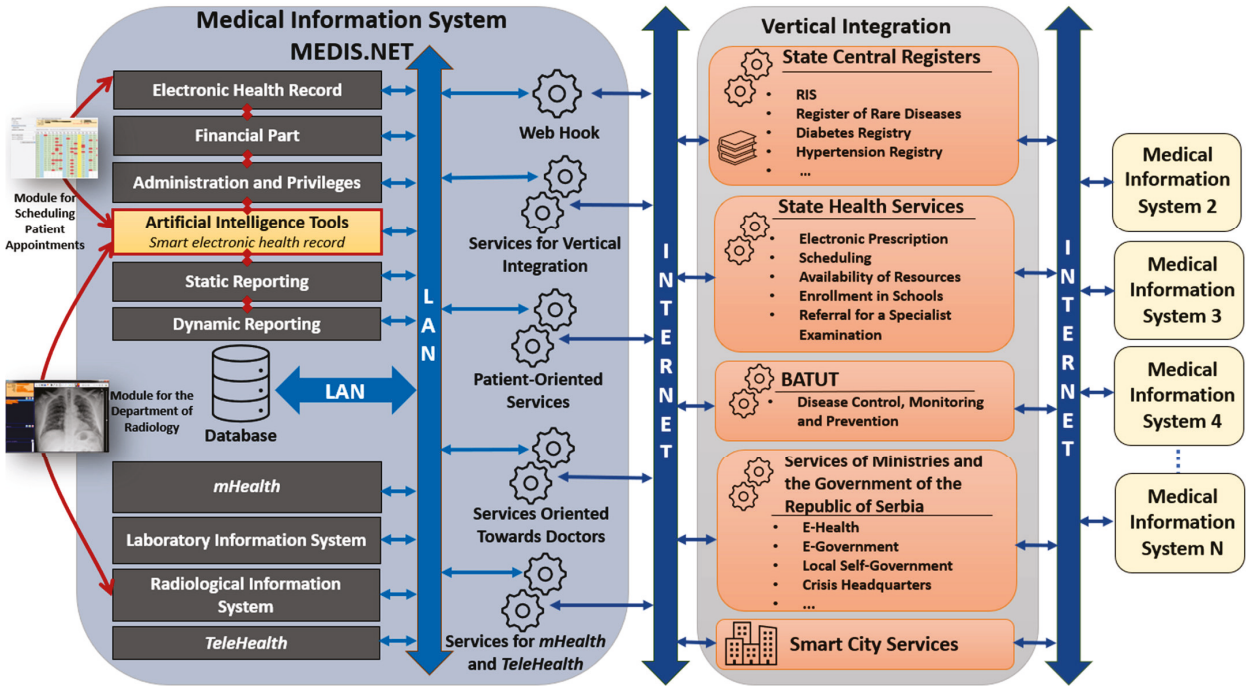



**Figure 16.** Architecture of MIS MEDIS.NET and position of artificial intelligence tools.

### *5.2. Physician-Centered Decision-Making Supported by AI*

Table 4 provides a comparative analysis of three possible approaches to decision-making in MIS: AI-assisted decision-making, traditional physician decision-making, and the AI-supported combined approach. Each option has its advantages and disadvantages, including accuracy, speed, risk of errors, and the level of physician intervention. This analysis facilitates the understanding of how AI can enhance MISes while simultaneously emphasizing the importance of the physician's role in final decision-making. This study, presented in this paper, is primarily focused on physician-centered decision-making supported by AI.

**Table 4.** Pros and cons of decision-making using AI vs. physician-centered decision-making vs. physician-centered decision-making supported by AI in MIS and EHR.

| Decision-Making Using AI | Physician-Centered Decision-Making | Physician-Centered Decision-Making Supported by AI |
|---|---|---|
| **Pros:** | **Pros:** | **Pros:** |
| High accuracy in identifying patterns in large datasets. | Leverages experience, expertise, and patient history. | Combines AI's data analysis with the physician's clinical judgment, improving decision quality. |
| Reduces human error, especially in complex diagnoses (e.g., cancer detection). | Personalized care based on the physician's knowledge and understanding of the patient. | AI enhances decision-making by providing additional insights and predictions. |
| Speeds up the diagnostic process, leading to quicker treatments. | Physicians can consider a wide range of factors, including non-quantifiable elements like intuition. | Reduces workload by automating routine tasks, allowing physicians to focus on critical decisions. |

**Table 4.** *Cont.*

| Decision-Making Using AI | Physician-Centered Decision-Making | Physician-Centered Decision-Making Supported by AI |
|---|---|---|
| **Cons:** | **Cons:** | **Cons:** |
| Requires large, high-quality datasets to work effectively. | Can be influenced by bias, fatigue, or external factors affecting judgment. | AI recommendations might be misunderstood or misinterpreted without a proper understanding from the physician. |
| Potential for over-reliance on AI, leading to reduced clinical skills over time. | Time-consuming process, especially for complex diagnoses or rare diseases. | AI may not always be trained on specific patient data, leading to inaccurate suggestions or bias in recommendations. |
| Lack of transparency in AI models (black-box problem) can hinder trust and interpretation of results. | May not always have access to the latest research or technological advancements, leading to suboptimal decisions. | Over-reliance on AI could reduce physician autonomy or lead to an "AI-driven" mindset, potentially sidelining human judgment. |
| Ethical concerns, especially with privacy, data usage, and algorithmic bias. | Risk of misdiagnosis or treatment errors due to human limitations or misinterpretation of symptoms. | AI models are only as good as the data they are trained on, and poor-quality data can affect the quality of support it offers. |

### *5.3. Protection of Personal Data*

MIS MEDIS.NET was implemented in accordance with the requirements of the Ministry of Health of the Republic of Serbia, which are based on the EuroRec standard. After implementation, it received a certificate from the Ministry of Health of the Republic of Serbia for use in health institutions in the Republic of Serbia.

We are committed to complying with national data protection laws and the EU General Data Protection Regulation (GDPR). To ensure data privacy and security, we implemented strict safeguards throughout all stages of data management. Data are collected securely, with explicit consent from individuals and clear privacy notices. All data are encrypted during storage and transmission, with access restricted to authorized personnel who use multi-factor authentication. Anonymization or pseudonymization is used where possible (e.g., personal patient and physician data, demographic data, and sensitive information that requires special attention under the Law on Personal Data Protection in Serbia). Anonymized and pseudonymized data are securely destroyed when no longer needed and retained only as long as necessary. Staff in health institutions received extensive lessons on data protection during the deployment of MIS MEDIS.NET. These measures ensure compliance with data protection laws and maintain high standards of privacy. For scientific research purposes and to improve MIS MEDIS.NET, which is used daily, the Faculty of Electronics in Nis and the Laboratory for Medical Informatics have been granted permission to use anonymized data from HCN.

Final decisions about treatment rest solely with the physician, who bears full responsibility and signs the patient's treatment documents. AI tools serve primarily as a support for physicians, streamlining their work routine and clinical workflow.

## 6. Conclusions

We use deep learning techniques to create a model for the classification of chest X-ray images and a model for optimization of the time slot scheduling. With these models, we conduct a study demonstrating the significant impact of deep learning in healthcare, especially during the COVID-19 pandemic. The CNN model efficiently classifies chest X-rays, aiding rapid diagnosis, while the DNN model optimizes appointment scheduling, enhancing resource utilization.

A CNN with a pre-trained VGG19 model classifies chest X-ray images into non-COVID-19 and COVID-19 categories, achieving an average accuracy of 91.83 ± 2.82%. A DNN model with four dense layers classifies patients who miss scheduled appointments, achieving an average accuracy of 77.51 ± 0.7%. Both models are integrated into the MEDIS.NET system and have significantly contributed to combating the COVID-19 pandemic.

Future research should focus on expanding datasets, conducting external validation, refining models, conducting real-world deployment studies, and integrating feedback to support model refinement. This includes creating new DNN models for clustering patients by therapy and risk factors, predicting diagnoses, selecting laboratory tests, and forecasting disease spread. Chest X-ray models could be trained with larger datasets to classify images into categories like COVID-19, SARS, MERS, and pneumonia. Models for time slot scheduling could incorporate additional features like weather, traffic, and chronic therapy expiration. The deep learning techniques used for COVID-19 can also be applied to other epidemics and daily healthcare operations, laying the foundation for improved efficiency and outcomes in healthcare systems.

This study emphasizes the importance of physician-centered decision-making in urgent situations like pandemics, leveraging AI tools that incorporate advanced DNN models. The main objective of this research is to integrate AI into current MISes and improve EHRs with AI, rather than to compete with already established models in the field. In general, these developed models have demonstrated outstanding accuracy and performance. Events like the COVID-19 outbreak highlight scenarios where fast, physician-centered decision-making, supported by AI tools, is essential. The authors acknowledge that AI presents a major challenge for both healthcare professionals and engineers in the field of medical informatics, highlighting the importance of automating physicians' processes and decision-making in daily practice, while ensuring that the physician remains at the center of final decision-making (physician-centered decision-making).

The method presented in the paper can also be applied to other diseases, such as 'coin syndrome' or lung tumors. Firstly, it is essential to collect high-quality data to build a robust dataset. The data from the dataset needs to be well-annotated by experienced physicians in their respective fields of practice. After that, it is necessary to carry out quality preprocessing of the collected data and select those that can be used as a final dataset for training the created model (this process must be carried out under the constant supervision of experts—physicians of the appropriate specialization). After creating the model, training it, and fine-tuning the hyperparameters, it is necessary to validate the model using new data that was not included in the previously created dataset. After that, it is necessary to implement the developed AI model in the existing electronic patient record within the medical information system. Ultimately, experts in the field (physicians of the relevant specialty) must validate the AI model.

The methodologies and results presented here serve as a foundation for integrating deep learning into healthcare systems for improved efficiency and patient outcomes.

**Author Contributions:** Conceptualization, A.M.; methodology, A.M.; software, A.M., P.R. and A.D.; validation, A.M., A.D. and K.E.; investigation, A.M. and A.D.; writing—original draft preparation, A.M.; writing—review and editing, A.M., D.J., K.E. and T.G.; visualization, A.M.; supervision, D.J., K.E. and T.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was partially supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia [grant number 451-03-137/2025-03/200102].

**Data Availability Statement:** Data are contained within the article. The dataset is available upon request from the authors.

**Acknowledgments:** Special thanks are extended to COST (European Cooperation in Science and Technology) Action CA23125—The Metamaterial Formalism Approach to Recognize Cancer.

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

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| AI | Artificial Intelligence |
| ANN | Artificial Neural Network |
| ARDS | Acute Respiratory Distress Syndrome |
| CNN | Convolutional Neural Network |
| DNN | Deep Neural Network |
| EHR | Electronic Health Record |
| FN | False Negatives |
| FP | False Positive |
| HCN | Health Center Nis |
| MIS | Medical Information Systems |
| ML | Machine Learning |
| NLP | Natural Language Processing |
| TN | True Negative |
| TNR | True Negative Rate |
| TP | True Positive |
| TPR | True Positive Rate |
| WHO | World Health Organization |

## References

1. Shneiderman, B. *Human-Centered AI*; Oxford Academic: Oxford, UK, 2022.
2. Chouhan, V.; Singh, S.K.; Khamparia, A.; Gupta, D.; Tiwari, P.; Moreira, C.; Damaševičius, R.; de Albuquerque, V.H.C. A Novel Transfer Learning Based Approach for Pneumonia Detection in Chest X-ray Images. *Appl. Sci.* **2020**, *10*, 559. [CrossRef]
3. Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Hasan, M.; Van Essen, B.C.; Awwal, A.A.S.; Asari, V.K. A State-of-the-Art Survey on Deep Learning Theory and Architectures. *Electronics* **2019**, *8*, 292. [CrossRef]
4. Abidoye, I.; Ikeji, F.; Coupland, C.A.; Calaminus, S.D.J.; Sander, N.; Sousa, E. Platelets Image Classification Through Data Augmentation: A Comparative Study of Traditional Imaging Augmentation and GAN-Based Synthetic Data Generation Techniques Using CNNs. *J. Imaging* **2025**, *11*, 183. [CrossRef] [PubMed]
5. Neha, S.; Vibhor, J.; Anju, M. An Analysis of Convolutional Neural Networks For Image Classification. *Procedia Comput. Sci.* **2018**, *132*, 377–384. [CrossRef]
6. Shaheed, K.; Szczuko, P.; Abbas, Q.; Hussain, A.; Albathan, M. Computer-Aided Diagnosis of COVID-19 from Chest X-ray Images Using Hybrid-Features and Random Forest Classifier. *Healthcare* **2023**, *11*, 837. [CrossRef] [PubMed]
7. Tsai, T.H.; Chen, C.M. Exploring passengers' choices in the event of denied boarding compensation. *Transp. Policy* **2025**, *166*, 166–178. [CrossRef]
8. Arnaud, F.; Brigitte, A.; Bruno, L.; Marie, P.K.; Salim, S.A.K.; Devi, S. SARS-CoV-2 variants and ending the COVID-19 pandemic. *Lancet* **2021**, *397*, 952–954.
9. Chen, W.; Peter, W.H.; Frederick, G.H.; George, F.G. A novel coronavirus outbreak of global health concern. *Lancet* **2020**, *395*, 470–473. [CrossRef] [PubMed]
10. Janet, V.D.; Margaret, H.; Silvia, B.; Hannah, E.D.; Tarun, D.; Charu, K.; John, C.M.; Maria, R.P.; Nathalie, S.W.; Joan, B.S. Towards a universal understanding of post COVID-19 condition. *Bull. World Health Organ.* **2021**, *99*, 901–903.
11. Xie, M.; Chen, Q. Insight into 2019 novel coronavirus—An updated intrim review and lessons from SARS-CoV and MERS-CoV. *Int. J. Infect. Dis.* **2020**, *94*, 119–124. [CrossRef]
12. European Data. COVID-19 Coronavirus Data 2020. Available online: https://data.europa.eu/euodp/en/data/dataset/covid-19-coronavirus-data/resource/55e8f966-d5c8-438e-85bc-c7a5a26f4863 (accessed on 30 April 2025).
13. Basant, A.; Valentina, E.B.; Lakhmi, C.J.; Ramesh, C.P.; Manisha, S. *Deep Learning Techniques for Biomedical and Health Informatics*, 1st ed.; Academic Press: Cambridge, MA, USA, 2020.

14. Pasa, F.; Golkov, V.; Pfeiffer, F.; Cremers, D.; Pfeiffer, D. Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. *Sci. Rep.* **2019**, *9*, 6268. [CrossRef]
15. Linda, W.; Alexander, W. COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images. *arXiv* **2020**, arXiv:2003.09871.
16. Schmitt, M. How to Fight COVID-19 with Machine Learning. 2020. Available online: https://medium.com/data-science/fight-covid-19-with-machine-learning-1d1106192d84 (accessed on 30 April 2025).
17. Ali, N.; Ceren, K.; Ziynet, P. Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks. *arXiv* **2020**, arXiv:2003.10849. [CrossRef]
18. Xu, X.; Jiang, X.; Ma, C.; Du, P.; Li, X.; Lv, S.; Yu, L.; Ni, Q.; Chen, Y.; Su, J.; et al. Deep Learning System to Screen Coronavirus Disease 2019 Pneumonia. *arXiv* **2020**, arXiv:2002.09334. [CrossRef] [PubMed]
19. Adam, J.; Michael, C.; Adam, B.; Corey, E. Portable chest X-ray in coronavirus disease-19 (COVID-19): A pictorial review. *Clin. Imaging* **2020**, *64*, 35–42. [CrossRef] [PubMed]
20. Lin, N.; Paul, R.; Guerra, S.; Liu, Y.; Doulgeris, J.; Shi, M.; Lin, M.; Engeberg, E.D.; Hashemi, J.; Vrionis, F.D. The Frontiers of Smart Healthcare Systems. *Healthcare* **2024**, *12*, 2330. [CrossRef] [PubMed]
21. Babukarthik, R.G.; Chandramohan, D.; Tripathi, D.; Kumar, M.; Sambasivam, G. COVID-19 identification in chest X-ray images using intelligent multi-level classification scenario. *Comput. Electr. Eng.* **2022**, *104*, 108405. [CrossRef] [PubMed]
22. Adalja, A.A.; Toner, E.; Inglesby, T.V. Priorities for the US Health Community Responding to COVID-19. *JAMA* **2020**, *3*, 1343–1344. [CrossRef] [PubMed]
23. Ramirez-Alcocer, U.M.; Tello-Leal, E.; Romero, G.; Macías-Hernández, B.A. A Deep Learning Approach for Predictive Healthcare Process Monitoring. *Information* **2023**, *14*, 508. [CrossRef]
24. Bhati, D.; Neha, F.; Amiruzzaman, M. A Survey on Explainable Artificial Intelligence (XAI) Techniques for Visualizing Deep Learning Models in Medical Imaging. *J. Imaging* **2024**, *10*, 239. [CrossRef]
25. Neha, F.; Bhati, D.; Shukla, D.K.; Amiruzzaman, M. ChatGPT: Transforming Healthcare with AI. *AI* **2024**, *5*, 2618–2650. [CrossRef]
26. TensorFlow. Available online: https://www.tensorflow.org/ (accessed on 5 December 2024).
27. Keras. Available online: https://keras.io/ (accessed on 30 April 2025).
28. NumPy. Available online: https://numpy.org/ (accessed on 30 April 2025).
29. Pandas. Available online: https://pandas.pydata.org/ (accessed on 30 April 2025).
30. Matplotlib. Available online: https://matplotlib.org/ (accessed on 30 April 2025).
31. OpenCV. Available online: https://opencv.org/ (accessed on 30 April 2025).
32. Scikit-Learn. Available online: https://scikit-learn.org/ (accessed on 30 April 2025).
33. PyCharm. Available online: https://www.jetbrains.com/pycharm/ (accessed on 30 April 2025).
34. Cohen, J.P.; Morrison, P.; Dao, L. COVID-19 image data collection. *arXiv* **2020**, arXiv:2003.11597. [PubMed]
35. Bos, L.D.J.; Ware, L.B. Acute respiratory distress syndrome: Causes, pathophysiology, and phenotypes. *Lancet* **2022**, *400*, 1145–1156. [CrossRef]
36. Memish, A.Z.; Perlman, S.; Kerkhove, D.V.M.; Zumla, A. Middle East respiratory syndrome. *Lancet* **2020**, *395*, 1063–1077. [CrossRef] [PubMed]
37. Bugl, P. *SARS in Context: Memory, History, Policy*; The Lancet Infectious Diseases; McGill-Queen's University Press: Montreal, QC, Canada, 2007; p. 7.
38. Mooney, P. Chest X-Ray Images (Pneumonia). Available online: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia (accessed on 6 June 2025).
39. Mienye, I.D.; Swart, T.G.; Obaido, G.; Jordan, M.; Ilono, P. Deep Convolutional Neural Networks in Medical Image Analysis: A Review. *Information* **2025**, *16*, 195. [CrossRef]
40. Salehi, A.W.; Khan, S.; Gupta, G.; Alabduallah, B.I.; Almjally, A.; Alsolai, H.; Siddiqui, T.; Mellit, A. A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope. *Sustainability* **2023**, *15*, 5930. [CrossRef]
41. Raptis, C.; Karavasilis, E.; Anastasopoulos, G.; Adamopoulos, A. Comparative Analysis of Conventional CNN v's ImageNet Pretrained ResNet in Medical Image Classification. *Information* **2024**, *15*, 806. [CrossRef]
42. Brownlee, J. Transfer Learning in Keras with Computer Vision Models. Available online: https://machinelearningmastery.com/how-to-use-transfer-learning-when-developing-convolutional-neural-network-models/ (accessed on 6 June 2025).
43. Nguyen, T.-H.; Nguyen, T.-N.; Ngo, B.-V. A VGG-19 Model with Transfer Learning and Image Segmentation for Classification of Tomato Leaf Disease. *AgriEngineering* **2022**, *4*, 871–887. [CrossRef]
44. Allgaier, J.; Pryss, R. Cross-Validation Visualized: A Narrative Guide to Advanced Methods. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1378–1388. [CrossRef]

45. Milenkovic, A.; Jankovic, D.; Rajkovic, P. Extensions and Adaptations of Existing Medical Information System in Order to Reduce Social Contacts During COVID-19 Pandemic. *Int. J. Med. Inform.* **2020**, *141*, 104224. [CrossRef]
46. Krishnan, M. Understanding the Classification Report Through Sklearn. Available online: https://muthu.co/understanding-the-classification-report-in-sklearn/ (accessed on 30 April 2025).
47. Jayaswal, V. Performance Metrics: Confusion Matrix, Precision, Recall, and F1 Score. Available online: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9 (accessed on 6 June 2025).
48. Sethy, P.K.; Behera, S.K.; Ratha, P.K.; Biswas, P. Detection of Coronavirus Disease (COVID-19) based on Deep Features and Support Vector Machine. *Int. J. Math. Eng. Manag. Sci.* **2020**, *5*, 643–651. [CrossRef]
49. Wang, S.; Kang, B.; Ma, J.; Zeng, X.; Xiao, M.; Guo, J.; Cai, M.; Yang, J.; Li, Y.; Meng, X.; et al. A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19). *Eur. Radiol.* **2021**, *31*, 6096–6104. [CrossRef]
50. Fard, F.A.; Sy, M.; Ivanov, D. Optimal overbooking strategies in the airlines using dynamic programming approach in continuous time. *Transp. Res. Part E Logist. Transp. Rev.* **2019**, *128*, 384–399. [CrossRef]
51. Imani, M.; Beikmohammadi, A.; Arabnia, H.R. Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels. *Technologies* **2025**, *13*, 88. [CrossRef]
52. Grzelak, M.; Owczarek, P.; Stoica, R.-M.; Voicu, D.; Vilău, R. Application of Logistic Regression to Analyze the Economic Efficiency of Vehicle Operation in Terms of the Financial Security of Enterprises. *Logistics* **2024**, *8*, 46. [CrossRef]
53. Zhou, Z. *Ensemble Methods: Foundations and Algorithms*; CRC press: Boca Raton, FL, USA, 2025.
54. Rajković, P.; Janković, D.; Milenković, A. Developing and Deploying Medical Information Systems for Serbian Public Healthcare—Challenges, Lessons Learned and Guidelines. *ComSIS* **2013**, *10*, 1429–1454. [CrossRef]
55. Rajković, P.; Janković, D.; Milenković, A.; Kocić, I. Analysis of the level of use and acceptance of the medical information system in primary health care. *Acta Medica Median.* **2018**, *57*, 122–136. [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.

computers

MDPI

Article

# A New AI Framework to Support Social-Emotional Skills and Emotion Awareness in Children with Autism Spectrum Disorder

**Andrea La Fauci De Leo, Pooneh Bagheri Zadeh *, Kiran Voderhobli and Akbar Sheikh Akbari**

James Graham Building, Leeds Beckett University, Church Wood Ave, Leeds LS6 3QS, UK; andrea.lafauci92@gmail.com (A.L.F.D.L.); k.voderhobli@leedsbeckett.ac.uk (K.V.); a.sheikh-akbari@leedsbeckett.ac.uk (A.S.A.)
* Correspondence: p.bagheri-zadeh@leedsbeckett.ac.uk

**Abstract:** This research highlights the importance of Emotion Aware Technologies (EAT) and their implementation in serious games to assist children with Autism Spectrum Disorder (ASD) in developing social-emotional skills. As AI is gaining popularity, such tools can be used in mobile applications as invaluable teaching tools. In this paper, a new AI framework application is discussed that will help children with ASD develop efficient social-emotional skills. It uses the Jetpack Compose framework and Google Cloud Vision API as emotion-aware technology. The framework is developed with two main features designed to help children reflect on their emotions, internalise them, and train them how to express these emotions. Each activity is based on similar features from literature with enhanced functionalities. A diary feature allows children to take pictures of themselves, and the application categorises their facial expressions, saving the picture in the appropriate space. The three-level minigame consists of a series of prompts depicting a specific emotion that children have to match. The results of the framework offer a good starting point for similar applications to be developed further, especially by training custom models to be used with ML Kit.

**Keywords:** Autism Spectrum Disorder; Serious Games; Android; AI; Cloud Vision

## 1. Introduction

Autism Spectrum Disorder (ASD) is a pervasive and prominent disorder that has been well documented and vastly studied in the past decade [1]. ASD is difficult to categorise and diagnose in its early stages because of its characteristics [2]. Currently, approximately 2.94% of UK children between the ages of 10 and 14 are diagnosed with ASD [3], representing an important figure considering the population. The DSM-V-TR, with the latest 2022 revision, categorises the disorder as a large spectrum rather than a single condition [4]. It identifies different levels of severity, but one common characteristic is the impairment of social-emotional skills, which consequently might affect the individual's quality of life in everyday social interactions.

Tools such as artificial intelligence (AI) are becoming more popular, and their introduction to the industry represents an important milestone identified as a Fourth Industrial Revolution [5]. AI tools are becoming a more predominant element in our lives [6], especially in the education field to help teachers and educators. Specifically, facial-recognition

technologies have a wide range of applications spanning from security to education [7,8]. Machines can be trained to recognise facial expressions, identities, and emotions.

Serious games are stimulating and interactive games, which usually allow the players to improve a specific set of skills. They accomplish this by presenting the player with pleasant activities that can teach skills through playing. These games are known to be useful in helping both ASD and neurotypical children develop skills, especially for their multisensory approach [9]. Specifically, they can help create ways to simulate natural environments and can be used to develop specific activities aimed at focused training. Several serious games are developed specifically for children with ASD [10]. Serious games offer a new perspective on how these tools can be used to support vulnerable children's development. AI tools can be integrated into such games to offer innovative ways to support children's development, and mobile applications can be the perfect medium to offer educators or parents the support children with ASD might need.

ASD is usually wrongly represented in the media, reinforcing a misconception of the challenges that Autism might pose [11]. Usually, social communication abilities and interactions with peers are impacted by these disorders [12]. Individuals with ASD tend to struggle with social interactions and especially with expressing and identifying emotions. Individuals with ASD tend to have a poorly developed sense of self, often causing them not to elaborate on their internal states of mind. Specifically, a component of the Theory of Mind (ToM), defined as the ability to understand and attribute mental states to the self and others [13], is affected. Such a deficit in this area is defined as "mind-blindness", usually implying difficulties in "seeing others' minds", intentions, and motives. This is also demonstrated by the "Sally-Anne test" [14,15]. Consequently, children with ASD struggle to recognise emotions, especially using facial clues [16], which is one of the most important ways for children to communicate in early stages [17].

These deficits can induce anxiety and depression, especially when individuals use constant "masking" behaviours as a result [18]. However, tools designed to help teach emotion recognition techniques have proved to have a positive impact [19]. Specifically, games using new technologies and robots can help with the challenges posed by ASD [20]. In particular, children with ASD might find it hard to work with people face to face due to their social-emotional deficits, and the use of tablets and smartphones can help [21].

The integration of "affective computing" technologies can help learners develop their skills [22]. New methodologies can be applied to education and teaching by enabling computers to recognise emotions. Using CNN (Convolutional Neural Network) AI models, which can learn by identifying patterns from the data provided [23], is currently one of the most popular approaches to teaching AI to recognise emotions from facial expressions [24–26]. Several APIs (Application Programming Interfaces) use such approaches for their pre-trained models. Google Cloud Vision API is one of the most popular ones, which enables the application to recognise emotions with an 85% accuracy [27].

These technologies can be utilised to support children with special needs in their development. For example, specific training can be implemented to train facial recognition skills [28], effectively translating behavioural or developmental therapy into digital practices. Other serious games that are proven to be effective with ASD individuals provide customised activities in a multisensory virtual environment. This approach could result in higher engagement and improved results in developing an aspect of the ToM, namely, the skills which involve emotion awareness [29].

Serious games designed for children with ASD can allow them to be more engaged with entertaining activities while improving their socio-emotional skills. This methodology is highly effective, as individuals with ASD tend to be visual learners, preferring mobile

technologies [30]. For this reason, several applications have been developed to aid children with ASD. An example is represented by Ying, a mobile Android application integrating ABA techniques in therapeutic settings [31].

Other applications exist to help such individuals develop their skills. An example is Emoface [32], with several activities to allow children to better develop deficient skills, and EmoTen [33], which also includes emotion-aware technologies. The latter is developed using Affectiva Software Development Kits (SDK) and enables a few activities focused on improving certain skills. The application is tested and evaluated against a small cohort of children with ASD, supported by a team of developmental psychologists. Although the application suffered from certain limitations, the results of the evaluation are positive, confirming the existing literature on the effectiveness of such an approach.

The main contribution of this paper is to propose a new framework that can have an impact on future works. The developed framework aims to improve existing applications by introducing facial-recognition technology to help create engaging and motivating activities. Two key features are presented and developed, each aimed at improving quality of life for children with ASD. This is achieved with the introduction of two functionalities that can help users to learn how to recognise emotions and express them through a serious game. Furthermore, it uses a sandbox environment, which automatically categorises emotions and offers an emotion diary to be consulted by users. The application, developed as a result of the designed framework, takes advantage of the power of the cloud-based API offered by Google Cloud Vision, which gives an example of how such cutting-edge technology can be applied and used by professionals in the developmental psychology field.

The paper is structured in the following way: Section 2 presents the developed new framework and how it can enhance the social emotional skills of children with ASD. It also highlights the gaps in the current literature and sets the basics for further implementation and applications. Section 3 illustrates how the framework is applied to create an experimental Android application. It highlights its design, the actual implementation, and the preparation of a custom AI model. Section 4 describes the results obtained from the development and testing of the application EmoPal. In Section 5, the results are discussed, highlighting the benefits of the presented approach, and the paper concludes with recommendations for further research and investigations.

## 2. AI Framework—Improve Socio-Emotional Skills Through AI Technologies

The proposed framework to help children with ASD takes inspiration from existing socio-emotional therapy, which also involves play therapy and emotion recognition techniques. The main focus of the framework is to take advantage of the advancements in AI technology, computer vision, and emotion-aware technology to create a new methodology to help children, educators, teachers, and legal guardians to overcome the challenges posed by such a condition. The field of study regarding ASD is well documented and rich in literature, especially in regard to play therapies and how emotional therapy can enhance socio-emotional skills in children with ASD. One of the characteristics of ASD is alexithymia, also known as emotional blindness. These individuals usually struggle to express and recognise emotions in themselves and others. This leads to masking behaviours and increased levels of stress that might lead to anxiety and depression.

The core feature of the framework is to somehow translate existing therapies and techniques into a digital format that is powered by AI technologies.

Play therapy has proven to be an effective way in improving emotion regulation in children with Special Educational Needs (SEN) [34]. This type of therapy involves using several

techniques, such as interaction with play activities, that encourage children to express feelings. This helps children to navigate their complex range of emotions without the strict need of using words, which can be challenging when verbal language is not enough. An example is found in The Stories of Ciro and Beba [35], a book which contains a set of short stories specifically designed to help children to learn and expand their emotional skills by building up different socio-emotional constructs, such as emotional expression, understanding, and regulation.

The idea of these games is to integrate the playful element of a game and emotional conversation, encouraging the user to not only express emotions but also to understand them better. The games are designed to incorporate such elements. For example, the games are highly portable as a mobile application, and they can be used anywhere; for example, they can be used during Floortime activities, which is one of the models used in play therapy, focused on improving emotional growth. Furthermore, the game presents the user with images depicting emotions, promoting their recognition, and mimicking and encouraging emotional talk between parents and children.

Advancements in AI, especially in face recognition, are allowing models to be built and tuned to recognise facial expressions like emotions. These models can be used to develop tools that can effectively help these individuals to develop the required skills.

The idea behind the framework is to have an AI companion that can recognise emotions and use recognition to give feedback to the user. This methodology enables the development of invaluable strategies. New technologies such as AI can be used to facilitate therapy and training by offering a valid and easy-to-use aid. This topic is currently not well documented in the literature, and the present framework is thought to fill the gap and highlight the importance of these tools.

AI models can be leveraged to enhance learning opportunities and are adaptable for use in virtually any environment. Current socio-emotional therapies involve play-therapy techniques, which can be digitally translated and combined with AI-driven emotion aware technologies. The presented framework aims to use socio-emotional therapy techniques and play therapy through serious games, combined with AI facial-recognition models that can recognise several emotions. Figure 1 shows the designed framework, unifying socio-emotional therapy and play therapy, with AI tools like emotion-aware technology and existing or customised AI models.

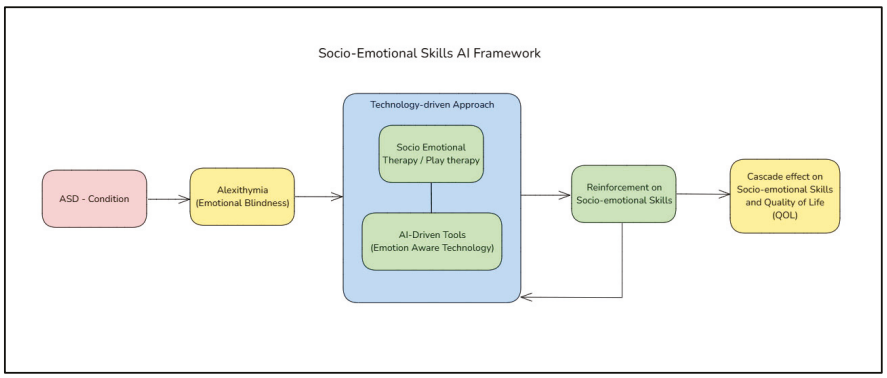



**Figure 1.** The proposed framework: socio-emotional therapy and play therapy, with AI tools.

The core engine of the framework is an AI tool which is used to recognise expressed emotions and give users feedback whilst playing. The main objective of this approach is to help the user learn how to recognise and express emotions. This is meant to reinforce their socio-emotional skills with the goal of supporting and developing them further through engaging and stimulating activities. The feedback provided from AI models and the outcome of successfully

progressing through a game can be used to reinforce the learning process, especially if used repeatedly through pleasant activities that can be engaged more than once in different settings.

This methodology allows children with ASD to learn while using new engaging technologies that can act as an effective companion and, at the same time, as a valid method for parents or therapists to help the user. Such reinforcement and training is expected to have a cascade effect on an individual's quality of life. By increasing their competence in identifying and expressing emotions, the likelihood of developing anxiety symptoms and depression decreases, including social anxiety and suicidal thoughts in later stages of life [36].

The proposed framework focuses on AI technology as an aid and not a complete substitute for the human factor, which is still needed. However, this approach would highly help the target audience to internalise the meanings of emotions, as well as understanding them and expressing their internal states. Training such deficient skills could help them reduce stress, anxiety, and depression, especially when caused by not recognizing social cues, or by the uncertainty caused by social interaction. Furthermore, it would give them the right tools to understand their internal states and be more aware of felt emotions. The framework is applicable to new tools and applications in order to promote an innovative way of supporting young individuals.

As a result of the study and design of the framework, a native Android application named EmoPal has been developed to implement the framework's principles and serve as an experimental tool, demonstrating the discussed features and implementations while highlighting their strengths and limitations.

## 3. Framework Implementation—Materials and Methods

The EmoPal application is developed primarily for the Android platform due to its open-source nature and its status, based on reviewed data, as one of the most widely used mobile operating systems [37]. Android is also developed by Google, which offers support for developers to implement compatible services and AI features. The application and its framework are developed for mobile devices running Android 13.0 (API 30—"R"). Such a decision is to allow the application to run on most devices available in the market. According to the Android documentation, the API is currently leveraged by 67.6% of its users [38] as also shown in Figure 2.

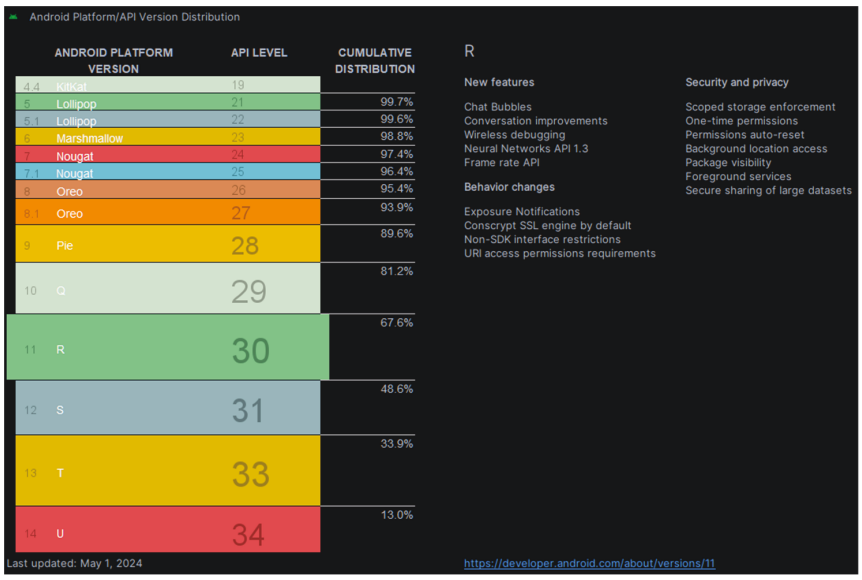



**Figure 2.** Android API distribution.

Table 1 summarises the technologies used to develop EmoPal. Such technologies allow for fine control of the development of the UI in a scalable way.

**Table 1.** Technical Details.

| **Technical Specifications** | |
|---|---|
| Target Audience | Young people affected by ASD |
| Target Devices | Tablets—Resolution 1024 × 768/768 × 1024 |
| Android SDK | API 30—"R"; Android 13.0 |
| Programming Language | Kotlin 1.9.22 |
| Compose Compiler | 1.5.10 |
| Build Automation Tool | Gradle Kotlin DSL v8.3.1/Plugin v4.2.1 |

The user interface is designed to be both effective and easy to use. An intuitive and easily understandable design is essential, especially for the target user. The design is inspired by the user-centred design paradigm, as it allows for focusing primarily on the user's needs, and it has previously been used to re-adapt mobile applications for children with ASD [39]. This approach would help reduce the user's cognitive fatigue while using the application and reduce cluttering and background stimuli [40].

The Jetpack Compose framework is chosen as the tool for designing and developing the UI. This framework supports the new features from Material3 themes and replaces the traditional XML approach. This is slowly becoming the new industry standard, as it gives developers full control over UI development [41]. Furthermore, it introduces the concept of composables, which are reusable UI elements that help to decouple code and allow the creation of multipurpose components that can be reused. An example is shown in Figure 3. The methodology used allows the framework to be reliable, scalable, and reusable.

```kotlin
@Composable
fun PictureCase(pictures: MutableList<BitmapMetadata>, category: String) {

    var activePicture by remember { mutableStateOf<Bitmap?>( value: null) }
    val haptics = LocalHapticFeedback.current
    var contextMenuPhoto by rememberSaveable { mutableStateOf<Bitmap?>( value: null) }

    if (pictures.isEmpty()) {
        Box(
            modifier = Modifier.padding(16.dp),
            contentAlignment = Alignment.Center
        ) { this: BoxScope
            CategoryHeader(category = category)
        }
        Divider()

        Text( text: "Feels Empty Here...Take a Picture")

    } else {
        CategoryHeader(category = category)
        Divider()
        LazyRow { this: LazyListScope
            items(pictures) { this: LazyItemScope  picture ->
                PictureBox(
                    picture.bitmap,
                    onClick = { activePicture = picture.bitmap },
                    onLongClick = {
                        haptics.performHapticFeedback(HapticFeedbackType.LongPress)
                        contextMenuPhoto = picture.bitmap
                    }
                )
            }
        }
    }
    // If the active picture is selected then it will enlarge it for the size of the parent component.
    if (activePicture != null) {
        val activePictureMetadata = pictures.firstOrNull { it.bitmap == activePicture }
        if (activePictureMetadata != null) {
            FullScreenImage(photo = activePictureMetadata.bitmap) {
                activePicture = null
            }
        }
}
/**
 * Category Header Composable which displays the category in a header style.
 * @param category the category to be displayed.
 */
 ♟ andrea.lafauci92@gmail.com <andrea.lafauci92@gmail.com>
@Composable
fun CategoryHeader(category: String) {
    Text(
        text = category,
        modifier = Modifier
            .padding(16.dp),
        fontSize = MaterialTheme.typography.headlineLarge.fontSize,
        fontFamily = MaterialTheme.typography.headlineLarge.fontFamily
    )
}
```

**Figure 3.** EmoPal picture case reusable composable.

Emotion-aware technologies are the core of the serious games developed to enforce the overlaying framework principles described previously in Section 2. Furthermore, the produced AI Model sets the base for a new way to use such a technology and expand it with further work. The games are designed to be entertaining and able to interact with the user by enabling emotion recognition through AI technology. This is to enable an active interaction between users and intelligent technology.

### *3.1. EmoPal Technical Design*

EmoPal is designed with an MVVM design pattern in mind as illustrated in Figure 4. This is common in mobile development, and it is widely used in the industry. It allows the application to be easily scalable and flexible [42]. This approach allows the use of different ViewModels to populate specific Views with data. This helps it to have a more loosely coupled code, easier to maintain and develop [43]. The goal is to apply a flexible solution that can help other developers to expand the use cases and adapt the framework in several possible ways.

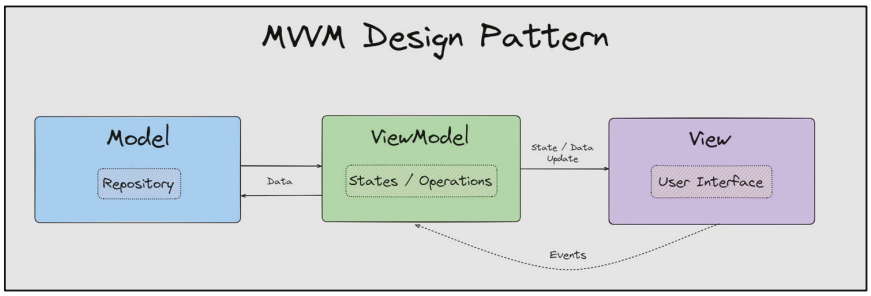



**Figure 4.** MVVM Design Pattern.

The application is structured in three distinct parts. The View is what the user sees, which is populated by the ViewModel, where the business logic is applied. Finally, the Model is where the data is stored.

In conjunction with this design pattern, the Dependency Injection (DI) pattern is also heavily used through the Dagger Hilt API. Consequently, this allows developers to decouple code, making the structure modular, and it allows it to reduce memory usage by creating each dependency once [44]. This enables the injection of ViewModels in different parts of the code, allowing a high flexibility and reusability of components.

EmoPal uses an internal SQLite database using Room ORM. It stores data used for the game and pictures taken using the diary features of the application. The tables use the path of the pictures instead of storing them as BLOBS. This design allows the user to take several pictures without the memory limitation of the SQLite database, which might cause issues in saving them through a TypeConverter. Furthermore, it allows the application to store other types of data for future advancements. Figure 5 shows the schema of the database and the tables' structure.

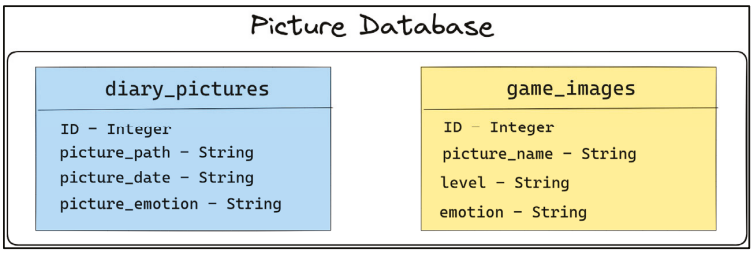



**Figure 5.** Database Schema.

EmoPal connects through a RESTful API to Google Cloud Vision, which elaborates and interprets the image transferred. Google's AI model is used in the first iteration of the experimental application to facilitate the development and test the model's capabilities. Its design is illustrated in Figure 6.

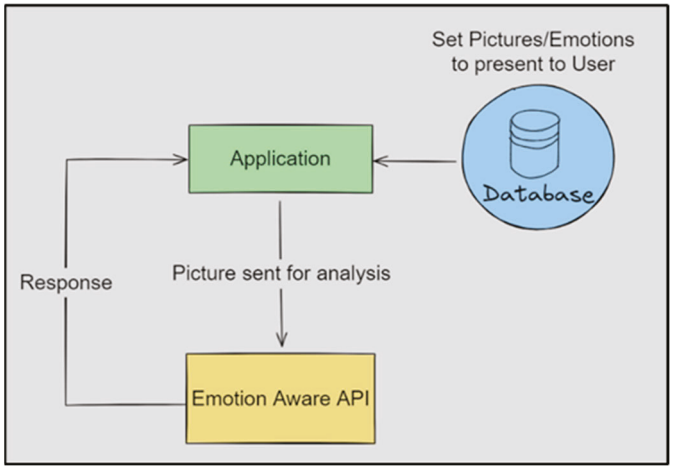



**Figure 6.** EmoPal design structure.

EmoPal and its framework aim to fill the gap set by the recent publications and similar applications currently present in the market, but also to define a standard in the application of such technologies designed to help individuals with ASD. Two key features are implemented to allow EmoPal to supplement and enrich the current literature with this reusable framework. The main purpose is to improve the minigame proposed in EmoTen and the diary functionality in Emoface by introducing an automated tool to make the learning process more interactive and innovative. Specifically, the minigame proposed works as illustrated in Figure 7.

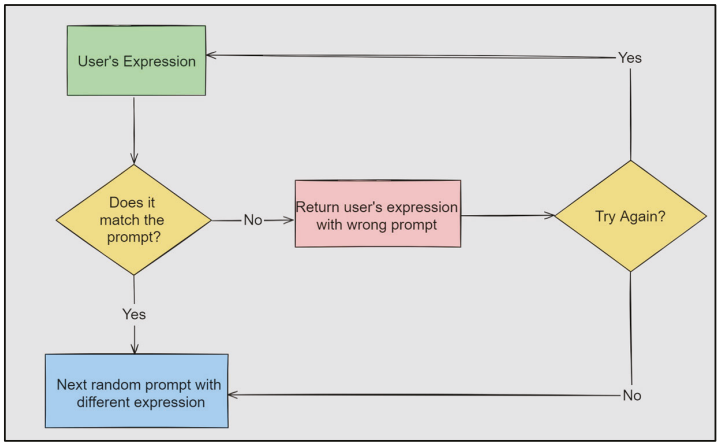



**Figure 7.** Minigame flowchart.

The game is structured to present a prompt to the user. The user is then asked to take a picture of themselves so that the application can compare their expression to the prompt. If the emotion is matched, the game will continue with the next prompt. Otherwise, a message will return the current user expression, and a message will appear to try again. Even if the emotion is not matched, it is important to state what emotion they used. This is to enforce the learning aspect of the framework and its reinforcement goals. This also

allows the user to think of how they expressed their emotion and to try again to match the current prompt and learn from errors.

The emotion diary feature allows users to have a sandbox environment where they can freely experiment with their emotions by taking pictures of themselves. This functionality offers a good way for children to play with their expressions and to check which one is associated with the relative emotion (Figure 8).

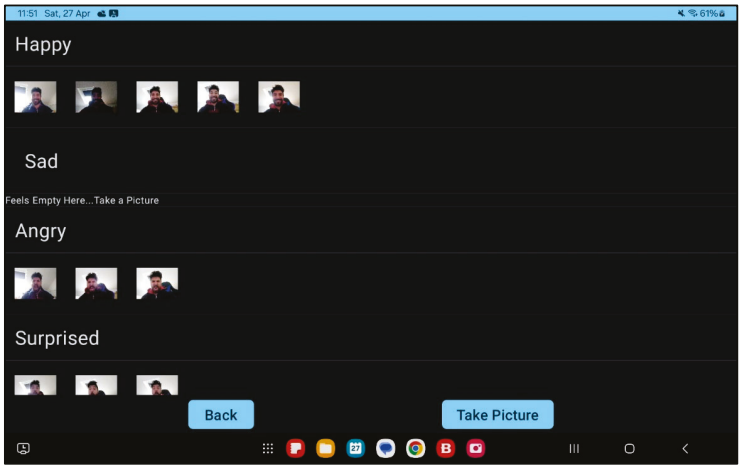



**Figure 8.** Diary feature.

A similar functionality is seen in another application called Emoface, sponsored by the National Centre for Scientific Research in France (CNRS). However, such an application does not use emotion-aware technology, as the user would manually add their pictures to the list.

EmoPal recognises the emotion and categorises it. Google Cloud Vision API is used for this specific operation. CameraX library is also used with Jetpack Compose, as it provides a consistent API which enables the use of the camera. It is also compatible with ML Kit for future developments with image analysis capabilities [45]. The camera allows the user to have immediate feedback, and once the picture is taken, it is then stored in the internal memory of the application, together with its unique URI in the Room database configured.

The picture is sent to the API, which analyses the facial features and sends back a response in JSON format, which is then interpreted by the customised parser. At the current stage, only four emotions are supported: happiness, anger, sadness, and surprise.

The second core feature is the emotion mimicking game. The game presents picture prompts expressing a specific facial expression. The goal of the user is to mimic the expression by taking a picture of themselves and progressing through the game. A response is displayed to give feedback to the user if the expressed emotion is correct or not. Also, if the expressed emotion is wrong, the user will be given a prompt with the emotion they expressed. The game has three levels, each of which focuses on specific training outcomes.

### *3.2. Implementation*

The first level is mainly focused on developing the ability to recognise facial expressions. The pictures shown display high-intensity emotions, which are more easily recognised by ASD individuals [46]. Random prompts are presented, which is a crucial aspect, as it can prevent the user from expecting a repeated sequence of presented pictures. This is to improve the effectiveness of the training, as ASD individuals tend to adopt mechanical learning strategies. Furthermore, the pictures presented depict real people in contrast with other applications reviewed in the literature (Figure 9).

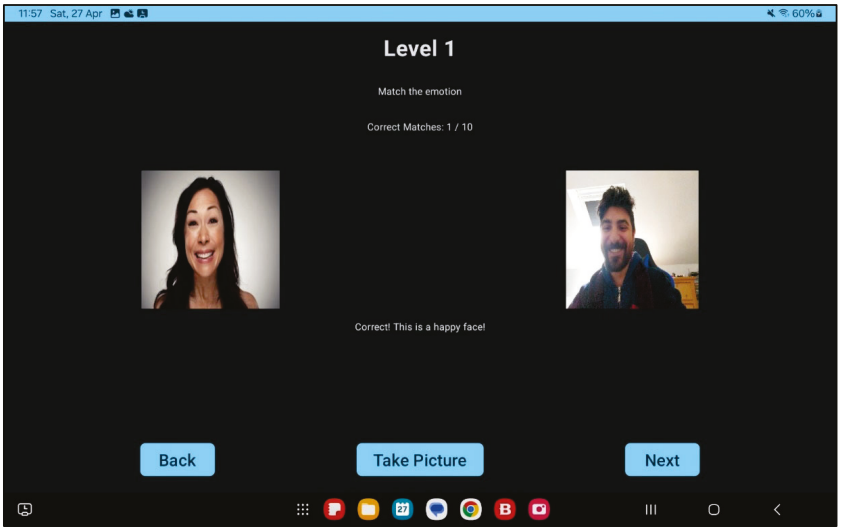



**Figure 9.** EmoPal main game.

Each level of the minigame focuses on a specific learning goal. The first level focuses only on facial expressions and emotions. The pictures are clear and set in a neutral environment. The background is usually white, and this is to reduce distractions and users' hyper-focussing on elements that are not part of the scope of the game. Such a problem is illustrated in research by using eye-tracking technologies [47].

The second level increases the difficulty by proposing pictures having a background and other elements of the body. The reasons are to help the user train their ability to recognise emotions through facial expressions and body language, an important medium of communication during social interactions.

The images are still focused on facial expressions; however, more complexity is added to the background, along with more subtle expressions and more complex emotions. This adds another layer of difficulty, which can be challenging for children with ASD. The aim of the second level is to represent clear pictures with more distracting elements. The emotions included in the sample include pictures depicting the emotions of surprise and the complex emotion of disgust. These are created according to Ekman's categorisation of emotions [48].

Lastly, the third level presents contextual situations where the user needs to infer the general emotion expressed by analysing the environmental elements. This is inspired by EmoTen [33], which presents a social situation as an animated scene with cartoon characters. This third level is improved by providing pictures with real people in several scenarios. The scenes are tailored to represent the range of emotions taken into consideration for this proof of concept. For example, to represent happiness, an image will display people cheering and laughing together, a scene in which the focus is the social situation rather than a single person's expression. This is important to teach the user how to abstract social situations and think about the kinds of emotions expressed in a more social context, facilitating abstraction and reasoning, as also illustrated in the literature [49]. The pictures displayed are part of a freely accessible repository and serve the proof of concept. For further development, different pictures or even short videos should be taken and prepared for the purposes of the game.

### *3.3. CNN Model*

This application is limited to having access to the network, as it uses the Google Cloud Vision API. This limits the capabilities of the application, as it requires constant internet access, as well as slowing the picture processing time and being bound to the Google AI

Model. As such, an experimental model is introduced to overcome these limitations and for future experiments and developments.

This model is trained using the FER-2013 dataset, freely available on Kaggle. The Keras API and a CNN algorithm at 96 Epoch are used. A Python (v3.13) script is written using Google Colab, which uses the provided dataset to train the model. A code snippet is shown in Figure 10.

Create CNN Model

```
# Initialising the CNN
model = Sequential()

# Convolution Block 1
model.add(Conv2D(64,(5,5), padding='same', activation='elu', input_shape=(48, 48,1)))
model.add(BatchNormalization())
model.add(Conv2D(64,(5,5), padding='same', activation='elu'))
model.add(BatchNormalization())
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Dropout(0.4))

# Convolution Block 2
model.add(Conv2D(128,(3,3), padding='same', activation='elu'))
model.add(BatchNormalization())
model.add(Conv2D(128,(3,3), padding='same', activation='elu'))
model.add(BatchNormalization())
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Dropout(0.4))

# Convolution Block 3
model.add(Conv2D(256,(3,3), padding='same', activation='elu'))
model.add(BatchNormalization())
model.add(Conv2D(256,(3,3), padding='same', activation='elu'))
model.add(BatchNormalization())
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Dropout(0.5))

# Flattening
model.add(Flatten())

# Fully Connected layer
model.add(Dense(128, activation='elu'))
model.add(BatchNormalization())
model.add(Dropout(0.6))

# Output layer
model.add(Dense(7, activation='softmax'))

# Model Compilation
opt = Adam(0.001)
model.compile(optimizer=opt, loss='categorical_crossentropy', metrics=['accuracy'])
model.summary()
```

**Figure 10.** CNN model script.

Because of the data structure, the training is performed using a CNN as opposed to an NNM algorithm or Support Vector Machine (SVM).

Bogush et al. [50] illustrated how a CNN achieved one of the highest levels of accuracy at 90.9% by training a model to recognise facial expressions. This exceeds other algorithms such as NNMs, which have achieved 82% as the highest in emotion recognition. Their systematic review shows a distribution of the various algorithms taken into consideration in the review (Figure 11).

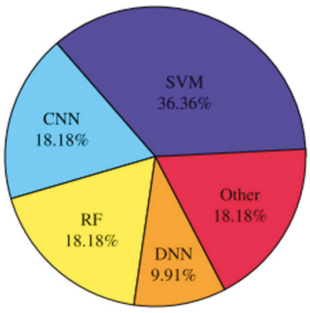



**Figure 11.** Distribution of machine learning algorithms by Bogush et al. [50].

It is also shown how accuracy is influenced by the quality and quantity of the data collected. Therefore, the effectiveness of the model is directly impacted by such variables.

For the present implementation, the produced custom model is then converted into a TensorFlowLite model. This model is compatible with ML Kit and CameraX, which suit the future scope of the project as, for example, the actual implementation and further developments to achieve the highest possible level of accuracy.

## 4. Results

The developed application is based on similar, but also different, applications present in the literature, but it also aims to set the standard on how such frameworks should be designed and built. According to the literature, not so many applications are present in the PlayStore with the mentioned capabilities. Also, most of the applications reviewed do not use facial-recognition technology, which in this case is used through Google Cloud API and also through a custom AI model.

The framework and application are evaluated against the features present in similar applications reviewed in the literature. EmoPal enhances the features implemented in EmoTea by using randomised pictures using real people. This is to improve the learning outcome. Also, including different levels with different training outcomes helps users to learn from different contexts. The minigame concept is designed to be engaging, including a level progression and a victory message once 10 prompts are guessed correctly.

The decision to use real people instead of emojis or drawn pictures is to ensure that the user would better recognise such emotions in different people. Individuals express emotions differently, especially when considering facial features. The feature improves the effectiveness of the application during the game. However, the implementation is not addressed in the reviewed and evaluated work.

The diary feature is compared and evaluated with the one present in Emoface. This activity is designed for the user to have a free sandbox environment to experiment with emotions. The diary helps keep track of how the user feels at a specific moment and automates the process of recognising their expressions. The implementation of such a feature is intended to prove how this technology can be used to promote self-directed and independent learning by using computer vision and AI, and applying the framework's core concept of using AI technology with social-emotional therapy techniques.

Both features use the same API to analyse the picture. The feasibility test was conducted with non-ASD control users to confirm the program was working to its specifications. During the test phase, it was noticed that Google Cloud Vision API fails to recognise the sad emotion in most cases, as shown in Figure 12.

The reason for this could have multiple implications which span between the light exposure, the quality of the camera, and the facial element coordinates. Google advertises the Emotion Recognition feature with 80% accuracy, which applies mainly to the other emotions. The emotions recognised are also limited to four.

Both the diary and the mimicking games rely on Google Cloud Vision API. Although the game works well after conducting a feasibility test, effectively digitalising a training activity from CBT, it fails at recognising emotions accurately in a few cases. This has an impact on the effectiveness of the application, as it needs a more accurate model. Also, the REST API used requires an internet connection. This limits the use of the application only when a network connection is available, which led to the early development of a model to be directly integrated into the application. The 80% accuracy advertised by Google is impressive, and it can only improve with further adaptation and training of models. The application developed should be considered a feasibility study, and such results are going to be used to improve further iterations and developments by using customised models with potential higher accuracy. The use of the application is also encouraged with the

guidance of an adult, which can help elaborate on complex emotions and give accurate feedback to the young user.

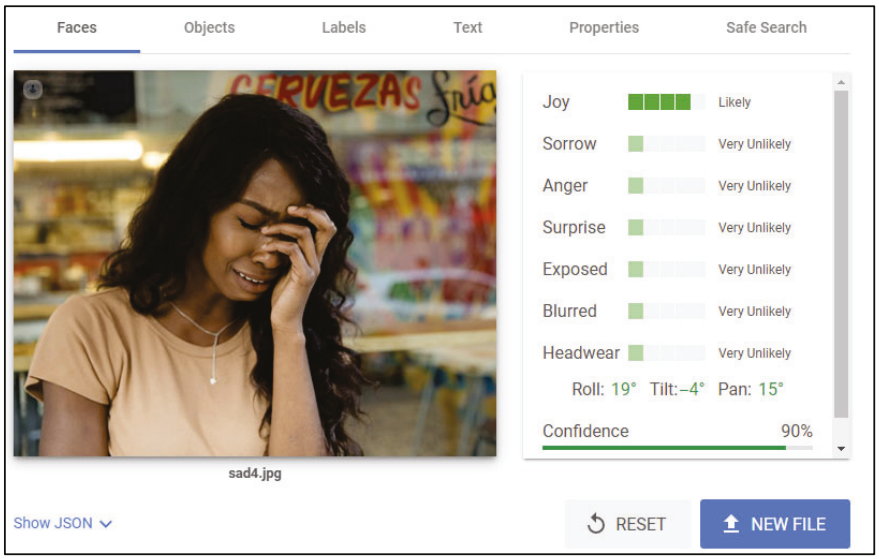



**Figure 12.** Google Cloud Vision API failing to recognise sorrow.

Future expansions should aim to test the application on students with ASD to analyse the effectiveness of the treatment and significant effects based on users' age and ASD severity.

*Custom NNMs Model Result Discussion*

The custom model is trained with the purpose of mitigating the limitations of the implementation. Therefore, it is trained with the FER-2013 dataset containing a large number of facial expressions categorised by the expressed emotion. The current categories include seven emotions: happiness, disgust, fear, surprise, anger, neutral, and sadness.

Test results indicate that with only 96 epochs and one hour of training, the model reaches an overall accuracy of ~70%, as shown in Figure 13.

This model is also trained to recognise more complex emotions, such as fear and disgust. However, its accuracy in detecting disgust is expected to be lower due to the limited number of training samples—only 436 images—representing this emotion included in the dataset. A detailed breakdown of the dataset by emotion is presented in Figure 14.

For ethical reasons, the model is not implemented and tested on actual ASD subjects. The research did not have a clinical focus, with control subjects and a large sample. The pictures representing real people used to train the custom AI model are from a public dataset. Using real humans on the pictures increases the accuracy of the model in recognising human expressions using CNN techniques, as well as giving the user a way to recognise themselves in the depicted facial expressions.

The accuracy of the model is based on the test dataset provided during the AI model training and results from the training script. However, the dataset can be expanded with more pictures as well as increasing the epochs and time spent training to achieve a higher accuracy. Overall, the product effectively improves the already-seen features in similar applications by expanding them using a new and different approach, which covers the gaps in previous research. However, it still lacks high accuracy, which has an impact on the efficiency of the developed framework.

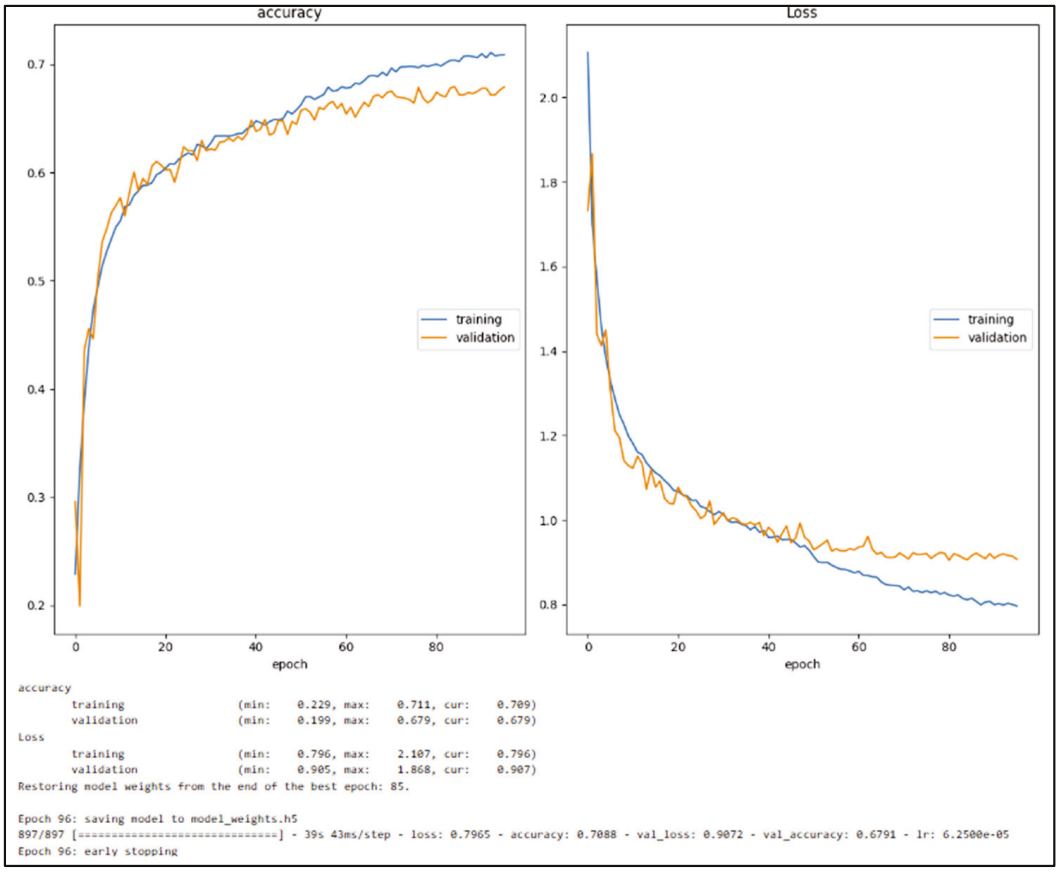



**Figure 13.** Custom CNN training.

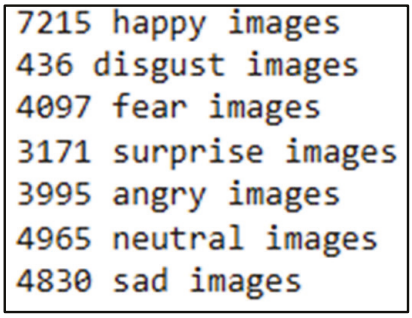



**Figure 14.** Categorised dataset.

## 5. Discussion

The results of this study are promising, as an increasing number of technologies today are leveraging AI to support the assessment and intervention processes for individuals with ASD [51]. Nevertheless, there remains a strong need for a structured and flexible framework. EmoPal serves as a valuable prototype that demonstrates the potential of applying a novel AI-based framework. Such tool can be effectively utilised by psychologists and educators to collaboratively support children with ASD and their families. This is particularly significant given the current scarcity of applications on the market that utilise AI-driven emotion recognition. Figure 15 demonstrates an overview of available applications that apply AI-driven technologies in mobile environments.

| Name | Platform | Camera | Facial recognition | Real pictures | Tactile interface |
|---|---|---|---|---|---|
| Happy baby | Android | No | No | No | Yes |
| Emotionalyser | Android | Yes | Yes | Yes | Yes |
| Camp Discovery | Android/iOS | No | No | No | Yes |
| Touch-Emotions | Android/iOS | No | No | No | Yes |
| Emotions for Children | Android | No | No | Yes | Yes |
| My Emotions | Android | No | No | No | Yes |
| Learning emotions | Android/iOS/Windows | No | No | No | Yes |
| Montessori | Android/iOS | No | No | No | Yes |
| Emotions Project | Android/Windows | No | No | Yes | Yes |
| PictoTEA | Android | No | No | No | Yes |
| MITA | Android/iOS | No | No | No | Yes |
| Autism-Discovering emotions | Android/iOS | No | No | Yes | Yes |
| How Do I Feel? | Android | No | No | Yes | Yes |
| Emotion Learning | Android | No | No | No | Yes |
| #Autism Emotions | Android/iOS | No | No | Yes | Yes |
| Toddler Feelings | Android | No | No | No | Yes |
| Jose' Aprende | Android/iOS | No | No | No | Yes |
| Even better | Web | No | No | No | Yes |

**Figure 15.** Similar applications table [33].

The proposed framework aims to ease future work and implementations by creating a new methodology to refer to, which is also capable of being dynamically expanded in the future. The idea behind the core functionalities is that it is designed to be used in a domestic environment and commonly used devices such as smartphones and tablets. This is to increase the accessibility of the application and allow the methodology to be applied on a larger scale. Furthermore, having in mind an accessible UI facilitates the operations to access the requested features with ease. The minigame has an easy-to-understand implementation, considering the challenges that children with ASD can face. This is an example of how it is possible to digitalise these activities by using AI technology and to make them highly interactive and easier to use.

The game represents an enhanced version of those described in the literature, offering greater potential for further development through the integration of new elements and improved human–computer interaction. The concepts of the minigame levels are used as the base to develop more functionalities to make EmoPal more engaging to use and to reinforce its use. An achievement system can be implemented in future versions, promoting a streak point gathering and achievements. This is to improve the reinforcement principle of the framework even further. The diary implementation is a great way for users to keep track of their emotions in a sandbox environment. Such functionality would help the user to recognise and name emotions expressed by themselves. The association between facial expressions and emotions is made clear to facilitate their identification. This would help children with ASD to give a name to an emotional state and keep track of it. Educators or legal guardians can use such functionality to discuss emotions and have training sessions to elaborate further on them.

Serious games are a very effective way to teach children with ASD, especially using new technologies. The current state of the application allows users to play with only one type of minigame, which can be easily expanded with new features in the future. However, the implementation, although in a prototype state of development, can prove effective to help children develop social-emotional skills by playing as illustrated in the reviewed literature. As also stated in the literature review, serious games can help such individuals recognise and express emotions effectively by experimenting in a playful and engaging environment.

This API can also be replaced with the offline model produced. However, more testing and research are needed to reach a satisfactory level of accuracy. The implementation would greatly improve the effectiveness of EmoPal, as it would remove the need for having an active internet connection. By working offline, the model also allows us to perform the analysis of the image provided for free and in less time. The current average time recorded for the image to be processed is ~300 ms, which will be reduced if the offline model is used.

The application's architecture and its underlying framework are designed to be expanded and improved in time for future development. The libraries used supporting TensorFlow Lite models, which means that the application can use such custom and trained AI models to further improve its accuracy. Further research can expand on the current work by effectively producing

a more accurate model to be used offline, as well as by testing the application with joint work with a professional developmental psychologist and assessing children with ASD, as seen in previous work.

## 6. Conclusions

In this paper, a new framework is designed and thoroughly studied to support the development of social-emotional skills in children with ASD. EmoPal and its framework allow for the development of an engaging application specifically tailored for children with ASD to play to improve their social-emotional skills as well as their emotion awareness. The adopted methodologies help stimulate creativity and abstraction, teaching the user how to recognise patterns seen in facial expressions and social contexts by leveraging AI capabilities.

New emotions can be implemented in further versions of the applications, along with more complex emotions, which might have a new game designed specifically for them. The current state of the application serves as a good prototype for further developments in a complex and interdisciplinary subject. Further studies could help extend the research and future developments with human tests using students with ASD, which was not possible during the present iterations.

The limitations highlighted from testing the AI model and the API are valuable insights for future work and developments. The following work will see the use of a more accurate AI custom model and a better refined interface. Further research is needed to analyse how the produced prototype would impact ASD children, as well as introducing several variables such as the severity of ASD and age of subjects.

Developmental psychology can greatly benefit from computer science solutions in such a thriving and ever-changing world. The findings aim to offer a novel approach to both subjects, shedding a revealing light on how to improve the quality of life of children with ASD by using new emerging technologies such as computer vision and AI, using current technologies at their fullest, and to set standards with the proposed framework, allowing different disciplines to collaborate to create new innovative solutions using an interdisciplinary approach.

**Author Contributions:** Conceptualization, A.L.F.D.L. and P.B.Z.; methodology, A.L.F.D.L.; software, A.L.F.D.L.; validation, A.L.F.D.L., P.B.Z. and K.V.; formal analysis, A.L.F.D.L.; investigation, A.L.F.D.L.; resources, A.L.F.D.L.; data curation, A.L.F.D.L.; writing—original draft preparation, A.L.F.D.L.; writing—review and editing, A.L.F.D.L. and P.B.Z.; visualization, A.L.F.D.L.; supervision, P.B.Z., K.V. and A.S.A.; project administration, P.B.Z.; funding acquisition, A.S.A. 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 used for training the offline model can be found at the following link: https://www.kaggle.com/datasets/msambare/fer2013; URL accessed on 3 April 2024. The developed application can be found at the following link: https://github.com/Bosurgi/EmoPal; URL accessed on 3 April 2024.

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

## References

1. Wu, X.; Deng, H.; Jian, S.; Chen, H.C.; Li, Q.; Gong, R.; Wu, J. Global Trends and Hotspots in the Digital Therapeutics of Autism Spectrum Disorders: A Bibliometric Analysis from 2002 to 2022. *Front. Psychiatry* **2023**, *14*, 1126404. [CrossRef] [PubMed]
2. Hus, Y.; Segal, O. Challenges Surrounding the Diagnosis of Autism in Children. *Neuropsychiatr. Dis. Treat.* **2021**, *17*, 3509–3529. [CrossRef] [PubMed]
3. O'Nions, E.; Petersen, I.; Buckman, J.E.J.; Charlton, R.A.; Cooper, C.; Corbett, A.; Happé, F.; Manthorpe, J.; Richards, M.; Saunders, R.; et al. Autism in England: Assessing Underdiagnosis in a Population-Based Cohort Study of Prospectively Collected Primary Care Data. *Lancet Reg. Health* **2023**, *29*, 100626. [CrossRef] [PubMed]

4. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Text Revision, DSM-5-TR*, 5th ed.; American Psychiatric Association: Washington, DC, USA, 2022; ISBN 9780890425763.
5. Shahroom, A.A.; Hussin, N. Industrial Revolution 4.0 and Education. *Int. J. Acad. Res. Bus. Soc. Sci.* **2018**, *8*, 314–319. [CrossRef] [PubMed]
6. Tahiru, F. AI in Education. *J. Cases Inf. Technol.* **2021**, *23*, 1–20. [CrossRef]
7. Rusia, M.K.; Singh, D.K. A Comprehensive Survey on Techniques to Handle Face Identity Threats: Challenges and Opportunities. *Multimed. Tools Appl.* **2022**, *82*, 1669–1748. [CrossRef] [PubMed]
8. Andrejevic, M.; Selwyn, N. Facial Recognition Technology in Schools: Critical Questions and Concerns. *Learn. Media Technol.* **2019**, *45*, 115–128. [CrossRef]
9. Zakari, H.M.; Ma, M.; Simmons, D. A Review of Serious Games for Children with Autism Spectrum Disorders (ASD). *Serious Games Dev. Appl.* **2014**, 93–106. [CrossRef]
10. Grossard, C.; Grynspan, O.; Serret, S.; Jouen, A.-L.; Bailly, K.; Cohen, D. Serious Games to Teach Social Interactions and Emotions to Individuals with Autism Spectrum Disorders (ASD). *Comput. Educ.* **2017**, *113*, 195–211. [CrossRef]
11. Draaisma, D. Stereotypes of Autism. *Philos. Trans. R. Soc. B Biol. Sci.* **2009**, *364*, 1475–1480. [CrossRef] [PubMed]
12. Kodak, T.; Bergmann, S. Autism Spectrum Disorder. *Pediatr. Clin. North Am.* **2020**, *67*, 525–535. [CrossRef] [PubMed]
13. Baron-Cohen, S. Theory of Mind and Autism: A Review. *Autism* **2000**, *23*, 169–184. [CrossRef]
14. Baron-Cohen, S.; Leslie, A.M.; Frith, U. Does the Autistic Child Have a "Theory of Mind"? *Cognition* **1985**, *21*, 37–46. [CrossRef] [PubMed]
15. Wimmer, H.; Perner, J. Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception. *Cognition* **1983**, *13*, 103–128. [CrossRef] [PubMed]
16. Lozier, L.M.; Vanmeter, J.W.; Marsh, A.A. Impairments in Facial Affect Recognition Associated with Autism Spectrum Disorders: A Meta-Analysis. *Dev. Psychopathol.* **2014**, *26*, 933–945. [CrossRef] [PubMed]
17. Tronick, E.; Als, H.; Adamson, L.; Wise, S.; Brazelton, T.B. The Infant's Response to Entrapment between Contradictory Messages in Face-To-Face Interaction. *J. Am. Acad. Child Psychiatry* **1978**, *17*, 1–13. [CrossRef] [PubMed]
18. Bargiela, S.; Steward, R.; Mandy, W. The Experiences of Late-Diagnosed Women with Autism Spectrum Conditions: An Investigation of the Female Autism Phenotype. *J. Autism Dev. Disord.* **2016**, *46*, 3281–3294. [CrossRef] [PubMed]
19. Kuusikko, S.; Haapsamo, H.; Jansson-Verkasalo, E.; Hurtig, T.; Mattila, M.-L.; Ebeling, H.; Jussila, K.; Bölte, S.; Moilanen, I. Emotion Recognition in Children and Adolescents with Autism Spectrum Disorders. *J. Autism Dev. Disord.* **2009**, *39*, 938–945. [CrossRef] [PubMed]
20. Nisiforou, E.A.; Zaphiris, P. Let Me Play: Unfolding the Research Landscape on ICT as a Play-Based Tool for Children with Disabilities. *Univers. Access Inf. Soc.* **2018**, *19*, 157–167. [CrossRef]
21. Manrique, A.L.; Brito Kozma, E.V.; Tadeu Dirani, E.A.; da Silva, M.L.; Frere, A.F. ICTs in the Classroom, Multiliteracy and Special Education: A Required Interface. *Creat. Educ.* **2016**, *7*, 963–970. [CrossRef]
22. Harley, J.M.; Lajoie, S.P.; Frasson, C.; Hall, N.C. Developing Emotion-Aware, Advanced Learning Technologies: A Taxonomy of Approaches and Features. *Int. J. Artif. Intell. Educ.* **2016**, *27*, 268–297. [CrossRef]
23. Raitoharju, J. Chapter 3—Convolutional Neural Networks. In *Deep Learning for Robot Perception and Cognition*; Iosifidis, A., Tefas, A., Eds.; Academic Press: Cambridge, MA, USA, 2022; pp. 35–69.
24. Di Luzio, F.; Rosato, A.; Panella, M. A Randomized Deep Neural Network for Emotion Recognition with Landmarks Detection. *Biomed. Signal Process. Control* **2023**, *81*, 104418. [CrossRef]
25. Santhoshkumar, R.; Geetha, M.K. Deep Learning Approach for Emotion Recognition from Human Body Movements with Feedforward Deep Convolution Neural Networks. *Procedia Comput. Sci.* **2019**, *152*, 158–165. [CrossRef]
26. Manewa, R.M.A.H.; Mayurathan, B. Emotion Recognition and Discrimination of Facial Expressions Using Convolutional Neural Networks. In Proceedings of the 2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC), Kuching, Malaysia, 1–3 December 2020. [CrossRef]
27. Al-Omair, O.M.; Huang, S. A Comparative Study on Detection Accuracy of Cloud-Based Emotion Recognition Services. In Proceedings of the SPML '18: Proceedings of the 2018 International Conference on Signal Processing and Machine Learning, Shanghai, China, 28 November 2018; pp. 142–148.
28. Daou, N.; Hady, R.T.; Poulson, C.L. Teaching Children with Autism Spectrum Disorder to Recognize and Express Emotion: A Review of the Literature. *Int. Electron. J. Elem. Educ.* **2016**, *9*, 419–432.
29. Francillette, Y.; Boucher, E.; Bouchard, B.; Bouchard, K.; Gaboury, S. Serious Games for People with Mental Disorders: State of the Art of Practices to Maintain Engagement and Accessibility. *Entertain. Comput.* **2020**, *37*, 100396. [CrossRef]

30. de Urturi, Z.S.; Zorrilla, A.M.; Zapirain, B.G. Serious Game Based on First Aid Education for Individuals with Autism Spectrum Disorder (ASD) Using Android Mobile Devices. In Proceedings of the 16th International Conference on Computer Games (CGAMES), Louisville, KY, USA, 25 August 2011; pp. 223–227.
31. Alharbi, M.; Huang, S. An Augmentative System with Facial and Emotion Recognition for Improving Social Skills of Children with Autism Spectrum Disorders. Available online: https://ieeexplore.ieee.org/abstract/document/9275659 (accessed on 21 March 2022).
32. Emoface EMOFACE—APPRENDRE LES EMOTIONS. Available online: https://www.emoface.fr/en/home/ (accessed on 31 May 2024).
33. Garcia-Garcia, J.M.; Penichet, V.M.R.; Lozano, M.D.; Fernando, A. Using Emotion Recognition Technologies to Teach Children with Autism Spectrum Disorder How to Identify and Express Emotions. *Univers. Access Inf. Soc.* **2021**, *21*, 809–825. [CrossRef]
34. Roghani, F.; Jadidi, M.; Peymani, J. The Effectiveness of Floortime Play Therapy on Improving Executive Functions and Cognitive Emotion Regulation in Children with Attention Deficit/Hyperactivity Disorder (ADHD). *Int. J. Educ. Cogn. Sci.* **2022**, *2*, 30–44. [CrossRef]
35. Ornaghi, V.; Agliati, A.; Grazzani, I. *THE STORIES of CIRO and BEBA How to Enhance Conversation with Toddlers on Emotions*; Fronteretro: "R. Massa" Department of Educational Human Sciences University of Milano-Bicocca: Milano, Italy, 2014.
36. White, S.W.; Simmons, G.L.; Gotham, K.O.; Conner, C.M.; Smith, I.C.; Beck, K.B.; Mazefsky, C.A. Psychosocial Treatments Targeting Anxiety and Depression in Adolescents and Adults on the Autism Spectrum: Review of the Latest Research and Recommended Future Directions. *Curr. Psychiatry Rep.* **2018**, *20*, 82. [CrossRef] [PubMed]
37. Statista Mobile OS Market Share | Statista. Available online: https://www.statista.com/statistics/272698/global-market-share-held-by-mobile-operating-systems-since-2009/ (accessed on 24 September 2024).
38. Google Android API Reference. Available online: https://developer.android.com/reference (accessed on 25 September 2024).
39. Prawira, G.S.A.; Kaburuan, E.R.; Effendy, V. User Interface (UI) Design of Scheduling Activity Apps for Autistic Children. In Proceedings of the 2017 International Conference on Orange Technologies (ICOT), Singapore, 8–10 December 2017. [CrossRef]
40. Shabbir, N.; Bhatti, Z.; Hakro, D.N. Serious Game User Interface Design Rules for Dyslexic Children. In Proceedings of the 2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS), Karachi, Pakistan, 14–15 December 2019. [CrossRef]
41. Milla, E.; Radonjić, M. Analysis of Developing Native Android Applications Using XML and Jetpack Compose. *Balk. J. Appl. Math. Inform.* **2023**, *6*, 167–178.
42. Paramadani, R.B.; Aminul Akbar, M.; Pinandito, A. User Interface Rendering Time in Android Applications: Revealing the Effects of Design Patterns. In Proceedings of the 8th International Conference on Sustainable Information Engineering and Technology, Bali, Indonesia, 24–25 October 2023. [CrossRef]
43. Anderson, C. The Model-View-ViewModel (MVVM) Design Pattern. *Pro Bus. Appl. Silverlight* **2012**, *5*, 461–499. [CrossRef]
44. Dagger Documentation. Available online: https://dagger.dev/ (accessed on 23 April 2024).
45. Google CameraX Architecture. Available online: https://developer.android.com/media/camera/camerax/architecture (accessed on 27 April 2024).
46. Jelili, S.; Halayem, S.; Taamallah, A.; Ennaifer, S.; Rajhi, O.; Moussa, M.; Ghazzei, M.; Nabli, A.; Ouanes, S.; Abbes, Z.; et al. Impaired Recognition of Static and Dynamic Facial Emotions in Children with Autism Spectrum Disorder Using Stimuli of Varying Intensities, Different Genders, and Age Ranges Faces. *Front. Psychiatry* **2021**, *12*, 693310. [CrossRef] [PubMed]
47. Black, M.H.; Chen, N.T.M.; Iyer, K.K.; Lipp, O.V.; Bölte, S.; Falkmer, M.; Tan, T.; Girdler, S. Mechanisms of Facial Emotion Recognition in Autism Spectrum Disorders: Insights from Eye Tracking and Electroencephalography. *Neurosci. Biobehav. Rev.* **2017**, *80*, 488–515. [CrossRef] [PubMed]
48. Ekman, P. Basic Emotions. *Handb. Cogn. Emot.* **1999**, 45–60. [CrossRef]
49. Golan, O.; Ashwin, E.; Granader, Y.; McClintock, S.; Day, K.; Leggett, V.; Baron-Cohen, S. Enhancing Emotion Recognition in Children with Autism Spectrum Conditions: An Intervention Using Animated Vehicles with Real Emotional Faces. *J. Autism Dev. Disord.* **2010**, *40*, 269–279. [CrossRef] [PubMed]
50. Bogush, D.; Costello, C.; Fong, E.; Sharmin, M.; Ahmed, S. Emotion Recognition Technology for People with Autism: A Systematic Literature Review. In Proceedings of the 48th IEEE International Conference on Computers, Software, and Applications (COMPSAC), Osaka, Japan, 2–4 July 2024; pp. 851–861. [CrossRef]
51. deLeyer-Tiarks, J.M.; Li, M.G.; Levine-Schmitt, M.; Andrade, B.; Bray, M.A.; Peters, E. Advancing Autism Technology. *Psychol. Sch.* **2022**, *60*, 495–506. [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.

computers

MDPI

Article

# From Patterns to Predictions: Spatiotemporal Mobile Traffic Forecasting Using AutoML, TimeGPT and Traditional Models

**Hassan Ayaz [1], Kashif Sultan [1], Muhammad Sheraz [2] and Teong Chee Chuah [2,*]**

[1] Department of Software Engineering, Bahria University, Islamabad 44000, Pakistan; sh_hassan_ayaz@outlook.com (H.A.); kashif.buic@bahria.edu.pk (K.S.)
[2] Centre for Smart Systems and Automation, CoE for Robotics and Sensing Technologies, Faculty of Artificial Intelligence and Engineering, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia; engr.msheraz@gmail.com
* Correspondence: tcchuah@mmu.edu.my

**Abstract:** Call Detail Records (CDRs) from mobile networks offer valuable insights into both network performance and user behavior. With the growing importance of data analytics, analyzing CDRs has become critical for optimizing network resources by forecasting demand across spatial and temporal dimensions. In this study, we examine publicly available CDR data from Telecom Italia to explore the spatiotemporal dynamics of mobile network activity in Milan. This analysis reveals key patterns in traffic distribution highlighting both high- and low-demand regions as well as notable variations in usage over time. To anticipate future network usage, we employ both Automated Machine Learning (AutoML) and the transformer-based TimeGPT model, comparing their performance against traditional forecasting methods such as Long Short-Term Memory (LSTM), ARIMA and SARIMA. Model accuracy is assessed using standard evaluation metrics, including root mean square error (RMSE), mean absolute error (MAE) and the coefficient of determination ($R^2$). Results show that AutoML delivers the most accurate forecasts, with significantly lower RMSE (2.4990 vs. 14.8226) and MAE (1.0284 vs. 7.7789) compared to TimeGPT and a higher $R^2$ score (99.96% vs. 98.62%). Our findings underscore the strengths of modern predictive models in capturing complex traffic behaviors and demonstrate their value in resource planning, congestion management and overall network optimization. Importantly, this study is one of the first to Comprehensively assess AutoML and TimeGPT on a high-resolution, real-world CDR dataset from a major European city. By merging machine learning techniques with advanced temporal modeling, this study provides a strong framework for scalable and intelligent mobile traffic prediction. It thus highlights the functionality of AutoML in simplifying model development and the possibility of TimeGPT to extend transformer-based prediction to the telecommunications domain.

**Keywords:** Mobile Network Traffic; TimeGPT; AutoML; Call Detail Records (CDRs); Time-Series forecasting; traffic classification

## 1. Introduction

The rapid growth of mobile technologies has extremely transformed the way we communicate, interact and perform our daily tasks. As a result, next generation networks (NGNs) have become the foundation of modern telecommunications. With the extensive adoption of smartphones, tablets, connected devices and the expansion of the Internet of Things (IoT) ecosystem, mobile connectivity has become fundamental to everyday life.

According to reports from Ericsson and Cisco, global mobile data traffic now exceeds 24 exabytes per month [1], exhibiting the enormous scale and intensity of network usage.

This explosive increase in traffic is particularly pronounced in densely populated urban areas, where demand often exceeds the capacity of existing frameworks. Urban network planning has therefore become a vital task for telecommunications operators and urban planners [2]. The heterogeneous nature of mobile services, including voice, video, messaging and data further complicates the challenge of traffic management. To maintain service quality and avoid congestion, advanced techniques for traffic forecasting and resource allocation are urgently needed.

Call Detail Records (CDRs) are a precious resource for understanding mobile network behavior. These records capture granular information about user activity, such as time, location and type of communication, enabling detailed spatiotemporal analysis. With the growth of mobile and IoT technologies, the volume and velocity of this data will continue to expand, requiring analytics solutions that adhere to the fundamental principles of big data: volume, variety, velocity, veracity and value [3].

Conventional forecasting models, such as ARIMA and SARIMA, although widely used, struggle to model the nonlinear and evolving dynamics of mobile traffic patterns. These models frequently fail to account for spatial variability and time dependent behaviors, resulting in lower forecast accuracy and poor resource planning, especially in high demand urban habitat [4].

Mobile network traffic forecasting is fundamentally complex due to its spatiotemporal nature. Traffic volumes vary not only over time but also across locations and are determined by factors such as human mobility, time of day, weekdays and weekends, public events and socioeconomic behaviors. Moreover, different network services contribute unequally to the total traffic load. This complexity requires forecasting solutions that can simultaneously handle spatial heterogeneity and temporal dynamics.

An successful initial step is to segment mobile traffic based on parameters such as service type, origin destination flow and usage trends. This helps identify restricted demand patterns and allows network operators to identify peak periods and high-demand regions. Forecasting builds on this proposition, allowing telecom operators to foresee future demand and make proactive decisions regarding infrastructure upgrades and traffic optimization.

To address these needs, advanced time-series forecasting models are gaining popularity. These include Time Series Pretrained Generative Transformer (TimeGPT) and Automated Machine Learning (AutoML). TimeGPT leverages transformer-based architectures to apprehend composite temporal dependencies, while AutoML automates key steps in the modeling process, such as feature engineering, model selection and hyperparameter tuning, easing operational arrangement.

This study aims to evaluate and compare the capabilities of TimeGPT and AutoML in forecasting mobile network traffic using real-world CDR data. While generative models like TimeGPT have demonstrated impressive performance in fields such as natural language processing, their application in mobile traffic forecasting remains relatively underexplored. Similarly, AutoML provides a practical solution by minimizing the need for manual intervention, yet its comparative advantage in this domain is still being established.

To this end, we propose a comprehensive framework that systematically benchmarks these two advanced approaches on a high-resolution spatiotemporal dataset. The CDR data used in this study is sourced from Telecom Italia and captures mobile activity in the city of Milan, segmented into a 100 × 100 spatial grid and aggregated at fine temporal intervals.

Unlike prior studies that focus solely on conventional statistical or deep learning models, this research integrates state-of-the-art transformer-based learning and automated

modeling within a unified framework. The findings not only demonstrate the superior accuracy of these advanced models but also offer practical insights into scalable forecasting strategies for next-generation mobile networks.

The main contributions of the study are as follows:

1. High-Resolution Spatiotemporal Analysis: We conduct a detailed exploration of traffic patterns across different regions and times, enabling insights into user behavior and network demand.
2. Comparative Forecasting using TimeGPT and AutoML: We apply both models independently to forecast mobile traffic. TimeGPT is used for its strength in modeling temporal dependencies, while AutoML is leveraged for its automation of the end-to-end modeling process.
3. Benchmarking Against Traditional Models: We compare the forecasting accuracy of TimeGPT and AutoML against baseline models, including ARIMA, SARIMA and LSTM, using RMSE, MAE and $R^2$ as evaluation metrics. Results show that AutoML achieves the highest accuracy ($R^2$ = 99.96%), followed by TimeGPT (98.62%).

## 2. Literature Review

This section discusses the main research on mobile network traffic forecasting and classification. Traditionally, this type of analysis has been based on traditional models that include ARIMA, SARIMA, Support Vector Machines (SVMs), Random Forest, etc. These models require significant computing power, extensive feature engineering and a lot of time to produce accurate results. Various methods, including clustering algorithms, time series forecasting and statistical modeling, have been widely used. However, recent advancements have introduced innovative solutions such as TimeGPT and AutoML that provide improved efficiency and scalability.

In [5], the authors explore the use of CDRs for time series forecasting and trend detection in call data. They highlight the role of CDRs in improving information on user behavior and network performance. The paper uses the hyperparameter tuning models for such methods as SVMs, Naïve Bayes and Decision Trees to provide the most suitable results. However, challenges such as data quality, computational requirements and the complexity of interpreting results can limit real-world applications.

The paper [6] presents Informer, an effective transformer-based model for long-sequence time-series forecasting that significantly improves scalability and accuracy by introducing the ProbSparse self-attention mechanism, reducing complexity to O(L log L), as well as a generative decoder that predicts entire sequences in a single step, resulting in faster inference. Among its main advantages are enhanced computational efficiency, superior dependency modeling and robust performance on large datasets, surpassing existing variants of transformers and RNN-based models. This approach can serve as a valuable foundation for studying the prospects of using Call Data Recorders (CDRs) in predicting long-term activities and mobility, as its ability to effectively model complex long-term dependencies makes it suitable for analyzing extensive temporal data. However, the model's dependence on scarcity assumptions may limit its effectiveness on data exhibiting vague dependencies and the lack of formal guarantees regarding approximation errors poses a challenge for certain applications. Future research could focus on adaptive parsimony techniques and theoretical analyses to address these limitations.

The comparison of the correctness of Multi-Layer perceptron (MLP), MLP with Dropout (MLPWD) and SVMs in forecasting mobile traffic using real-world data is presented in [7]. The results show that SVM works best on complex traffic while MLPWD is more accurate on one-dimensional data. However, the study is limited to a specific

business network due to a lack of generalizability between different networks and external factors such as user behavior and network events.

In the study [8], various statistical prediction models for real IP network traffic are used to improve 5G network management, especially traffic forecasting, which enables better resource allocation and load balancing. It allows for self-adaptive networks, but depends on the quality of the data for its accuracy. Implementation requires significant investments in infrastructure and skills, with adapting to different grid conditions remaining a challenge.

The article [9] presents the results of training a neural network for prediction and looks at neural networks for mobile traffic prediction in 5G to improve accuracy and resource allocation. It highlights ongoing approaches, research opportunities and the importance of continuously updating the radio heatmaps. One of the main challenges is the reliance on high-quality training data that limits the effectiveness of the model under network conditions.

A machine learning-based approach to mobile traffic classification and forecasting is incorporated in [10]. It combined the Naïve Bayesian classifier with Holt–Winters prediction to improve traffic pattern detection, improve resource management, reduce data loss and stabilize traffic fluctuations to improve network performance and energy efficiency.

The study [11] introduces a multitask learning (MTL) model to improve the classification and prediction of mobile traffic at the edge. Unlike traditional methods that manage tasks separately and rely on data-level inputs, the proposed approach improves data extraction in the control channel and improves resource allocation, classification accuracy and computational efficiency.

The article focuses [12] on data preprocessing and highlights robust techniques for dealing with outliers and missing values to improve forecast accuracy. It also discusses incremental learning for seamless adaptation to changing data. Parallel computing is studied with respect to the efficiency of large datasets. However, the complexity of the model can reduce accuracy and transferability, thus limiting its use in practice. In addition, advanced methods can struggle to manage dependencies over long distances and require significant computational resources.

The examination of deep learning models for time series forecasting highlights hybrid models that include parametric model selection and traditional machine learning methods such as kernel regression, support vector regression and Gaussian process are discussed in [13]. It proposes future research on hierarchical structures, interpretation methods and counterfactual forecasting. A central challenge is the need for regularly spaced data, which complicates forecasting for noisy datasets and underscores the necessity for continuous time models.

The article [14] presents the STGCN, an innovative deep learning framework that effectively models both spatial and temporal dependencies in traffic data using graph convolutional layers and temporal convolutional layers with gates. Its main advantages include faster training, fewer parameters and enhanced accuracy compared to traditional methods, particularly in complex traffic networks. The STGCN's ability to integrate spatial topology significantly improves the performance of short-, medium- and long-term forecasts. However, the model's dependence on precise graphical structures and its current approach to traffic prediction may limit its generalization to other domains or to less structured data. Overall, the STGCN offers an effective and scalable solution for real-time traffic forecasting, with potential applications extending beyond transportation systems.

The work classifies the initial methods of time series classification into four categories: prefix-based, shapelet-based, model-based and others in [15]. It emphasizes interpretability, particularly in the healthcare domain, as well as the challenge of reconciling accuracy

with timely forecasts. Applications in medical diagnosis and industrial monitoring are extensively examined. Future research focuses on probabilistic classifiers and multivariate relationships of time series. The study highlights limitations, such as ensuring reliable forecasts, managing the tradeoff between accuracy and speed, interpretation gaps and the risk of overlooking emerging methods.

The examination of spatiotemporal graph neural networks (GNNs) in time series analysis and analyses of over 150 specialized journal articles is conducted in [16]. It highlights applications in global networks, particularly for renewable energy forecasting while addressing significant limitations such as comparability, reproducibility, explainability and scalability. A central issue is the lack of transparency in GNN models, which are often regarded as "black boxes," limiting their application in fields requiring interpretability.

The article [17] examines how big data analytics enhance mobile networks through the use of CDR data for anomaly detection and traffic forecasting. Unlike traditional methods that utilize incomplete data, big data ensures more accurate insights. K-means clustering detects anomalies for better resource allocation, while clean data improves model performance. The use of ARIMA for forecasting emphasizes that high-quality data increases accuracy. However, challenges include the need for high-quality CDR data and high computational requirements. Ultimately, big data analysis enhances network efficiency through better anomaly detection and accurate forecasting.

The paper [18] highlights that large language models (LLMs) do not outperform simpler, more efficient methods in time series forecasting, as evidenced by ablation studies showing no significant advantage in leveraging their sequence modeling capabilities. It emphasizes that straightforward approaches like patching and attention-based encoders achieve comparable performance with much less computational cost, challenging the prevalent use of LLMs for such tasks. However, its limitation is that it mainly focuses on datasets with uniform intervals and traditional forecasting, leaving open whether LLMs could be more effective in multimodal or complex reasoning scenarios involving time series. Overall, the study encourages shifting focus toward more specialized and efficient models for time series analysis.

## 3. Proposed Framework

The proposed framework for mobile network traffic analysis and forecast begins with a dataset that undergoes preprocessing, including handling missing values, outlier detection and data cleaning. Following preprocessing, exploratory data analysis (EDA) is conducted to examine spatial and temporal variations in the data. This helps recognize the traffic forecasting patterns, where two predictive models, TimeGPT and AutoML, are applied. The models' performance is compared based on accuracy. The final result is then obtained, ensuring an optimized and reliable traffic forecasting system. The proposed framework is represented in Figure 1.

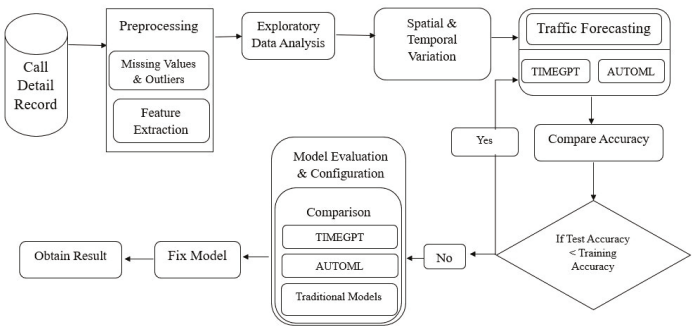



**Figure 1.** Proposed framework.

Dataset Description: This study utilizes a publicly available dataset from Telecom Italia comprising Call Detail Records (CDRs) that capture mobile network activity across the city of Milan. The dataset aggregates mobile communication events including voice calls, SMS and data usage into 10 min intervals, each tagged with both a timestamp and spatial coordinates. This granular spatiotemporal data enables detailed analysis of usage patterns across time and geographic regions. Key attributes in the dataset include the following:

- DateTime: Timestamp for each 10 min aggregation interval.
- Country Code: Identifier for the country (removed during preprocessing).
- Grid ID: Spatial identifier; Milan is divided into a 100 × 100 grid (10,000 cells), with each grid cell covering approximately 0.3 $km^2$.
- Call In/Out: Number of incoming and outgoing calls in a given grid.
- SMS In/Out: Volume of incoming and outgoing SMSs per grid.
- Data Usage: Total data consumption recorded per grid.

Preprocessing Steps: To ensure the dataset's reliability and suitability for modeling, several preprocessing steps were applied.

- Column Removal and Aggregation: The Country Code column, being noninformative for intracity analysis, was excluded. Additionally, the data was aggregated to an hourly resolution to balance granularity with computational efficiency, preserving key usage trends while reducing dimensionality.
- Handling Missing Values: Null entries were identified using standard pandas operations. Given the nature of CDR aggregation, missing values were rare. Where they occurred (mostly in low-traffic grids), they were imputed using forward-fill and backward-fill methods to maintain temporal continuity.
- Outlier Detection and Treatment: We applied the Interquartile Range (IQR) method to detect outliers in each of the five key features: call_in, call_out, sms_in, sms_out and data_all. Any value falling below Q11.5 × IQR or above Q3 + 1.5 × IQR was considered an outlier. These were either capped at the nearest threshold or were excessive are removed to maintain statistical validity.
- Feature Engineering: A new composite feature, avg_use, was created by averaging the five core activity features. This helped standardize network activity across all grid cells and time intervals, simplifying the modeling input while retaining essential patterns.
- Normalization: The average usage values were normalized with MinMaxScaler to scale the data into the range [0, 1]. This step resulted in faster convergence and consistent performance between TimeGPT and AutoML models.

This carefully structured preprocessing pipeline improved the dataset's consistency and enhanced its utility for downstream forecasting tasks.

Exploratory Data Analysis (EDA): Exploratory data analysis was conducted to extract meaningful spatial and temporal patterns from the CDR dataset. Visualizations across different times of day and locations helped uncover usage peaks in high-density urban areas and lower activity in suburban zones. These insights are critical for interpreting demand distribution and understanding behavioral patterns over time.

Traffic Forecasting Strategy: To forecast future mobile network traffic, two advanced approaches were adopted: TimeGPT and AutoML. These models were chosen for their distinct strengths, TimeGPT for its deep temporal modeling capabilities and AutoML for its ability to automate the entire model-building pipeline. Both were evaluated against traditional forecasting methods (ARIMA, SARIMA, LSTM) to benchmark their performance comprehensively.

TimeGPT Architecture: The state-of-the-art transformer-based model specifically designed for time series forecasting models addresses the temporal dependencies and provides accurate network traffic forecasts. Figure 2 describes the TimeGPT architecture, which is a transformer-based model for time series forecasting specifically applied to cellular network traffic analysis. It processes Call Detail Records and integrates temporal features such as timestamp, date and grid ID for structured learning. It is built on a GPT-based foundation and uses an integration layer and a self-attention mechanism to capture temporal dependencies. The decoder generates short- and long-term forecasts, providing adaptability. With scalability and performance optimization, the model evaluates forecasting using robust metrics. Lastly, output integration transforms the results into meaningful forecasts, enabling efficient traffic forecasting.

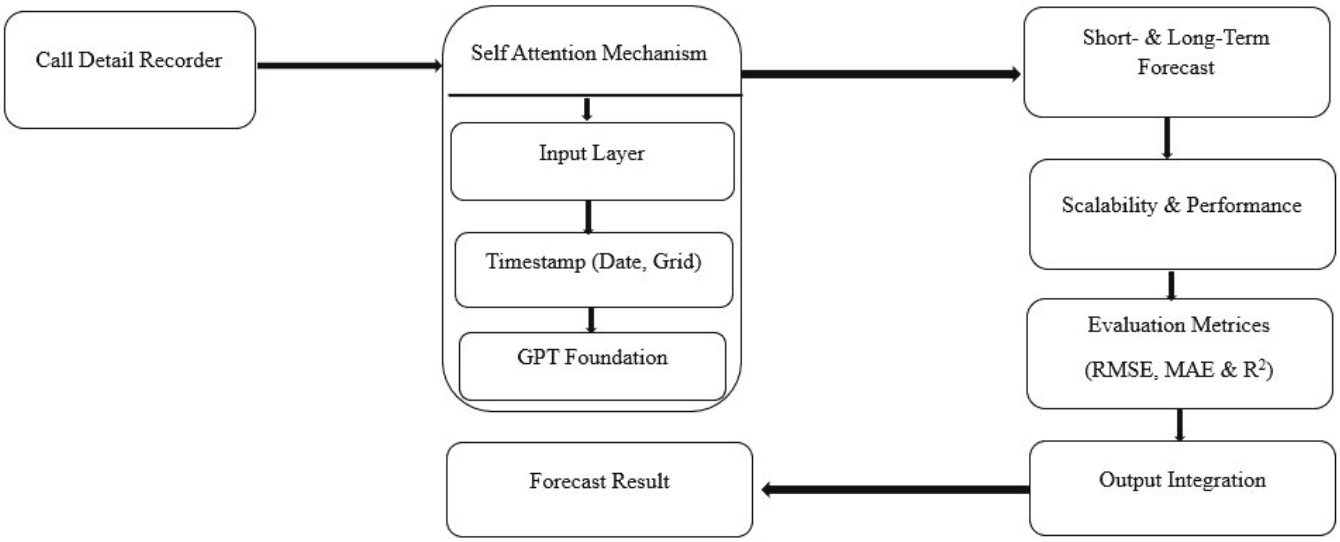



**Figure 2.** TimeGPT architecture.

In this study, the architecture is adapted to learn mobile network usage patterns from CDRs. It includes several key steps.

- Input encoding: Raw features such as timestamps, grid IDs and usage metrics are converted into high-dimensional embeddings. Positional encodings are added to preserve the temporal sequence of the inputs.
- Temporal integration layer: Periodic patterns, such as daily or weekly cycles, are captured through additional integrations, allowing the model to learn the seasonality of network behavior.
- Transformer structure: A stacked self-attention mechanism allows the model to learn long-term dependencies over specific time intervals. Multi-head attention further improves its ability to model parallel input relationships.
- Feedback layer: The transformed features are passed through a nonlinear feedback network to enrich the representation.
- Decoder: The model predicts future usage values across a defined forecast horizon (e.g., 24 h), using the contextual embeddings generated in the earlier layers.
- Evaluation Block: The model's accuracy is validated using standard metrics: RMSE, MAE and $R^2$, ensuring both reliability and generalization.

The final output is seamlessly integrated into forecasting pipelines for real-time traffic monitoring, anomaly detection and strategic planning. Each component in TimeGPT plays a distinct role in ensuring that temporal dependencies, seasonal variations and

contextual usage shifts are accurately captured. The combination of ProbSparse attention, unidirectional decoding and transformer-based scaling makes this model highly suitable for large-scale mobile network forecasting.

AutoML simplifies and accelerates the model development process by automating key stages such as feature engineering, hyperparameter tuning and model selection. For this study, the H2O AutoML platform was selected due to its strong compatibility with time series data like Call Detail Records (CDRs), which capture temporal and spatial variations in mobile network activity.

The AutoML-based forecasting pipeline, shown in Figure 3, provides a fully automated and scalable approach to processing CDR data, from raw data ingestion to model deployment. The workflow begins with data preparation, which involves cleaning, structuring and formatting raw CDRs. This includes parsing timestamps, normalizing features and encoding spatial identifiers to ensure consistency and interpretability.

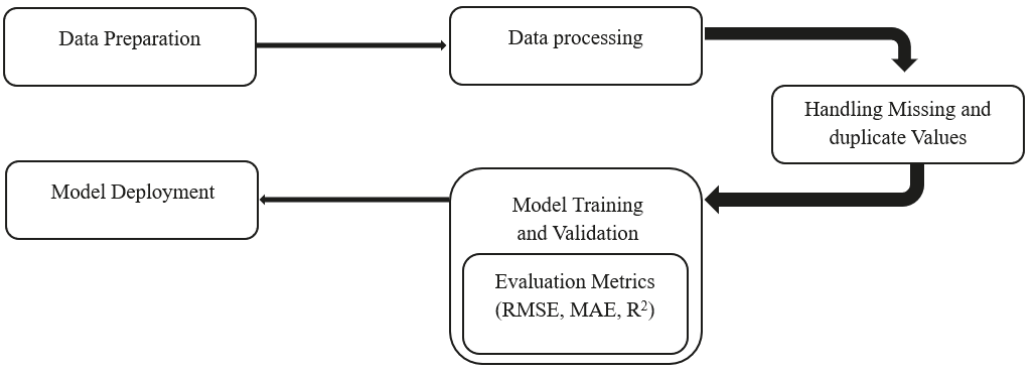



**Figure 3.** AUTOML Architecture.

A dedicated preprocessing module addresses data quality issues by removing duplicates, filling in missing values and generating important temporal features such as time of day, day of the week and lag indicators. Once preprocessing is complete, the data enters the model training and validation phase. The AutoML engine then runs multiple candidate models in parallel, including Gradient Boosting Machines (GBMs), XGBoost, Random Forests, deep neural networks and stacked ensembles.

During training, H2O AutoML automatically optimizes hyperparameters using advanced techniques such as random grid search and Bayesian optimization to improve model performance. To rigorously evaluate these models, a time series-based cross-validation strategy (step-forward validation) is applied, preserving the temporal sequence and simulating real-world forecasting conditions. Each candidate model is evaluated using standardized performance metrics, including root mean square error (RMSE), mean absolute error (MAE) and $R^2$. The model with the best RMSE across all validation criteria is ranked first.

After validation, the best-performing model is deployed for real time forecasting. The pipeline also includes a monitoring mechanism that continuously evaluates forecast performance. If signs of model deviation or degradation, such as increased error rates or changing data patterns, are detected, the system automatically triggers model retraining with the most recent data to ensure high forecast accuracy.

To support this end-to-end workflow, the AutoML pipeline includes several specialized modules.

- Feature Engineering Module: Automatically generates informative predictors, including lag variables, moving averages and spatiotemporal encodings.

- Model Selector: Compares a wide range of algorithms such as Gradient Boosting Machines, XGBoost, deep learning models and stacked ensembles using a ranking mechanism to identify the optimal candidate.
- Hyperparameter Optimizer: Employs techniques like grid search, random search, or Bayesian optimization to fine-tune models for best performance.
- Cross-Validation Engine: Applies time-aware validation methods (e.g., walk-forward validation) to ensure the robustness and generalizability of the selected model.
- Ensemble Module: Optionally blends the predictions of top-performing models to enhance accuracy and reduce variance.
- Forecast Generator: Produces real-time traffic forecasts and supports adaptive learning by triggering model updates based on changing data conditions.

By leveraging H2O AutoML, the entire forecasting pipeline becomes both efficient and interpretable, making it particularly well suited for large-scale, high-frequency, spatiotemporal datasets, such as those used in modern mobile network environments.

Model Evaluation: The forecasting performance of TimeGPT and AutoML was assessed using standard evaluation metrics: mean absolute error (MAE), mean squared error (MSE), root mean square error (RMSE) and the coefficient of determination ($R^2$). To provide a well-rounded analysis, both models were benchmarked against conventional forecasting approaches, including ARIMA, SARIMA and LSTM. These metrics offer a quantifiable measure of predictive accuracy and serve as a consistent basis for comparing the strengths and limitations of each model in capturing the dynamic nature of mobile network traffic.

Model Configuration: To ensure a fair comparison across all models, we standardized the experimental conditions, including data preprocessing, evaluation protocols and parameter tuning. The following configurations were applied:

- TimeGPT (Chronos Library): TimeGPT was implemented using a 24 h sliding window to predict the subsequent 24 h period. Training was conducted with a learning rate of 0.001, a batch size of 32 and a maximum of 50 epochs. Early stopping was applied after five epochs without improvement. Pretrained weights were used to simulate real-world deployment and reduce the need for extensive retraining.
- AutoML (H2O): The AutoML pipeline was configured with a runtime cap of 30 min per run and limited to a maximum of 20 model candidates. RMSE was used as the leader board metric and 5-fold cross-validation ensured the robustness of the results. All preprocessing, feature engineering, model selection and hyperparameter tuning were handled automatically by the H2O AutoML engine, with default settings used for early stopping and random seed to maintain reproducibility.
- Traditional Models:
  ARIMA: Configured with parameters (p = 5, d = 1, q = 0), selected based on AIC minimization and residual analysis.
  SARIMA: Tuned to (p = 1, d = 1, q = 1) (P = 1, D = 1, Q = 0) [using seasonal decomposition and grid search.
  LSTM: Implemented with one LSTM layer (50 units), followed by a dense output layer. The model was trained using the Adam optimizer and MSE loss, with a batch size of 32, dropout rate of 0.2 and 100 epochs.

Fix Model (Model Refinement Strategy): In cases where a model exhibited high training accuracy but underperformed on validation or test data, further optimization was conducted to improve generalization. This included the following:

- Hyperparameter Tuning: Grid search and random search were used to explore various parameter settings for each model.

- Advanced Feature Engineering: Additional features were created based on domain knowledge, including interaction terms and polynomial transformations, to better capture complex patterns.
- Cross-Validation: Models were reevaluated using multiple data splits to ensure consistent performance across different subsets.
- Ensemble Techniques: Where beneficial, ensemble methods such as bagging, boosting, or stacking were applied to combine the strengths of multiple models.

This iterative refinement process continued until the models achieved balanced performance across training and testing phases, indicating readiness for practical deployment in mobile traffic forecasting scenarios.

## 4. Results and Experimentation

### *4.1. Correlation Analysis*

A correlation map visually represents the strength and direction of relationships between numerical variables. Figure 4 shows the correlation map for the dataset at the specific time interval. It highlights the relationships between different attributes such as incoming/outgoing SMSs, incoming/outgoing calls and data usage. Color-coded cells indicate positive, negative, or no correlation between variables.

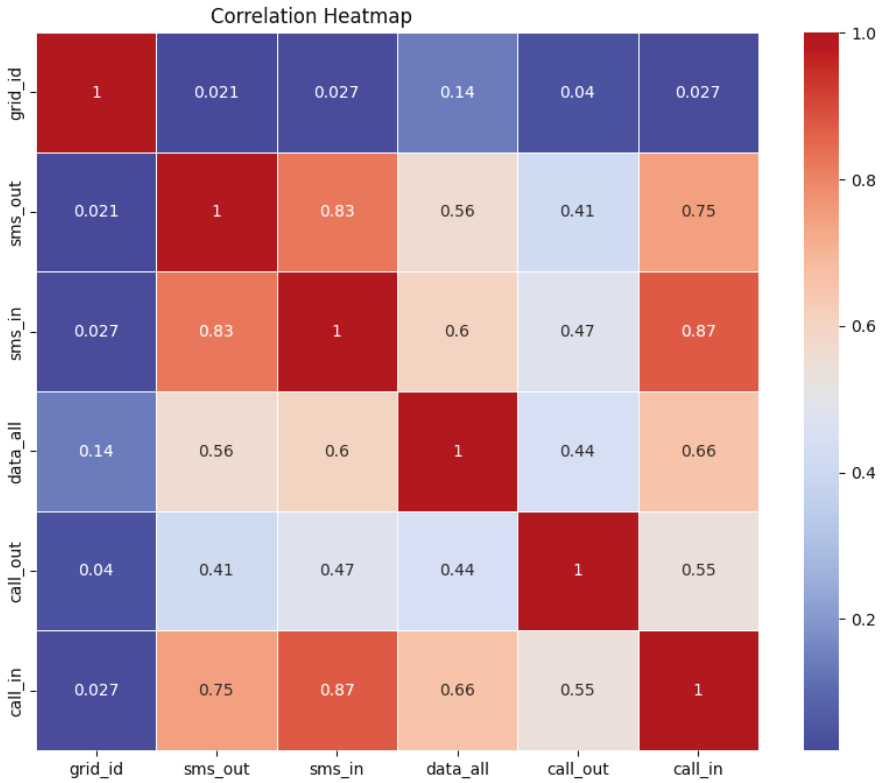



**Figure 4.** Correlation heatmap showing relationships between SMSs, calls and data usage.

Figure 4 presents a correlation matrix generated using Pearson correlation analysis on the preprocessed CDR (Call Detail Record) data. It is important to clarify that this matrix was not derived from a neural network weight matrix but rather constructed to quantify linear dependencies among key telecommunication variables: sms_in, sms_out, call_in, call_outand data_all.

The correlation map in Figure 4 provides a visual representation of the strength and direction of these relationships. Each cell is color-coded to reflect whether the correlation

is positive, negative, or negligible. For example, a strong positive correlation (0.83) is observed between sms_in and sms_out, indicating synchronized texting behavior areas with high outgoing SMS activity also tend to receive a high volume of incoming messages. Additionally, the variable data_all shows a moderate positive correlation (0.66) with both call_in and call_out. This suggests that regions with higher mobile data usage also tend to exhibit increased voice call activity, pointing to potential convergence in communication preferences across different services in densely active grid areas.

This correlation analysis helps establish foundational insights into user behavior patterns and supports further modeling and feature selection for predictive tasks.

### *4.2. Spatial Analysis*

After analyzing the correlations, we conducted a spatial analysis of the data by fixing the time and examining the entire city of Milan using CDRs. Figure 5 illustrates the spatial distribution of aggregated CDR activities at a specific timestamp. To simplify the visualization and enable clearer interpretation, we introduced a derived metric termed Combined CDR Activity, calculated by summing all normalized communication metrics: incoming and outgoing calls, SMSs and data usage. The data is plotted on a 100 × 100 spatial grid, with the X and Y axes representing geographic coordinates. Activity levels are represented using a heatmap where color intensity corresponds to usage volume: darker purple shades indicate low activity, while lighter yellow areas signify regions of peak telecom usage. This spatial distribution reveals pronounced activity hotspots in central urban zones likely due to higher population densities and more intensive mobile activity while activity levels diminish in suburban and peripheral areas, reflecting lower population density and reduced communication demand. This visualization offers valuable insight for urban planners and network operators to identify zones requiring infrastructure investment or optimization.

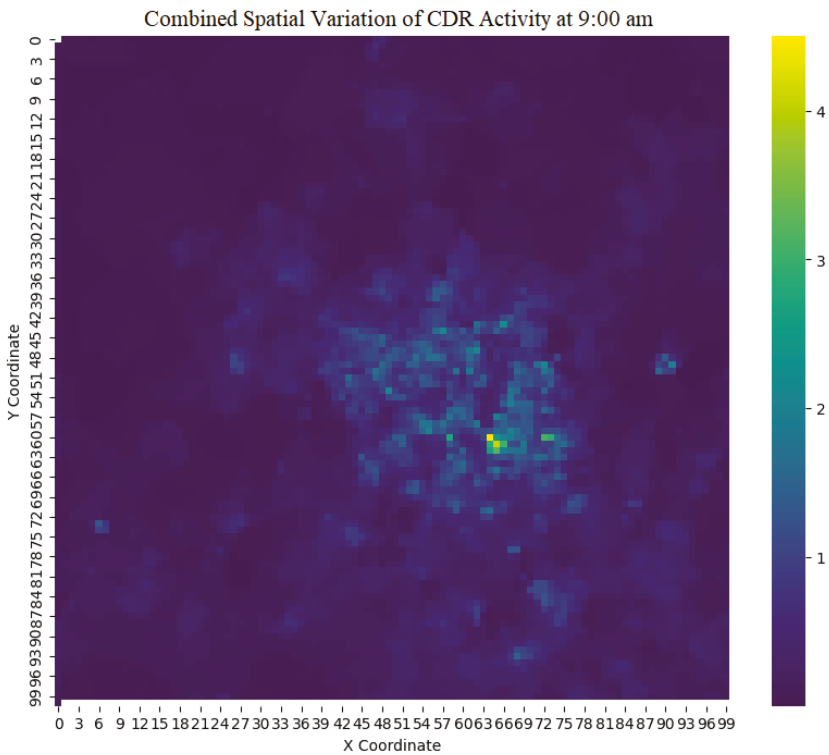



**Figure 5.** Spatial distribution of combined CDR activities across Milan at a fixed time, highlighting central areas of high mobile network usage.

Figure 6 illustrates the temporal progression of spatial CDR usage throughout a single day, using snapshots taken every three hours (00:00, 03:00, 06:00, . . . , 21:00). This series of heatmaps provides insight into how network activity shifts across the city during daily cycles. The persistence of higher intensity in central regions aligns with working hours and business zones, while peripheral areas show periodic increases, possibly tied to residential routines or commuting.

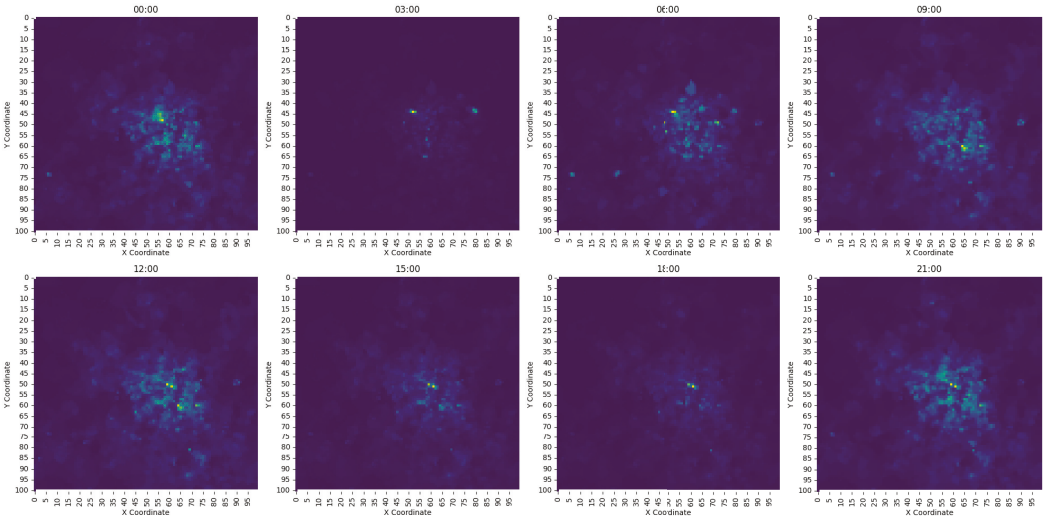



**Figure 6.** Spatiotemporal variation in mobile network activity at quarterly time intervals.

We further extended our spatial analysis to examine variations across different time slots throughout the day, as shown in Figure 6. These variation maps depict how network activity changes over time, providing insight into usage dynamics throughout the day. Figure 6 presents activity during the complete day at quarterly intervals: 00:00, 03:00, 06:00, 09:00, 12:00, 15:00, 18:00 and 21:00. The analysis shows that central regions consistently experience higher traffic, confirming their status as urban and commercial hubs. In contrast, other regions show fluctuating activity levels depending on the time of day, revealing shifting communication demands.

On the heat maps in Figure 6, dark areas represent grids with the lowest CDR utilization, indicating low network activity. Light areas indicate regions with the highest utilization, often in urban or commercial centers. Moderate or medium colors reflect areas with average or fluctuating activity throughout the day. This color gradient helps visualize the dynamics of communication patterns in different regions and at different times.

### *4.3. Temporal Analysis*

For the temporal analysis, we examined the CDR activity of a specific grid/location over the course of a single day. Figure 7 illustrates the temporal variation in CDR activity. The x-axis represents hourly intervals, while the y-axis shows the level of CDR activity. The figure reveals a distinctive temporal pattern in mobile network usage. Activity is moderate around midnight and gradually declines, reaching its lowest point between 4 a.m. and 6 a.m., a period typically associated with minimal user engagement. A sharp increase in usage begins around 7 a.m., aligning with the start of daily routines, commuting and business communications. The first peak is observed in the early afternoon (around 2–3 p.m.), likely driven by midday internet usage and ongoing work-related activities. The blue line in the figure represents the average usage values per hour and the circular markers indicate the exact data points. Because this figure contains only one data series,

no additional color differentiation is applied. The legend field is displayed by default, but does not indicate multiple categories.

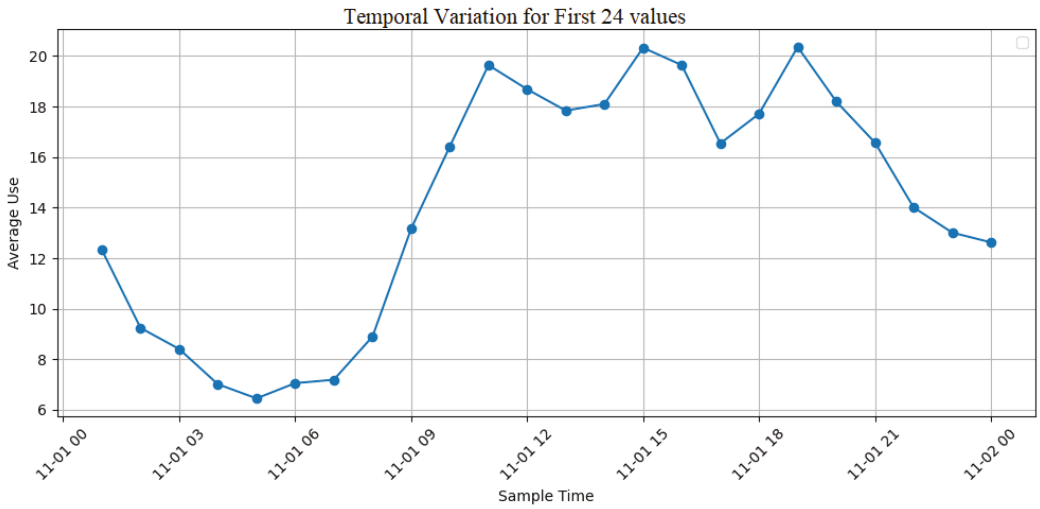



**Figure 7.** Hourly temporal variation in CDR activity in a specific grid location, illustrating daily usage trends.

To explore longer-term trends, we analyzed the CDR activity over a one-week period for the same grid. Figure 8 shows the temporal variation in network activity throughout the week. The x-axis represents time in days, while the y-axis shows CDR activity levels. This representation effectively captures usage fluctuations over time. Careful observation reveals a recurring 24 h cycle, indicating that activity closely mirrors human behavioral patterns of higher usage during the day and lower activity at night. The amplitude of these peaks varies across the week, with notable surges on November 3 and 6, possibly due to specific events or anomalies. These trends and outliers are essential for understanding user behavior, optimizing network resources and guiding infrastructure planning.

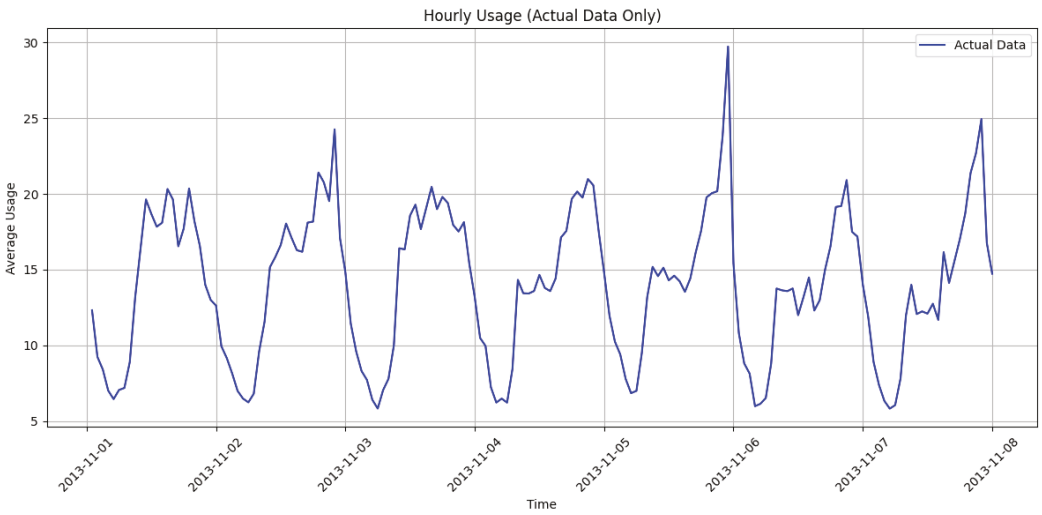



**Figure 8.** Weekly temporal variation in CDR activity in a fixed grid location, capturing daily cycles and anomalous spikes.

### *4.4. Network Traffic Forecasting*

After understanding the spatiotemporal variation of the CDR actives, we proceeded to forecast them. For CDR traffic forecasting, we used two different state-of-the-art models: TimeGPT and AutoML.

#### 4.4.1. Traffic Forecasting with TimeGPT

Figure 9 presents the forecasting results of CDR activities generated by the TimeGPT model, using one week of data as the input window. The blue line with circular markers represents the actual average usage, while the dashed red line with cross-markers indicates the predicted values produced by TimeGPT. This side-by-side visualization provides a clear reference point for evaluating the model's performance in capturing real-world mobile network usage patterns.

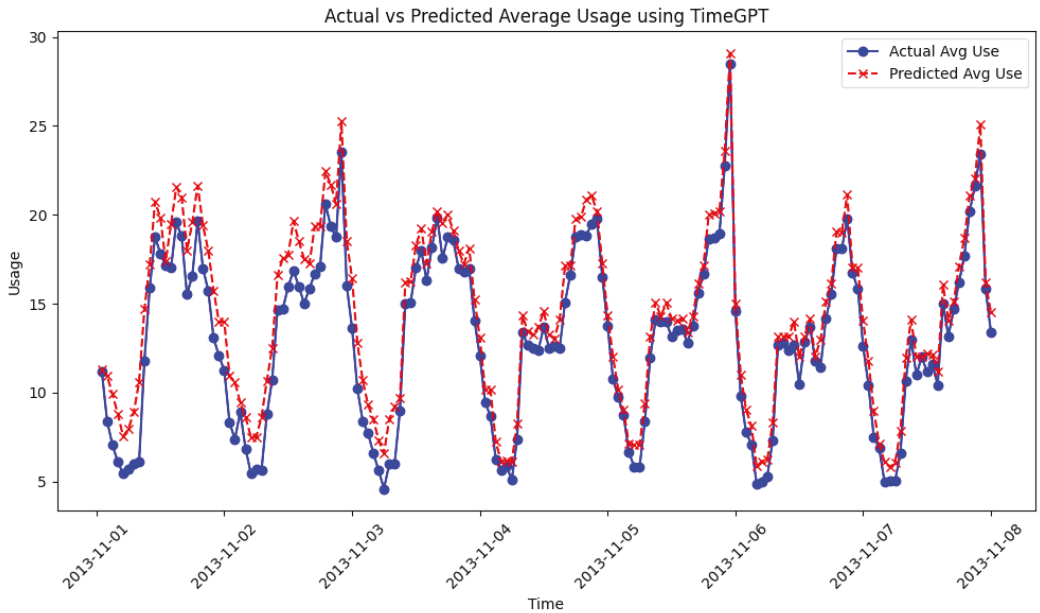



**Figure 9.** TimeGPT-based forecast of mobile network usage over one week, showing high alignment between actual and predicted CDR activity.

The forecasting process involved several key steps to prepare the data and generate predictions. First, the CDR dataset was preprocessed by hourly resampling, handling missing values and normalizing the average application metric to stabilize the learning process. A sliding window approach was then applied to transform the sequential data into a supervised learning framework, where each 24 h input window was used to forecast the next 24 h period. The transformed data was then input into TimeGPT using the Chronos library, which applies a transformer-based temporal architecture to model complex dependencies and periodic patterns. Finally, the results were postprocessed to reverse the normalization and align them with the scale of the original data.

Overall, Figure 9 shows strong agreement between actual and forecasted values throughout the forecast week. Both curves show highly synchronized patterns, with forecasted values closely tracking the peaks and troughs observed during actual usage. This reflects the model's ability to learn and generalize temporal behavior, capturing the cyclical nature of mobile network usage patterns, such as increased activity during the day and reduced usage in the late evening.

These results highlight the robustness of TimeGPT in modeling dense, high-frequency time series data. The close correspondence between actual and predicted usage indicates

that the model effectively internalizes usage dynamics without requiring extensive manual manipulation. While some peak values could benefit from further refinement through external regressors or anomaly correction techniques, the basic forecasting capability remains statistically reliable and operationally efficient.

The evaluation measurements of the TimeGPT model, as shown in Table 1, demonstrate its high accuracy and effectiveness in forecasting mobile network traffic. The root mean square error (RMSE) is 14.8229, indicating minimal deviation between actual and forecasted values. The mean absolute error (MAE) is 7.7788, which further reflects the model's strong predictive accuracy. Additionally, the $R^2$ value of 98.62% reflects the excellent model fit that captures nearly all the variability in the data. Overall, these metrics confirm that TimeGPT provides accurate and reliable forecasts, making it a robust tool for mobile network traffic forecasting.

**Table 1.** Evaluation metrics for the TimeGPT model.

| Evaluation Metrix | TIMEGPT |
|---|---|
| RMSE | 14.8229 |
| MAE | 7.7788 |
| $R^2$ | 98.62% |

4.4.2. Traffic Forecasting with AutoML

Figure 10 illustrates the forecasting results of CDR activities using the AutoML framework. This prediction covers one week of average mobile network usage. The blue line with circular markers represents the actual observed average usage, while the dashed red line with crosses denotes the predicted values generated by the AutoML pipeline. This visual comparison facilitates a straightforward assessment of how well the automated model aligns with real-world usage trends.

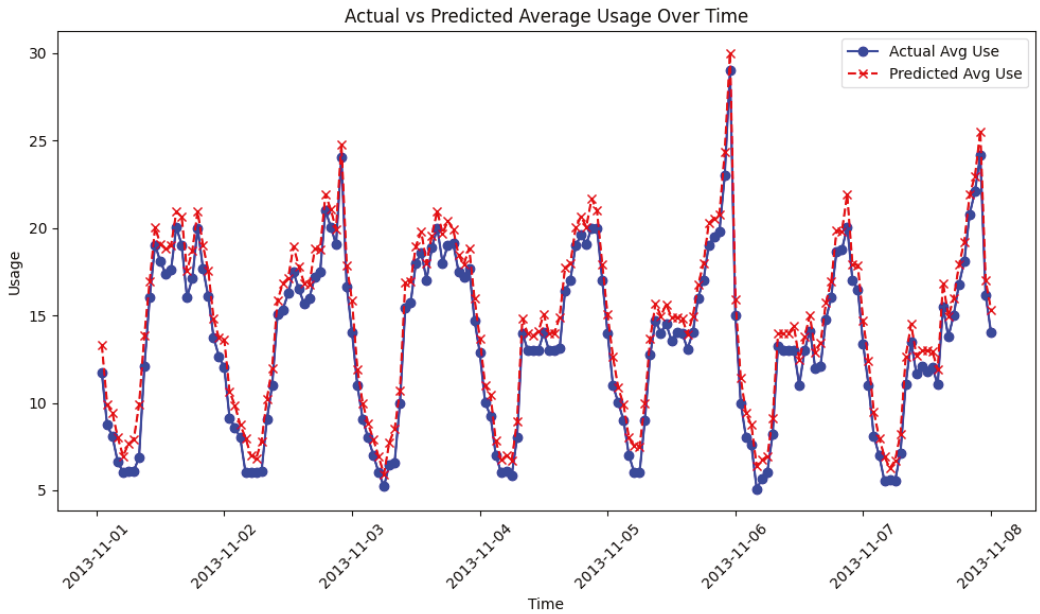



**Figure 10.** AutoML-based forecast of mobile network usage over one week, highlighting strong predictive accuracy and pattern consistency.

The AutoML forecasting process followed a structured workflow to ensure optimal model selection and tuning. The raw CDR dataset was first aggregated hourly and then

cleaned to correct for missing values. The H2O AutoML framework then automated feature engineering steps, including the creation of lag variables, temporal features (such as time of day and day of the week) and trend indicators. The data was split into training and validation sets using a sliding-window cross-validation method to preserve temporal structure. H2O AutoML explored different models, including gradient boosting, deep learning and stacked ensembles. The best performing model was selected based on the root mean square error (RMSE) and used to generate the forecasts. The final predictions were scaled back to their initial values for comparison with actual usage.

As shown in Figure 10, the AutoML model shows a strong correlation between actual average usage and predicted average usage throughout the week. The predicted values effectively track the observed fluctuations, capturing daily peaks and troughs with remarkable accuracy. The alignment of the curves demonstrates AutoML's ability to automatically extract relevant temporal features and learn usage patterns without manual intervention.

These results reinforce the usefulness of AutoML frameworks for large-scale time-series prediction, particularly in telecom scenarios where data is abundant and rapidly evolving. Despite slight deviations during peak usage, the model displays consistent performance across most time series. This underscores the potential of AutoML as a scalable solution for network operators seeking to anticipate demand and efficiently allocate resources.

The AutoML model evaluation metrics, summarized in Table 2, reveal outstanding forecasting performance. The RMSE score of 2.4990 indicates a small average deviation between the actual and forecasted values. The MAE of 1.0284 reflects an even smaller average error size. In addition, the $R^2$ value of 99.96% shows an almost perfect model fit, demonstrating its ability to capture virtually all variability in the dataset. These evaluation metrics confirm the strength and reliability of the AutoML model in forecasting mobile network traffic utilization.

**Table 2.** Performance metrics for the AutoML model.

| **Evaluation Metrix** | **AUTOML** |
|---|---|
| RMSE | 2.4990 |
| MAE | 1.0284 |
| $R^2$ | 99.96% |

An additional analysis revealed that the TimeGPT and AutoML models tend to slightly overestimate periods of high and low usage. This behavior could indicate that they are overfitting to the dominant patterns in the training data, leading to increased sensitivity to extreme values. Furthermore, potential strategies to improve the accuracy of forecasts in extreme cases are discussed that include the development of hybrid modeling approaches. These improvements aim to enhance the robustness of the models and provide promising approaches for future research. This recognition not only increases the transparency of the model evaluation process but also indicates practical ways to improve the overall performance of forecasts.

#### 4.4.3. Hyperparameter Configuration and Tuning Strategy

To ensure optimal performance and a fair comparison, both TimeGPT and AutoML models were configured with clearly defined strategies tailored to their respective architectures. While TimeGPT relies on a pretrained transformer network via the Chronos framework, AutoML leverages H2O's automated pipeline to find the best model and parameters with minimal manual intervention.

Table 3 presents a consolidated overview of the main hyperparameter configurations for both models. TimeGPT was configured with an input window and a forecast horizon of 24 h. A batch size of 32 and a maximum of 50 training periods were used, with early stopping after five consecutive periods without improvement. The learning rate was kept at the default value of 0.001 and transfer learning was performed with pretrained weights to minimize the need for manual tuning. Chronos handles most of the optimization internally, eliminating the need for manual hyperparameter tuning.

**Table 3.** Comparison of TimeGPT (Chronos) and AutoML (H2O framework).

| Parameter | TimeGPT (Chronos) | AutoML (H2O Framework) |
|---|---|---|
| Input Window | 24 h | 24 h |
| Forecast Horizon | 24 h | 24 h |
| Learning Rate | 0.001 (default) | Random grid search |
| Batch Size | 32 | Auto Selected per Model |
| Epochs | 50 | Determined per Model |
| Pretrained Weights | Used | Not Applicable |
| Cross-Validation | Not applied | Five-Fold Time-Series-Aware CV |
| Model Candidates | Single | Max 20 models |
| Hyperparameter Tuning | Skipped | Random Grid Search |
| Ensemble Methods | Not used | Stacked Ensembles |
| Feature Engineering | Positional Encoding | Automated |

A more dynamic and fully automated workflow was introduced for AutoML. The H2O framework autonomously performed feature engineering, model selection and hyperparameter tuning using random grid search strategies. Each candidate model was given a maximum run time of 120 s and the system evaluated up to 20 models using five fold time series cross-validation. Different algorithms were explored, including Gradient Boosting Machines, XGBoost, Random Forests, deep learning and stacked ensembles, with automatic selection of the best performing model based on RMSE.

As summarized in Table 3, the two models differ significantly in their handling of feature tuning and engineering. TimeGPT's pipeline emphasizes the efficiency of the pretrained architecture, while AutoML optimizes predictive performance through extensive automated exploration.

## 5. Discussion and Comparison

TimeGPT and AutoML offer distinct capabilities for mobile network traffic forecasting, each excelling in different areas. TimeGPT is specifically designed for time series data, making it highly effective at capturing and forecasting complex patterns in large, dynamic datasets. However, in terms of absolute accuracy, TimeGPT exhibits slightly higher error values compared to AutoML.

Figure 11 illustrates a comparative analysis of the average 24 h usage profiles for two predictive models: AutoML and TimeGPT. The x-axis represents the time in hourly intervals since November 1st with the time change, while the y-axis shows the average usage values. The AutoML model is represented by a solid blue line and circle markers, while the TimeGPT model is represented by a dashed orange line and x markers. Both models show a similar usage pattern throughout the day, with a decrease in usage in the early morning

(around 5:00 to 6:00 AM), followed by a steady increase from 8:00 AM, reaching maximum values between 1:00 PM and 4:00 PM and gradually decreasing around midnight. Notably, TimeGPT shows slightly higher average usage during peak hours, suggesting better responsiveness or adoption during high demand. The close alignment of the two curves during most of the day indicates that both models perform similarly in tracking and predicting usage patterns, with TimeGPT showing marginal advantages during peak hours. This comparison underscores the reliability and effectiveness of both models in time-sensitive applications while also highlighting TimeGPT's potential advantage in handling more intense usage scenarios.

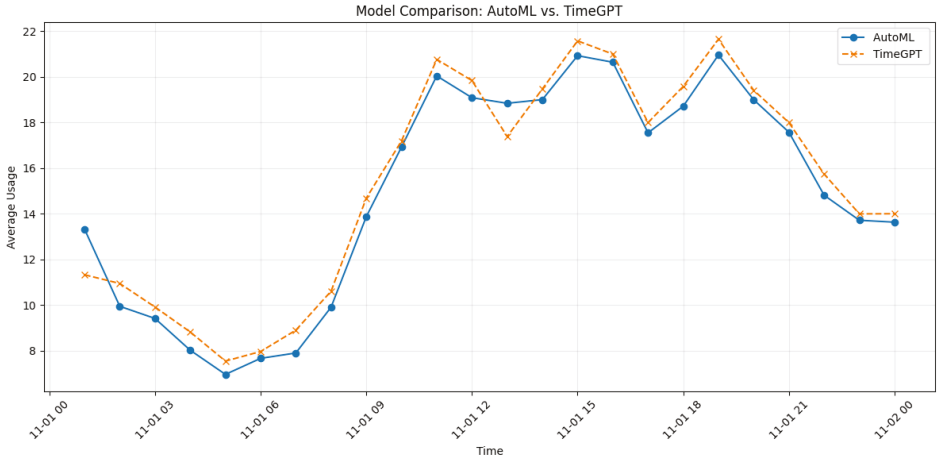



**Figure 11.** Model comparison AutoML vs. TimeGPT (1-day record).

Figure 12 illustrates a time series comparison of the average utilization predictions of the AutoML and TimeGPT models from 1 to 8 November 2013. The x-axis represents the dates, while the y-axis shows the average utilization values. Both models display similar temporal patterns, capturing daily cycles and fluctuations in utilization. It is notable that they respond effectively to rapid changes, such as the sharp spike observed on November 6. TimeGPT sometimes exhibits slightly higher or earlier spikes than AutoML, indicating slightly stronger sensitivity to changes. Despite these slight differences, the overall trends produced by the two models are closely aligned. This similarity suggests that both models are reliable for forecasting or modeling utilization over time. TimeGPT's subtle advantage in responsiveness might make it slightly more suitable for scenarios involving rapid or frequent variations.

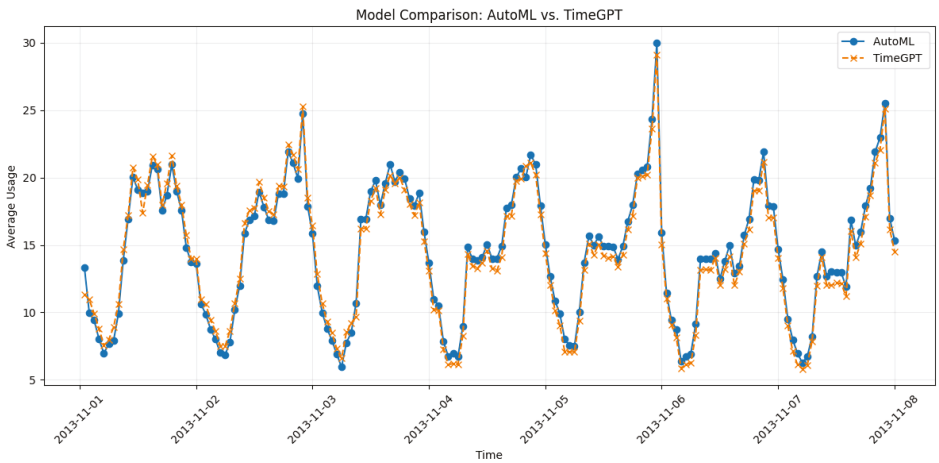



**Figure 12.** Model comparison AutoML vs. TimeGPT (weekly record).

We compared TimeGPT and AutoML to traditional statistical models such as ARIMA, SARIMA and LSTM, and TimeGPT and AutoML significantly outperformed them. Figure 13, titled "Mobile network usage: actual forecasts and model forecasts (day-1 records)," illustrates a comparative analysis of several forecasting models using actual mobile network usage data on November 1. The actual usage trend is represented by a solid black line, which serves as a benchmark to evaluate the accuracy of five predictive models: ARIMA (blue dotted line), SARIMA (green dotted line), LSTM (red dotted line), AutoML (blue dotted line) and TimeGPT (orange dotted line). Over 24 h, all models attempt to capture the daytime usage pattern with varying degrees of accuracy. Notably, TimeGPT and AutoML closely follow the actual utilization curve, effectively capturing morning troughs and evening peaks, suggesting high adaptability to short-term fluctuations. In contrast, ARIMA and SARIMA exhibit more pronounced deviations, especially during transition hours, indicating limitations in handling nonlinear and seasonal trends. The LSTM model exhibits smoother transitions but tends to underestimate expectations during peak hours, likely due to its reliance on historical smoothing. This comparison highlights TimeGPT and AutoML as the most reliable models for real-time mobile network demand forecasting, demonstrating their potential usefulness in network management and resource allocation strategies.

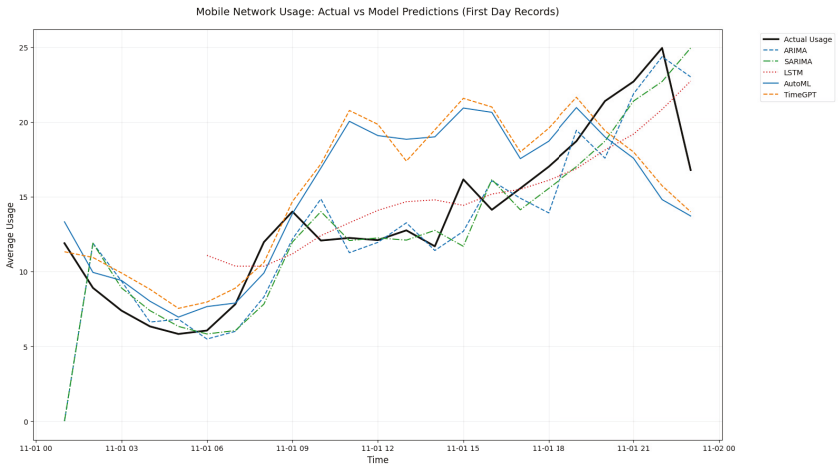



**Figure 13.** Model comparison: Actual vs. AutoML vs. TimeGPT vs. ARIMA vs. SARIMA vs. LSTM (1-day Record).

Figure 14, "Mobile network usage: models comparison (actual vs. TimeGPT vs. AutoML vs. ARIMA vs. SARIMA vs. LSTM)", provides a detailed comparative analysis of six forecasting models in predicting average mobile network usage over the period from 1 to 8 November 2013. The actual observed data is represented by a bold black line, serving as the benchmark against which the predictive accuracy of the models: ARIMA (blue dashed), SARIMA (orange dashdot), LSTM (green dotted), AutoML (red solid) and TimeGPT (purple dashed) are assessed. All models generally succeed in capturing the underlying daily patterns and periodic surges in network usage, reflecting a shared ability to model cyclical behavior. However, TimeGPT and AutoML exhibit the closest alignment with the actual data, especially during peak demand periods and abrupt fluctuations, suggesting superior sensitivity and forecasting precision. In contrast, ARIMA and SARIMA tend to underperform during rapid changes, often over or underestimating peak values. The LSTM model provides smoother forecasts but occasionally lags in responsiveness, potentially missing finer temporal variations. Notably, the consistent proximity of TimeGPT and AutoML to the actual usage line underscores their robustness and adaptability in complex, time-variant environments. This comparative visualization highlights TimeGPT

and AutoML as the most effective models among those evaluated, making them strong candidates for real-time network demand forecasting applications.

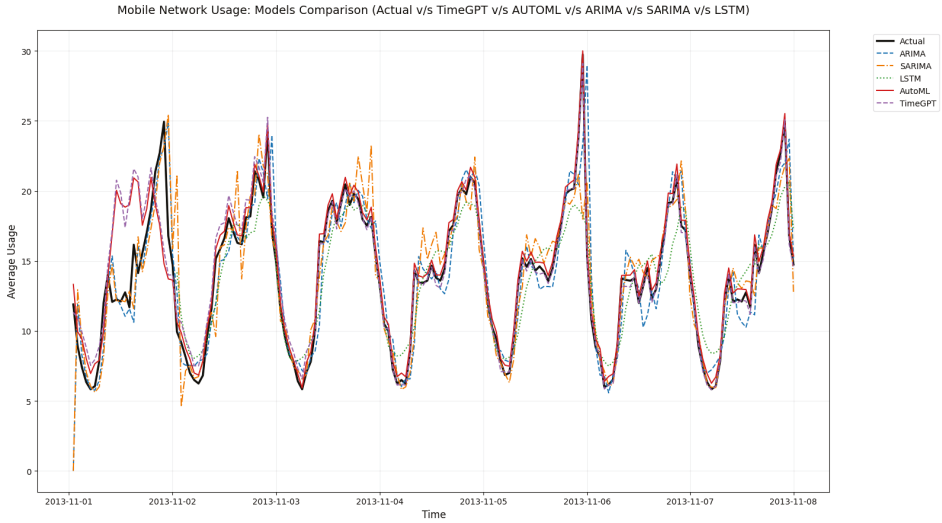



**Figure 14.** Model comparison: actual vs. AutoML vs. TimeGPT vs. ARIMA vs. SARIMA vs. LSTM (1-week Record).

As shown in Table 4, the RMSE for TimeGPT is 14.8226 and the MAE is 7.7789, while AutoML performs significantly better with an RMSE of 2.4990 and an MAE of 1.0284. This suggests that AutoML produces more accurate forecasts with lower deviations from actual values. Additionally, the $R^2$ value, which measures how well the model explains the variance in the data, shows a crucial difference: AutoML achieves an impressive $R^2$ of 99.96%, indicating near-perfect predictive power, while TimeGPT scores 98.62%, which is strong but slightly lower, as shown in Table 4. This reinforces AutoML's superior forecasting accuracy.

**Table 4.** Comparison of forecasting performance metrics across various models.

| Evaluation Metrix | RMSE | MAE | $R^2$ |
|---|---|---|---|
| AutoML | 2.4990 | 1.0284 | 99.96% |
| TimeGPT | 14.8226 | 7.7789 | 98.62% |
| ARIMA | 10.8205 | 9.1472 | 83.80% |
| SARIMA | 15.1946 | 13.1610 | 60.76% |
| LSTM | 9.6017 | 9.6017 | 79.3% |

This study sets a new benchmark in mobile traffic forecasting by demonstrating that the proposed models AutoML and TimeGPT significantly outperform traditional forecasting approaches such as ARIMA, SARIMA and LSTM. As detailed in Table 4, conventional statistical models exhibit notable limitations: ARIMA yields a moderate $R^2$ of 83.80% with RMSE and MAE scores of 10.8205 and 9.1472, respectively; SARIMA performs even more poorly with an $R^2$ of just 60.76%, with RMSE and MAE scores of 10.8205 and 9.1472, respectively, highlighting its inability to capture the complex spatiotemporal dynamics of mobile network data. Although LSTM shows slight improvement with an $R^2$ of 79.3%, its results lack consistency and robustness.

By contrast, the integration of TimeGPT and AutoML represents a novel methodological leap in forecasting accuracy and model efficiency. TimeGPT, with its transformer-based

architecture, is purpose-built for time series forecasting and excels at learning intricate temporal dependencies, seasonalities and irregular fluctuations. AutoML, on the other hand, delivers a fully automated modeling pipeline, including feature engineering, algorithm selection and hyperparameter optimization, enabling high-performance forecasting with minimal human intervention.

The empirical results clearly favor AutoML as the best-performing model, with a near-perfect $R^2$ of 99.96%, the lowest RMSE 2.4990 and the lowest MAE 1.0284. TimeGPT follows closely with a solid $R^2$ of 98.62%, demonstrating its robustness in capturing temporal trends. The novelty of this work lies not only in the systematic application of these advanced tools to mobile traffic prediction but also in setting a new standard through rigorous comparative evaluation. Moreover, the study proposes a promising hybrid strategy that combines the temporal modeling advantages of TimeGPT with the automation of AutoML to provide a scalable, accurate and generalizable prediction framework, laying the foundation for future research in intelligent network planning.

In conclusion, both TimeGPT and AutoML offer unique advantages for time series forecasting: AutoML excels in its high overall accuracy, while TimeGPT is particularly effective at capturing complex temporal patterns. A hybrid approach leveraging both could result in a robust forecasting framework, combining predictive accuracy with nuanced temporal understanding.

## 6. Conclusions

This study proposed a data-driven mobile traffic analysis and forecasting framework using Telecom Italia's Call Detail Records (CDRs). By dividing Milan into a detailed $100 \times 100$ grid and categorizing traffic into four intensity levels, we captured the spatial and temporal dynamics of network demand across the city. The analysis revealed that high traffic concentrations are mainly located in central urban areas, highlighting the influence of population density and mobility patterns. To evaluate the prediction performance, AutoML and TimeGPT were used to predict traffic trends. While TimeGPT effectively captured complex temporal dependencies, AutoML consistently achieved higher accuracy and demonstrated greater adaptability. Its automated pipeline, including feature selection and model tuning, made it a more practical solution for real-world telecom applications. These results highlight the considerable potential of AutoML for intelligent and scalable network management, particularly in the context of smart city infrastructures. Future research could benefit from incorporating additional contextual data (such as weather conditions, special events, or travel patterns) to improve forecast accuracy. Testing the proposed models in multiple cities and over longer time periods would assess their robustness and generalizability. Furthermore, integrating privacy preserving techniques, such as federated learning, could enable broader collaborative use without compromising user data. A hybrid approach combining the temporal modeling strengths of TimeGPT with AutoML optimization could also offer a promising direction. Finally, evaluating the model's performance in various urban environments will be essential to validate its adaptability and ensure its practical deployment in various real-world scenarios.

**Author Contributions:** Conceptualization, K.S. and M.S.; Methodology, H.A. and M.S.; Software, H.A.; Writing—original draft, H.A. and K.S.; Writing—review & editing, K.S., M.S. and T.C.C.; Supervision, K.S., Fund acquisition, T.C.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by Multimedia University under the Research Fellow Grant MMUI/250008, and in part by Telekom Research & Development Sdn Bhd under Grant RDTC/241149.

**Data Availability Statement:** The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/marcodena/mobile-phone-activity (accessed on 5 July 2025).

**Conflicts of Interest:** The authors declare that they have no conficts of interest.

## References

1. Yew, H.T.; Dasuki, A.; Asmat, A.E.; Japarudin, M.S.; Herifian, M.N.H.; Mamat, M. Internet-of-Things: Role of 5G Wireless Communications. In Proceedings of the 2023 12th International Conference on Awareness Science and Technology (iCAST), Taichung, Taiwan, 9–11 November 2023; pp. 335–339.
2. Ghahramani, M.; Zhou, M.; Wang, G. Urban sensing based on mobile phone data: Approaches, applications, and challenges. *IEEE/CAA J. Autom. Sin.* **2020**, *7*, 627–637. [CrossRef]
3. Rad, M.M.; Fouli, K.; Fathallah, H.A.; Rusch, L.A.; Maier, M. Passive optical network monitoring: Challenges and requirements. *IEEE Commun. Mag.* **2011**, *49*, s45–S52. [CrossRef]
4. Kocaman, H. Forecasting User Behaviors in Call Detail Records Using Lstm Models. Master's Thesis, Middle East Technical University, Ankara, Turkey, 2021.
5. Barrett, J.; Legg, P.; Smith, J.; Barnes, T.; Boyle, C. Trends and Challenges in Data Analytics and Machine Learning Using Call Detail Records in Telecom Systems. 2024. Available online: https://www.preprints.org/frontend/manuscript/b0030ae119f582bcb7315acb5a2b37c1/download_pub (accessed on 5 July 2025).
6. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 11106–11115. [CrossRef]
7. Nikravesh, A.Y.; Ajila, S.A.; Lung, C.H.; Ding, W. Mobile network traffic prediction using MLP, MLPWD, and SVM. In Proceedings of the 2016 IEEE international congress on big data (BigData Congress), San Francisco, CA, USA, 27 June–2 July 2016; pp. 402–409.
8. Perez-Guirao, M.D.; Weisshaupt, T.; Wilzeck, A. DECT NR+: Unveiling the Essentials of a new non-cellular 5G Standard for Verticals. In Proceedings of the Mobile Communication-Technologies and Applications; 26th ITG-Symposium, Osnabrueck, Germany, 18–19 May 2022; pp. 1–6.
9. Selvamanju, E.; Shalini, V.B. Machine learning based mobile data traffic prediction in 5G cellular networks. In Proceedings of the 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2–4 December 2021; pp. 1318–1324.
10. Clemente, D.; Soares, G.; Fernandes, D.; Cortesao, R.; Sebastiao, P.; Ferreira, L. S. Traffic forecast in mobile networks: Classification system using machine learning. In Proceeding of the 2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall), Honolulu, HI, USA, 22–25 September 2019.
11. Rago, A.; Piro, G.; Boggia, G.; Dini, P. Multi-task learning at the mobile edge: An effective way to combine traffic classification and prediction. *IEEE Trans. Veh. Technol.* **2020**, *69*, 10362–10374. [CrossRef]
12. Liu, Z.; Zhu, Z.; Gao, J.; Xu, C. Forecast methods for time series data: A survey. *IEEE Access* **2021**, *9*, 91896–91912. [CrossRef]
13. Lim, B.; Zohren, S. Time-series forecasting with deep learning: A survey. *Philos. Trans. R. Soc. A* **2021**, *379*, 20200209. [CrossRef] [PubMed]
14. Yu, B.; Yin, H.; Zhu, Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv* **2017**, arXiv:1709.04875.
15. Gupta, A.; Gupta, H.P.; Biswas, B.; Dutta, T. Approaches and applications of early classification of time series: A review. *IEEE Trans. Artif. Intell.* **2020**, *1*, 47–61. [CrossRef]
16. Corradini, F.; Gori, M.; Lucheroni, C.; Piangerelli, M.; Zannotti, M. A Systematic Literature Review of Spatio-Temporal Graph Neural Network Models for Time Series Forecasting and Classification. *arXiv* **2024**, arXiv:2410.22377.
17. Sultan, K.; Ali, H.; Zhang, Z. Call detail records driven anomaly detection and traffic prediction in mobile cellular networks. *IEEE Access* **2018**, *6*, 41728–41737. [CrossRef]
18. Tan, M.; Merrill, M.; Gupta, V.; Althoff, T.; Hartvigsen, T. Are language models actually useful for time series forecasting? *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 60162–60191.

**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.

computers

MDPI

Article

# EMGP-Net: A Hybrid Deep Learning Architecture for Breast Cancer Gene Expression Prediction

**Oumeima Thâalbi and Moulay A. Akhloufi ***

Perception, Robotics, and Intelligent Machines (PRIME), Department of Computer Science, Université de Moncton, Moncton, NB E1A 3E9, Canada; eot1123@umoncton.ca
* Correspondence: moulay.akhloufi@umoncton.ca

**Abstract:** *Background*: The accurate prediction of gene expression is essential in breast cancer research. However, spatial transcriptomics technologies are usually too expensive. Recent studies have used whole-slide images combined with spatial transcriptomics data to predict breast cancer gene expression. To this end, we present EMGP-Net, a novel hybrid deep learning architecture developed by combining two state-of-the-art models, MambaVision and EfficientFormer. *Method*: EMGP-Net was first trained on the HER2+ dataset, containing data from eight patients using a leave-one-patient-out approach. To ensure generalizability, we conducted external validation and alternately trained EMGP-Net on the HER2+ dataset and tested it on the STNet dataset, containing data from 23 patients, and vice versa. We evaluated EMGP-Net's ability to predict the expression of 250 selected genes. EMGP-Net mixes features from both models, and uses attention mechanisms followed by fully connected layers. *Results*: Our model outperformed both EfficientFormer and MambaVision, which were trained separately on the HER2+ dataset, achieving the highest PCC of 0.7903 for the PTMA gene, with the top 14 genes having PCCs greater than 0.7, including other important breast cancer biomarkers such as GNAS and B2M. The external validation showed that it also outperformed models that were retrained with our approach. *Conclusions*: The results of EMGP-Net were better than those of existing models, showing that the combination of advanced models is an effective strategy to improve performance in this task.

**Keywords:** breast cancer; gene expression; whole-slide images; spatial transcriptomics; deep learning

## 1. Introduction

Breast cancer remains one of the most common and potentially most deadly forms of cancer, with hundreds of thousands of deaths and millions of women diagnosed each year, as of 2018 [1]. Despite important advances in early detection, the molecular complexity and heterogeneity of breast cancer pose serious challenges to the accurate diagnosis of the disease. Histopathology, particularly the use of hematoxylin and eosin (H&E)-stained tissue slides, has been one of the main bases for diagnosing breast cancer. These images provide detailed visual information about the structure of the tissue, making it easier for pathologists to determine the type, grade, and stage of cancer. Histopathological images were initially employed to determine cancer types [2], and several studies have utilized datasets with histopathological samples, such as the BreakHis dataset, the CIAR 2018 dataset, the BreCaHAD dataset, and

others. These studies have contributed to classifying tumors into different subtypes. For example, Al-Jabbar et al. [3], Obayya et al. [4], and Clement et al. [5] developed various approaches to detect benign and malignant classes. Other studies progressed further by identifying additional subclasses of breast cancer, such as those proposed by Bagchi et al. [6], Bhausaheb and Kashyap [7], and Yu et al. [8]. However, these methods primarily focused on morphology and did not capture the molecular details that could better inform treatment decisions, such as gene expression levels, which are critical to understanding tumor behavior. ST technologies have recently emerged as strong tools that enable researchers to profile gene expression while preserving the spatial context of the tissue and allow for high-resolution mapping of gene expression across tissue sections, providing valuable insights into the molecular architecture of tumors. However, the high cost of these technologies limits their large-scale use, particularly in clinical settings. This challenge has led to the exploration of alternative approaches, such as predicting gene expression directly from histopathological images, especially whole-slide images (WSIs). In the literature, past studies have shown that WSIs contain biologically rich information that can be leveraged to predict gene expression. Many methods have successfully used deep learning models to predict gene expression patterns from WSIs. ST-Net [9], which uses DenseNet-121 to predict 250 genes, was one of the first transfer learning-based approaches to make marked progress in breast cancer gene expression prediction. Afterwards, BrST-Net [10] adapted CNN models, such as EfficientNet-b0, and incorporated an auxiliary network to predict the expression of 250 genes. GeNetFormer [11], our previous approach, integrated several transformer models, including EfficientFormer, to predict 250 genes. HisToGene [12], on the other hand, used attention mechanisms and ViT to predict the expression of 785 genes. Other models such as SEPAL [13], Hist2ST [14], and THItoGene [15] used graph neural networks (GNNs) and showed interesting results. SEPAL [13], Hist2ST [14], and THItoGene [15] not only capture spatial information from individual spots, but also use neighboring spots to gain a global perspective. Similarly, PH2ST [16] and HGGEP [17] incorporate hypergraphs, allowing both local and global integration of spatial context. In the studies reviewed, methods such as ST-Net [9], BrST-Net [10], GeNetFormer [11], EGN [18], and EGGN [19] were used to focus primarily on local predictions, whereas HisToGene [12] involves an adapted global prediction approach. However, methods such as SEPAL [13] and THItoGene [15] combine both local and global predictions. Although transformers were initially designed for natural language processing, their ability to model global dependencies has made them highly valuable in computer vision [20]. CNNs are good at detecting local patterns, but they have trouble with long-range interactions, while transformers, on the other hand, use attention mechanisms to capture relationships across the whole image [20]. This makes them effective at analyzing complex spatial structures in histopathological slides, where fine-grained detail and large-scale tissue architecture are both important for gene expression prediction. Recently, state-space models (SSMs) have emerged as a valuable alternative to attention mechanisms, offering competitive modeling capabilities at a lower computational cost. MambaVision [21], built on this principle, combines the efficiency of SSMs with the global awareness of attention. It delivered a strong performance on large image datasets while being much faster and more memory-efficient than traditional transformers. These properties make MambaVision well suited for high-resolution tasks, such as predicting gene expression from histopathological images. The goal of this study is to explore the application of advanced hybrid architectures combining MambaVision and EfficientFormer to predict breast cancer gene expression from WSIs. While convolutional neural networks (CNNs) have long been the architecture of choice for medical image analysis, they tend to excel at capturing local features, and struggle to model long-range spatial relationships and global patterns. Building on advances in deep learning-based gene expression prediction from histopathological

images, and to address the limitations of older state-of-the-art (SOTA) approaches, we propose a novel approach that combines two SOTA architectures, EfficientFormer and MambaVision. MambaVision, one of the most recent models in the field, has demonstrated important potential in visual recognition tasks. By leveraging MambaVision's advanced feature extraction capabilities, our approach aims to capture the complicated patterns in histopathological images that can contribute to more accurate gene expression prediction. EfficientFormer, a transformer model known for its efficiency and performance in gene expression prediction tasks, can complement MambaVision by providing strong spatial dependency processing capabilities. Our hybrid model, called EMGP-Net, integrates these two powerful architectures to improve gene expression prediction from WSIs. This combination of both models provides a robust approach. For this work, we used two widely used datasets: the HER2+ dataset and the STNet dataset. The HER2+ dataset consists of 36 sections from 8 patients, while the STNet dataset contains 68 sections from 23 patients. Both datasets have associated ST data. To ensure the generalizability of our model, we employed two validation strategies. First, we used the HER2+ dataset for internal validation. Then, for external validation, we evaluated the model using two approaches: one test performed on the HER2+ dataset and another on the STNet dataset. In both cases, the model was trained on the remaining dataset. All tests aimed to predict the expression of 250 genes, as in other methods [9–11]. This innovative hybrid architecture motivated us to further explore the potential of combining these two SOTA models for gene expression prediction. Our contributions are as follows:

- **Proposing EMGP-Net:** We propose a hybrid model combining MambaVision and EfficientFormer to predict gene expression more effectively from WSIs.
- **Performing exhaustive validation:** We perform internal and external validation on the HER2+ and STNet datasets to ensure model robustness and generalizability.
- **Demonstrating benefits of hybrid deep learning:** We demonstrate the benefits of combining the latest powerful SOTA models and evaluate them on medical tasks, contributing to advances in breast cancer research, particularly in gene expression prediction for diagnosis.

## 2. Related Work

### *2.1. CNN-Based Approaches*

Recently, several studies have focused on gene expression prediction from WSIs with associated ST data to mitigate the high cost of ST technologies by using the histopathological images that are widely available. For instance, ST-Net [9] was the first approach that combined ST data and WSIs of breast cancer tissues to predict gene expression, and a CNN-based method was developed using DenseNet-121. It uses H&E tissue images from the STNet dataset, first introduced in [9], and employs cross-validation. The model achieved the highest median PCC of 0.3400 for the GNAS gene. ST-Net was trained to predict 250 genes and was tested on the external 10x Genomics dataset, as well as on the Cancer Genome Atlas (TCGA) dataset. BrST-Net [10], also a CNN-based approach, has an integrated primary network and auxiliary network (AuxNet). It was evaluated using 10 SOTA models, and the best performing was EfficientNet-b0 with AuxNet. This model was trained and tested on the STNet dataset and achieved the highest PCC of 0.6325 for the B2M gene. The TRIPLEX approach [22], designed to predict the expression of 250 genes, consists of three parts: a target encoder, a global encoder, and a neighbor encoder. The target encoder processes the specific region of interest (target spot) in the tissue using ResNet18 and outputs predictions through a predictor. The global encoder, which uses multiple layers of transformer blocks and an atypical position encoding generator (APEG),

captures a broader context by analyzing the entire tissue, while the neighbor encoder focuses on the regions around the target spot, also using ResNet18, with a final fusion layer integrating attention modules. The model was validated on three internal datasets, STNet, HER2+, and cSCC, and externally validated on the Visium dataset of breast cancer patients. On the HER2+ dataset, the model achieved a mean PCC of 0.3140 for all genes and 0.4970 for highly predictive genes. On the STNet dataset, it achieved a mean PCC of 0.3520 for all genes and 0.2060 for highly predictive genes.

### *2.2. Transformer-Based Approaches*

Other methods apply transformer-based architectures, such as the GeNetFormer framework [11], which evaluated eight advanced transformer models, including EfficientFormer, FasterViT, BEiT v2, Swin Transformer v2, PyramidViT v2, MobileViT v2, MobileViT, and EfficientViT. These models were applied to predict 250 genes using the STNet dataset. The framework was trained using different image resolutions (224 $\times$ 224, 256 $\times$ 256) and loss functions (MSELoss, SL1Loss). The approach integrates these models into a comprehensive pipeline. The highest PCC was obtained with the configuration with the MSELoss function, 224 $\times$ 224 resolution, and EfficientFormer, which gave 9 out of the top 10 genes with the highest PCC values. HisToGene [12] focuses on modeling the spatial dependencies of gene expression using multi-head attention and a modified ViT architecture that allows the model to handle heterogeneity. In terms of results, HisToGene's predictions on the HER2+ dataset achieved the highest mean R of 0.3200 for the GNAS gene, and it was also evaluated on the human cutaneous squamous cell carcinoma (cSCC) dataset.

### *2.3. Hybrid Transformer and GNN Approaches*

Zeng et al. [14] proposed Hist2ST, based on transformer and GNN architectures, which starts the process by dividing the images around each spot into patches. These patches are then processed by three components. The convmixer extracts internal visual features within the image patches, while the transformer captures global spatial dependencies between spots. The GNN is then used to capture the local relationships between neighboring spots. The model was trained on the HER2+ and cSCC datasets using leave-one-out cross-validation. Hist2ST achieved an average PCC of 0.3900 for the top gene FN1. SEPAL [13] predicts gene expression by analyzing tissue images in two stages: local learning and spatial learning. First, it processes each image patch to extract visual features and predict 256 genes. In the second step, a graph is constructed for each patch and its neighbors, allowing the model to learn from both the local features and the spatial context. The GNN is used to refine the initial predictions by incorporating spatial relationships between patches. It was trained on the STNet dataset and the 10x Genomics dataset. The THItoGene [15] architecture consists of three parts: First, the image segmentation and position embedding step divides the image into patches corresponding to the location of each spot, embedding the spatial coordinates and aggregating the data for processing. Second, feature extraction is performed using a dynamic convolution module that captures deep molecular features from the patches, with the efficient capsule network improving the model by using self-attention mechanisms to adjust the convolution kernels based on the spatial context. Finally, global modeling is performed by the ViT module, which integrates the position embeddings and image features. The graph attention network (GAT) module is then used to learn the relationships between adjacent spots. The HER2+ dataset, with 785 selected genes, along with the cSCC dataset, was used for analysis, and leave-one-out cross-validation was applied. On the HER2+ dataset, PCC values of 0.7470, 0.7110, 0.6720, and 0.4520 were achieved for the genes FN1, SCD, IGKC, and FASN, respectively.

### *2.4. Graph-Based and Relational Modeling Approaches*

Approaches such as ErwaNet [23] consist of two modules: the edge relational module (ERM) and the window attention module (WAM). The ERM captures local information by constructing a heterogeneous graph where each window in the slide image is treated as a node, and the relationships between windows are represented by three different types of edges: K-nearest neighbor (KN) edges, percent similarity (PS) edges, and K-nearest similarity (KS) edges. The WAM, on the other hand, captures global information by using an attention mechanism that aggregates feature representations from the entire tissue slide. The model was validated on two datasets, the STNet dataset and the 10x Genomics dataset, using cross-validation. A total of 250 targeted genes were selected for prediction. ErwaNet achieved PCC@F, PCC@S, and PCC@M values of 2.3600, 3.5200, and 3.3300 on the STNet dataset and 8.2800, 8.6900, and 8.3600 on the 10x Genomics dataset, respectively. Some other approaches are based on hypergraphs: PH2ST [16] is a prompt-based framework designed to handle multi-scale histological features by integrating dual-scale hypergraphs and ViT for both global and local features. This method uses a set of known ST values to guide over unmeasured spots. For feature extraction, PH2ST uses UNI, a universal histology image encoder pre-trained on a large corpus of WSIs. It captures both local and neighboring spatial context through dual-scale hypergraph-based spot representations, where each spot is connected to neighboring regions via hypergraph convolution. A cross-attention mechanism is then used to refine these representations for the final prediction. The model was applied to the HER2+ dataset containing 785 genes and also to the cSCC datasets. The highest PCC values were obtained for the genes TMSB10, CISD3, CD74, and COL6A2 from the HER2+ dataset: 0.5095, 0.2911, 0.5603, and 0.4091, respectively. The HGGEP architecture [17] consists of a gradient enhancement module (GEM) that enhances the gradients and captures cell morphological information, a ShuffleNet V2 backbone that extracts latent features from histology images, and a convolutional block attention module (CBAM) and ViT to refine the extracted features. The model includes a hypergraph association module (HAM) that captures spatial relationships between different regions in the tissue and uses long short-term memory (LSTM) to model dependencies between features. The HER2+ dataset with 785 genes and cSCC datasets were used. HGGEP achieved PCC values of 0.637 for GNAS, 0.564 for FASN, 0.652 for MR12B, and 0.649 for SCD.

### *2.5. Exemplar-Guided Approaches*

Other approaches employ exemplar guidance learning, such as EGN [18] and its enhanced version EGGN [19]. EGN uses an exemplar bridging (EB) block and ViT as a backbone. The framework first retrieves the nearest exemplars for each tissue image window and constructs a graph to model spatial relations. It then updates window features using information from the exemplars to predict genes. EGGN constructs visual similarity graphs with the exemplars, which are then processed by a GraphSAGE-based backbone. Both models are based on the idea that images with similar visual features will have similar gene expression patterns, regardless of their location within the tissue.

## 3. Materials and Methods

### *3.1. Dataset*

Two datasets were used in this work:

- HER2+ is the HER2 (human epidermal growth factor receptor)-positive breast cancer dataset that was investigated in [24]. It was collected from eight patients (A-H). A

total of 36 sections comprise the samples in this dataset. Three or, alternatively, six replicates (sections from the same patient) were assigned to each patient and stained with H&E. Each sample is in JPG format and comes with associated ST data. The dataset may represent various tissue types, including invasive cancer, breast glands, immune infiltrate, cancer in situ, connective tissue, and adipose tissue.

- STNet is the fifth edition of the human breast cancer in situ capturing transcriptomics dataset, referred to as the STNet dataset, as was presented in [9]. It was obtained from 23 patients. It contains a total of 68 sections, with 3 sections per patient (except for 2 sections for 1 patient). The images are also stained with H&E and each sample is in JPG format and has corresponding ST data. The subtypes represented in the STNet dataset are luminal A, luminal B, triple negative, HER2 luminal, and non-luminal HER2.

Both datasets include files with the spot coordinates, the count matrices, and the gene names or symbols. Some examples of the datasets are shown in Figure 1.

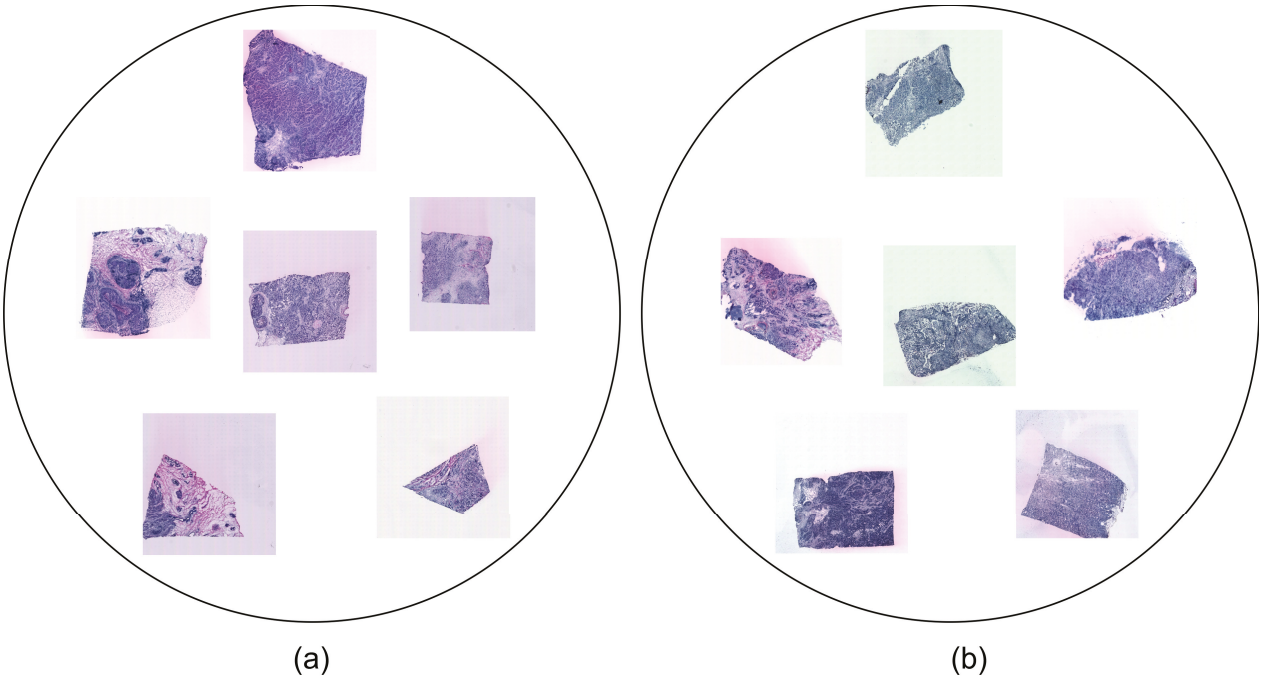



**Figure 1.** Representative whole-slide images from the HER2+ and STNet datasets: (**a**) Sample WSIs from the HER2+ dataset. (**b**) Sample WSIs from the STNet dataset.

### *3.2. Data Pre-Processing and Augmentation*

Before our data were used by the model, we prepared them by applying various pre-processing techniques previously used in our earlier work [11] to both datasets. First, we filtered out spots with a total count of less than 1000. This is the same approach used in [10,11]. This resulted in a final number of spots of about 28,792 for the STNet dataset and 11,666 for the HER2+ dataset. We then normalized the gene expression count, and we added 1 as pseudo-count to avoid zero issues. We applied a log1p transformation ($\log(1 + x)$) to the normalized counts to stabilize variance, reduce skewness, and ensure coherence between studies. Patches were generated according to the coordinates of each spot. In addition, to adjust the input dimension to fit with the model, the extracted patches were fixed to $224 \times 224 \times 3$. For the STNet dataset, the list of genes with Ensembl identifiers (IDs) was converted to gene symbols using the HUGO Gene Nomenclature Committee (HGNC) database to ensure the use of unique gene symbols, thus simplifying electronic

data retrieval and minimizing ambiguity. Classical data augmentation was also used to diversify the training data to help generalize and avoid overfitting. We applied random horizontal flipping, random vertical flipping, and random 90-degree rotation for each image patch. These techniques are important for learning from features independently of the direction or the rotation of the images. During testing, we computed the average of the eight symmetries obtained from the rotations and reflection.

### *3.3. Proposed Approach*

#### 3.3.1. Overview of the EMGP-Net Architecture

In this study, we introduce EMGP-Net, a hybrid deep learning architecture named after its components, EfficientFormer (E) and MambaVision (M), and its purpose, gene expression prediction (GP). The aim is to improve gene expression prediction from WSIs by using recent and robust SOTA models, and to improve the feature representations by exploiting the capabilities of both models. We followed three steps: the preparation and preprocessing of the data, the training of the model, and, finally, the evaluation. We extracted patches from the WSIs, centering on spots according to the positions provided for them in the corresponding ST data. This was performed to standardize the inputs before they were fed into the model for training. Initially, EMGP-Net uses MambaVision, an SOTA model known for its significant capacity in image tasks, especially when compared with other SOTA models [21]. In parallel, it integrates EfficientFormer, an efficient vision transformer that has shown good performance in gene expression prediction tasks, and achieved the best scores among seven evaluated transformer models implemented in the GeNetFormer framework that was introduced in [11]. We first adapted both models to produce 1024-dimensional feature vectors by replacing their original classification heads with custom linear layers. After extracting the features from both branches, they were stacked and passed on a multi-head attention mechanism. The attention layer takes these stacked vectors as input and learns how to weigh them based on their importance. This allows the model to focus on the most important features of each branch and to summarize the information in an appropriate way. Unlike simple concatenation, which increases dimensionality, or basic averaging without using attention, this approach allows the attention mechanism to influence the final combination, improving the fusion of information by capturing relevant relationships between the outputs of MambaVision and EfficientFormer. The fused features are then passed through a layer normalization (LN) layer. This helps to stabilize the training and improve the generalizability of the model. Then, we applied the Gaussian error linear unit (GeLU) activation function, which introduces non-linearity and improves learning capacity compared to the traditional ReLU. Finally, the normalized and activated features are passed through two fully connected linear layers. The first layer reduces the dimensionality from 1024 to 512, and the second layer gives predictions for the 250 targeted genes. The final output is a 250-dimensional vector representing the predicted gene expression values. To guarantee the generalizability of our model, the training was conducted with the leave-one-patient-out approach using the HER2+ dataset for the first evaluation. Then, we applied the first external validation using the STNet dataset with the model trained on the whole HER2+ dataset and the second external validation using the HER2+ dataset with the model trained on the whole STNet dataset. More details are mentioned below, and the EMGP-Net architecture is shown in Figure 2.

#### 3.3.2. MambaVision

MambaVision: Introduced in [21], it is a hybrid architecture comprising Mamba [25] and transformers designed for improving feature learning. It presents a new SOTA ap-

proach trained on the ImageNet dataset. MambaVision combines Mamba blocks with transformer and self-attention layers. The network is divided into four stages. Two stages are based on CNN layers which integrate the GeLU activation and the batch normalization (BN) as follows:

$$\begin{aligned} \hat{z} &= \mathrm{GeLU}(\mathrm{BN}(\mathrm{Conv}_{3\times3}(z))), \\ z &= \mathrm{BN}(\mathrm{Conv}_{3\times3}(\hat{z})) + z. \end{aligned} \tag{1}$$

The other two stages come with MambaVision and transformer blocks. This novel approach showed its capacity in capturing the global context and long-range spatial dependencies and handling high-resolution images with faster training and inference than standard transformers. MambaVision provided a better accuracy–throughput balance compared to other SOTA models such as FasterViT and Swin Transformer v2 and the two models that were evaluated on gene expression prediction in [11]. Also, it outperformed other SOTA models such as NextViT, VMamba, FastViT, and Vim using the ImageNet dataset. It achieved higher accuracy than transformer and Mamba models. Figure 3 shows the architecture of the MambaVision block.

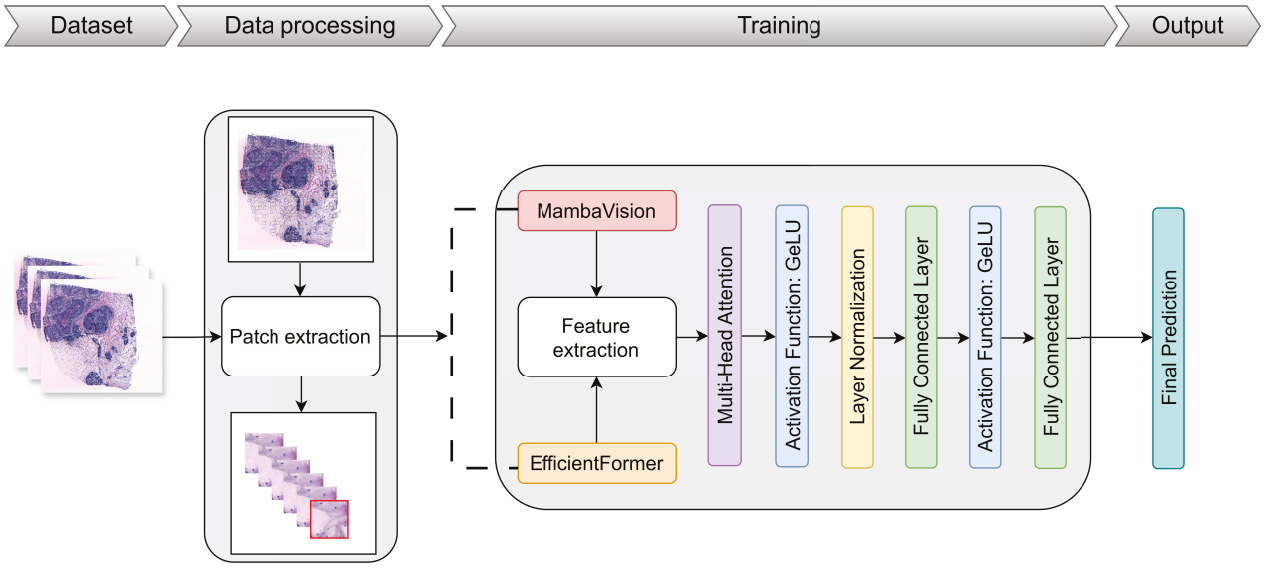



**Figure 2.** Overview of the EMGP-Net architecture. The workflow comprises multiple stages: extracting patches from WSIs (the red frame shows an extracted patch and its location in the original image), extracting features using MambaVision and EfficientFormer backbones, and utilizing multi-head attention followed by activation functions (GeLU), layer normalization, and fully connected layers to output the final gene expression predictions.

### 3.3.3. EfficientFormer

EfficientFormer: Presented in [26], it is efficient for low-latency applications and is completely based on a pure transformer architecture. Like many other SOTA models, it was trained on the ImageNet dataset. Within each of the four network stages, a number of meta transformer blocks (MB) are implemented to avoid the use of MobileNet components. The following equation shows the relationship between the two components, the token mixer (TokenMixer) and the multi-layer perceptron (MLP), that construct the meta transformer block:

$$\mathcal{X}_{i+1} = \mathrm{MB}_i(\mathcal{X}_i) = \mathrm{MLP}(\mathrm{TokenMixer}(\mathcal{X}_i)). \tag{2}$$

EfficientFormer achieved the best results compared to other SOTA models [26]. It outperformed several ViT-based architectures, such as DeiT-Small, LeViT-256, and PoolFormer-S24, in terms of accuracy and latency using the ImageNet dataset. Furthermore, GeNet-

Former, a framework for gene expression prediction task, showed the best scores when implementing the EfficientFormer model, and it outperformed FasterViT, BEiT v2, Swin Transformer v2, PyramidViT v2, MobileViT v2, MobileViT, and EfficientViT, seven models evaluated in the GeNetFormer framework. Figure 4 shows the architecture of the meta transformer block.

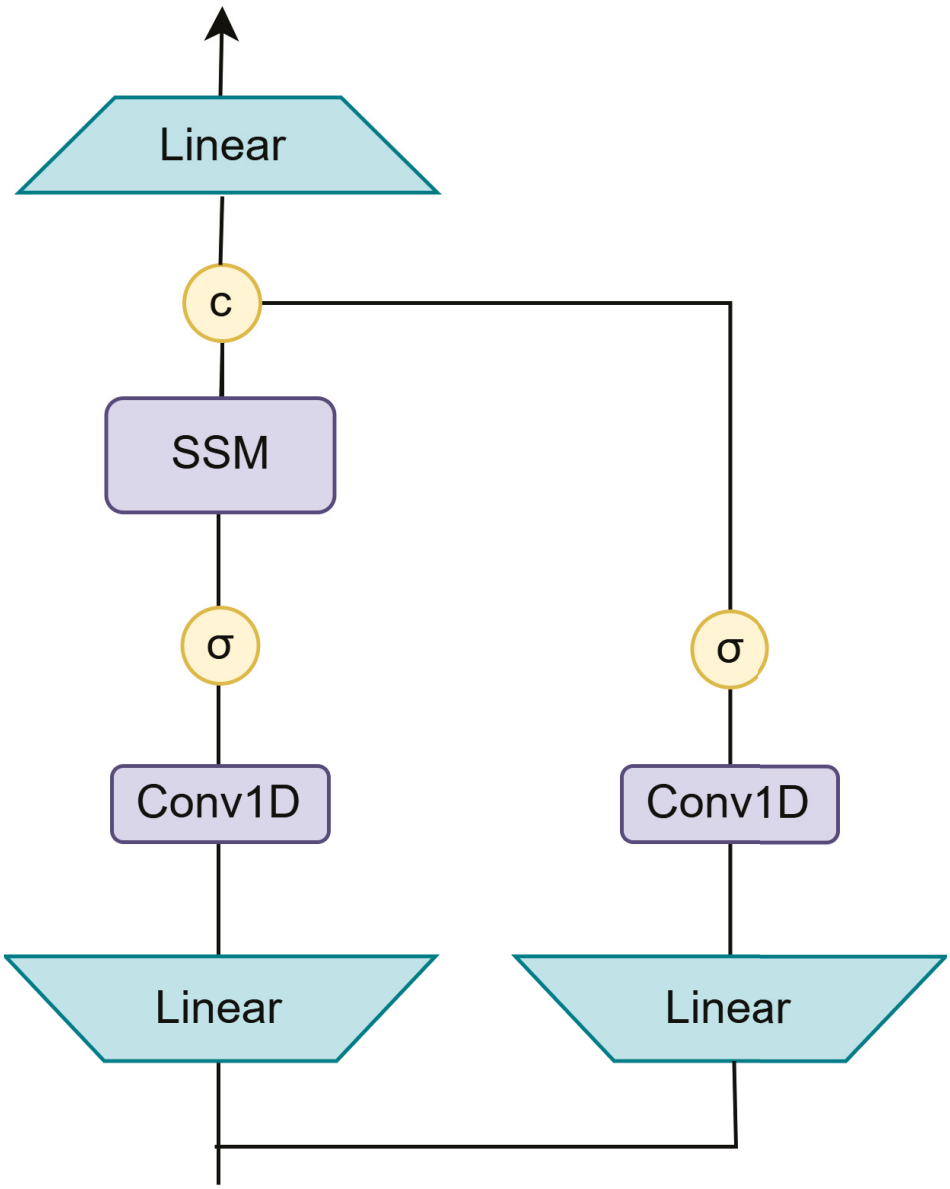



**Figure 3.** Architecture of the MambaVision block. The block integrates two parallel branches. One branch contains a linear layer and applies a selective SSM and a 1D convolutional layer. The other branch also includes a linear layer and applies only a 1D convolution. The outputs are then combined and passed through a last linear layer.

#### 3.3.4. Multi-Head Attention Mechanism

Multi-head attention: shown in Figure 5, in multi-head attention, instead of using one attention function with the same size for queries, keys, and values, several attention functions in parallel are used. Each function works with smaller versions of the queries, keys, and values. After each function processes its version, the results are combined, allowing the model to look at the input from different perspectives at the same time. In the

transformer model, there are eight such attention functions, called heads, each looking at the input in different ways [27].

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(head_1, \ldots, head_h)W^O \\ \text{where} \quad head_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{3}$$

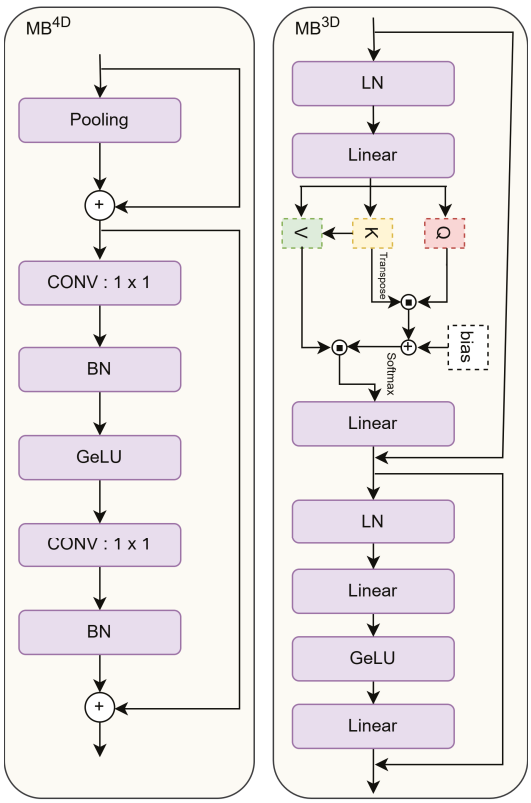



**Figure 4.** Architecture of the meta transformer block, including both $MB^{4D}$ and $MB^{3D}$ variants. The $MB^{4D}$ variant uses pooling followed by two convolutional layers with $1 \times 1$ kernels, batch normalization, and a GeLU activation function. In contrast, the $MB^{3D}$ variant includes a sequence of linear layers alternating with layer normalization and GeLU. In this variant, the input is projected into a query (Q), key (K), and value (V) before final integration through linear transformations.

The implementation of our proposed approach was based on two SOTA models, MambaVision and EfficientFormer. We used the pre-trained versions on the ImageNet dataset. We added a set of layers to improve the feature learning mentioned in the previous part.

### *3.4. Evaluation Metrics*

To evaluate the performance of the proposed EMGP-Net architecture, we used three metrics commonly applied in regression tasks. These metrics are described as follows:

- *MAE* (Mean Absolute Error): Measures the average of the absolute differences between the observed and predicted values with the following equation:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{4}$$

- *RMSE* (Root Mean Squared Error): Measures the square root of the average of the squared differences between the observed and predicted values with the following equation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

- *PCC* (Pearson Correlation Coefficient): A measure of the strong correlation between the observed and predicted values. It is defined by the following equation:

$$PCC = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{6}$$

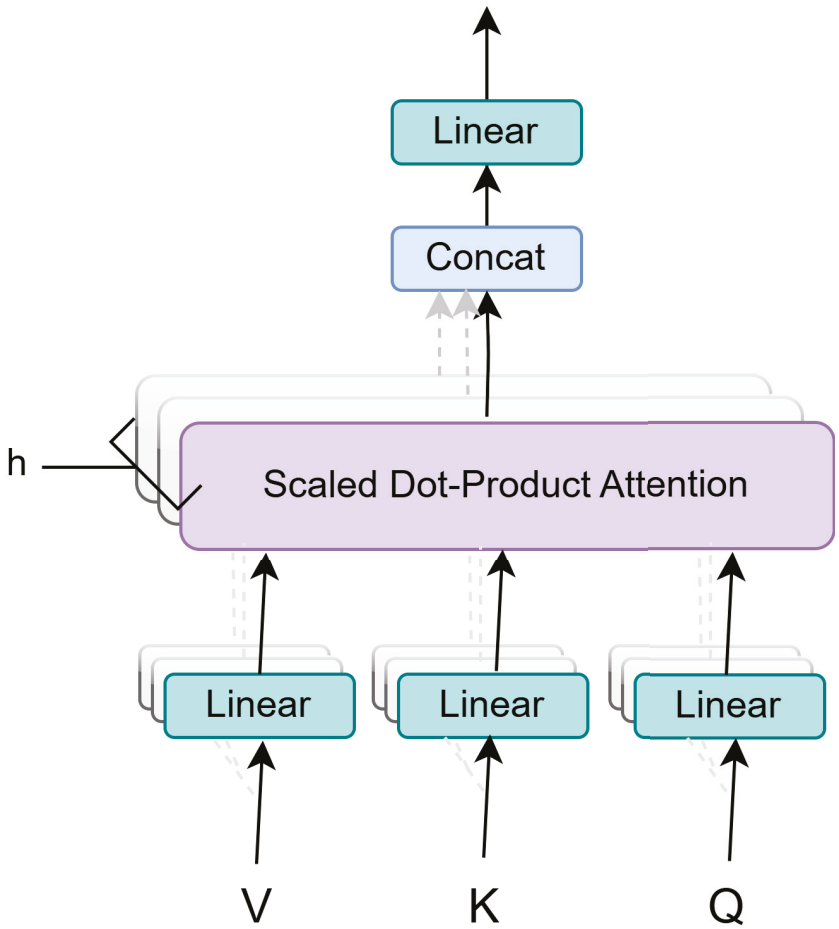



**Figure 5.** Architecture of the multi-head attention mechanism. The figure illustrates the multi-head attention structure. First, the inputs are passed through linear layers, and then the outputs are processed in parallel by several attention heads (shown as h). Then, the outputs from all heads are combined and passed through another linear layer.

## 4. Experimental Results

The EMGP-Net architecture proposed in this research was designed to predict the expression of 250 genes, which were initially set and selected based on their highest mean expression in the dataset. (The lists of the 250 genes from the HER2+ and STNet datasets can be found in Appendix A). In this study, we focus on presenting the top 14 predicted genes among the 250. First, we evaluated our model using a leave-one-patient-out approach with the HER2+ dataset. To assess the model's ability to generalize, we performed external validation in two ways: In the first test, the model was trained on the entire HER2+ dataset and tested on the entire STNet dataset. In the second test, the model was trained on the

entire STNet dataset and tested on the entire HER2+ dataset. Each evaluation was repeated for all 8 patients in the HER2+ dataset and 23 patients in the STNet dataset and then we selected the highest PCC value across all patients for each gene. Image dimensions were fixed at 224 × 224 pixels as input to the model, and we applied the MSELoss function along with the MAE and RMSE metrics to refine the results. The equation for the MSELoss is as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{7}$$

The training was performed on eight NVIDIA GeForce RTX 3090 GPUs with 24 GB of memory each (https://www.nvidia.com/, accessed on 20 May 2025). We selected a batch size of 768, and the entire script was implemented using the PyTorch library. The learning rate and weight decay were set to $10^{-4}$. In the following section, we present the results of our EMGP-Net model.

We applied the Wilcoxon signed-rank test to evaluate whether the observed performance improvements of EMGP-Net over the different models were reliable. This non-parametric test is well suited to our study because it handles paired data. The Wilcoxon signed-rank test statistic W is computed as follows:

$$W = \sum_{i=1}^{n} R_i \cdot \text{sign}(d_i), \tag{8}$$

where $d_i = x_i - y_i$ is the difference in PCC between EMGP-Net and the other model for gene $i$, $R_i$ is the rank of $|d_i|$ among all non-zero differences, and sign($d_i$) indicates the direction of the difference.

The $p$-values resulting from this analysis are reported in the last rows of Tables 1–3. Values below the common threshold of 0.05 were considered statistically meaningful, indicating that EMGP-Net achieved a better gene prediction score for most of the genes tested.

**Table 1.** Overview of performance comparison between EMGP-Net, EMGP-Net-noAttn, EfficientFormer, and MambaVision. The highest PCC values across the different models for the top 14 predicted genes are in bold and underlined. The $p$-values at the bottom show the statistical differences between each model and EMGP-Net. Values < 0.05 indicate significance.

| Genes | EfficientFormer | MambaVision | EMGP-Net-noAttn | EMGP-Net |
|---|---|---|---|---|
| Gene 1 | 0.7777 (PTMA) | **0.8049 (B2M)** | 0.7791 (PTMA) | 0.7903 (PTMA) |
| Gene 2 | 0.7746 (B2M) | 0.7763 (GNAS) | 0.7768 (B2M) | **0.7843 (GNAS)** |
| Gene 3 | 0.7661 (GNAS) | 0.7674 (PTMA) | 0.7700 (GNAS) | **0.7777 (B2M)** |
| Gene 4 | 0.7266 (HNRNPA2B1) | 0.7363 (TPT1) | 0.7356 (TPT1) | **0.7532 (HNRNPA2B1)** |
| Gene 5 | 0.7245 (TPT1) | 0.7198 (HNRNPA2B1) | 0.7331 (HNRNPA2B1) | **0.7360 (TPT1)** |
| Gene 6 | 0.7075 (ACTG1) | 0.7089 (HLA-DRA) | 0.7271 (ACTG1) | **0.7339 (XBP1)** |
| Gene 7 | 0.6965 (XBP1) | 0.7042 (ACTG1) | 0.7237 (XBP1) | **0.7318 (ACTG1)** |
| Gene 8 | 0.6964 (HLA-DRA) | 0.7032 (HLA-B) | 0.7005 (HLA-B) | **0.7228 (HLA-B)** |
| Gene 9 | 0.6938 (CD24) | 0.7010 (XBP1) | 0.6959 (HLA-DRA) | **0.7122 (TMSB10)** |
| Gene 10 | 0.6929 (HLA-B) | 0.6921 (COX6C) | 0.6951 (ACTB) | **0.7085 (DDX5)** |
| Gene 11 | 0.6868 (TMSB10) | 0.6832 (VMP1) | 0.6873 (TMSB10) | **0.7056 (HLA-DRA)** |
| Gene 12 | 0.6859 (DDX5) | 0.6809 (ACTB) | 0.6826 (TUBA1B) | **0.7020 (ACTB)** |
| Gene 13 | 0.6839 (ACTB) | 0.6789 (PSMB4) | 0.6826 (COX6C) | **0.7016 (S100A11)** |
| Gene 14 | 0.6834 (S100A11) | 0.6780 (NACA) | 0.6799 (DDX5) | **0.7002 (CD24)** |
| ***p*-value** | 0.0001 (<0.05) | 0.0009 (<0.05) | 0.0001 (<0.05) | N/A |

**Table 2.** Overview of EMGP-Net performance vs. ST-Net and GeNetFormer performance. The highest PCC values across the different models for the top 14 predicted genes are in bold and underlined. The *p*-values at the bottom show the statistical differences between each model and EMGP-Net. Values < 0.05 indicate significance.

| Genes | ST-Net | GeNetFormer | EMGP-Net |
|---|---|---|---|
| Gene 1 | 0.6708 (GNAS) | 0.7069 (DDX5) | **0.7145 (ERBB2)** |
| Gene 2 | 0.6592 (RPL23) | 0.6510 (ACTG1) | **0.7051 (ACTG1)** |
| Gene 3 | 0.6503 (PTPRF) | 0.6384 (CPB1) | **0.7047 (CALR)** |
| Gene 4 | 0.6460 (ACTG1) | 0.6235 (PTMA) | **0.6973 (RPL23)** |
| Gene 5 | 0.6406 (DDX5) | 0.6130 (RPL23) | **0.6962 (GNAS)** |
| Gene 6 | 0.6325 (PRDX1) | 0.5974 (PTPRF) | **0.6894 (PSMD3)** |
| Gene 7 | 0.6274 (TAGLN2) | 0.5943 (GNAS) | **0.6867 (PTPRF)** |
| Gene 8 | 0.6273 (CALR) | 0.5864 (CALR) | **0.6842 (TMSB10)** |
| Gene 9 | 0.6235 (HSPB1) | 0.5840 (HSPB1) | **0.6835 (GAPDH)** |
| Gene 10 | 0.6201 (PTMA) | 0.5701 (TMSB10) | **0.6814 (TAGLN2)** |
| Gene 11 | 0.6144 (CPB1) | 0.5638 (TAGLN2) | **0.6724 (DDX5)** |
| Gene 12 | 0.6027 (NHEERF1) | 0.5344 (P4HB) | **0.6645 (HSPB1)** |
| Gene 13 | 0.5908 (ENSG00000203812) | 0.5307 (PRDX1) | **0.6588 (PTMA)** |
| Gene 14 | 0.5749 (HLA-DRA) | 0.5250 (KRT19) | **0.6563 (KRT19)** |
| ***p*-value** | 0.0001 (<0.05) | 0.0001 (<0.05) | N/A |

**Table 3.** Overview of EMGP-Net performance vs. ST-Net and GeNetFormer performance. The highest PCC values across the different models for the top 14 predicted genes are in bold and underlined. The *p*-values at the bottom show the statistical differences between each model and EMGP-Net. Values < 0.05 indicate significance.

| Genes | ST-Net | GeNetFormer | EMGP-Net |
|---|---|---|---|
| Gene 1 | 0.6719 (ATP5E) | 0.6746 (ATP5E) | **0.7285 (ERBB2)** |
| Gene 2 | 0.6620 (ERBB2) | 0.6434 (S100A11) | **0.6686 (S100A11)** |
| Gene 3 | 0.6374 (S100A11) | 0.6141 (ERBB2) | **0.6650 (ATP5E)** |
| Gene 4 | 0.6227 (PTPRF) | 0.6115 (PTPRF) | **0.6404 (HSP90B1)** |
| Gene 5 | 0.5918 (LGALS3) | 0.5986 (HSP90B1) | **0.6347 (LGALS3)** |
| Gene 6 | 0.5903 (HSP90B1) | 0.5967 (CST3) | **0.6262 (CD24)** |
| Gene 7 | 0.5880 (CST3) | 0.5572 (ACTG1) | **0.6049 (PTPRF)** |
| Gene 8 | 0.5812 (KRT19) | 0.5449 (MYH9) | **0.5927 (FN1)** |
| Gene 9 | 0.5750 (PSMB4) | 0.5400 (PSMB4) | **0.5905 (PTMA)** |
| Gene 10 | 0.5662 (GNAS) | 0.5393 (LGALS3) | **0.5832 (FTH1)** |
| Gene 11 | 0.5353 (EEF2) | 0.5384 (KRT19) | **0.5763 (PSMB4)** |
| Gene 12 | 0.5317 (ACTG1) | 0.5310 (CD24) | **0.5726 (ACTB)** |
| Gene 13 | 0.5301 (IGLC2) | 0.5293 (FTH1) | **0.5642 (MYH9)** |
| Gene 14 | 0.5287 (LASP1) | 0.5185 (COL1A2) | **0.5465 (GNAS)** |
| ***p*-value** | 0.0001 (<0.05) | 0.0001 (<0.05) | N/A |

### *4.1. Model Trained on the HER2+ Dataset*

First, we present the results obtained using the leave-one-patient-out approach with the HER2+ dataset. Table 1 shows the top 14 genes with the highest PCC. We compared four models: the EfficientFormer, MambaVision, and two versions of our proposed model, EMGP-Net, which includes an attention mechanism, and EMGP-Net-noAttn, which does not. EMGP-Net and EMGP-Net-noAttn are both based on a combination of EfficientFormer and MambaVision, along with the set of added layers mentioned previously. EMGP-Net gave better results in most cases.

Comparing EMGP-Net-noAttn with EMGP-Net demonstrated the impact of the attention mechanism. Although they have the same architecture, EMGP-Net uses attention to weigh and combine features, while EMGP-Net-noAttn simply concatenates them and

applies fully connected layers. EMGP-Net outperformed EMGP-Net-noAttn on nearly all genes. This confirmed that the attention-based fusion enhances the model's ability to focus on more representations, thus improving prediction.

4.1.1. Comparison of Architectural Components by PCC for Top-Ranked Genes:

EMGP-Net was able to predict 13 out of 14 genes with PCC scores greater than 0.7, which is higher than the PCC scores of the MambaVision and EfficientFormer models separately, as well as that of EMGP-Net-noAttn. The top 14 genes predicted by EMGP-Net included PTMA, GNAS, B2M, HNRNPA2B1, and TPT1, with PCC values of 0.7903, 0.7843, 0.7777, 0.7532, and 0.7360, respectively. On the other hand, MambaVision predicted only 1 gene, the B2M gene, out of the 14 genes with the highest PCC score, equal to 0.8049, and scored 0.7763, 0.7674, 0.7363, and 0.7198 for GNAS, PTMA, TPT1, and HNRNPA2B1, respectively. For EfficientFormer, all PCCs were lower than those of EMGP-Net, with values of 0.7777, 0.7746, 0.7661, 0.7266, and 0.7245 for PTMA, B2M, GNAS, HNRNPA2B1, and TPT1, respectively. EMGP-Net outperformed EfficientFormer on 14 genes, MambaVision on 13 genes out of 14 genes, and EMGP-Net-noAttn on all 14 genes. These results demonstrate that combining features from both backbone models and refining them with the attention mechanism greatly improved the PCC scores.

4.1.2. Comparison of Architectural Components by PCC for Common Genes

The common genes among the selected top 14 genes evaluated in all three models included PTMA, GNAS, B2M, HNRNPA2B1, TPT1, XBP1, ACTG1, HLA-B, HLA-DRA, and ACTB. As shown in Table 4, the results indicated that EMGP-Net outperformed EMGP-Net-noAttn, MambaVision, and EfficientFormer in predicting gene expression for the majority of common genes. Specifically, EMGP-Net achieved higher PCC values than MambaVision for 7 out of the 10 common genes. The genes where MambaVision outperformed EMGP-Net were B2M, TPT1, and HLA-DRA, with MambaVision achieving higher PCC scores of 0.8049, 0.7363, and 0.7089, respectively, compared to the 0.7777, 0.7360, and 0.7056 of EMGP-Net. However, when compared to EMGP-Net-noAttn and EfficientFormer, EMGP-Net showed better performance across all genes. Figure 6 shows the visualization of the predictions of six genes with the highest PCCs.

**Table 4.** Overview of performance comparison between EMGP-Net, EMGP-Net-noAttn, EfficientFormer, and MambaVision. The highest PCC values for the common genes among the top 14 genes predicted by the different models are in bold and underlined.

| Gene | EfficientFormer | MambaVision | EMGP-Net-noAttn | EMGP-Net |
|---|---|---|---|---|
| PTMA | 0.7777 | 0.7674 | 0.7791 | **0.7903** |
| GNAS | 0.7661 | 0.7763 | 0.7700 | **0.7843** |
| B2M | 0.7746 | **0.8049** | 0.7768 | 0.7777 |
| HNRNPA2B1 | 0.7266 | 0.7198 | 0.7331 | **0.7532** |
| TPT1 | 0.7245 | **0.7363** | 0.7356 | 0.7360 |
| XBP1 | 0.6965 | 0.7010 | 0.7237 | **0.7339** |
| ACTG1 | 0.7075 | 0.7042 | 0.7271 | **0.7318** |
| HLA-B | 0.6929 | 0.7032 | 0.7005 | **0.7228** |
| HLA-DRA | 0.6964 | **0.7089** | 0.6959 | 0.7056 |
| ACTB | 0.6839 | 0.6809 | 0.6951 | **0.7020** |

### *4.2. Quantitative Analysis of the Results*

Our method aims to predict the expression of 250 breast cancer-related genes, selected based on their highest mean expression in the dataset.

#### 4.2.1. Analysis of EMGP-Net Results

To evaluate the effectiveness of our model, we categorized the gene expression values into ranges. The results showed that there were 14 genes with PCC values between 0.7 and 0.8, while 56 genes had PCC values between 0.6 and 0.7. Additionally, 75 genes were in the range of 0.5 to 0.6, 54 genes had PCC values between 0.4 and 0.5, and 32 genes were in the range of 0.3 to 0.4. Furthermore, 15 genes had PCC values between 0.2 and 0.3, and 4 genes had values between 0.1 and 0.2. Notably, all genes were predicted with positive PCC values, indicating that our model successfully predicted gene expression for all selected genes.

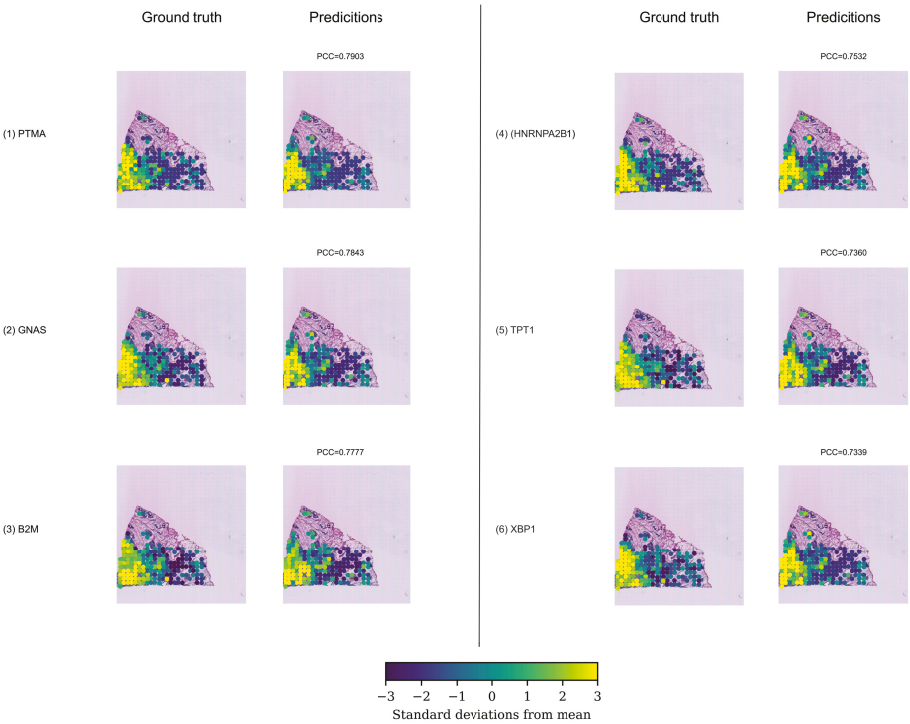



**Figure 6.** Visualization of the top 6 genes predicted by EMGP-Net using the HER2+ dataset with a leave-one-patient-out approach. Each pair of images shows the ground truth on the left and the corresponding prediction on the right for one gene. The color indicates gene expression levels as standard deviations from the mean. The corresponding PCC values are shown for each gene.

#### 4.2.2. Analysis of MambaVision Results

The MambaVision results show the following distribution: 1 gene with a PCC value higher than 0.8, 8 genes with a score between 0.7 and 0.8, and 39 genes with PCC values between 0.6 and 0.7. In addition, 65 genes were in the range of 0.5 to 0.6, 73 genes had PCC values between 0.4 and 0.5, and 39 genes were in the range of 0.3 to 0.4. In addition, 20 genes had PCC values between 0.2 and 0.3, 4 genes had PCC values between 0.1 and 0.2, and 1 gene had a PCC value between 0.0 and 0.1.

#### 4.2.3. Analysis of EfficientFormer Results

The EfficientFormer PCC values were distributed as follows: 6 genes with PCC values between 0.7 and 0.8, and 45 genes with PCC values between 0.6 and 0.7. In addition, 68 genes were in the range of 0.5 to 0.6, 70 genes had PCC values between 0.4 and 0.5, and 35 genes were in the range of 0.3 to 0.4. In addition, 18 genes had PCC values between 0.2 and 0.3, 7 genes had values between 0.1 and 0.2, and 1 gene had a PCC value between 0.0 and 0.1.

#### 4.2.4. Analysis of EMGP-Net-noAttn Results

The EMGP-Net-noAttn variant, which eliminated the attention mechanism and employed simple concatenation, achieved positive PCC values for all genes. Of the genes, 8 had PCC values between 0.7 and 0.8, 49 had a PCC value between 0.6 and 0.7, and 72 had a PCC value between 0.5 and 0.6. Additionally, 66 genes had a PCC that fell between 0.4 and 0.5, 32 had a value between 0.3 and 0.4, 15 had a value between 0.2 and 0.3, and 8 had a value between 0.1 and 0.2. Despite performing slightly worse than the full EMGP-Net, this model still showed a clear improvement over the individual backbone models.

EMGP-Net predicted a total of 145 genes with a PCC value greater than 0.5, outperforming EMGP-Net-noAttn which predicted 129, MambaVision which predicted 113, and EfficientFormer which predicted 119. In this context, EMGP-Net had the highest number of predicted genes in the PCC intervals of 0.7 to 0.8, 0.6 to 0.7, and 0.5 to 0.6, indicating its improved performance in these ranges compared to the other models, as shown in Figure 7.

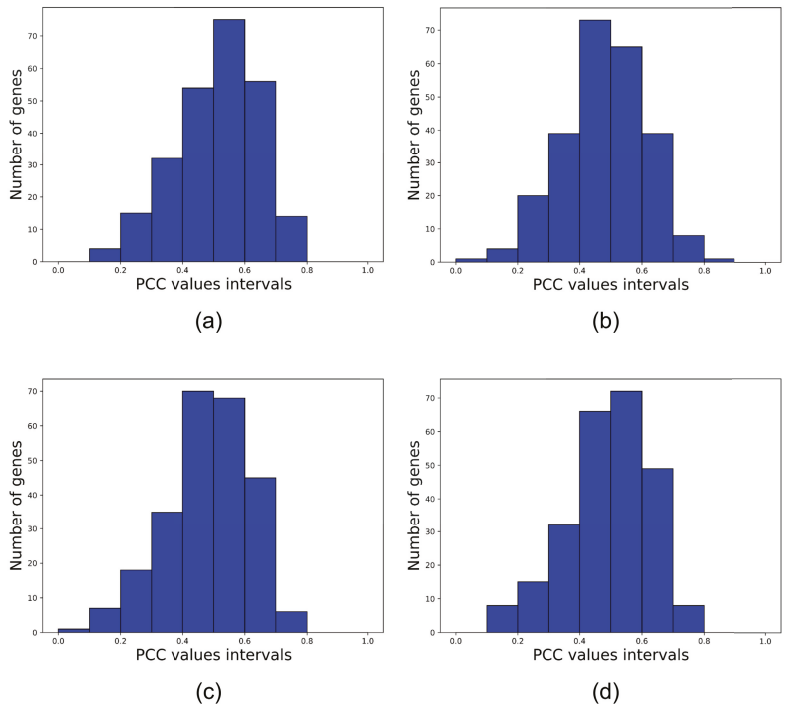



**Figure 7.** Distribution of PCC values for gene expression predictions from each model: (**a**) PCC distribution for EMGP-Net with attention mechanism. (**b**) PCC distribution for MambaVision. (**c**) PCC distribution for EfficientFormer. (**d**) PCC distribution for EMGP-Net without attention mechanism. Each histogram shows the number of genes that fell within specific PCC value intervals. The x-axis represents the PCC ranges, and the y-axis shows the number of genes in each range.

### *4.3. External Validation*

In this section, we discuss the external validation process, where we evaluated our model's generalizability on a dataset different from the one used for training. This approach allowed us to evaluate the performance of the model on a new set of patients and validate its ability to make accurate predictions on different datasets. Our approach was compared to GeNetFormer [11] and ST-Net [9] after it was retrained.

#### 4.3.1. Model Evaluation on the STNet Dataset

This part presents the results obtained when the model was trained on the entire HER2+ dataset and tested on all patients from the STNet dataset, selecting the best PCC

for each gene across all patients. Table 2 shows that EMGP-Net had the best PCC values for the set of the top 14 genes predicted compared to GeNetFormer and ST-Net. The set of the top 14 genes predicted by EMGP-Net included ERBB2, ACTG1, CALR, RPL23, and GNAS, with PCC values of 0.7145, 0.7051, 0.7047, 0.6973, and 0.6962, respectively. For GeNetFormer, the genes were DDX5, ACTG1, CPB1, PTMA, and RPL23, with PCC values of 0.7069, 0.6510, 0.6384, 0.6235, and 0.6130, respectively. For ST-Net, the genes were GNAS, RPL23, PTPRF, ACTG1, and DDX5, with PCC values of 0.6708, 0.6592, 0.6503, 0.6460, and 0.6406, respectively.

We present a comparison of the performance of the EMGP-Net, GeNetFormer, and ST-Net models based on common genes. Table 5 shows the PCC values for the 9 common genes out of the top 14 genes for all three models. The results show that EMGP-Net outperformed GeNetFormer and ST-Net. We can see that EMGP-Net predicted eight out of nine common genes with higher PCC values compared to GeNetFormer and nine out of nine genes with higher PCC values compared to ST-Net. Specifically, EMGP-Net achieved the highest PCC scores for genes such as ACTG1, CALR, RPL23, GNAS, and PTPRF, with PCC values of 0.7051, 0.7047, 0.6973, 0.6962, and 0.6867, respectively, while GeNetFormer showed the best PCC in predicting only DDX5, with a PCC value of 0.7069.

**Table 5.** Overview of EMGP-Net performance vs. ST-Net and GeNetFormer performance. The highest PCC values for common genes among the top 14 genes predicted by the different models are in bold and underlined.

| Gene | ST-Net | GeNetFormer | EMGP-Net |
|---|---|---|---|
| ACTG1 | 0.6460 | 0.6510 | **0.7051** |
| CALR | 0.6273 | 0.5864 | **0.7047** |
| RPL23 | 0.6592 | 0.6130 | **0.6973** |
| GNAS | 0.6708 | 0.5943 | **0.6962** |
| PTPRF | 0.6503 | 0.5974 | **0.6867** |
| TAGLN2 | 0.6274 | 0.5638 | **0.6814** |
| DDX5 | 0.6406 | **0.7069** | 0.6724 |
| HSPB1 | 0.6235 | 0.5840 | **0.6645** |
| PTMA | 0.6201 | 0.6235 | **0.6588** |

4.3.2. Model Evaluation on the HER2+ Dataset

This part presents the results obtained when the model was trained on the entire STNet dataset and tested on all patients from the HER2+ dataset, selecting the best PCC for each gene across all patients. Table 3 shows that EMGP-Net had the best PCC values for the set of the top 14 genes predicted compared to GeNetFormer and ST-Net. The set of the top 14 genes predicted by EMGP-Net included ERBB2, S100A11, ATP5E, HSP90B1, and LGALS3, with PCC values of 0.7285, 0.6686, 0.6650, 0.6404, and 0.6347, respectively. For GeNetFormer, the top 14 genes included ATP5E, S100A11, ERBB2, PTPRF, and HSP90B1, with PCC values of 0.6746, 0.6434, 0.6141, 0.6115, and 0.5986, respectively. For ST-Net, the top 14 genes included ATP5E, ERBB2, S100A11, PTPRF, and LGALS3, with PCC values of 0.6719, 0.6620, 0.6374, 0.6227, and 0.5918, respectively.

We compared the performance of EMGP-Net, GeNetFormer, and ST-Net on the set of the common genes among the top 14 predicted genes. Table 6 shows the PCC values for ERBB2, S100A11, ATP5E, HSP90B1, LGALS3, PTPRF, and PSMB4. The results show that EMGP-Net outperformed both GeNetFormer and ST-Net for most of these genes. It is clear that EMGP-Net showed the best performance, especially for ERBB2, where it achieved a PCC value of 0.7285, compared to the 0.6620 of ST-Net and the 0.6141 of GeNetFormer. For S100A11, EMGP-Net again outperformed the other models, with a PCC of 0.6686,

higher than the 0.6374 of ST-Net and the 0.6434 of GeNetFormer. Similarly, for HSP90B1, EMGP-Net achieved a PCC of 0.6404, higher than both the 0.5903 of ST-Net and the 0.5986 of GeNetFormer. GeNetFormer performed better on one gene, ATP5E, with a PCC value of 0.6746, and ST-Net outperformed on one gene, PTPRF, with a PCC value of 0.6227.

**Table 6.** Overview of EMGP-Net performance vs. ST-Net and GeNetFormer performance. The highest PCC values for common genes among the top 14 genes predicted by the different models are in bold and underlined.

| Gene | ST-Net | GeNetFormer | EMGP-Net |
|---|---|---|---|
| ERBB2 | 0.6620 | 0.6141 | **0.7285** |
| S100A11 | 0.6374 | 0.6434 | **0.6686** |
| ATP5E | 0.6719 | **0.6746** | 0.6650 |
| HSP90B1 | 0.5903 | 0.5986 | **0.6404** |
| LGALS3 | 0.5918 | 0.5393 | **0.6347** |
| PTPRF | **0.6227** | 0.6115 | 0.6049 |
| PSMB4 | 0.5750 | 0.5400 | **0.5763** |

## 5. Discussion

Obviously, the increasing interest in gene expression prediction using WSIs combined with ST data shows the importance of improving predictive models in the field of cancer research, especially breast cancer research, diagnosis, and treatment, and this study has contributed towards addressing the limitation of the high cost of ST technologies.

Several studies have proposed different architectures. He et al. [9] used DenseNet-121. BrST-Net [10] is a CNN-based framework with EfficientNet-b0 as the best-performing model. GeNetFormer [11] is a transformer-based model, with EfficientFormer being the best performing compared to seven other transformer models. TRIPLEX [22] adapts another method and uses a three-encoder approach based on ResNet18 in which the global encoder also integrates transformer blocks. HisToGene [12] integrates multi-head attention and vision transformers. GNNs have been widely used; for example, Zeng et al. [14] proposed an architecture based on a GNN and transformer as well as the architecture introduced in [13], while Jia et al. [15] used GAT to introduce THItoGene. ErwaNet [23] follows a different strategy and integrates different modules like an ERM and a WAM and different edges like K-nearest neighbor, K-nearest similarity, and percent similarity edges. Hypergraphs were also used in several studies, such as PH2ST [16], which uses ViT and a universal histology image encoder, and the HGGEP architecture [17], which includes a hypergraph association module, gradient enhancement module, ShuffleNet V2 backbone, convolutional block attention module, and ViT. EGN [18] and EGGN [19] were based on exemplar guidance learning. All previous studies have followed different strategies and implemented different training approaches. Some have used a leave-one-patient-out approach, others have used cross-validation, and others have simply divided datasets into training, validation, and test sets. Also, different datasets have been used in different studies, such as the HER2+, STNet, and 10x Genomics datasets with different numbers of targeted genes. To maintain the coherence of research, we introduced EMGP-Net in this paper, a hybrid deep learning model that combines MambaVision and EfficientFormer, designed to predict 250 genes. By leveraging features from both models and utilizing an attention-based fusion layer, we aimed to improve the prediction of breast cancer gene expression. We employed the leave-one-patient-out approach in both internal and external validation.

Our results showed that EMGP-Net outperformed both individual models, MambaVision and EfficientFormer, in internal validation using the HER2+ dataset. It had the highest PCC scores ranging from 0.7002 (CD24) to 0.7903 (PTMA) for the top 14 genes compared to EfficientFormer, which generally dropped behind our model with scores ranging from 0.6834 (S100A11) to 0.7777 (PTMA). It also outperformed MambaVision on 13 out of 14 genes, with MambaVision's top PCC score being 0.8049 (B2M). Of the 14 genes, we took the common genes across all models, comprising 10 genes, and compared their PCCs. Our model outperformed EfficientFormer on all genes including PTMA, GNAS, B2M, HNRNPA2B1, and TPT1. MambaVision outperformed EMGP-Net only on three genes, namely, B2M, TPT1, and HLA-DRA.

External validation was used to evaluate the performance of EMGP-Net on two datasets: the HER2+ and STNet, representing different breast cancer subtypes. It was trained on one dataset and tested on the other. The model achieved the highest PCC for most of the top 14 genes and showed the best external validation results compared to other retrained models from other studies.

- When trained on the HER2+ dataset and tested on the STNet dataset, our model outperformed GeNetFormer and ST-Net on all the 14 genes, with PCC values ranging from 0.6563 (KRT19) to 0.7145 (ERBB2) compared to the PCC values of GeNetFormer, which ranged from 0.5250 (KRT19) to 0.7069 (DDX5), and the PCC values of ST-Net, which ranged from 0.5749 (HLA-DRA) to 0.6708 (GNAS). On 9 common genes out of the 14 genes, our model outperformed ST-Net, including ACTG1, CALR, RPL23, GNAS, and PTPRF, while GeNetFormer only outperformed our model on 1 gene, which was DDX5.
- When trained on the STNet dataset and tested on the HER2+ dataset, our model outperformed GeNetFormer and ST-Net on all 14 genes, with PCC scores ranging from 0.5465 (GNAS) to 0.7285 (ERBB2) compared to the PCC values of GeNetFormer, which ranged from 0.5185 (COL1A2) to 0.6746 (ATP5E), and the PCC values of ST-Net, which ranged from 0.5287 (LASP1) to 0.6719 (ATP5E). Of the 7 common genes out of 14 genes, our model outperformed GeNetFormer on 6 genes, namely, ERBB2, S100A11, HSP90B1, LGALS3, PTPRF, and PSMB4, and only 1 gene, ATP5E, was well predicted with GeNetFormer. Our model also outperformed ST-Net on six other genes, namely, ERBB2, S100A11, ATP5E, HSP90B1, LGALS3, and PSMB4, and only PTPRF was well predicted by ST-Net.

When compared to other studies that applied different approaches, EMGP-Net achieved the highest PCC value of 0.7903 (PTMA) using the HER2+ dataset: THItoGene reached a top PCC value of 0.7470 (FN1), HGGEP achieved a top PCC value of 0.6520 (MYL12B), and Hist2ST achieved a top PCC value of 0.7310 (FN1) on the same dataset. Using the STNet dataset, BrST-Net reached a top PCC value of 0.6325 (B2M), while SEPAL had a top PCC value of 0.6390 (ENSG00000145824). EMGP-Net, tested on the STNet dataset as an external validation, achieved a top PCC value of 0.7145 (ERBB2).

Notably, EMGP-Net performed well in predicting the expression of genes such as PTMA, GNAS, B2M, HNRNPA2B1, TPT1, and XBP1 out of 250 targeted genes, which are important biomarkers in breast cancer prognosis. When compared to the results of MambaVision and EfficientFormer, EMGP-Net had the best PCC values for most of the genes evaluated. The model's ability to integrate features from both MambaVision and EfficientFormer using a multi-head attention mechanism contributed to its improved performance.

Although our results are encouraging, this study has some limitations. First, the HER2+ and STNet datasets that we used are smaller than other datasets that are commonly

used in computer vision. This small size may make the model more difficult to apply to other patient groups. Differences in how tissue slides are stained, the type of scanner used, and the unique characteristics of each patient could affect the model's performance on new data. While we tested the model on both HER2+ and STNet datasets to ensure its effectiveness in various cases, using larger, more diverse datasets would give more reliable results. Additionally, the datasets used in this study are specific to particular types of breast cancer. This could make it difficult for the model to perform well on some other cancer types. For instance, what the model learned from the HER2+ samples may not be directly applicable to triple-negative or luminal A types. We may need to create specialized versions of the model for different subtypes or develop a more general model. Another limitation is that the model is not easily interpretable from a biological point of view. Even though EMGP-Net makes accurate predictions, it does not explain which genes are important for its decisions. In the future, we could use techniques like Grad-CAM or others to better understand how the model works and help doctors trust its predictions.

## 6. Conclusions

In this study, we introduced EMGP-Net, a hybrid deep learning architecture that combines MambaVision and EfficientFormer to predict gene expression from breast cancer WSIs. By leveraging the capabilities of both models and incorporating a multi-head attention mechanism, EMGP-Net successfully improved gene expression predictions and outperformed models in internal and external validation tests. Our results showed that EMGP-Net achieved a high performance compared to MambaVision and EfficientFormer, with high PCC values. It showed an improvement in the prediction of 250 genes, with EMGP-Net surpassing both models in most cases. External validation on two datasets, HER2+ and STNet, demonstrated the reliability of EMGP-Net. It outperformed both GeNetFormer and ST-Net in most cases. Despite the high performance of EMGP-Net, our study has some limitations. Future work will focus on increasing the number of predicted genes and exploring the integration of explainable AI techniques to improve interpretability. Synthetic data generated by generative models may also enhance training diversity and performance. EMGP-Net provides a solid foundation for advancing gene expression prediction from histopathological images in breast cancer research.

**Author Contributions:** Conceptualization, O.T. and M.A.A.; methodology, O.T. and M.A.A.; validation, O.T. and M.A.A.; formal analysis, O.T. and M.A.A.; writing—original draft preparation, O.T.; writing—review and editing, M.A.A.; funding acquisition, M.A.A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was enabled in part by support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), funding reference number RGPIN-2024-05287, and by the AI in Health Research Chair at the Université de Moncton.

**Institutional Review Board Statement:** This research did not require Institutional Review Board (IRB) approval as it exclusively utilized publicly available data.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** This work uses two public datasets: 1. HER2+ dataset (https://www.synapse.org/Synapse:syn52503858/files/, accessed on 20 May 2025). 2. STNet dataset (https://data.mendeley.com/datasets/29ntw7sh4r/5, accessed on 20 May 2025).

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

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| APEG | Atypical Position Encoding Generator |
| AuxNet | Auxiliary Network |
| BN | Batch Normalization |
| CBAM | Convolutional Block Attention Module |
| CNN | Convolutional Neural Network |
| EB | Exemplar Bridging |
| EMGP-Net | EfficientFormer (E), MambaVision (M), gene expression prediction (GP) network (Net). |
| EMGP-Net-noAttn | EfficientFormer (E), MambaVision (M), gene expression prediction (GP) network (Net), no attention (noAtt). |
| ERM | Edge Relational Module |
| GAT | Graph Attention Network |
| GeLU | Gaussian Error Linear Unit |
| GEM | Gradient Enhancement Gradient |
| GNN | Graph Neural Network |
| HAM | Hypergraph Association Module |
| H&E | Hematoxylin and Eosin |
| LN | Layer Normalization |
| MAE | Mean Absolute Error |
| PCC | Pearson Correlation Coefficient |
| RMSE | Root Mean Squared Error |
| SOTA | State Of The Art |
| SSM | State-Space Model |
| ST | Spatial Transcriptomics |
| ViT | Vision Transformer |
| WAM | Window Attention Module |
| WSIs | Whole-Slide Images |

## Appendix A. Lists of the 250 Genes Included in This Study Across the Two Datasets: HER2+ and STNet

This appendix provides complete lists of the 250 genes used in this study. Appendix A.1 presents the genes selected from the HER2+ dataset, and Appendix A.2 presents the genes selected from the STNet dataset.

### *Appendix A.1. List of the 250 Genes from the HER2+ Dataset*

This list contains the 250 genes selected from the HER2+ dataset that were used to train and evaluate the gene expression prediction models described in this study.

**Table A1.** List of the 250 genes from the HER2+ dataset. The last row uses the asterisk (*) to fill empty cells and complete the seven-column layout.

| Gene Name | Gene Name | Gene Name | Gene Name | Gene Name | Gene Name | Gene Name |
|---|---|---|---|---|---|---|
| PTMA | GNAS | B2M | HNRNPA2B1 | TPT1 | XBP1 | ACTG1 |
| HLA-B | TMSB10 | DDX5 | HLA-DRA | ACTB | S100A11 | CD24 |
| HSP90B1 | PSMB4 | COX6C | TUBA1B | EIF4G2 | PRDX1 | HLA-C |
| HLA-A | LAPTM4A | VMP1 | HSP90AA1 | UBC | ATP5E | CALM2 |
| SCGB2A2 | NACA | FTH1 | COX7C | CALR | CCT3 | FASN |
| PEBP1 | HSPB1 | PSAP | SPINT2 | BEST1 | PFN1 | PLXNB2 |

**Table A1.** *Cont.*

| Gene Name | Gene Name | Gene Name | Gene Name | Gene Name | Gene Name | Gene Name |
|---|---|---|---|---|---|---|
| ATP5B | SERF2 | LGALS3 | P4HB | MYH9 | CRIP2 | CHCHD2 |
| ATP1A1 | ERBB2 | KRT19 | CD74 | FN1 | GAPDH | HSP90AB1 |
| HSPA8 | PTPRF | FTL | LSM4 | KDELR1 | CFL1 | VCP |
| MIDN | PPP1CA | SLC9A3R1 | PABPC1 | APOE | GRB7 | RACK1 |
| EEF2 | TUBB | JTB | SH3BGRL3 | TXNIP | SCD | OAZ1 |
| LASP1 | ATG10 | SPDEF | SEPW1 | VIM | MDK | CTSB |
| SEC61A1 | GRINA | IDH2 | UBE2M | COPS9 | MMACHC | MZT2B |
| JUP | UBA52 | PSMD8 | SLC2A4RG | MLLT6 | SSR2 | DBI |
| TAPBP | CIB1 | PPDPF | CST3 | TSPO | CD63 | COL1A1 |
| PTBP1 | AES | TAGLN2 | ATP5G2 | MYL6 | NUCKS1 | GNAI2 |
| PLD3 | GNB2 | LMAN2 | HM13 | RALY | SNRPB | SDC1 |
| ENO1 | COPE | PHB | GRN | HLA-E | STARD10 | COL1A2 |
| A2M | ALDOA | NUPR1 | LAPTM5 | EIF3B | EDF1 | MAPKAPK2 |
| SERINC2 | FLNA | MIEN1 | SYNGR2 | MUC1 | COX4I1 | EIF4G1 |
| C3 | PERP | H1FX | GPX4 | C1QB | APOC1 | DHCR24 |
| PRSS8 | COX6B1 | IGLC2 | KRT18 | ERGIC1 | GUK1 | PGAP3 |
| IGLC3 | IGHG3 | FAU | UQCRQ | UQCR11 | ZYX | CLDN4 |
| CD81 | CD99 | NDUFA3 | CISD3 | RRBP1 | COX5B | S100A6 |
| LGALS3BP | PCGF2 | TYMP | TIMP1 | NDUFB9 | ATP6V0B | AP2S1 |
| COX8A | FNBP1L | COL3A1 | STARD3 | PTMS | IFI27 | KRT7 |
| PFKL | CTSD | RABAC1 | PSMB3 | PSMD3 | LMNA | H2AFJ |
| ARHGDIA | SPARC | EEF1D | SLC25A6 | INTS1 | ACTN4 | IGHA1 |
| CHPF | ELOVL1 | SSR4 | ATP6AP1 | CYBA | TAGLN | C1QA |
| PRRC2A | RHOC | IGHG1 | MMP14 | PPP1R1B | CALML5 | BSG |
| CLDN3 | AEBP1 | LY6E | TRAF4 | IGKC | BGN | NBL1 |
| FKBP2 | AP000769.1 | ROMO1 | COL6A2 | IGHM | C12orf57 | MYL9 |
| BCAP31 | SCAND1 | TCEB2 | PFDN5 | BST2 | KIAA0100 | NDUFB7 |
| MUCL1 | LGALS1 | POSTN | TFF3 | MGP | COL18A1 | NDUFA11 |
| IGFBP2 | KRT81 | SUPT6H | ORMDL3 | S100A9 | MUC6 | AZGP1 |
| S100A14 | S100A8 | IGHG4 | ADAM15 | ISG15 | * | * |

### *Appendix A.2. List of the 250 Genes from the STNet Dataset*

This list contains the 250 genes selected from the STNet dataset that were used to train and evaluate the gene expression prediction models described in this study.

**Table A2.** List of the 250 genes from the STNet dataset. Gene names marked as N/A indicate names that were originally ambiguous. The last row uses the asterisk (*) to fill empty cells and complete the seven-column layout.

| Gene Name | Gene Name | Gene Name | Gene Name | Gene Name | Gene Name | Gene Name |
|---|---|---|---|---|---|---|
| ERBB2 | ACTG1 | CALR | RPL23 | GNAS | PSMD3 | PTPRF |

**Table A2.** *Cont.*

| Gene Name | Gene Name | Gene Name | Gene Name | Gene Name | Gene Name | Gene Name |
|---|---|---|---|---|---|---|
| TMSB10 | GAPDH | TAGLN2 | DDX5 | HSPB1 | PTMA | KRT19 |
| P4HB | PRDX1 | PFN1 | HLA-C | S100A11 | RPL28 | ENSG00000203812 |
| B2M | HLA-DRA | CPB1 | NHERF1 | RPLP0 | S100A9 | RPL19 |
| HLA-B | C4B | CALML5 | ACTB | S100A8 | RPLP2 | TMSB4X |
| APOE | GRINA | ENO1 | RPL35 | MGP | TIMP1 | HLA-A |
| RPS11 | IGLL5 | PRSS8 | ENSG00000272196 | COX6C | ATP1A1 | CYBA |
| RPS19 | RPLP1 | RPS28 | RPS18 | JUP | RPS2 | UBA52 |
| TUBA1B | SELENOW | IFI27 | ELF3 | FTL | N/A | RPL13 |
| RPL9 | ATP5F1E | N/A | RPL10 | CST3 | RPS4X | RPL38 |
| TAPBP | SYNGR2 | RPS20 | CD74 | SERF2 | FASN | C1QA |
| CLDN3 | N/A | SPDEF | RACK1 | UBC | BCAP31 | PABPC1 |
| RPS6 | N/A | FLNA | RPS13 | H1-10 | SDC1 | EIF4G1 |
| FTH1 | RPS9 | CRIP2 | RPS27A | AEBP1 | CLU | S100A6 |
| RPL8 | FN1 | SEC61A1 | MYL6 | RPL15 | RPS17 | PPP1CA |
| GPX4 | RPS7 | BGN | RPL13A | ATP6V0B | BSG | TPT1 |
| A2M | BST2 | PPDPF | MYL9 | VIM | RPS15A | XBP1 |
| COL1A1 | RPS14 | STARD10 | RPS12 | RPS3 | ISG15 | RPS15 |
| ENSG00000169100 | MZT2A | HSP90AB1 | CD81 | LY6E | IFITM3 | MZT2B |
| EIF4A1 | PFDN5 | RPS8 | COX8A | UBB | LGALS3BP | RPL23A |
| EEF2 | RPL29 | N/A | TAGLN | EVL | N/A | RPL3 |
| MUC1 | SPARC | N/A | APOC1 | H3-3B | RPS23 | N/A |
| KRT8 | RPS21 | UQCR11 | TSPO | RPL27 | UQCRQ | GNB2 |
| RPL34 | ARHGDIA | LAPTM5 | SNHG25 | RPL5 | N/A | N/A |
| RHOC | TUFM | RPL35A | RPL14 | EDF1 | N/A | CFL1 |
| RPL18A | HLA-E | SSR2 | FXYD3 | H2AJ | FAU | AZGP1 |
| BEST1 | COL1A2 | LMNA | RPL12 | GUK1 | COX4I1 | OAZ1 |
| RPL37A | PLXNB2 | ELOB | GAS5 | N/A | GRN | MALAT1 |
| RPS24 | IGFBP2 | COX6B1 | CTSB | TFF3 | RPL24 | ALDOA |
| RPL32 | RPS16 | PRDX2 | EEF1D | RPL4 | RPL31 | CCND1 |
| NDUFA13 | RPL7A | RPL11 | RPL36 | NBEAL1 | EIF5A | PLD3 |
| RPL27A | CD63 | SH3BGRL3 | ATP6AP1 | PSAP | ZNF90 | TLE5 |
| RPS29 | RPL7 | RPS25 | KRT18 | RPS5 | NDUFA11 | CTSD |
| NDUFB9 | SSR4 | C3 | RPS27 | N/A | ENSG00000279274 | RPL37 |
| RPS3A | ENSG00000255823 | POLR2L | IFI6 | ENSG00000269028 | RPS10 | RPL30 |
| ENSG00000279483 | C12orf57 | GNAI2 | TFF1 | RPL18 | * | * |

## References


1. Obeagu, E.I.; Obeagu, G.U. Breast cancer: A review of risk factors and diagnosis. *Medicine* **2024**, *103*, e36905. [CrossRef]
2. Thaalbi, O.; Akhloufi, M.A. Deep learning for breast cancer diagnosis from histopathological images: Classification and gene expression: Review. *Netw. Model. Anal. Health Inform. Bioinform.* **2024**, *13*, 52. [CrossRef]
3. Al-Jabbar, M.; Alshahrani, M.; Senan, E.M.; Ahmed, I.A. Multi-Method Diagnosis of Histopathological Images for Early Detection of Breast Cancer Based on Hybrid and Deep Learning. *Mathematics* **2023**, *11*, 1429. [CrossRef]
4. Obayya, M.; Maashi, M.S.; Nemri, N.; Mohsen, H.; Motwakel, A.; Osman, A.E.; Alneil, A.A.; Alsaid, M.I. Hyperparameter Optimizer with Deep Learning-Based Decision-Support Systems for Histopathological Breast Cancer Diagnosis. *Cancers* **2023**, *15*, 885. [CrossRef] [PubMed]

5. Clement, D.; Agu, E.; Obayemi, J.; Adeshina, S.; Soboyejo, W. Breast Cancer Tumor Classification Using a Bag of Deep Multi-Resolution Convolutional Features. *Informatics* **2022**, *9*, 91. [CrossRef]
6. Bagchi, A.; Pramanik, P.; Sarkar, R. A Multi-Stage Approach to Breast Cancer Classification Using Histopathology Images. *Diagnostics* **2022**, *13*, 126. [CrossRef]
7. Bhausaheb, D.P.; Kashyap, K.L. Detection and classification of breast cancer availing deep canid optimization based deep CNN. *Multimed. Tools Appl.* **2023**, *82*, 18019–18037. [CrossRef]
8. Yu, D.; Lin, J.; Cao, T.; Chen, Y.; Li, M.; Zhang, X. SECS: An effective CNN joint construction strategy for breast cancer histopathological image classification. *J. King Saud Univ.-Comput. Inf. Sci.* **2023**, *35*, 810–820. [CrossRef]
9. He, B.; Bergenstråhle, L.; Stenbeck, L.; Abid, A.; Andersson, A.; Borg, Å.; Maaskola, J.; Lundeberg, J.; Zou, J. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Eng.* **2020**, *4*, 827–834. [CrossRef]
10. Rahaman, M.M.; Millar, E.K.; Meijering, E. Breast cancer histopathology image-based gene expression prediction using spatial transcriptomics data and deep learning. *Sci. Rep.* **2023**, *13*, 13604. [CrossRef]
11. Thaalbi, O.; Akhloufi, M.A. GeNetFormer: Transformer-Based Framework for Gene Expression Prediction in Breast Cancer. *AI* **2025**, *6*, 43. [CrossRef]
12. Pang, M.; Su, K.; Li, M. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *BioRxiv* **2021**. [CrossRef]
13. Mejia, G.; Cárdenas, P.; Ruiz, D.; Castillo, A.; Arbeláez, P. SEPAL: Spatial Gene Expression Prediction from Local Graphs. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 2294–2303. [CrossRef]
14. Zeng, Y.; Wei, Z.; Yu, W.; Yin, R.; Yuan, Y.; Li, B.; Tang, Z.; Lu, Y.; Yang, Y. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings Bioinform.* **2022**, *23*, bbac297. [CrossRef] [PubMed]
15. Jia, Y.; Liu, J.; Chen, L.; Zhao, T.; Wang, Y. THItoGene: A deep learning method for predicting spatial transcriptomics from histological images. *Briefings Bioinform.* **2024**, *25*, bbad464. [CrossRef]
16. Niu, Y.; Liu, J.; Zhan, Y.; Shi, J.; Zhang, D.; Machado, I.; Crispin-Ortuzar, M.; Li, C.; Gao, Z. ST-Prompt Guided Histological Hypergraph Learning for Spatial Gene Expression Prediction. *arXiv* **2025**, arXiv:2503.16816. [CrossRef]
17. Li, B.; Zhang, Y.; Wang, Q.; Zhang, C.; Li, M.; Wang, G.; Song, Q. Gene expression prediction from histology images via hypergraph neural networks. *Briefings Bioinform.* **2024**, *25*, bbae500. [CrossRef]
18. Yang, Y.; Hossain, M.Z.; Stone, E.A.; Rahman, S. Exemplar guided deep neural network for spatial transcriptomics analysis of gene expression prediction. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 5039–5048. [CrossRef]
19. Yang, Y.; Hossain, M.Z.; Stone, E.; Rahman, S. Spatial transcriptomics analysis of gene expression prediction using exemplar guided graph neural network. *Pattern Recognit.* **2024**, *145*, 109966. [CrossRef]
20. Liu, Z.; Qian, S.; Xia, C.; Wang, C. Are transformer-based models more robust than CNN-based models? *Neural Netw.* **2024**, *172*, 106091. [CrossRef]
21. Hatamizadeh, A.; Kautz, J. Mambavision: A hybrid mamba-transformer vision backbone. *arXiv* **2024**, arXiv:2407.08083. [CrossRef]
22. Chung, Y.; Ha, J.H.; Im, K.C.; Lee, J.S. Accurate Spatial Gene Expression Prediction by Integrating Multi-Resolution Features. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 11591–11600. [CrossRef]
23. Chen, C.; Zhang, Z.; Tang, P.; Liu, X.; Huang, B. Edge-relational window-attentional graph neural network for gene expression prediction in spatial transcriptomics analysis. *Comput. Biol. Med.* **2024**, *174*, 108449. [CrossRef]
24. Andersson, A.; Larsson, L.; Stenbeck, L.; Salmén, F.; Ehinger, A.; Wu, S.Z.; Al-Eryani, G.; Roden, D.; Swarbrick, A.; Borg, Å.; et al. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nat. Commun.* **2021**, *12*, 6012. [CrossRef] [PubMed]
25. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752v2. [CrossRef]
26. Li, Y.; Yuan, G.; Wen, Y.; Hu, J.; Evangelidis, G.; Tulyakov, S.; Wang, Y.; Ren, J. Efficientformer: Vision transformers at mobilenet speed. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 12934–12949. [CrossRef]
27. Vaswani, A. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762. [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.

computers

MDPI

Article

# Adoption Drivers of Intelligent Virtual Assistants in Banking: Rethinking the Artificial Intelligence Banker

**Rui Ramos [1,*], Joaquim Casaca [1] and Rui Patrício [2]**

[1] Universidade Europeia, IADE, Faculdade de Design, Tecnologia e Comunicação, UNIDCOM/IADE, Unidade de Investigação em Design e Comunicação, Universidade Europeia, Av. D. Carlos I, 4, 1200-649 Lisboa, Portugal; joaquim.casaca@universidadeeuropeia.pt

[2] GOVCOPP and DEGEIT, Department of Economics, Management, Industrial Engineering and Tourism, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal; rui.patricio@ua.pt

* Correspondence: ruip.pramos@gmail.com

**Abstract:** The adoption of Intelligent Virtual Assistants (IVAs) in the banking sector presents new opportunities to enhance customer service efficiency, reduce operational costs, and modernize service delivery channels. However, the factors driving IVA adoption and usage, particularly in specific regional contexts such as Portugal, remain underexplored. This study examined the determinants of IVA adoption intention and actual usage in the Portuguese banking sector, drawing on the Technology Acceptance Model (TAM) as its theoretical foundation. Data were collected through an online questionnaire distributed to 154 banking customers after they interacted with a commercial bank's IVA. The analysis was conducted using Partial Least Squares Structural Equation Modeling (PLS-SEM). The findings revealed that perceived usefulness significantly influences the intention to adopt, which in turn significantly impacts actual usage. In contrast, other variables—including trust, ease of use, anthropomorphism, awareness, service quality, and gendered voice—did not show a significant effect. These results suggest that Portuguese users adopt IVAs based primarily on functional utility, highlighting the importance of outcome-oriented design and communication strategies. This study contributes to the understanding of technology adoption in mature digital markets and offers practical guidance for banks seeking to enhance the perceived value of their virtual assistants.

**Keywords:** intelligent virtual assistant (IVA); chatbots; automation; customer experience; finance; banking

## 1. Introduction

The growing expansion of e-commerce is partially justified by using Intelligent Virtual Assistants (IVAs) to enhance the customer experience, as these can improve the whole efficiency of the process. In e-commerce, IVAs improve the customer experience by providing personalized recommendations and assistance at every contact with a brands' customer journey [1,2].

The application of IVAs has been exponential in many sectors, addressing the consumers' requirements to access information in the blink of an eye. The banking sector has long invested worldwide in technologies like IVAs, increasing their performance and technical efficiency [3,4]. The banking companies in Portugal are also diving into this new technology, offering IVAs in their app to respond to clients' needs and lifestyles but still trying to understand the drivers of adoption intention fully.

To study the adoption of IVAs in banking, we used the Technology Acceptance Model (TAM). This model has proven its value in studying the adoption of technology for banking, proving its worth and contribution to this type of study [3,5–8]. TAM has been used in various areas, such as education [9–13], health [14–16] and tourism [17,18].

### *1.1. Perceived Usefulness and Perceived Ease of Use*

The perceived usefulness and the perceived ease of use have been used in numerous adoption theories and models, such as TAM, and were found to have great implications for the usage intentions regarding adoption in the m-banking sectors [19–21]. The user's intention concerning new technologies is impacted by the user's conviction of the perceived usefulness and the solution's ease of use [22]. The interactivity with the IVAs is affected by the user's perceived ease of use and usefulness of the assistant [23]. Thus, we hypothesize the following:

**H1.** *Perceived usefulness affects the adoption intention of chatbots for banking.*

**H2.** *Perceived ease of use affects the adoption intention of chatbots for banking.*

### *1.2. Perceived Trust*

Trust can not only be established between people but also between a person and an automated system. The process of forming trust incorporates people's emotions and feelings. The perceived trust toward an IVA can be defined as the user's willingness to be susceptible to the assistant [1,24]. The trust in IVAs can also be described as the IVA's being capable, reliable and credible [25].

**H3.** *Perceived Trust affects the adoption intention of chatbots for banking.*

### *1.3. Anthropomorphism*

The presence of anthropomorphism on an IVA indicates that some human characteristics or behaviors have been implemented in the system. With this presence can also come a negative side effect, where a high level of anthropomorphism can reach the point of being uncanny, where humans are deeply repulsed by the IVA [26]. A negative effect can also occur when users realize that they are communicating with these AI systems while they were under the impression that they are interacting with humans [27].

**H4.** *Anthropomorphism affects the adoption intention of chatbots for banking.*

### *1.4. Awareness of the Service*

The awareness of the service, which is the knowledge that a service exists, is often linked with familiarity with the IVA and together presents a vital role to decrease the privacy concerns toward these assistants [28]. The low awareness of these assistants in the banking sector and their undeniable benefits relate to the customers' low acceptance and usage of this technology [29,30]. Additionally, a lack of familiarity with the assistant is responsible for the users' uncomfortable feelings concerning the IVA, as they have reservations about the accuracy of the errands completed with the solution [31].

**H5.** *Awareness of the service affects the adoption intention of chatbots for banking.*

### *1.5. Service Quality*

Service quality can be defined as a personal evaluation of the provided service given to the solution user [32]. Service quality is significantly impacted by the functional capabilities and the sensorial components of the service [33]. Implementing IVAs is responsible for altering consumers' understanding of service quality as parameters like response time present a significant change between humans and IVA-based services. For this reason, the previously known scales to measure service quality had to be adapted to this new reality [27,34].

**H6.** *Service quality affects the adoption intention of chatbots for banking.*

### *1.6. Gendered Voice*

The introduction of voice features to IVAs has been getting traction, resulting in the increased use of voice assistants offered to help users [35]. These types of assistants provide a more humanized and strong service thanks to the presence of anthropomorphism seen in characteristics such as voice pitch and voice tone [36]. The attribution of gender to the IVA, and consequently to their voice, impacts the perceived human likeness of the assistant, as studies state that female IVAs have an advantage when compared with male IVAs in this capacity [37]. Additionally, ref. [38] stated that there is a female gender favoritism when designing an IVA, resulting in assistants with female features having a prominent presence in the areas of sales, customer service and branded conversations. To try to overcome this favoritism, genderless voices have been introduced to IVAs to overcome the impact this type of voice has on users [39]. Thus, we hypothesized the following:

**H7.** *A gendered voice moderates the relationship between the adoption intention and the actual usage of a chatbot for banking.*

### *1.7. Adoption Intention*

While the adoption intention, which is the subjective likelihood of having a specific behavior, is a convincing indication that the actual adoption will take place, it does not ensure that it will happen despite the high level of influence between the actual usage and the intention to use [18,40,41]. Previous research about the adoption intention of IVAs in the banking sector found that IVAs' adoption intention is dependent on factors such as a positive intention toward these types of technology, perceived usefulness of the solution, perceived ease of use and anthropomorphism [3,42,43]. Thus, we hypothesize the following:

**H8.** *Adoption intention affects the actual usage of the chatbot for banking.*

The study goal is to overcome the research gap in adopting IVAs for banking in Portugal, where the potential for this type of technology is massive, by answering the research question: What influences the experience with IVAs in the banking sector? Since these chatbots are still underdeveloped, their potential to provide users with the best experience was not fully explored. The study's findings shed light on the main variables that influence IVA adoption intentions and actual usage of this technology in the banking sector in Portugal.

Taking into consideration the above, the study model that follows (Figure 1) is suggested, expanding upon the TAM model by analyzing the determinants of perceived usefulness, perceived ease of use, perceived trust and awareness of the service that were adapted from [3]. The proposed model aims to analyze whether the user's perceived

usefulness, perceived ease of use, perceived trust, anthropomorphism, awareness of the service and service quality affects the intention to adopt the chatbot and the actual usage of this technology. It also aims to identify if a gendered voice moderates the relationships between the adoption intention and the actual usage of the chatbot for banking.

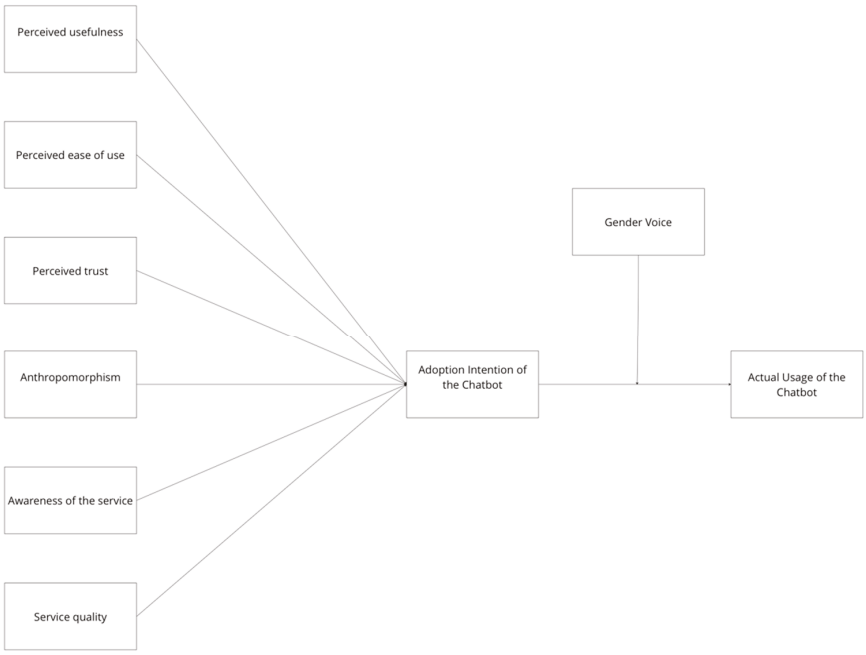



**Figure 1.** Proposed model.

## 2. Materials and Methods

### *2.1. Sample, Data Collection and Analysis*

To validate the comprehension of the questions, we tested the questionnaire using a small sample of 30 people. Based on the results of this pilot study, we adjusted some of the items to ensure that all of them had clarity and that the participants could understand each question. After that, the questionnaire was completed using Microsoft Forms. The collection of data was conducted through an online survey during a period of 1 month. The survey was made available on the chatbot of Bank X, a Portuguese commercial bank, with more than 10% of the market share and a large client base, by a link to Microsoft Forms after the respondent used the virtual assistant. A total of 154 questionnaire responses were collected. The content of the questionnaire was translated from English into Portuguese as the survey was applied in Portugal. The participants were informed that the survey was anonymous and would not collect any personal information, such as names or any form of contact, to ensure the anonymity of the contributors. The participants had no exclusion criteria for this study to be classified as valid. A total of 154 questionnaire responses were collected. The data collected did not present any missing values, so a sample of 154 questionaries was used to research the model's analysis. A normality test was performed on the dataset to assess if the sample followed a normal distribution and if any of the variables had a normal distribution. Non-parametric Mann–Whitney tests were performed to check if the responses collected earlier in the process presented significant differences from the ones collected closer to the end of the publication of the questionnaire. From those tests, no significant differences were found, demonstrating that there is no non-response bias. Analyzing the skewness of the variables, all the values were between the $-2/+2$ interval, meaning that the variables do not present non-normality [44].

### 2.2. Sample Characteristics

As far as the characteristics of the sample (Table 1), the proportion of men (51%) was comparable to women (49%) with only one person preferring not to disclose their gender, and age ranged from 18 to 76 years old, with a mean age of 41 years old. Concerning educational qualification, 66% of the respondents had an undergraduate degree, 29% had a graduate degree, 5% had a master's degree and 1% had a doctoral degree.

**Table 1.** Sample demographics.

| | Frequency | Percentage |
|---|---|---|
| Gender | | |
| Male | 78 | 51% |
| Female | 75 | 49% |
| Not disclosed | 1 | 1% |
| Education | | |
| Undergraduate | 102 | 66% |
| Graduation | 44 | 29% |
| Master | 7 | 5% |
| PhD | 1 | 1% |
| | Mean | Min–Max |
| Age (years) | 41 | 18–76 |

### 2.3. Measurement Instruments

Firstly, four demographic questions were asked. Secondly, three questions regarding the experience with the solution were answered using a ten-point Likert scale ranging from "Bad" (1) to "Excellent" (10). Since the goal was to continuously record perceived experience, a neutral choice was excluded. Based on the literature, the variables were evaluated from 33 items. Answers to all items were given on 5-point scales (strongly disagree–strongly agree) (Tables 2–8). Thirdly, answers about a gendered voice were given on a 3-point scale (male voice, female voice, without humanization). Table A1 in the Appendix A presents all the measurement items and survey questions and their respective scales.

Perceived usefulness (PU) was rated on three items as follows, adapted from [3] (Table 2). Answers were provided on a 5-point scale (strongly disagree–strongly agree).

**Table 2.** Perceived usefulness scale.

| Items | |
|---|---|
| PU1 | I find the banking chatbot useful in my daily life. |
| PU2 | Using the banking chatbot helps me accomplish things more quickly. |
| PU3 | Using the banking chatbot increases my productivity. |

Perceived ease of use (PEU) was rated on three items as follows, adapted from [3] (Table 3). Answers were provided on a 5-point scale (strongly disagree–strongly agree).

**Table 3.** Perceived ease of use scale.

| Items | |
|---|---|
| PEU1 | Learning how to use the banking chatbot is easy for me. |
| PEU2 | My interaction with the banking chatbot is clear and understandable. |
| PEU3 | I find the banking chatbot easy to use. |

Perceived trust (PT) was rated on four items as follows, adapted from [18] (Table 4). Answers were provided on a 5-point scale (strongly disagree–strongly agree).

**Table 4.** Perceived trust scale.

| Items | |
|---|---|
| PT1 | I feel that the information provided by the chatbot for banking is honest and authentic. |
| PT2 | I feel that chatbots for banking have clarity of services provided and frank opinions that are reliable. |
| PT3 | I feel that chatbots in banking are trustworthy. |
| PT4 | I feel that chatbots for banking have the necessary ability to provide financial counseling. |

Anthropomorphism (ANM) was rated on four items as follows, adapted from [18] (Table 5). Answers were provided on a 5-point scale (strongly disagree–strongly agree).

**Table 5.** Anthropomorphism scale.

| Items | |
|---|---|
| ANM1 | Chatbots for banking have their own mind. |
| ANM2 | Chatbots for banking can experience emotions. |
| ANM3 | I feel that chatbots for banking are inanimate: living. |
| ANM4 | I feel that chatbots for banking are computer-animated: real. |

Awareness of the service (AS) was rated on three items as follows, adapted from [3] (Table 6). Answers were provided on a 5-point scale (strongly disagree–strongly agree).

**Table 6.** Awareness of the service scale.

| Items | |
|---|---|
| AS1 | My bank has communicated a banking chatbot usage policy to me. |
| AS2 | My bank has a strategy regarding the usage of the banking chatbot. |
| AS3 | I have received sufficient information from my bank regarding the usage of the banking chatbot. |

Service quality (SQ) was rated on eleven items as follows, adapted from [45] (Table 7). Answers were provided on a 5-point scale (strongly disagree–strongly agree).

**Table 7.** Service quality scale.

| Items | |
|---|---|
| SQ1 | When I have problems, this chatbot is sympathetic and reassuring. |
| SQ2 | I felt that I could rely on the chatbot's services to fulfill my needs. |
| SQ3 | The chatbot responds promptly to my requests. |
| SQ4 | I trust the chatbot. |
| SQ5 | I feel safe and reassured when having a conversation with the chatbot. |
| SQ6 | The chatbot has adequate knowledge to answer my questions. |

**Table 7.** *Cont.*

| Items | |
|---|---|
| SQ7 | I can be in control of my personal needs through the chatbot. |
| SQ8 | The chatbot gives me the opportunity to respond. |
| SQ9 | I feel valued by the brand through my conversations with the chatbot. |
| SQ10 | I feel empathetically understood through conversations with the chatbot. |
| SQ11 | I feel that the chatbot was developed to meet my personal needs. |

Actual usage (AU) was rated on four items as follows, adapted from [18] (Table 8). Answers were provided on a 5-point scale (strongly disagree–strongly agree).

**Table 8.** Actual usage scale.

| Items | |
|---|---|
| AU1 | I use the chatbot to clarify my doubts. |
| AU2 | I use the chatbot to manage my banking account. |
| AU3 | I use the chatbot to gain knowledge about financial products. |
| AU3 | I use the chatbot to get financial advice. |

Adoption intention (ADI) was rated on one item as follows, as used previously in articles [46–49], adapted from [18] (Table 9). Answers were provided on a 10-point scale (strongly disagree–strongly agree).

**Table 9.** Adoption intention scale.

| Items | |
|---|---|
| ADI1 | There is a possibility that I will encourage my family and friends to use the banking chatbot. |

## 3. Results

The investigation model was analyzed through PLS structural equation modeling (PLS-SEM), using SmartPLS software, version 4.0.9.6 [44]. In the analysis, we also explored the moderating effect of the gender of the voice on the Adoption Intention → Actual Usage.

### *3.1. Measurement Models*

Firstly, we estimated the relationships between the constructs and their indicators. Table 10 displays the convergent validity, the internal consistency reliability, and the discriminant validity. All of the outer loadings of the reflective constructs are well above the threshold value of 0.708, which suggests sufficient levels of the reliability indicator [44]. The indicator ANM3 has the most negligible indicator reliability with a value of 0.549 $(0.741)^2$, whereas ADI and GV had the highest indicator reliabilities with a value of 1 $(1)^2$ as these constructs are unidimensional. All the AVE values are also above the required minimum of 0.50, meaning that the seven reflective constructs have high levels of convergent validity [44]. As the $\rho_A$ from all the constructs is above the threshold value of 0.70, all of them have a high level of internal consistency reliability [44]. The Cronbach's alpha and the $\rho_c$ values for the constructs all exceed the threshold value of 0.70, meaning all the constructs analyzed have a high level of consistency and reliability [44].

**Table 10.** Results of the measurement models.

| Latent Variable | Indicators | Convergent Validity | | | Internal Consistency Reliability | | | Discriminant Validity |
|---|---|---|---|---|---|---|---|---|
| | | Loadings | Indicator Reliability | AVE | Cronbach's Alpha | Reliability $\rho_A$ | Composite Reliability $\rho_C$ | HTMT Significantly Lower than 0.90? |
| ADI | ADI | 1 | 1 | | | | | |
| ANM | ANM1 | 0.914 | 0.835 | 0.728 | 0.877 | 0.910 | 0.914 | YES |
| | ANM2 | 0.911 | 0.830 | | | | | |
| | ANM3 | 0.741 | 0.549 | | | | | |
| | ANM4 | 0.835 | 0.697 | | | | | |
| AS | AS1 | 0.95 | 0.903 | 0.898 | 0.944 | 0.945 | 0.964 | YES |
| | AS2 | 0.951 | 0.904 | | | | | |
| | AS3 | 0.942 | 0.887 | | | | | |
| AU | AU1 | 0.802 | 0.643 | 0.809 | 0.920 | 0.927 | 0.944 | YES |
| | AU2 | 0.933 | 0.870 | | | | | |
| | AU3 | 0.938 | 0.880 | | | | | |
| | AU4 | 0.919 | 0.845 | | | | | |
| GV | GV | 1 | 1 | | | | | |
| PEU | PEU1 | 0.96 | 0.922 | 0.926 | 0.960 | 0.965 | 0.974 | YES |
| | PEU2 | 0.957 | 0.916 | | | | | |
| | PEU3 | 0.971 | 0.943 | | | | | |
| PT | PT1 | 0.961 | 0.924 | 0.904 | 0.965 | 0.972 | 0.974 | NO |
| | PT2 | 0.953 | 0.908 | | | | | |
| | PT3 | 0.959 | 0.920 | | | | | |
| | PT4 | 0.93 | 0.865 | | | | | |
| PU | PU1 | 0.953 | 0.908 | 0.927 | 0.961 | 0.962 | 0.974 | YES |
| | PU2 | 0.97 | 0.941 | | | | | |
| | PU3 | 0.964 | 0.929 | | | | | |
| SQ | SQ1 | 0.949 | 0.901 | 0.889 | 0.988 | 0.989 | 0.989 | YES |
| | SQ2 | 0.961 | 0.924 | | | | | |
| | SQ3 | 0.943 | 0.889 | | | | | |
| | SQ4 | 0.958 | 0.918 | | | | | |
| | SQ5 | 0.921 | 0.848 | | | | | |
| | SQ6 | 0.947 | 0.897 | | | | | |
| | SQ7 | 0.95 | 0.903 | | | | | |
| | SQ8 | 0.926 | 0.857 | | | | | |
| | SQ9 | 0.931 | 0.867 | | | | | |
| | SQ10 | 0.945 | 0.893 | | | | | |
| | SQ11 | 0.939 | 0.882 | | | | | |
| GV × ADI | GV × ADI | 1 | 1 | | | | | |

After that, we analyzed the discriminant validity of the constructs through the *Heterotrait–monotrait ratio* (HTMT). Table 11 represents the HTMT values for all the pairs of constructs as well as the bootstrap confidence intervals. To exam if the HTMT values are significantly different from the threshold value, we ran a bootstrapping procedure to compute the bootstrap confidence intervals. This test focuses on the right tail of the bootstrap distribution to show that the HTMT values are significantly lower than the corre-

sponding threshold value of 0.90 with a 5% probability of error [44]. Since the threshold for the HTMT values is 0.90, only PT, PEU, SQ and AU are above the threshold, meaning that a discriminant validity has not been established between the two reflective constructs. For the other constructs, discriminant validity has been established between the reflective constructs [44].

**Table 11.** Heterotrait–monotrait (HTMT) values.

| | **ADI** | **ANM** | **AS** | **AU** | **GV** | **PEU** | **PT** | **PU** | **SQ** |
|---|---|---|---|---|---|---|---|---|---|
| ADI | | | | | | | | | |
| ANM | 0.488 [0.327; 0.620] | | | | | | | | |
| AS | 0.565 [0.404; 0.696] | 0.795 [0.676; 0.880] | | | | | | | |
| AU | 0.577 [0.417; 0.705] | 0.760 [0.638; 0.856] | 0.844 [0.740; 0.919] | | | | | | |
| GV | 0.211 [0.040; 0.373] | 0.156 [0.040; 0.321] | 0.182 [0.035; 0.340] | 0.153 [0.040; 0.310] | | | | | |
| PEU | 0.584 [0.437; 0.708] | 0.763 [0.665; 0.839] | 0.842 [0.742; 0.907] | 0.811 [0.718; 0.879] | 0.180 [0.037; 0.349] | | | | |
| PT | 0.584 [0.437; 0.708] | 0.830 [0.737; 0.894] | 0.845 [0.733; 0.920] | 0.834 [0.743; 0.902] | 0.200 [0.041; 0.367] | 0.926 [0.876; 0.963] | | | |
| PU | 0.688 [0.547; 0.787] | 0.778 [0.681; 0.851] | 0.813 [0.710; 0.888] | 0.830 [0.732; 0.901] | 0.246 [0.077; 0.405] | 0.898 [0.843; 0.938] | 0.886 [0.824; 0.932] | | |
| SQ | 0.639 [0.498; 0.747] | 0.806 [0.663; 0.894] | 0.907 [0.821; 0.962] | 0.836 [0.695; 0.921] | 0.215 [0.056; 0.371] | 0.834 [0.699; 0.911] | 0.883 [0.757; 0.953] | 0.875 [0.792; 0.928] | |

### *3.2. Structural Model*

To evaluate the structural model, we first needed to check the structural model for collinearity issues by examining the Variance Inflation Factors (VIF) values. Table 12 presents the significance testing results of the structural model path coefficients. Some of the VIF values were above the threshold of five [44], meaning that some of the constructs (PEU → ADI, PT → ADI, PU → ADI and SQ → ADI) presented collinearity among them. Then, we assessed the significance and relevance of the structural model relationships by examining the path coefficients. All the path coefficient values were within the threshold of −1/+1 [44], with the strongest positive relationship between PU and ADI and the strongest negative relationship between ANM and ADI. Then, we ran the bootstrapping procedure to calculate the *t* values, *p* values, 95% BC confidence interval and the significance of the relationships in the structural model. By analyzing Table 12, only two relationships have significance. These relationships were ADI → AU, meaning that the adoption intention affects the actual usage of the chatbot for banking, which confirms H8, and PU → ADI,

meaning that the perceived usefulness affects the adoption intention of the chatbot for banking, which confirms H1.

**Table 12.** Significance testing results of the structural model path coefficients.

| | VIF | Path Coefficient | *t* Values | *p* Values | 95% BC Confidence Interval (with Bias Correction) | Significance ($p < 0.05$)? |
|---|---|---|---|---|---|---|
| ADI → AU | 1.147 | 0.575 | 8.306 | 0.000 | [0.423; 0.697] | YES |
| ANM → ADI | 2.812 | −0.103 | 1.135 | 0.256 | [−0.280; 0.067] | NO |
| AS → ADI | 4.857 | −0.037 | 0.297 | 0.766 | [−0.299; 0.205] | NO |
| GV → AU | 1.056 | −0.039 | 0.515 | 0.607 | [−0.188; 0.106] | NO |
| PEU → ADI | 6.415 | −0.061 | 0.391 | 0.696 | [−0.385; 0.226] | NO |
| PT → ADI | 7.258 | −0.057 | 0.260 | 0.795 | [−0.436; 0.417] | NO |
| PU → ADI | 5.602 | 0.576 | 4.302 | 0.000 | [0.294; 0.827] | YES |
| SQ → ADI | 7.292 | 0.355 | 1.850 | 0.064 | [−0.065; 0.692] | NO |
| GV × ADI → AU | 1.123 | −0.079 | 1.063 | 0.288 | [−0.232; 0.056] | NO |

Then, we turned our attention to the assessment of the model's explanatory power. Table 13 presents the explanatory power of the model. The $R^2$ value of ADI can be considered moderate, whereas the $R^2$ value of AU is rather weak [44]. From Table 8, we can see that ADI has a large effect size of 0.421 on AU, PU has a small effect size of 0.113 on ADI, SQ has a small effect size of 0.033 on ADI, and none of the other predictors have an effect on the outcome with their $f^2$ being less than 0.02 [44,50].

**Table 13.** Explanatory power.

| Predictor | Outcome | $f^2$ | $R^2$ |
|---|---|---|---|
| ADI | | 0.421 | |
| ANM | | 0.007 | |
| AS | ADI | 0.001 | 0.477 |
| GV | | 0.002 | |
| PEU | | 0.001 | |
| PT | AU | 0.001 | 0.315 |
| PU | | 0.113 | |
| SQ | | 0.033 | |
| GV × ADI | | 0.011 | |

To summarize, H1 is confirmed, meaning that perceived usefulness affects the adoption intention of chatbots for banking, agreeing with [7,14,16,18] and contradicting [11]. H2 is not confirmed, meaning that perceived ease of use does not affect the adoption intention of chatbots for banking, agreeing with [11] and contradicting [14,16,18]. H3 is not confirmed, meaning that perceived trust does not affect the adoption intention of chatbots for banking, agreeing with [7] and contradicting [18,51]. H4 is not confirmed, meaning that anthropomorphism does not affect the adoption intention of chatbots for banking, which contradicts [18]. H5 is not confirmed, meaning that the awareness of the service does not affect the adoption intention of chatbots for banking. H6 is not confirmed, meaning that

service quality does not affect the adoption intention of chatbots for banking. H7 is not confirmed, meaning that a gendered voice does not moderate the relationship between the adoption intention and the actual usage of the chatbot for banking. Finally, H8 is confirmed, meaning that the adoption intention affects the actual usage of the chatbot for banking, agreeing with [3,7,17].

### *3.3. Mediation*

To perform the mediation analysis, we tested the significance of the indirect effect of perceived usefulness (PU) via adoption intention (ADI) to actual usage (AU), which is the product of the path coefficients from PU to ADI and from ADI to AU. We ran the bootstrapping routine to test the significance of these path coefficients' product [44]. Table 14 represents the analysis of the direct and indirect effects of the ADI. By analyzing the indirect effect, we conclude that the indirect effect is significant, as the 95% confidence interval does not include zero [44]. The direct effect of ADI exerts a pronounced (0.575) and significant ($p < 0.05$) effect on AU. We then concluded that ADI partially mediates the relationship since both the direct and the indirect effects are meaningful. To further substantiate the type of partial mediation, we computed the product of the direct and indirect effects. Since both are positive, the sign of their product is also positive (i.e., 0.575 * 0.0311 = 0.0179) [44]. Our findings provide empirical support for the mediating role of adoption intention in the studied model. The adoption intention serves as a complementary mediator for the relationship between perceived usefulness and actual usage.

**Table 14.** Analysis of the indirect effects.

| | **Direct Effect** | **95% Confidence Interval (with Bias Correction) of the Direct Effect** | **Significance ($p < 0.05$)?** | **Indirect Effect (via ADI)** | **95% Confidence Interval (with Bias Correction) of the Indirect Effect** | **Significance ($p < 0.05$)?** |
|---|---|---|---|---|---|---|
| ADI → AU | 0.575 | [0.423; 0.697] | YES | | | |
| PU → AU | | | | 0.0331 | [0.159, 0.499] | YES |

### *3.4. Moderation*

To conduct the moderation analysis of the GV in relation to ADI and AU, we first checked the size of the moderating effect. Figure 2 tests whether GV moderates (i.e., changes the strength or direction of) the relationship between ADI and AU. In other words, it examines whether the type or prominence of the chatbot's gendered voice influences how likely people are to actually use the assistant once they intend to adopt it.

The base path coefficient from adoption intention to actual usage (main effect of ADI → AU) is 0.575, which is positive and statistically significant ($p < 0.001$). This indicates that higher adoption intention clearly leads to more actual use of the IVA.

In terms of the moderating effect of GV × ADI, the interaction term (GV × ADI) has a path coefficient of −0.079. This means that when GV increases (i.e., more gendered or human-like voices), the strength of the relationship between ADI and AU slightly decreases. However, the effect is not statistically significant ($p = 0.288$), and the confidence interval includes zero. So, no real moderating effect is confirmed.

Therefore, we can conclude that although there is a small interaction effect, where more gendered voices slightly weaken the influence of adoption intention on usage, the

moderation is not statistically significant. Therefore, a gendered voice does not significantly alter user behavior, and adoption intention remains the dominant predictor of actual usage.

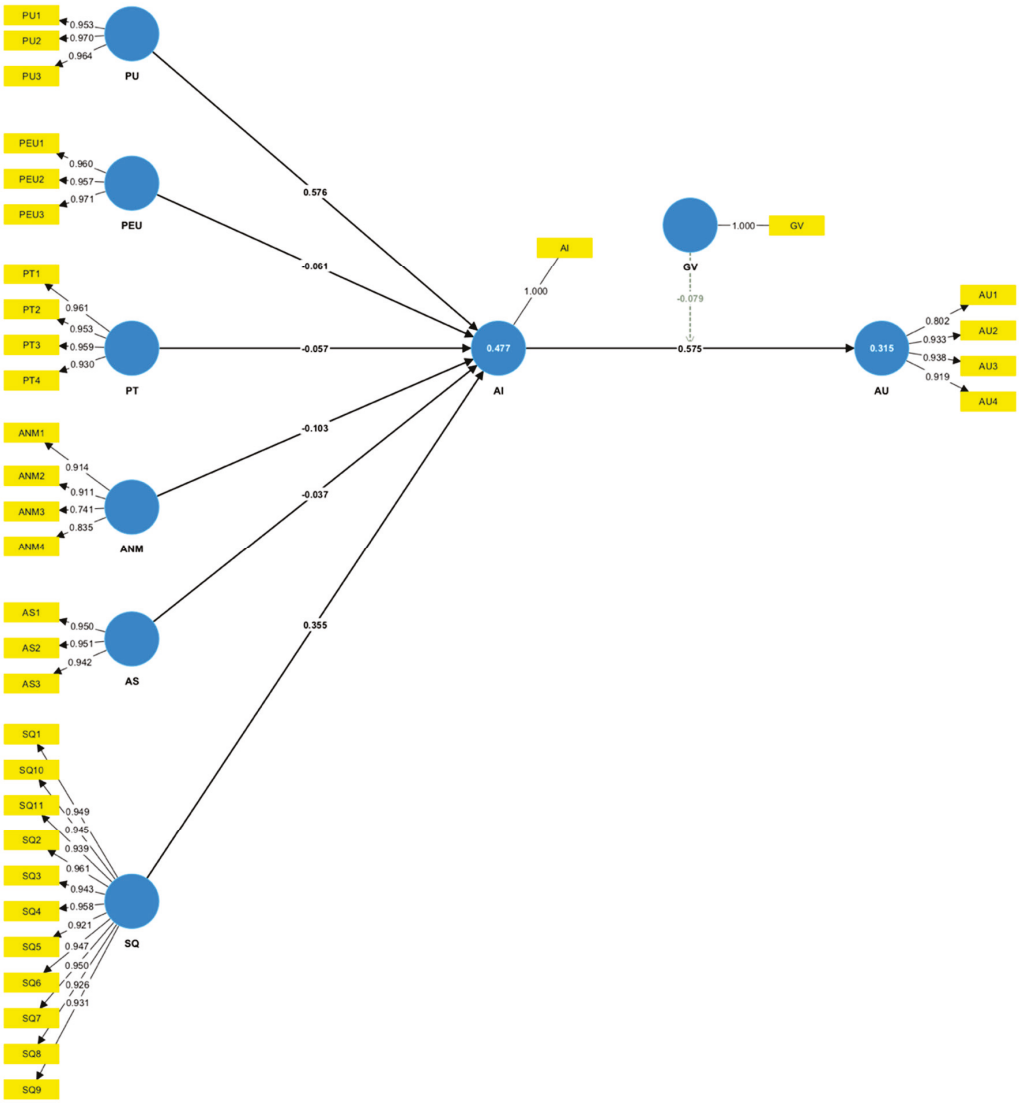



**Figure 2.** Moderator Analysis Results.

Figure 3 illustrates how the strength of the relationship between adoption intention and actual usage varies according to the level of the moderating variable—gendered voice. The graph includes three regression lines, each representing a different level of GV: middle line: average (mean) level of gendered voice; upper line: one standard deviation below the mean (i.e., "lower" GV—possibly gender-neutral or less salient voice); lower line: one standard deviation above the mean (i.e., "higher" GV—more pronounced gendered voice).

All three lines have a positive slope, confirming that higher adoption intention is associated with higher actual usage, regardless of the level of GV. However, the steepness of the lines varies slightly:

- The steepest line (GV below the mean) indicates that when the gendered voice is less prominent, the impact of adoption intention on actual usage is stronger.
- The flattest line (GV above the mean) shows that when the gendered voice is more noticeable or gendered, the impact of adoption intention on usage is slightly weaker.
- The middle line represents the effect at the average GV level.

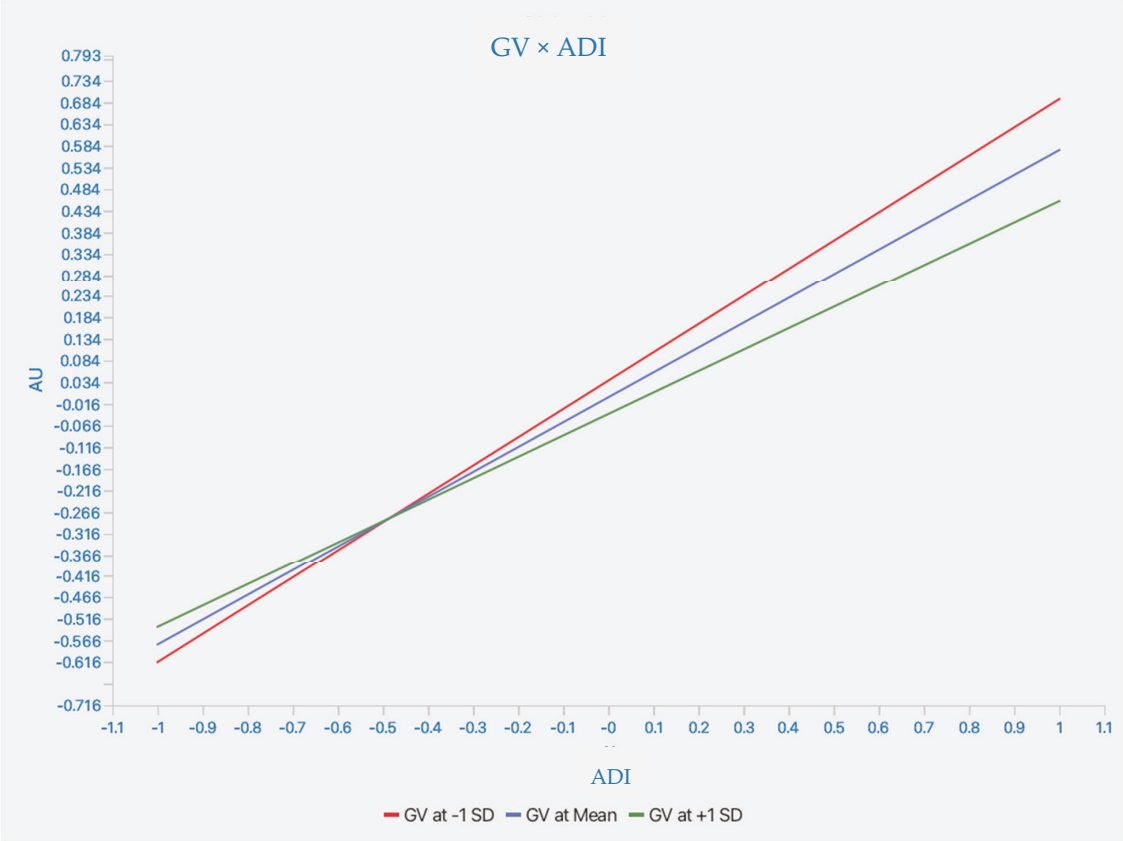



**Figure 3.** Relationship between ADI and AU.

The results of the moderation analysis are provided in Table 15. The moderation effect is insignificant as the $p$ value is more significant than 0.05. Similarly, as the 95% percentile bootstrap bias-corrected confidence interval of the interaction term's effect includes zero, we conclude that the effect is insignificant [44]. This means that a gendered voice does not influence the relationship between the adoption intention and the actual usage of the chatbot for banking.

**Table 15.** Analysis of the moderation.

| | $p$ Value | 95% Confidence Interval (with Bias Correction) of the Indirect Effect | Significance ($p < 0.05$)? |
|---|---|---|---|
| GV × ADI → AU | 0.288 | [−0.232, 0.056] | NO |

## 4. Discussion

This study proposed eight hypotheses to investigate the determinants influencing the adoption intention and actual usage of Intelligent Virtual Assistants (IVAs) in the Portuguese banking sector. Rooted in the Technology Acceptance Model (TAM), the research hypothesized that perceived usefulness (H1), ease of use (H2), trust (H3), anthropomorphism (H4), awareness of the service (H5), and service quality (H6) would affect adoption intention, while gendered voice (H7) would moderate the relationship between adoption intention and actual usage, and that adoption intention would influence actual usage (H8). Using PLS-SEM on data from 154 banking customers, the analysis confirmed only two hypotheses: H1, showing that perceived usefulness significantly impacts adoption intention, and H8, demonstrating that adoption intention positively influences actual usage.

The remaining hypotheses (H2 to H7) were not supported. The novel contribution of this manuscript lies in its contextual focus on a digitally mature market (Portugal) where traditional adoption drivers such as trust, ease of use, and anthropomorphism are no longer significant, and perceived usefulness alone emerges as the central determinant. This insight highlights a shift toward a pragmatic, outcome-oriented evaluation of IVAs, advancing existing adoption models by demonstrating how maturity in user behavior and market infrastructure can reconfigure technology acceptance patterns.

According to the findings, in the context of Portuguese banking, variables such as trust, ease of use, anthropomorphism, awareness, service quality, and gendered voice do not significantly influence the adoption or usage of IVAs. This outcome is largely attributed to the high digital literacy and familiarity with banking technologies among Portuguese consumers, who evaluate these tools mainly based on their practical benefits rather than emotional or aesthetic characteristics [52,53]. Trust in IVAs appears to be derived from the broader trust in established banking institutions, while features like anthropomorphism and gendered voice are either underutilized or less culturally significant in Portugal [54]. Additionally, because many users had prior exposure to IVAs, awareness of the service did not substantially impact usage, and service quality is still strongly associated with human interaction, reducing its perceived relevance in automated contexts [53].

These findings reflect a pragmatic, utility-driven approach to technology adoption in Portugal, where perceived usefulness is the primary determinant of user behavior [52,53]. Culturally, Portuguese users adopt a task-focused attitude toward digital banking, prioritizing efficiency and usefulness over emotional or design features such as anthropomorphism or gendered voices. Technologically, Portugal's high digital literacy and widespread experience with online banking make ease of use and awareness less decisive, as users already expect digital tools to be accessible and available [55]. From a market perspective, the Portuguese banking sector is digitally mature, with IVAs increasingly seen as standard features rather than innovations. In this environment, users assess IVAs primarily based on their utility and ability to improve task performance, making perceived usefulness the only factor with a significant impact [52,53].

To enhance the usefulness of IVAs, Portuguese banks should focus on increasing their practical value in customers' daily financial activities. This includes integrating IVAs with a broader range of services, such as personalized financial advice, transaction automation, account management, and real-time assistance for complex tasks [56]. Ensuring that IVAs are responsive, accurate, and capable of handling nuanced interactions will further improve their effectiveness. Highlighting features that save time or simplify banking processes can also strengthen their perceived value [55,56]. Additionally, continuous learning and adaptation based on user preferences and history can result in more efficient and personalized interactions [54]. Marketing efforts should clearly communicate these advantages, emphasizing how IVAs contribute to convenience and productivity. Ultimately, IVAs should evolve from basic tools to proactive, multifunctional assistants that deliver tangible benefits in banking [52,53].

The findings of this study suggest that banks need to refine both the design and communication strategies of their IVAs to emphasize functionality and value. With perceived usefulness as the principal driver of adoption, design efforts should center on facilitating streamlined interactions and automating tasks [56]. IVAs must be equipped to manage complex banking activities, such as bill payments and budget planning, before they can be regarded as indispensable tools. Communication strategies should focus on clearly demonstrating the capabilities of IVAs, avoiding excessive emphasis on emotional or an-

thropomorphic features, and positioning the IVA as a smart assistant that enhances user control and productivity in banking [52,53].

## 5. Conclusions

This paper performed empirical research focusing on the Portuguese market in the banking sector from where we found the answer for this investigation regarding the research question, "What influences the experience with IVAs in the banking sector?" From this analysis, we highlight the importance of the perceived usefulness of the IVAs adoption intention, supporting H1 and complementing other studies [7,14,16,18]. This demonstrates that other banking companies should focus on the usability of their assistants and provide solutions that help facilitate their customers to complete daily tasks, aiming to give a positive customer experience in every small task. With this, once again, IVAs prove their value by assisting individuals with their daily tasks, increasing their productivity, and freeing them to do other personal actions that add more value to their lives. Together with that, the adoption intention of IVAs for banking also influences the actual usage of this type of technology, supporting H8 and backing up other investigations [3,7,17]. The findings prove that the users are ready and want to use these assistants that enhance their daily lives. These robust IVAs provide another layer for banking customers, giving them advice when needed, offering services in the palm of their hands, and serving as an essential connecting point with the bank. As the banking industry is sometimes seen as complex, developing this type of assistant can also break that assumption and provide a more people-centered experience, offering a way to easily manage their users' finances.

### *5.1. Contributions to Theory and Practice*

This study suggests some theoretical implications and practical contributions for several industries.

In theory, it presents a model to analyze the adoption intention and actual usage of IVAs for banking, serving as a basis for further studies in this area. It also provides the first analysis of the banking sector in Portugal, opening the way for more researchers to investigate this field. Furthermore, we would also like to think that by analyzing the Portuguese sample, more investigators will think of investigating this sample, not only on the topic of IVAs but also on other topics.

For management actions, this paper provides a guide for banking employees to analyze their IVAs and understand what the Portuguese consumer values more for this type of technological solution.

### *5.2. Limitations and Future Studies*

This study is not exclusive of its limitations. To begin, identifying the papers relevant for this study was a significant challenge as the topic is evolving quickly. The next limitation is the lack of many articles focusing on the application of IVAs in the banking sector, which proved to be a challenge as there were not very many terms of comparison for the empirical investigation. Another limitation is that this study was restricted to a single bank in Portugal, which could be mitigated by expanding the research to encompass additional banks across diverse demographics.

This paper identified multiple open questions that could lead to further research on this topic.

To start with the implementation of IVAs in Portugal, future investigators could focus on the research question: How can the experience with IVAs in the retail sector improve the experience with brands in Portugal? For this topic, it would be fundamental for brands

to understand what they could gain from implementing these assistants and to have an investigation that provides a guide on what the essential features are for their consumers. After that, as education has been an area where IVAs have been implemented lately, future research could analyze what influences the users' experience by studying various types of education, whether it is online or remote, at different levels of the academic system and comparing the experience that these different groups of students have with this growing technology. This investigation could focus on the research question: What influences the experience with IVAs in education?

**Author Contributions:** The authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Restrictions apply to the availability of these data. Data were obtained from a third party (banking institution) and are available from the authors with the permission of the third party.

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

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| ADI | Adoption Intention |
| ANM | Anthropomorphism |
| AS | Awareness of the Service |
| AU | Actual Usage |
| GV | Gendered Voice |
| IVA | Intelligent Virtual Assistant |
| PEU | Perceived Ease of Use |
| PLS-SEM | Partial Least Squares Structural Equation Modeling |
| PT | Perceived Trust |
| PU | Perceived Usefulness |
| SQ | Service Quality |
| TAM | Technology Acceptance Model |

## Appendix A

**Table A1.** Measurement items and survey questions.

| Items | Items | Scale |
|---|---|---|
| PU1 | I find the banking chatbot useful in my daily life. | 5-point scale (strongly disagree–strongly agree) |
| PU2 | Using the banking chatbot helps me accomplish things more quickly. | 5-point scale (strongly disagree–strongly agree) |
| PU3 | Using the banking chatbot increases my productivity. | 5-point scale (strongly disagree–strongly agree) |
| PEU1 | Learning how to use the banking chatbot is easy for me. | 5-point scale (strongly disagree–strongly agree) |
| PEU2 | My interaction with the banking chatbot is clear and understandable. | 5-point scale (strongly disagree–strongly agree) |
| PEU3 | I find the banking chatbot easy to use. | 5-point scale (strongly disagree–strongly agree) |
| PT1 | I feel that the information provided by the chatbot for banking is honest and authentic. | 5-point scale (strongly disagree–strongly agree) |
| PT2 | I feel that chatbots for banking have clarity of services provided and frank opinions that are reliable. | 5-point scale (strongly disagree–strongly agree) |

**Table A1.** *Cont.*

| Items | Items | Scale |
|---|---|---|
| PT3 | I feel that chatbots in banking are trustworthy. | 5-point scale (strongly disagree–strongly agree) |
| PT4 | I feel that chatbots for banking have the necessary ability to provide financial counseling. | 5-point scale (strongly disagree–strongly agree) |
| ANM1 | Chatbots for banking have their own mind. | 5-point scale (strongly disagree–strongly agree) |
| ANM2 | Chatbots for banking can experience emotions. | 5-point scale (strongly disagree–strongly agree) |
| ANM3 | I feel that chatbots for banking are inanimate: living. | 5-point scale (strongly disagree–strongly agree) |
| ANM4 | I feel that chatbots for banking are computer-animated: real. | 5-point scale (strongly disagree–strongly agree) |
| AS1 | My bank has communicated a banking chatbot usage policy to me. | 5-point scale (strongly disagree–strongly agree) |
| AS2 | My bank has a strategy regarding the usage of the banking chatbot. | 5-point scale (strongly disagree–strongly agree) |
| AS3 | I have received sufficient information from my bank regarding the usage of the banking chatbot. | 5-point scale (strongly disagree–strongly agree) |
| SQ1 | When I have problems, this chatbot is sympathetic and reassuring. | 5-point scale (strongly disagree–strongly agree) |
| SQ2 | I felt that I could rely on the chatbot's services to fulfill my needs. | 5-point scale (strongly disagree–strongly agree) |
| SQ3 | The chatbot responds promptly to my requests. | 5-point scale (strongly disagree–strongly agree) |
| SQ4 | I trust the chatbot. | 5-point scale (strongly disagree–strongly agree) |
| SQ5 | I feel safe and reassured while having a conversation with the chatbot. | 5-point scale (strongly disagree–strongly agree) |
| SQ6 | The chatbot has adequate knowledge to answer my questions. | 5-point scale (strongly disagree–strongly agree) |
| SQ7 | I can be in control of my personal needs through the chatbot. | 5-point scale (strongly disagree–strongly agree) |
| SQ8 | The chatbot gives me the opportunity to respond. | 5-point scale (strongly disagree–strongly agree) |
| SQ9 | I feel valued by the brand through my conversations with the chatbot. | 5-point scale (strongly disagree–strongly agree) |
| SQ10 | I feel empathetically understood through the conversations with the chatbot. | 5-point scale (strongly disagree–strongly agree) |
| SQ11 | I feel that the chatbot was developed to meet my personal needs. | 5-point scale (strongly disagree–strongly agree) |
| AU1 | I use the chatbot to clarify my doubts. | 5-point scale (strongly disagree–strongly agree) |
| AU2 | I use the chatbot to manage my banking account. | 5-point scale (strongly disagree–strongly agree) |
| AU3 | I use the chatbot to gain knowledge about financial products. | 5-point scale (strongly disagree–strongly agree) |
| AU3 | I use the chatbot to get financial advice. | 5-point scale (strongly disagree–strongly agree) |
| ADI1 | There is a possibility that I will recommend the use of the banking chatbot to my family and friends. | 5-point scale (strongly disagree–strongly agree) |
| GV1 | What kind of voice do you prefer in your personal assistant? | (Male, Female, Without humanization) |
| SQ1 | What is your gender? | (Male, Female, Not Disclosed) |
| SQ2 | What is your educational background? | (Undergraduate, Graduation, Master, PhD) |
| SQ3 | How old are you? | Numerical |
| SQ4 | What is your profession? | Open Answer |
| SQ5 | How would you rate your experience with this virtual assistant from Bank X? | 10-point scale (bad–excellent) |

**Table A1.** *Cont.*

| Items | Items | Scale |
|---|---|---|
| SQ6 | How would you rate your experience with virtual assistants in general? | 10-point scale (bad–excellent) |
| SQ7 | How often do you use the Bank X Contact Center? | 10-point scale (bad–excellent) |

## References


1. Cheng, X.; Bao, Y.; Zarifis, A.; Gong, W.; Mou, J. Exploring consumers' response to text-based chatbots in e-commerce: The moderating role of task complexity and chatbot disclosure. *Internet Res.* **2022**, *32*, 496–517. [CrossRef]
2. Cui, R.M.; Li, M.; Zhang, S.C. Al and Procurement. *Manuf. Serv. Oper. Manag.* **2022**, *24*, 691–706. [CrossRef]
3. Alt, M.A.; Ibolya, V. Identifying Relevant Segments of Potential Banking Chatbot Users Based on Technology Adoption Behavior. *Market-Trziste* **2021**, *33*, 165–183. [CrossRef]
4. Mor, S.; Gupta, G. Artificial intelligence and technical efficiency: The case of Indian commercial banks. *Strateg. Chang.* **2021**, *30*, 235–245. [CrossRef]
5. Alenizi, A.S. Understanding the subjective realties of social proof and usability for mobile banking adoption: Using triangulation of qualitative methods. *J. Islam. Mark.* **2022**, *14*, 2027–2044. [CrossRef]
6. Al-Khasawneh, M.H. A Mobile Banking Adoption Model in the Jordanian Market: An Integration of TAM with Perceived Risks and Perceived Benefits. *J. Internet Bank. Commer.* **2015**, *20*, 1–35.
7. Marakarkandy, B.; Yajnik, N.; Dasgupta, C. Enabling internet banking adoption: An empirical examination with an augmented technology acceptance model (TAM). *J. Enterp. Inf. Manag.* **2017**, *30*, 263–294. [CrossRef]
8. Mutahar, A.M.; Daud, N.M.; Ramayah, T.; Isaac, O.; Aldholay, A.H. The effect of awareness and perceived risk on the technology acceptance model (TAM): Mobile banking in Yemen. *Int. J. Serv. Stand.* **2018**, *12*, 180–204. [CrossRef]
9. Alismaiel, O.A.; Cifuentes-Faura, J.; Al-Rahmi, W.M. Social Media Technologies Used for Education: An Empirical Study on TAM Model During the COVID-19 Pandemic. *Front. Educ.* **2022**, *7*, 882831. [CrossRef]
10. Almogren, A.S.; Aljammaz, N.A. The integrated social cognitive theory with the TAM model: The impact of M-learning in King Saud University art education. *Front. Psychol.* **2022**, *13*, 1050532. [CrossRef]
11. Mailizar, M.; Almanthari, A.; Maulina, S. Examining teachers' behavioral intention to use e-learning in teaching of mathematics: An extended tam model. *Contemp. Educ. Technol.* **2021**, *13*, 1–16. [CrossRef] [PubMed]
12. Prieto, J.C.S.; Miguelánez, S.O.; García-Peñalvo, F.J. Mobile learning adoption from informal into formal: An extended TAM model to measure mobile acceptance among teachers. In Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM '14), Salamanca, Spain, 1–3 October 2014; Association for Computing Machinery: New York, NY, USA, 2014; pp. 595–602. [CrossRef]
13. Sánchez-Prieto, J.C.; Olmos-Migueláñez, S.; García-Peñalvo, F.J. MLearning and pre-service teachers: An assessment of the behavioral intention using an expanded TAM model. *Comput. Hum. Behav.* **2017**, *72*, 644–654. [CrossRef]
14. Dutta, B.; Peng, M.H.; Sun, S.L. Modeling the adoption of personal health record (PHR) among individual: The effect of health-care technology self-efficacy and gender concern. *Libyan J. Med.* **2018**, *13*, 1500349. [CrossRef]
15. Melas, C.D.; Zampetakis, L.A.; Dimopoulou, A.; Moustakis, V. Modeling the acceptance of clinical information systems among hospital medical staff: An extended TAM model. *J. Biomed. Inform.* **2011**, *44*, 553–564. [CrossRef] [PubMed]
16. Tung, F.C.; Chang, S.C.; Chou, C.M. An extension of trust and TAM model with IDT in the adoption of the electronic logistics information system in HIS in the medical industry. *Int. J. Med. Inform.* **2008**, *77*, 324–335. [CrossRef] [PubMed]
17. Mathew, V.; Soliman, M. Does digital content marketing affect tourism consumer behavior? An extension of technology acceptance model. *J. Consum. Behav.* **2021**, *20*, 61–75. [CrossRef]
18. Pillai, R.; Sivathanu, B. Adoption of AI-based chatbots for hospitality and tourism. *Int. J. Contemp. Hosp. Manag.* **2020**, *32*, 3199–3226. [CrossRef]
19. Dahlberg, T.; Guo, J.; Ondrus, J. A critical review of mobile payment research. *Electron. Commer. Res. Appl.* **2015**, *14*, 265–284. [CrossRef]
20. Selamat, M.A.; Windasari, N.A. Chatbot for SMEs: Integrating customer and business owner perspectives. *Technol. Soc.* **2021**, *66*, 101685. [CrossRef]
21. Shaikh, A.A.; Karjaluoto, H. Mobile banking adoption: A literature review. *Telemat. Inform.* **2015**, *32*, 129–142. [CrossRef]
22. Davis, F.D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q. Manag. Inf. Syst.* **1989**, *13*, 319–339. [CrossRef]

23. Orús, C.; Ibáñez-Sánchez, S.; Flavián, C. Enhancing the customer experience with virtual and augmented reality: The impact of content and device type. *Int. J. Hosp. Manag.* **2021**, *98*, 103019. [CrossRef]
24. Yagoda, R.E.; Gillan, D.J. You Want Me to Trust a ROBOT? The Development of a Human-Robot Interaction Trust Scale. *Int. J. Soc. Robot.* **2012**, *4*, 235–248. [CrossRef]
25. You, S.; Kim, J.H.; Lee, S.H.; Kamat, V.; Robert, L.P. Enhancing perceived safety in human–robot collaborative construction using immersive virtual environments. *Autom. Constr.* **2018**, *96*, 161–170. [CrossRef]
26. Bartneck, C.; Kulić, D.; Croft, E.; Zoghbi, S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* **2009**, *1*, 71–81. [CrossRef]
27. Noor, N.; Rao Hill, S.; Troshani, I. Recasting Service Quality for AI-Based Service. *Australas. Mark. J.* **2022**, *30*, 297–312. [CrossRef]
28. Bouhia, M.; Rajaobelina, L.; PromTep, S.; Arcand, M.; Ricard, L. Drivers of privacy concerns when interacting with a chatbot in a customer service encounter. *Int. J. Bank Mark.* **2022**, *40*, 1159–1181. [CrossRef]
29. Al-Somali, S.A.; Gholami, R.; Clegg, B. An investigation into the acceptance of online banking in Saudi Arabia. *Technovation* **2009**, *29*, 130–141. [CrossRef]
30. Sathye, M. Adoption of Internet Banking by Australian Consumers: An Empirical Investigation. *Int. J. Bank Mark.* **1999**, *17*, 324–334. [CrossRef]
31. Lishomwa, L.; Phiri, J. Adoption of Internet Banking Services by Corporate Customers for Forex Transactions Based on the TRA Model. *Open J. Bus. Manag.* **2020**, *08*, 329–345. [CrossRef]
32. Chen, Q.; Gong, Y.; Lu, Y.; Tang, J. Classifying and measuring the service quality of AI chatbot in frontline service. *J. Bus. Res.* **2022**, *145*, 552–568. [CrossRef]
33. Kushwaha, A.K.; Kumar, P.; Kar, A.K. What impacts customer experience for B2B enterprises on using AI-enabled chatbots? Insights from Big data analytics. *Ind. Mark. Manag.* **2021**, *98*, 207–221. [CrossRef]
34. Tran, A.D.; Pallant, J.I.; Johnson, L.W. Exploring the impact of chatbots on consumer sentiment and expectations in retail. *J. Retail. Consum. Serv.* **2021**, *63*, 102718. [CrossRef]
35. Malodia, S.; Islam, N.; Kaur, P.; Dhir, A. Why Do People Use Artificial Intelligence (AI)-Enabled Voice Assistants. *IEEE Trans. Eng. Manag.* **2021**, *71*, 491–505. [CrossRef]
36. Fan, H.; Han, B.; Gao, W.; Li, W. How AI chatbots have reshaped the frontline interface in China: Examining the role of sales–service ambidexterity and the personalization–privacy paradox. *Int. J. Emerg. Mark.* **2022**, *17*, 967–986. [CrossRef]
37. Borau, S.; Otterbring, T.; Laporte, S.; Fosso Wamba, S. The most human bot: Female gendering increases humanness perceptions of bots and acceptance of AI. *Psychol. Mark.* **2021**, *38*, 1052–1068. [CrossRef]
38. Feine, J.; Gnewuch, U.; Morana, S.; Maedche, A. Gender Bias in Chatbot Design. In *Chatbot Research and Design. CONVERSATIONS 2019*; Følstad, A., Araujo, T., Papadopoulos, S., Law, E.L.-C., Granmo, O.-C., Luger, E., Brandtzaeg, P.B., Eds.; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Cham, Switzarland, 2020; Volume 11970, pp. 79–93. [CrossRef]
39. Mooshammer, S.; Etzrodt, K. Social Research with Gender-Neutral Voices in Chatbots—The Generation and Evaluation of Artificial Gender-Neutral Voices with Praat and Google WaveNet. In *Chatbot Research and Design Workshop; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Cham, Switzarland, 2022; Volume 13171, pp. 176–191. [CrossRef]
40. Gu, J.C.; Lee, S.C.; Suh, Y.H. Determinants of behavioral intention to mobile banking. *Expert Syst. Appl.* **2009**, *36*, 11605–11616. [CrossRef]
41. Mutambik, I.; Lee, J.; Almuqrin, A.; Zhang, J.Z.; Baihan, M.; Alkhanifer, A. Privacy Concerns in Social Commerce: The Impact of Gender. *Sustainability* **2023**, *15*, 12771. [CrossRef]
42. Abdennebi, H.B. M-banking adoption from the developing countries perspective: A mediated model. *Digit. Bus.* **2023**, *3*, 100065. [CrossRef]
43. Dhiman, N.; Jamwal, M.; Kumar, A. Enhancing value in customer journey by considering the (ad)option of artificial intelligence tools. *J. Bus. Res.* **2023**, *167*, 114142. [CrossRef]
44. Hair, J.F.; Hult, G.T.M.; Ringle, C.M.; Sarstedt, M. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*; SAGE Publications: Thousand Oaks, CA, USA, 2022.
45. Yun, J.; Park, J. The Effects of Chatbot Service Recovery with Emotion Words on Customer Satisfaction, Repurchase Intention, and Positive Word-of-Mouth. *Front. Psychol.* **2022**, *13*, 922503. [CrossRef] [PubMed]
46. Caycho-Rodríguez, T.; Tomás, J.M.; Yupanqui-Lorenzo, D.E.; Valencia, P.D.; Carbajal-León, C.; Vilca, L.W.; Ventura-León, J.; Paredes-Angeles, R.; Arias Gallegos, W.L.; Reyes-Bossio, M.; et al. Relationship Between Fear of COVID-19, Conspiracy Beliefs About Vaccines and Intention to Vaccinate Against COVID-19: A Cross-National Indirect Effect Model in 13 Latin American Countries. *Eval. Health Professions.* **2023**, *46*, 371–383. [CrossRef] [PubMed]

47. Charles-Leija, H.; Castro, C.G.; Toledo, M.; Ballesteros-Valdés, R. Meaningful Work, Happiness at Work, and Turnover Intentions. *Int. J. Environ. Res. Public Health* **2023**, *20*, 3565. [CrossRef] [PubMed]
48. Fishman, J.; Lushin, V.; Mandell, D.S. Predicting implementation: Comparing validated measures of intention and assessing the role of motivation when designing behavioral interventions. *Implement. Sci. Commun.* **2020**, *1*, 81. [CrossRef]
49. Matteau, L.; Toupin, I.; Ouellet, N.; Beaulieu, M.; Truchon, M.; Gilbert-Ouimet, M. Nursing students' academic conditions, psychological distress, and intention to leave school: A cross-sectional study. *Nurse Educ. Today* **2023**, *129*, 105877. [CrossRef]
50. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Routledge: New York, NY, USA, 1988. [CrossRef]
51. Tanihatu, M.C.; Suryanto; Syahchari, D.H. H. Technologies (ICICT), Lalitpur, Nepal, 26–28 April 2023; pp. 846–851. [CrossRef]
52. Afonso, V.O. The adoption of conversational assistants in the banking industry. *Heliyon* **2023**, *9*, e17428.
53. Silva, M.; Pereira, J.; Costa, P. Factors influencing the adoption of Intelligent Virtual Assistants in Portuguese banking. *J. Digit. Bank. Res.* **2023**, *12*, 45–61.
54. Bhanu, P.; Vivek, S. Exploring users' adoption intentions of intelligent virtual assistants in financial services: An anthropomorphic perspectives and socio-psychological perspectives. *Comput. Hum. Behav.* **2023**, *146*, 107912.
55. Foulds, S. *How Virtual Assistants Take Mobile Banking Apps to the Next Level*; The Financial Brand: Gig Harbor, WA, USA, 2023.
56. James, C. *The CX Benefits of Implementing an Intelligent Virtual Agent (IVA)*; CMSWire: San Francisco, CA, USA, 2022.


**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.

computers

MDPI

Article

# Test-Time Training with Adaptive Memory for Traffic Accident Severity Prediction

**Duo Peng and Weiqi Yan ***

Department of Computer and Information Science, Auckland University of Technology, Auckland 1010, New Zealand; pnt7686@autuni.ac.nz
* Correspondence: weiqi.yan@aut.ac.nz

**Abstract:** Traffic accident prediction is essential for improving road safety and optimizing intelligent transportation systems. However, deep learning models often struggle with distribution shifts and class imbalance, leading to degraded performance in real-world applications. While distribution shift is a common challenge in machine learning, Transformer-based models—despite their ability to capture long-term dependencies—often lack mechanisms for dynamic adaptation during inferencing. In this paper, we propose a TTT-Enhanced Transformer that incorporates Test-Time Training (TTT), enabling the model to refine its parameters during inferencing through a self-supervised auxiliary task. To further boost performance, an Adaptive Memory Layer (AML), a Feature Pyramid Network (FPN), Class-Balanced Attention (CBA), and Focal Loss are integrated to address multi-scale, long-term, and imbalance-related challenges. Our experimental results show that our model achieved an overall accuracy of 96.86% and a severe accident recall of 95.8%, outperforming the strongest Transformer baseline by 5.65% in accuracy and 9.6% in recall. The results of our confusion matrix and ROC analyses confirm our model's superior classification balance and discriminatory power. These findings highlight the potential of our approach in enhancing real-time adaptability and robustness under shifting data distributions and class imbalances in intelligent transportation systems.

**Keywords:** test-time training; traffic accident prediction; transformer network; self-supervised learning; adaptive memory; class-balanced learning

## 1. Introduction

Traffic accidents remain a critical public safety concern, resulting in significant human and economic losses worldwide. According to the World Health Organization (WHO), more than 1.35 million people lose their lives in road accidents annually, with economic damages exceeding USD 800 billion in the United States alone. The increasing complexity of urban transportation systems, driven by rising population densities, evolving infrastructure, and unpredictable human behavior, necessitates advanced, data-driven solutions for improving traffic safety. Deep learning models, particularly Transformer architectures, have demonstrated remarkable performance in modeling sequential data [1–3]. However, their effectiveness in real-world applications is often hindered by distribution shifts and class imbalance, two fundamental challenges that degrade prediction reliability [4,5].

Most existing traffic accident prediction models are trained on static datasets and struggle to generalize when deployed in dynamic, real-world scenarios. As a result, their predictions deteriorate under distribution shifts, where unseen environmental conditions

lead to significant accuracy degradation. To address this, Test-Time Training (TTT) has been introduced, enabling Transformer-based models to dynamically refine their parameters during inferencing using self-supervised auxiliary tasks [6]. This approach has shown success in various applications, such as handwritten document recognition, where an auxiliary branch continuously updates model parameters for enhanced adaptability [7]. Recent advancements like Test-Time Self-Training (TeST) have further improved test-time adaptation by employing a student–teacher framework, allowing models to learn robust and invariant representations under distribution shifts [8]. Comprehensive studies have emphasized the significance of TTT for handling such shifts across multiple domains [9].

Traffic accident prediction models often struggle with class imbalance, where severe but infrequent accidents are underrepresented in training data, leading to biased predictions that favor more common, less severe events while failing to recognize high-risk scenarios [5,10]. An analysis of the dataset in this paper reveals a highly skewed long-tail distribution of accident severity, as illustrated in Figure 1. The data show that moderate accidents (Level 2) constitute 79.6% of all cases, whereas severe accidents (Level 4) represent only 2.6%. The maximum-to-minimum imbalance ratio is as high as 93.2:1, highlighting the extreme disparity between frequent minor incidents and rare but critical severe accidents. To further characterize the Severity (levels 1–4) distribution and rule out any inadvertent train–test shifts, we computed skewness, kurtosis, and Jensen–Shannon divergence on the labels in both splits, as summarized in Table 1. These statistics confirm that the training and testing severity labels share virtually identical higher-order moments (JS divergence $\approx 0.00085$), indicating that the imbalance challenge stems from the inherent long-tail distribution rather than any split mismatch. This imbalance causes conventional models to prioritize majority classes, resulting in poor recall and frequent misclassification of severe accidents—the most critical category for traffic safety interventions [11,12]. Addressing this issue requires models that can adapt to the underlying distribution shifts while ensuring fair representation of minority accident types.

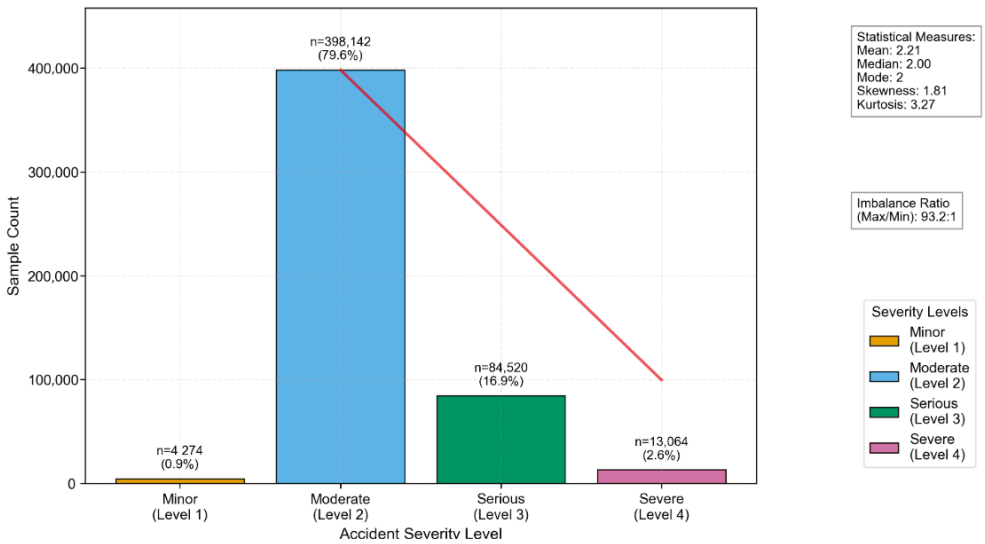



**Figure 1.** Distribution of accident severity levels in our dataset.

Previous studies address this challenge using Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, which improved temporal modeling but suffered from vanishing gradient issues and failed to capture long-range dependencies effectively [13]. The emergence of Transformer models revolutionized sequence modeling by leveraging self-attention mechanisms, enabling better long-term dependency modeling

in compared to RNNs and LSTMs [2,14]. However, even Transformer-based models remain vulnerable to class imbalance, leading to biased predictions toward frequent accident categories [5,11,15].

**Table 1.** Severity label distribution statistics.

| Feature | Skew Train | Skew Test | Kurt Train | Kurt Test | JS Divergence |
|---|---|---|---|---|---|
| Severity (1–4) | 1.812 | 1.808 | 6.271 | 6.26 | 0.000085 |

Recent advancements in self-supervised learning and adaptive training provide promising solutions to addressing distribution shifts and class imbalance [5,6,16]. While Test-Time Training (TTT) has demonstrated significant potential to dynamically adapt models during inferencing, the existing research has mainly focused on structured environments and standard classification tasks, with limited exploration of real-time traffic accident prediction [16–18]. To date, there is no unified framework explicitly integrating TTT with advanced memory mechanisms, multi-scale feature extraction, and class-imbalance-aware strategies tailored specifically to traffic accident severity prediction. Such integration is crucial, as existing models continue to struggle to accurately differentiate between frequent minor accidents and rare, severe cases, resulting in biased predictions and insufficient recall on critical accident scenarios [5,11,15]. Therefore, there is still an open research gap regarding a comprehensive approach that simultaneously addresses these challenges.

It should be noted that this study focuses on predicting the severity level of reported accidents rather than forecasting accident occurrences. This formulation is consistent with prior work on probabilistic risk modeling and supports more actionable safety interventions based on the severity of traffic incidents.

To address these challenges, in this paper, we introduce a unified Transformer-based framework that incorporates Test-Time Training (TTT) as the core adaptation mechanism for traffic accident severity prediction. TTT dynamically refines model parameters during inferencing, improving real-time adaptability and robustness to distribution shifts. Additionally, we integrate an Adaptive Memory Layer (AML) to retain long-term dependencies [19,20] and a Feature Pyramid Network (FPN) to enhance multi-scale feature extraction [21,22]. To specifically mitigate class imbalance, the proposed model combines Class-Balanced Attention (CBA), class-weighted cross-entropy, Focal Loss, and SMOTE-based oversampling [15,23]. Unlike prior approaches, in which these strategies are treated independently, our cohesive architecture simultaneously addresses distribution shifts and class imbalance at t data level, loss level, and model level. Our experimental results and ablation studies confirm that our model achieved significant improvements in robustness, inference efficiency, and recall for severe and rare accident scenarios, highlighting its suitability for real-time intelligent transportation systems.

## 2. Materials and Methods

We designed our study according to a structured methodology inspired by established workflows used in traffic accident prediction and machine learning research [1,24], aiming to develop a scalable and adaptive prediction framework. Specifically, our approach integrates deep learning models within a systematic workflow to enhance predictive performance. The subsequent subsections detail our research design, data-processing pipeline, model architecture, training procedures, evaluation metrics, and ablation studies conducted to assess the contributions of the different components of the model.

### 2.1. Data Collection

The dataset, sourced from Kaggle (DOI:10.34740/kaggle/ds/199387, available at https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents/data (accessed on 1 January 2025)), comprises traffic accident records from 49 states across the United States, spanning the years 2016 to 2023. A subset of 500,000 records was selected for model training and evaluation; these records included information on meteorological conditions, traffic density, road types, and accident severity. Accident severity, the target variable, is categorized into four levels: minor (minimal impact), moderate (traffic delays without major disruption), severe (significant congestion and possible injuries), and extreme (major disruptions with serious injuries or fatalities). As illustrated in Figure 1, the dataset exhibits a highly imbalanced distribution, where moderate accidents dominate, while severe and extreme cases are significantly underrepresented. This predefined classification serves as the foundation for training models to distinguish varying levels of accident severity.

### 2.2. Data Preprocessing

A structured data pre-processing pipeline was developed to ensure consistency and enhance predictive performance. The features with more than 30% missing values were removed, while those with lower missing rates were imputed using mean or median values. To improve feature representation, temporal attributes such as hour, weekday/weekend, and seasonality were extracted to capture variations in accident risk, while geospatial factors, including proximity to highways, intersections, and traffic signals, were incorporated to identify accident hotspots. Interaction terms, such as temperature–visibility and humidity–wind speed, were introduced to account for environmental dependencies. Numerical attributes were standardized using Min–Max Scaling, and categorical variables, including weather conditions and accident locations, were encoded using one-hot encoding to prevent artificial ordinal relationships. Additionally, composite features, such as traffic density and weather impact scores, were derived to better capture patterns associated with accident severity. The complete pre-processing workflow, covering data cleaning, feature engineering, standardization, and class balancing, is illustrated in Figure 2.

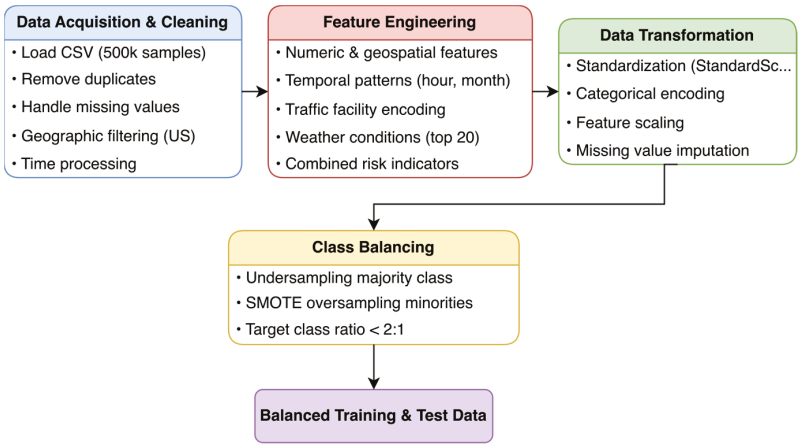



**Figure 2.** Overview of the data pre-processing pipeline, detailing data cleaning, filtering, feature engineering, and balancing steps.

### 2.3. Model Architecture

The proposed model extends the baseline Transformer by incorporating multi-scale feature extraction, adaptive memory, and TTT, as shown in Figure 3. It consists of four key components: Feature Pyramid Network (FPN) for capturing hierarchical traffic patterns,

Adaptive Memory Layer (AML) for retaining long-term dependencies, Class-Balanced Attention (CBA) for mitigating class imbalance, and TTT for real-time adaptation.

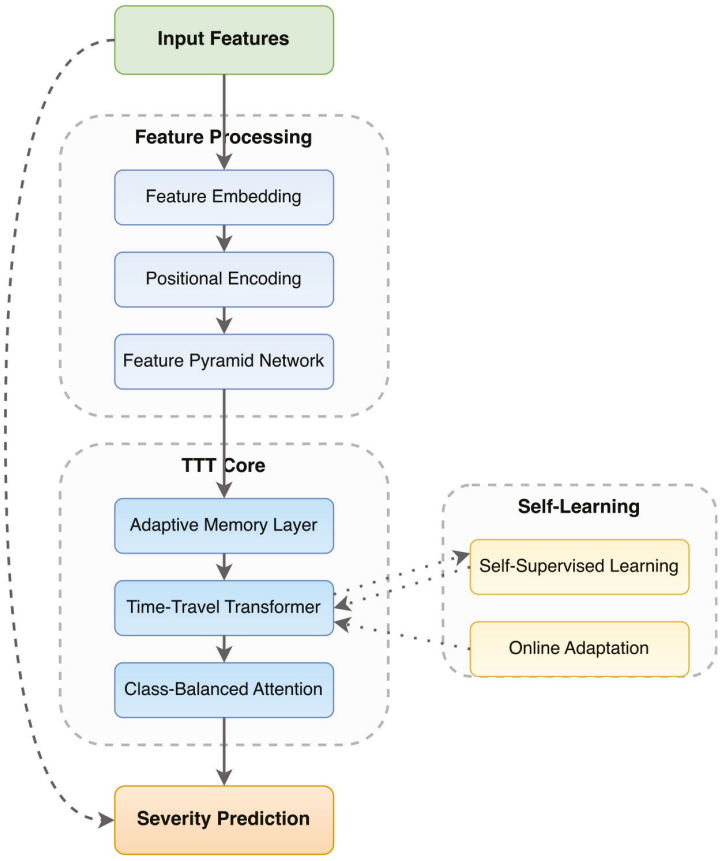



**Figure 3.** Model architecture of and training optimization process for the TTT-Enhanced Transformer for traffic accident prediction.

2.3.1. Multi-Scale Representation Learning via Feature Pyramid Network (FPN)

Traffic patterns exhibit hierarchical structures, wherein localized accident features interact with broader contextual influences. Standard Transformer models operate at a fixed resolution, potentially overlooking critical multi-scale dependencies. To address this limitation, a Feature Pyramid Network (FPN) is integrated to aggregate features across multiple spatial and temporal resolutions [21,25,26].

An FPN processes input feature maps at different scales—small, medium, and large—capturing both fine-grained accident characteristics and broader traffic patterns. Given an input feature map $X$ at three different scales $X_{small}$, $X_{medium}$, and $X_{large}$, the fused representation is computed as

$$X_{FPN} = W_1 X_{small} + W_2 X_{medium} + W_3 X_{large} \tag{1}$$

where $W_1$, $W_2$, and $W_3$ are learnable attention weights. Each scale-specific feature map is processed through 1D Convolutional Layers to refine temporal dependencies, enabling the model to retain both fine-grained and high-level accident patterns. Figure 4 presents the detailed structure of the Feature Pyramid Network (FPN) used in this study.

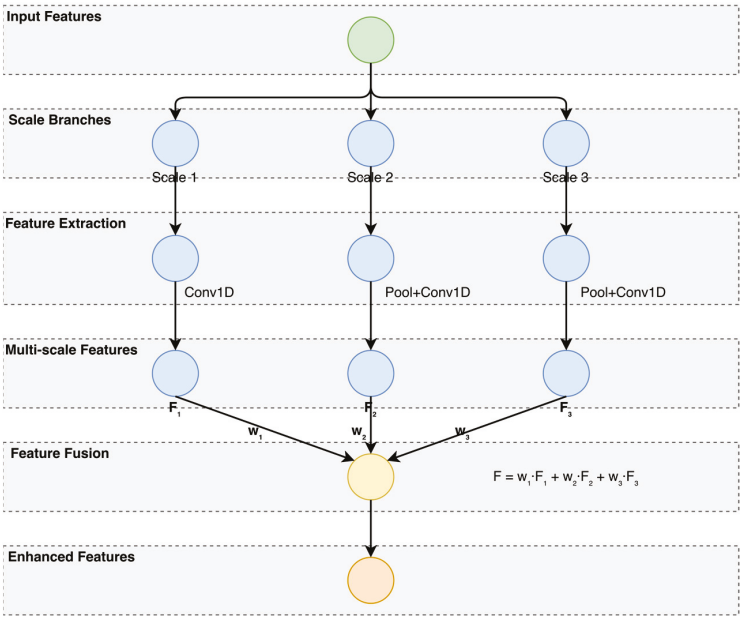



**Figure 4.** Schematic representation of the Feature Pyramid Network (FPN), showing the multi-scale processing pipeline, mathematical formulation, and theoretical guarantees.

### 2.3.2. Adaptive Memory Layer (AML) for Long-Term Dependency Modeling

Standard Transformers struggle to maintain long-term dependencies due to their fixed-length context windows, constituting a particularly problematic shortcoming for traffic accident prediction, where past incidents influence future risks [3,19,20]. To address this, an Adaptive Memory Layer (AML) introduces an external memory module that dynamically retains and updates contextual information, ensuring that historical patterns are effectively incorporated into inferencing [19,20]. At each timestep $t$, the memory state $M_t$ is updated recursively to maintain temporal continuity:

$$M_t = \gamma M_{t-1} + (1 - \gamma) MLP(X_t) \tag{2}$$

where $\gamma$ is a learnable decay factor that controls the balance between retaining past memory and incorporating new information, while $MLP(X_t)$ is a non-linear transformation extracting key accident-related features from the current input $X_t$.

$$A_{\text{AML}} = \text{softmax}(W_m M_t + b) \tag{3}$$

$$X_{\text{AML}} = A_{\text{AML}} \cdot M_t \tag{4}$$

where $W_m$ and $b$ are trainable parameters that determine which memory components are most relevant. The attention weight $A_{\text{AML}}$ selectively emphasizes critical historical patterns while filtering out less significant information [20].

Unlike standard self-attention, which primarily captures short-range dependencies [3], the proposed Adaptive Memory Layer (AML) is designed to maintain a dedicated memory state so that essential historical information remains accessible during inferencing [19]. The learnable decay $\gamma$ allows the model to adapt dynamically to traffic conditions, balancing recent and historical accident data [27,28]. We expect this mechanism will help the model recognize recurring traffic patterns and potentially improve prediction reliability, especially in accident-prone areas, where past incidents can offer useful predictive cues. Figure 5 de-

picts the AML architecture, including its three key components—Memory Representation, Multi-Head Attention for Retrieval, and the Memory Update Mechanism [19,20,29].

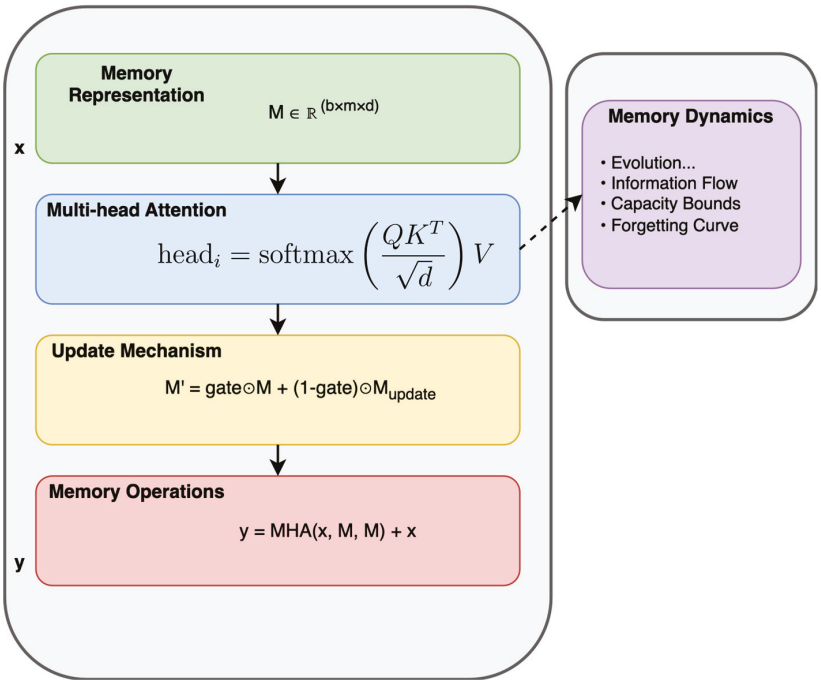



**Figure 5.** Adaptive Memory Layer (AML) architecture.

#### 2.3.3. Class-Balanced Attention (CBA) for Class Imbalance Mitigation

Accident severity levels exhibit a long-tail distribution (Figure 1), where severe accidents are significantly underrepresented [30]. Conventional Transformers tend to focus on frequent accident types, leading to biased predictions. To counteract this, we introduce Class-Balanced Attention (CBA), which dynamically adjusts attention weights based on class importance [23]. For each accident class $c$, the attention weight is computed as

$$A_c = (exp(W_c X))/\left(\sum_{j=1}^{N} exp(W_j X)\right) \tag{5}$$

where $W_c$ is the learnable class importance weight, and $N$ is the total number of classes. This formulation ensures that underrepresented accident categories receive more attention, thereby improving the model's robustness against class imbalance [15]. The computed attention weights $A_c$ are then applied in the Transformer decoder to reweight accident severity predictions, ensuring that the model focuses adequately on severe accidents despite their lower rates of occurrence.

#### 2.3.4. Test-Time Training (TTT) for Online Adaptation

Deep learning models often fail to adapt to dynamic traffic conditions due to their reliance on static training data [6,16]. Traditional Transformers assume a fixed data distribution, making them vulnerable to distribution shifts in real-world traffic scenarios. Test-Time Training (TTT) addresses this challenge by enabling real-time model updates through an auxiliary self-supervised learning (SSL) task [8,9]. Unlike conventional models, which remain unchanged after training, TTT continuously refines model parameters during inferencing, mitigating distribution shifts and enhancing predictive robustness [18]. The

optimization objective consists of classification loss $L_{CE}$ and self-supervised loss $L_{SSL}$, encouraging better generalization beyond the training set [16]:

$$L_{TTT} = L_{CE} + \lambda L_{SSL} \tag{6}$$

During inferencing, the model continuously refines its parameters based on incoming traffic data:

$$\theta_{t+1} = \theta_t - \alpha \nabla_\theta L_{TTT}(X_t) \tag{7}$$

where $\alpha$ is the adaptive learning rate, $X_t$ represents the current accident data input, and $\theta$ denotes the Transformer's parameters [8]. This iterative update mechanism enables continuous adaptation to evolving traffic patterns, ensuring robustness in highly dynamic environments. To further enhance adaptability, TTT prioritizes recent accident data, adjusting feature importance weights as follows:

$$w_f^{(t)} = w_f^{(t-1)} + \eta \cdot \nabla_w L_f(X_t) \tag{8}$$

where $\eta$ is the learning rate for feature weight updates [16]. This ensures the model focuses on the most relevant and time-sensitive accident indicators while filtering outdated information.

Additionally, TTT integrates an online memory retention mechanism, allowing the model to store and retrieve historical accident patterns [29]. By leveraging this memory, the model improves predictive accuracy in non-stationary environments where traffic risks evolve over time [25].

Figure 6 illustrates the TTT framework, detailing the interaction between self-supervised learning, online parameter updates, and memory-based adaptation. The diagram highlights how the Transformer encoder, in conjunction with a self-supervised prediction module, iteratively refines model parameters until convergence, ensuring optimal real-time adaptation [6,16]. For reproducibility, the pseudocode of the proposed method is provided in Appendix A (Algorithm A1).

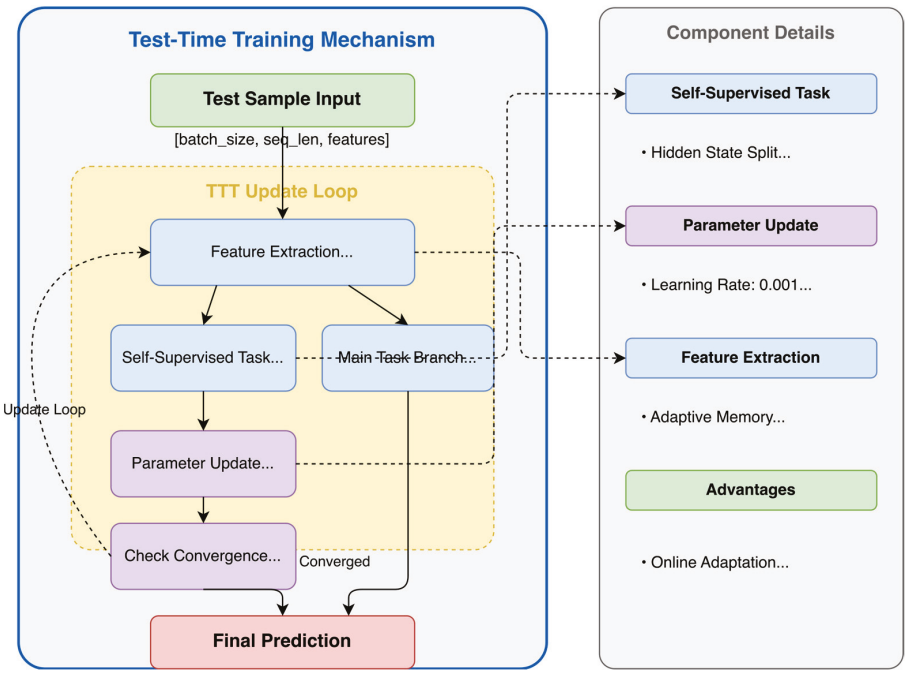



**Figure 6.** TTT framework for online adaptation.

2.3.5. Loss Function for Imbalanced Classification

In real-world accident data, severe accidents (Levels 3 and 4) are underrepresented. Traditional loss functions treat all samples equally, leading to a bias toward majority classes (minor accidents) [5,11,15]. This model adopts Focal Loss, defined as follows:

$$L_{\text{Focal}} = -\alpha_t(1 - p_t)^{\gamma} log(p_t) \tag{9}$$

where $\alpha_t$ is a class-dependent weighting factor, $p_t$ is the model's predicted probability for the correct class, and $\gamma$ is the focusing parameter, which reduces the impact of well-classified examples.

*2.4. Our Experiments*

To rigorously assess the effectiveness of the proposed TTT-Enhanced Transformer, a comprehensive experimental evaluation was conducted. In these experiments, we compared the proposed approach against existing models, investigated the contribution of key components through ablation studies, and analyzed performance using multiple evaluation metrics [6,16,31].

2.4.1. Baseline Model Comparisons

The evaluation framework includes multiple baselines for comparison. A Long Short-Term Memory (LSTM) network was selected due to its strong ability to model temporal dependencies in sequential data [32–34], serving as a traditional deep learning benchmark. Additionally, we implemented a Standard Transformer baseline to quantify the impact of our proposed enhancements [1,2,14]. To ensure this baseline was sufficiently competitive, we applied a comprehensive three-step class imbalance mitigation pipeline comprising (i) random undersampling of the majority class (Severity 2), (ii) SMOTE-based oversampling of minority classes (Severities 1, 3 and 4), and (iii) cost-sensitive training using class-weighted cross-entropy loss with label smoothing. This setup ensures that the baseline addresses class imbalance at the data level, loss level, and training strategy level, providing a strong foundation for fair comparison with our TTT-Enhanced framework.

To further isolate the contributions of each individual mitigation technique, we introduced two additional variants: Transformer + Only SMOTE, which applies synthetic oversampling without undersampling or weighted loss [5], and Transformer + Only Class-Weighted CE, which implements cost-sensitive learning without data-level balancing [35]. These ablation models help quantify the incremental benefit of each component.

In addition, we compare the results against those obtained using TabM, a recent state-of-the-art tabular deep learning architecture designed for structured data [36]. All models are evaluated using standard metrics such as overall accuracy, weighted precision, weighted recall, and weighted F1-score [11,12].

2.4.2. Ablation Studies

Ablation studies were conducted to quantify the contribution of individual components and training strategies within the proposed TTT-Enhanced Transformer [18,31]. In addition to systematically removing key modules—such as Test-Time Training (TTT), Adaptive Memory Layer (AML), Feature Pyramid Network (FPN), Class-Balanced Attention (CBA), and Focal Loss—we included two additional configurations to assess the isolated impact of CBA under data imbalance conditions, specifically without applying any sampling methods. All ablation variants were trained and evaluated under consistent experimental settings to enable fair comparison.

2.4.3. Evaluation Metrics

The performance of all models is assessed using multiple evaluation metrics to ensure a comprehensive understanding of predictive capabilities [6,29]. Overall accuracy measures the proportion of correct predictions [6,37], while weighted precision accounts for class imbalance by ensuring performance is evaluated fairly across different accident severity levels [5]. Weighted recall evaluates a model's ability to correctly identify severe accidents, adjusting for class imbalance [5,11]. The weighted F1-score, as the harmonic mean of precision and recall, ensures there is a balanced assessment of model performance across different severity levels [5,11]. The ROC-AUC score provides insight into a model's discrimination ability across multiple severity categories [11,12]. A confusion matrix further analyzes classification errors and misclassifications, highlighting the challenges posed by imbalanced accident severity distributions [5].

## 3. Results

The experimental results provide a comprehensive evaluation of the TTT-Enhanced Transformer, demonstrating its superior performance compared to traditional deep learning models. The model's effectiveness is assessed through a comparison with baseline architectures, a detailed analysis of classification accuracy across accident severity levels, and an ablation study to quantify the contributions of key components.

### *3.1. Model Performance Comparison*

To comprehensively evaluate the effectiveness of the proposed TTT-Enhanced Transformer, we compared its performance against several baseline models and the recent state-of-the-art (SOTA) TabM architecture.

As shown in Table 2, the TTT-Enhanced Transformer consistently outperforms all baseline and comparison models across every evaluation metric. Compared to the LSTM baseline, which suffers from limited temporal modeling capabilities and heightened sensitivity to class imbalance (overall accuracy: 0.4798, F1-score: 0.55), the TTT-Enhanced Transformer achieved substantial improvements—particularly in recall (0.96 vs. 0.47)—demonstrating its effectiveness in identifying rare and severe accident cases [5,33].

**Table 2.** Comparison of the performances of deep learning baselines, imbalance-aware variants, a SOTA tabular model (TabM), and the proposed TTT-Enhanced Transformer.

| Model | Overall Accuracy | Weighted Precision | Weighted Recall | Weighted F1-Score |
|---|---|---|---|---|
| LSTM | 0.4798 | 0.82 | 0.47 | 0.55 |
| Standard Transformer (Undersampling + SMOTE + Weighted CE) | 0.9120 | 0.93 | 0.91 | 0.92 |
| Transformer + Only SMOTE | 0.905 | 0.91 | 0.93 | 0.915 |
| Transformer + Only Class-Weighted CE | 0.920 | 0.94 | 0.89 | 0.915 |
| TabM (SOTA) | 0.925 | 0.94 | 0.92 | 0.93 |
| TTT-Enhanced Transformer | 0.9686 | 0.97 | 0.96 | 0.96 |

The Transformer (Undersampling + SMOTE + Weighted CE) performed markedly better than LSTM due to its superior sequence modeling and multi-level class imbalance handling (overall accuracy: 0.9120, F1-score: 0.92), yet it lacks adaptation during inferencing time, making it vulnerable to real-world distributional shifts [3,5,17]. The ablated variants—Transformer + Only SMOTE and Transformer + Only Class-Weighted CE—exhibited partial improvements but fell short of the balanced performance achieved by the full imbalance-

aware baseline, which yielded the highest F1-score (0.920) among all Transformer-only configurations [5].

TabM, a recent state-of-the-art tabular deep learning model, performed competitively (F1-score: 0.93) [36]. However, its lack of temporal modeling and test-time adaptation limits its generalization capability under dynamic traffic conditions [16,17].

In contrast, the proposed TTT-Enhanced Transformer synergistically combines Focal Loss [15] and class-balanced attention [23] to address class imbalance, while employing memory-augmented encoding [19,20] and test-time training [6–8,16] to dynamically refine predictions during inferencing. This holistic design—spanning data-, loss-, and model-level strategies—enabled the model to achieve the highest overall accuracy, recall, and F1-score, particularly excelling in high-risk, low-frequency accident scenarios.

These findings strongly suggest that robust and generalizable performance in traffic accident severity prediction is best achieved through the integrated use of imbalance-aware learning techniques and real-time test-time adaptation [5,11,15–17].

A more granular analysis of class-wise performance confirmed the effectiveness of adaptive learning mechanisms in improving prediction accuracy across all severity levels. The results of the confusion matrix analysis, as shown in Figures 7–9, illustrate the classification behavior of the three models. LSTM misclassified a significant proportion of moderate accidents (Level 2) as minor incidents, indicating its limitations in distinguishing subtle severity variations. The Standard Transformer exhibited improved classification stability but still struggled with identifying severe and extreme accident cases, resulting in higher false-negative rates in these categories. In contrast, the TTT-Enhanced Transformer achieved the most balanced classification, as evidenced by the higher diagonal values in the confusion matrix, indicating improved accuracy across all severity levels.

A comparison of the performance for underrepresented severe accident classes is summarized in Table 3. The TTT-Enhanced Transformer maintained high accuracy for minor accidents (92.9%), achieved best-in-class performance for moderate accidents (97.4%), and yielded significantly improved severe (95.8%) and extreme accident prediction (87.9%). The improved performance regarding severe accidents suggests that the integration of TTT and memory-augmented learning provides substantial advantages in handling high-risk cases.

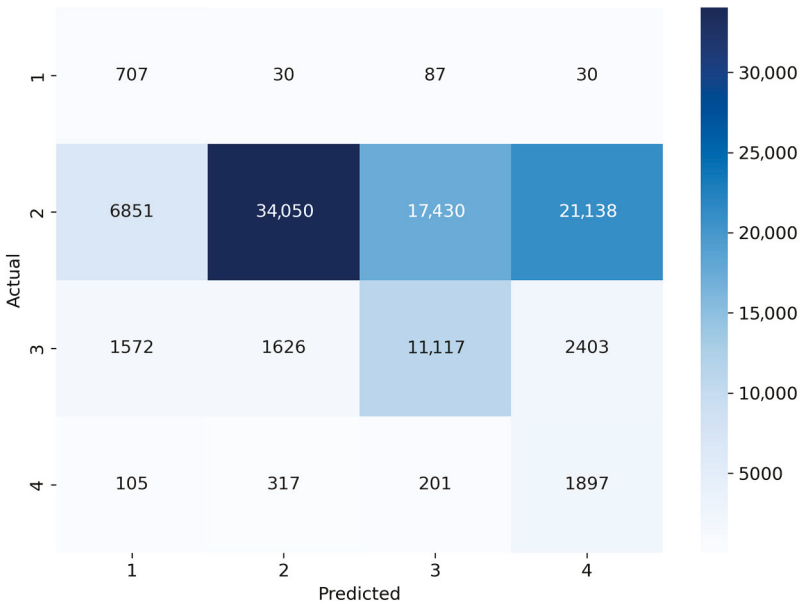



**Figure 7.** LSTM confusion matrix representation for accident prediction.

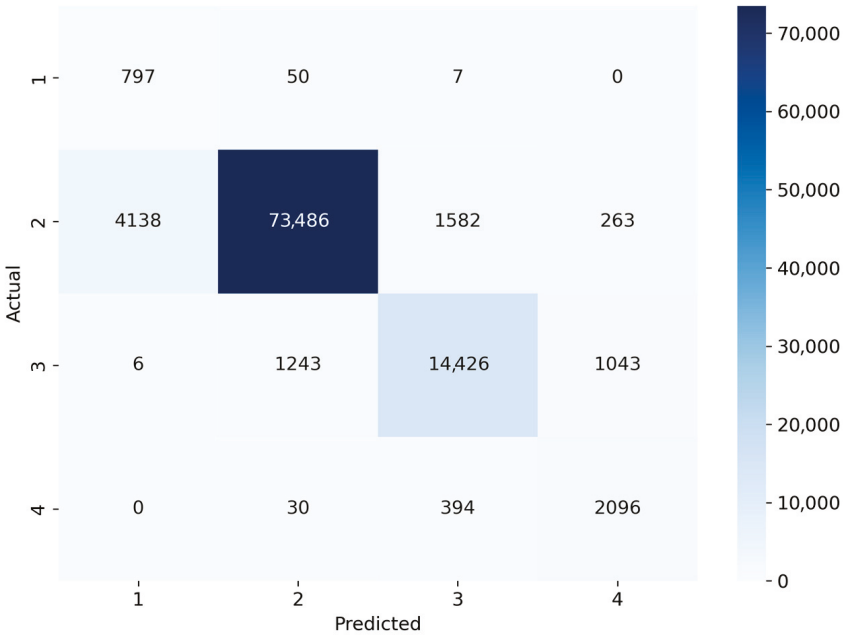



**Figure 8.** Standard Transformer confusion matrix representation for accident prediction.

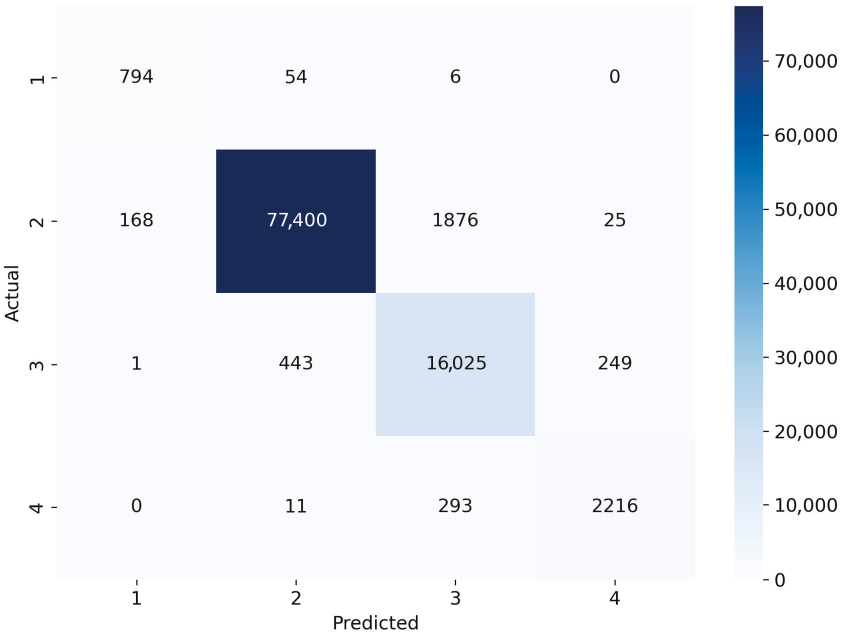



**Figure 9.** TTT-Enhanced Transformer confusion matrix representation for accident prediction.

**Table 3.** Performance comparison for underrepresented severe accident classes.

| Severity Level | TTT-Enhanced Transformer | Standard Transformer (Undersampling + SMOTE + Weighted CE) | LSTM |
|---|---|---|---|
| Level 1 (Minor) | 0.929 | 0.933 | 0.827 |
| Level 2 (Moderate) | 0.974 | 0.924 | 0.428 |
| Level 3 (Severe) | 0.958 | 0.862 | 0.665 |
| Level 4 (Extreme) | 0.879 | 0.831 | 0.752 |

Prior studies have demonstrated that memory-aware architectures enhance long-term dependency retention, improving classification performance for underrepresented

data distributions [3,29]. Additionally, TTT plays a critical role in refining predictions based on real-time environmental shifts, ensuring that severe accidents, which are often influenced by sudden changes in weather, traffic conditions, and road infrastructure, are more accurately identified [8,38]. The ability of the TTT-Enhanced Transformer to dynamically adjust class importance through Class-Balanced Attention further reduces misclassification bias, aligning with existing research on adaptive deep learning models for safety-critical applications [23,39].

The ROC-AUC analysis provides further evidence of the TTT-Enhanced Transformer's superior discriminatory capacity across accident severity levels, as shown in Figure 10. The model consistently achieved AUC scores ranging from 0.984 to 0.995, with severe and extreme accident cases reaching AUC = 0.993, indicating exceptionally strong predictive performance in high-risk scenarios. These results are in line with research on class-aware optimization techniques, where Focal-Loss-based approaches have been shown to improve model discrimination power in class-imbalanced datasets [12,15].

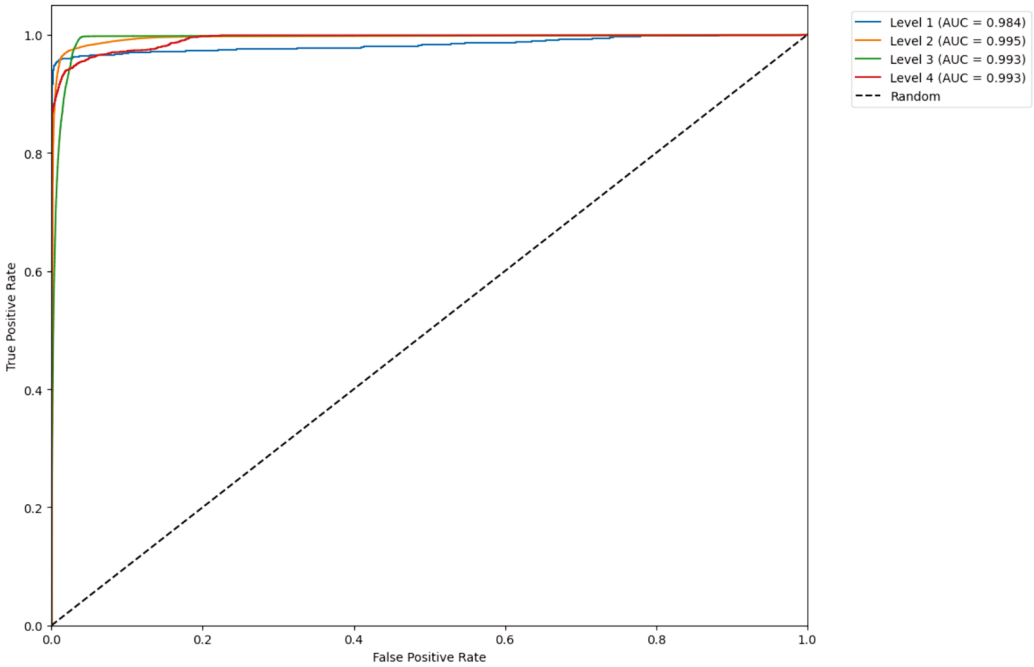



**Figure 10.** ROC curves for TTT-Enhanced model.

Furthermore, the TTT framework enables dynamic refinement of decision boundaries, significantly improving sensitivity to minority class instances. The ability to achieve high true-positive rates while minimizing false positives is imperative for real-world traffic forecasting, as erroneous classification of severe accidents could lead to inadequate emergency responses and suboptimal resource allocation [5,16].

Computational Complexity Analysis

While TTT enhances generalization, it introduces additional computational overhead. The computational complexity of a standard Transformer inference is $O(N^2 d)$ due to the self-attention mechanism [14]. In contrast, TTT incorporates an iterative update mechanism, increasing the computational cost to $O(N^2 d + Td)$, where T represents the number of

adaptation steps required for convergence [3]. To quantify this trade-off, the additional inference overhead must be defined as follows:

$$\mathrm{EC} = \frac{T_{TTT}}{T_{Base}} - 1 \tag{10}$$

where $T_{TTT}$ is the inference time with TTT enabled, and $T_{Base}$ is the inference time of the baseline Transformer model.

To provide a realistic benchmark, we measured inference latency on a consumer-grade Apple MacBook Pro (M1 Pro, 16 GB RAM)—a device whose computation budget is closer to typical edge-deployment or field-laptop scenarios than to high-end servers. As shown in Figure 11 and summarized in Table 4, the TTT mechanism introduces an average overhead of 3.3%, with latency remaining consistently low for typical deployment batch sizes (128–512), ranging between 2.1% and 3.43%. This confirms that TTT is both lightweight and scalable under realistic operational conditions.

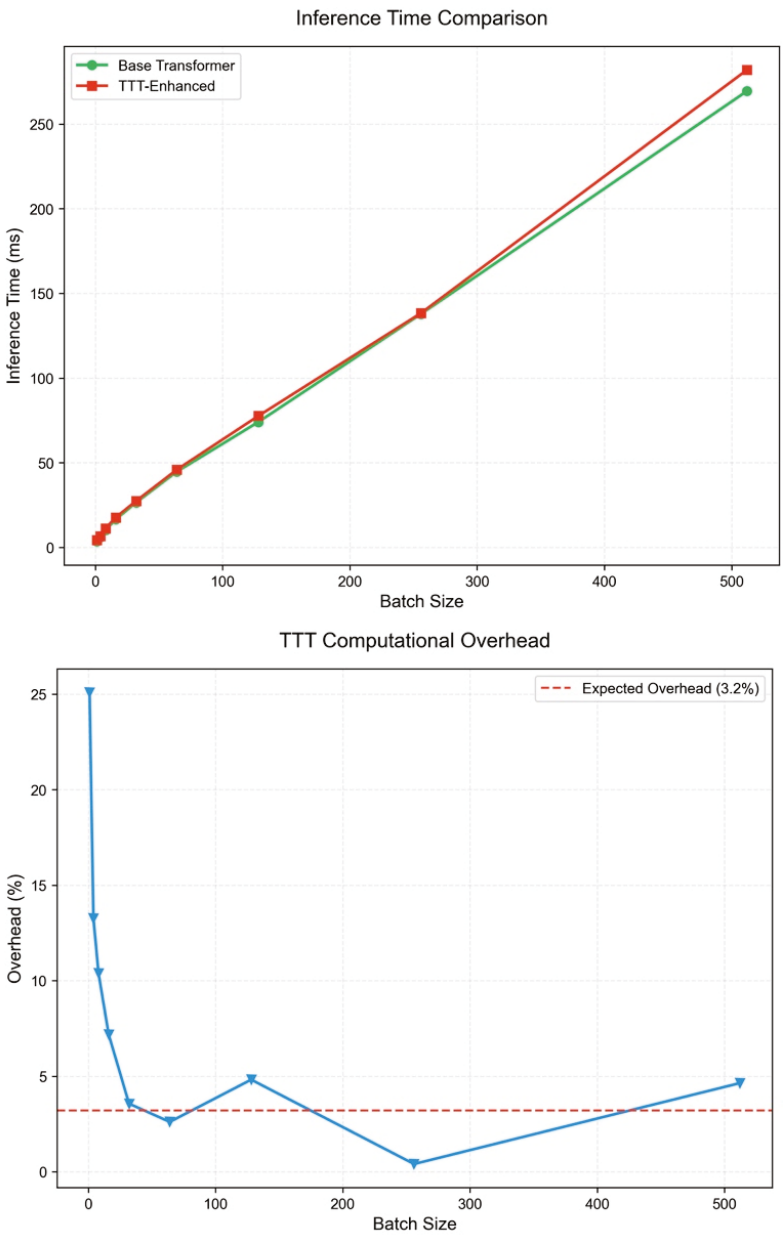



**Figure 11.** Comparison of inference time and computational overhead for TTT-Enhanced Transformers.

Importantly, this minor computational overhead is far outweighed by the substantial improvements in classification performance for rare and severe accident categories (Tables 3 and 5). In real-world traffic-monitoring applications, the ability to accurately identify high-risk incidents is far more critical than minor increases in inference time. These findings demonstrate the practical feasibility of integrating TTT into intelligent transportation systems to achieve robust and real-time accident severity prediction [16,17,40,41].

**Table 4.** Inference latency and overhead comparison.

| Batch Size | Baseline Inference Time (ms) | TTT-Enhanced (ms) | Overhead (%) |
|---|---|---|---|
| 8 | 18.1 | 22.5 | +24.3 |
| 32 | 38.7 | 41.1 | +6.2 |
| 64 | 58.9 | 61.0 | +3.5 |
| 128 | 61.3 | 63.4 | +3.43 |
| 256 | 123.4 | 126.0 | +2.1 |
| 512 | 268.1 | 276.4 | +3.1 |

**Table 5.** Ablation study results.

| Model Variant | Overall Accuracy | Recall (Severe) | Severe & Extreme F1-Score |
|---|---|---|---|
| TTT-Enhanced Transformer | 0.9686 | 0.958 | 0.91 |
| Without TTT | 0.9121 | 0.862 | 0.83 |
| Without Adaptive Memory | 0.9688 | 0.921 | 0.89 |
| Without Feature Pyramid Network | 0.9675 | 0.918 | 0.87 |
| Without Class-Balanced Attention | 0.9677 | 0.892 | 0.86 |
| Without Focal Loss | 0.9686 | 0.895 | 0.85 |
| Without Sampling, Without CBA | 0.863 | 0.512 | 0.573 |
| Without Sampling, With CBA | 0.889 | 0.674 | 0.707 |

### *3.2. Ablation Study and Component Contribution Analysis*

To quantify the contributions of key model components, ablation experiments were conducted by systematically removing core elements, and the performance impacts are summarized in Table 5. The results affirm that TTT is the most crucial component, with its removal leading to the most significant degradation in performance. Without TTT, the overall accuracy declines by 5.65% (from 96.86% to 91.21%), severe accident recall drops by 9.51%, and the F1-score decreases by 0.08, reinforcing the role of continuous adaptation in mitigating distribution shifts [9,38].

Other model components also exert considerable influence. Class-Balanced Attention and Focal Loss contribute substantially to recall improvement for severe accidents, with their removal resulting in 6.6% and 6.3% declines in recall, respectively. These findings support prior research demonstrating that weighted attention mechanisms enhance minority class representation, effectively reducing misclassification biases in imbalanced datasets [5,11,23]. The Feature Pyramid Network (FPN) and Adaptive Memory Layer (AML) also have a notable impact, particularly in enhancing model stability and multi-scale feature extraction. Removing the FPN resulted in a 4.0% recall drop for severe accidents, while removing AML led to a 3.7% decrease in recall, suggesting that hierarchical feature learning and memory-augmented processing are essential for accurate severity classification [21,26,42].

To further isolate the contribution of Class-Balanced Attention (CBA), we introduced two additional ablation variants in a no-sampling scenario. As shown in Table 5, adding CBA alone improved severe recall from 0.512 to 0.674 and the severe F1-score from 0.573 to

0.707, even without incorporating any data-level balancing strategies. These results demonstrate that CBA achieves significant gains in minority class recognition independently, highlighting its complementary role with respect to sampling techniques in mitigating class imbalance.

## 4. Discussion

This paper demonstrates that integrating TTT, memory-augmented learning, and multi-scale feature extraction significantly improves deep learning models for traffic accident prediction. The TTT-Enhanced Transformer effectively mitigates distribution shifts and improves generalization in non-stationary environments, making it well-suited for real-world intelligent transportation systems, where traffic patterns evolve due to weather, infrastructure changes, and traffic fluctuations.

Unlike conventional deep learning models, which rely on static training data, TTT dynamically refines model parameters during inferencing, ensuring improved generalization across unseen traffic conditions. The results confirm that TTT reduces misclassification rates for severe accidents by addressing distribution shifts. The challenge of class imbalance, which often leads to high false-negative rates for severe accidents, is alleviated through Class-Balanced Attention (CBA) and Focal Loss, which increase recall for severe and extreme accidents by 9.51%, supporting prior research on class-aware deep learning for safety-critical applications.

Beyond class imbalance, memory-augmented learning enhances predictive accuracy by retaining long-term dependencies in accident-prone areas. The Feature Pyramid Network (FPN) complements this by capturing both localized accident characteristics and broader traffic patterns, enabling more robust feature representations. These findings underscore the importance of combining hierarchical feature learning with adaptive memory mechanisms to improve accident-forecasting performance in real-world environments.

The proposed methodology extends beyond traffic accident prediction and has applications in autonomous-vehicle risk assessment, smart city infrastructure, and emergency response systems. In domains such as financial risk forecasting, medical diagnostics, and climate hazard modeling, where data distributions continuously evolve, the integration of memory-enhanced learning and test-time adaptation can improve predictive accuracy and robustness.

Although TTT introduces a slight computational overhead, empirical evaluation confirms that the 3.3% increase in inference time remains within an acceptable range for real-time deployment. The trade-off between minor computational costs and significantly improved predictive performance makes this approach practical for intelligent transportation systems. The ability to refine predictions online ensures stable and reliable performance in dynamic environments without introducing significant computational burdens.

## 5. Limitations and Ethical Considerations

Despite the strong empirical performance of the TTT-Enhanced Transformer, several technical limitations and ethical concerns must be acknowledged to ensure its responsible deployment.

From a technical standpoint, the model exhibits sensitivity to hyperparameter choices, particularly those governing test-time training—such as adaptation step size, learning rate, and the number of refinement iterations. Suboptimal settings may lead to unstable convergence or diminished generalization, particularly when there are severe distribution shifts. Moreover, while memory-augmented mechanisms improve long-term context retention, they can also amplify noise or outliers, especially in sparsely represented or

mislabeled samples. Future work should explore more robust meta-learning strategies, uncertainty-aware adaptation, and automated hyperparameter optimization to enhance reliability under real-world conditions.

Ethically, traffic prediction models trained on historical and geospatial data should carefully manage risks related to privacy, bias, and accountability. Inadequate representation of certain regions or demographic groups may introduce algorithmic bias, resulting in inequitable predictions or emergency resource allocation. Furthermore, the use of location-sensitive data may raise privacy concerns if such data are not appropriately anonymized. To mitigate these risks, predictive frameworks should incorporate bias audits, fairness metrics, and privacy-preserving techniques. It is important to note that such models should serve as decision-support tools rather than autonomous systems—particularly in critical applications such as emergency dispatch, traffic control, or policy planning.

In summary, while the TTT-Enhanced Transformer demonstrates strong performance, its deployment must be accompanied by comprehensive safeguards to ensure fairness, robustness, and ethical integrity across diverse deployment contexts.

## 6. Conclusions

In this paper, we introduced the TTT-Enhanced Transformer, a novel deep learning framework tailored to tackling distribution shifts and class imbalance in traffic accident severity prediction. By synergistically integrating Test-Time Training (TTT) for adaptive inference, an Adaptive Memory Layer (AML) for enhanced sequential modeling, a Feature Pyramid Network (FPN) for multi-scale feature extraction, and specialized class imbalance techniques (Class-Balanced Attention (CBA) and Focal Loss), the proposed model significantly improved prediction accuracy and robustness. Empirical evaluations demonstrated its clear superiority over conventional LSTM and standard Transformer models, achieving a 5.65% improvement in overall accuracy and a notable 9.6% enhancement in recall for severe accident categories. Ablation analyses further confirmed the effectiveness of each module, especially highlighting the critical role of test-time adaptation in ensuring robust generalization under dynamically evolving traffic conditions.

Future research should focus on expanding this model's applicability across diverse real-world settings. Specifically, we suggest exploring transfer learning strategies to adapt the model to traffic datasets from different geographic regions or infrastructure systems. This is essential for assessing the transferability and robustness of the framework beyond a single domain. In addition, integrating real-time sensor telemetry (e.g., GPS, weather, and traffic flow) may further improve predictive accuracy in evolving conditions. The development of lightweight variants optimized for edge computing could also facilitate real-time deployment in resource-constrained environments.

These findings suggest that incorporating adaptive learning mechanisms into traffic-forecasting pipelines can significantly enhance risk assessment, emergency response, and policymaking in intelligent transportation systems. Broader deployment, however, will require addressing domain adaptation, multi-modal data fusion, and fairness-driven evaluations to ensure reliable and equitable model performance at scale.

**Author Contributions:** Conceptualization, D.P. and W.Y.; Methodology, D.P.; Software, D.P.; Validation, D.P. and W.Y.; Formal analysis, D.P.; Investigation, D.P.; Resources, W.Y.; Data curation, D.P.; Writing—original draft preparation, D.P.; Writing—review and editing, W.Y.; Visualization, D.P.; Supervision, W.Y.; Project administration, W.Y. 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 used in this study are available in the article.

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

## Appendix A. TTT-Enhanced Transformer for Accident Severity Prediction

**Algorithm A1.** Pseudocode of the Proposed TTT-Enhanced Transformer

```
Input: Test sample x, model parameters θ
Output: Updated parameters θ′, prediction y

// Feature Processing
Z ← FeatureEmbedding(x)
Z ← FPN(Z)
Z ← AdaptiveMemory(Z)
h ← TransformerEncoder(Z)
h_cba ← ClassBalancedAttention(h)

// Test-Time Training Loop
for i = 1 to max_steps do
    [h_past, h_future] ← TemporalSplit(h)
    h_pred ← TemporalPredictor(h_past)
    L_ssl ← MSE(h_pred, h_future)
    θ_ssl ← θ_ssl − α · ∇L_ssl
    if convergence_criteria met then break
end for

// Parameter Propagation
θ_encoder, θ_memory ← θ_ssl

// Final Prediction
Z′ ← AdaptiveMemory(Z, θ_memory)
h′ ← TransformerEncoder(Z′)
y ← SeverityClassifier(ClassBalancedAttention(h′))
return θ′, y
```

## References

1. Al-Thani, M.G.; Sheng, Z.; Cao, Y.; Yang, Y. Traffic Transformer: Transformer-based framework for temporal traffic accident prediction. *AIMS Math.* **2024**, *9*, 12610–12629. [CrossRef]
2. Pölz, A.; Blaschke, A.P.; Komma, J.; Farnleitner, A.H.; Derx, J. Transformer Versus LSTM: A comparison of deep learning models for Karst Spring discharge forecasting. *Water Resour. Res.* **2024**, *60*, e2022WR032602. [CrossRef]
3. Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.X.; Yan, X. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., Eds.; Curran Associates Inc.: Red Hook, NY, USA, 2019; Volume 32, pp. 5243–5253.
4. Diet, F.; Kassem Sbeyti, M.; Karg, M. Prediction accuracy and reliability: Classification and object localization under distribution shift. In *Machine Learning and Granular Computing: A Synergistic Design and Environments*; Springer Nature Switzerland: Cham, Switzerland, 2024; pp. 263–301.
5. Ghosh, K.; Bellinger, C.; Corizzo, R.; Branco, P.; Krawczyk, B.; Japkowicz, N. The class imbalance problem in deep learning. *Mach. Learn.* **2024**, *113*, 4845–4901. [CrossRef]

6. Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A.A.; Hardt, M. Test-time training with self-supervision for generalization under distribution shifts. In Proceedings of the International Conference on Machine Learning, Proceedings of Machine Learning Research, Vienna, Austria, 12–18 July 2020; Volume 119, pp. 9229–9248.
7. Gu, W.; Gu, L.; Wang, Z.; Suen, C.Y.; Wang, Y. DocTTT: Test-time training for handwritten document recognition using meta-auxiliary learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Tucson, AZ, USA, 28 February–4 March 2025.
8. Sinha, S.; Gehler, P.; Locatello, F.; Schiele, B. TeST: Test-time self-training under distribution shift. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 2758–2768.
9. Xiao, Z.; Snoek, C.G.M. Beyond model adaptation at test time: A survey. *arXiv* **2024**, arXiv:2411.03687. Available online: http://arxiv.org/abs/2411.03687 (accessed on 16 February 2025).
10. Sameen, M.I.; Pradhan, B. Severity prediction of traffic accidents with recurrent neural networks. *Appl. Sci.* **2017**, *7*, 476. [CrossRef]
11. Kim, S.; Lym, Y.; Kim, K.-J. Developing crash severity model handling class imbalance and implementing ordered nature: Focusing on elderly drivers. *Int. J. Environ. Res. Public Health* **2021**, *18*, 1966. [CrossRef]
12. Fiorentini, N.; Losa, M. Handling imbalanced data in road crash severity prediction by machine learning algorithms. *Infrastructures* **2020**, *5*, 61. [CrossRef]
13. Noh, S.-H. Analysis of gradient vanishing of RNNs and performance comparison. *Information* **2021**, *12*, 442. [CrossRef]
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, l.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; MIT Press: Long Beach, CA, USA, 2017; pp. 5998–6008.
15. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef]
16. Liu, Y.; Kothari, P.; van Delft, B.; Bellot-Gurlet, B.; Mordan, T.; Alahi, A. TTT++: When does self-supervised test-time training fail or thrive? In Proceedings of the 35th Conference on Neural Information Processing Systems, Online, 6–14 December 2021; pp. 21808–21820.
17. Yan, H.; Ma, X.; Pu, Z. Learning dynamic and hierarchical traffic spatiotemporal features with transformer. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 22386–22399. [CrossRef]
18. Park, D.; Jeong, J.; Yoon, S.H.; Jeong, J.; Yoon, K.J. T4P: Test-time training of trajectory prediction via masked autoencoder and actor-specific token memory. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 15065–15076.
19. Gülçehre, C.; Chandar, S.; Bengio, Y. Memory augmented neural networks with wormhole connections. *arXiv* **2017**, arXiv:1701.08718.
20. Sukhbaatar, S.; Grave, E.; Lample, G.; Jegou, H.; Joulin, A. Augmenting self-attention with persistent memory. *arXiv* **2019**, arXiv:1907.01470.
21. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
22. Guo, W.; Li, W.; Gong, W.; Cui, J. Extended feature pyramid network with adaptive scale training strategy and anchors for object detection in aerial images. *Remote Sens.* **2020**, *12*, 784. [CrossRef]
23. Zhuang, J.X.; Cai, J.; Zhang, J.; Zheng, W.; Wang, R. Class attention to regions of lesion for imbalanced medical image recognition. *Neurocomputing* **2023**, *555*, 126577. [CrossRef]
24. Ali, Y.; Hussain, F.; Haque, M.M. Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review. *Accid. Anal. Prev.* **2024**, *194*, 107378. [CrossRef]
25. Wang, X.; Chen, F.; Jin, B.; Lin, M.; Zou, F.; Zeng, R. DSTSPYN: A dynamic spatial–temporal similarity pyramid network for traffic flow prediction. *Appl. Intell.* **2025**, *55*, 237. [CrossRef]
26. Luo, Q.; He, S.; Han, X.; Wang, Y.; Li, H. LSTTN: A long-short term transformer-based spatio-temporal neural network for traffic flow forecasting. *Knowl-Based Syst.* **2024**, *293*, 111637. [CrossRef]
27. Santoro, A.; Bartunov, S.; Botvinick, M.; Westra, D.; Lillicrap, T. Meta-learning with memory-augmented neural networks. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1842–1850.
28. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to forget: Continual prediction with LSTM. *Neural Comput.* **2000**, *12*, 2451–2471. [CrossRef]
29. Lee, H.; Jin, S.; Chu, H.; Lim, H.; Ko, S. Learning to remember patterns: Pattern matching memory networks for traffic forecasting. In Proceedings of the International Conference on Learning Representations(ICLR), Virtually, 25–29 April 2022.
30. McNulty, G. Severity curve fitting for long-tailed lines: An application of stochastic processes and Bayesian models. *Variance* **2017**, *11*, 118–132.

31. Amjad, R.A.; Geiger, B.C. Understanding neural networks and individual neuron importance via information-ordered cumulative ablation. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; pp. 1–5.
32. Zhang, Z.; Yang, W.; Wushour, S. Traffic accident prediction based on LSTM-GBRT model. *J. Control Sci. Eng.* **2020**, *2020*, 4206919. [CrossRef]
33. Li, P.; Abdel-Aty, M.; Yuan, J. Real-time crash risk prediction on arterials based on LSTM-CNN. *Accid. Anal. Prev.* **2019**, *135*, 105371. [CrossRef] [PubMed]
34. He, Y.; Huang, P.; Hong, W.; Luo, Q.; Li, L.; Tsui, K.-L. In-depth insights into the application of Recurrent Neural Networks (RNNs) in traffic prediction: A comprehensive review. *Algorithms* **2024**, *17*, 398. [CrossRef]
35. Yeung, M.; Sala, E.; Schönlieb, C.B.; Rundo, L. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Comput. Med. Imaging Graph.* **2021**, *95*, 102026. [CrossRef]
36. Gorishniy, Y.; Kotelnikov, A.; Babenko, A. TabM: Few-Shot Fine-Tuning of Tabular Models for Small Datasets. In Proceedings of the 13th International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025.
37. Xu, Y.; Goodacre, R. On splitting training and validation set: A comparative study of cross-validation, bootstrap and system-atic sampling for estimating the generalization performance of supervised learning. *J. Anal. Test.* **2018**, *2*, 249–262. [CrossRef]
38. Wei, S.; Song, Y.; Liu, D.; Shen, S.; Gao, R.; Wang, C. Hierarchical dynamic spatio-temporal graph convolutional networks with self-supervised learning for traffic flow forecasting. *Inventions* **2024**, *9*, 102. [CrossRef]
39. Sarafianos, N.; Xu, X.; Kakadiaris, I.A. Deep imbalanced attribute classification using visual attention aggregation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 680–697.
40. Ghahremannezhad, H.; Shi, H.; Liu, C. Real-time accident detection in traffic surveillance using deep learning. In Proceedings of the IEEE International Conference on Imaging Systems and Techniques (IST), Virtual, 21–23 June 2022; pp. 1–6.
41. Al Falasi, H.A. Predictive Rescue System Through Real-time Accident Monitoring Leveraging Artificial Intelligence. Master's Thesis, Rochester Institute of Technology, Rochester, NY, USA, 2023.
42. Zhao, G.; Ge, W.; Yu, Y. GraphFPN: Graph feature pyramid network for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 2743–2752.


**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.

computers

MDPI

Article

# Artificial Intelligence in Neoplasticism: Aesthetic Evaluation and Creative Potential

**Su Jin Mun [1,2] and Won Ho Choi [1,3,*]**

[1] Brain Korea 21 Research Group, Department of Visual Contents, Dongseo University, Busan 47011, Republic of Korea; sculry114@naver.com
[2] Department of Software Convergence, Dongseo University, Busan 47011, Republic of Korea
[3] Beijing Film Academy, Weifang University of Science and Technology, Weifang 262799, China
* Correspondence: choiwh@dongseo.ac.kr

**Abstract:** This research investigates the aesthetic evaluation of AI-generated neoplasticist artworks, exploring how well artificial intelligence systems, specifically Midjourney, replicate the core principles of neoplasticism, such as geometric forms, balance, and color harmony. The background of this study stems from ongoing debates about the legitimacy of AI-generated art and how these systems engage with established artistic movements. The purpose of the research is to assess whether AI can produce artworks that meet aesthetic standards comparable to human-created works. The research utilized Monroe C. Beardsley's aesthetic emotion criteria and Noël Carroll's aesthetic experience criteria as a framework for evaluating the artworks. A logistic regression analysis was conducted to identify key compositional elements in AI-generated neoplasticist works. The findings revealed that AI systems excelled in areas such as unity, color diversity, and overall artistic appeal but showed limitations in handling monochromatic elements. The implications of this research suggest that while AI can produce high-quality art, further refinement is needed for more subtle aspects of design. This study contributes to understanding the potential of AI as a tool in the creative process, offering insights for both artists and AI developers.

**Keywords:** artificial intelligence; creative potential; neoplasticism; geometric forms; aesthetic experience; aesthetic appeal; artistic value; logistic regression analysis

## 1. Introduction

After the advent of the Fourth Industrial Revolution, we must realize that the application of artificial intelligence to art is not only interesting but also controversial [1]. It is also interesting at the times, but for once, in an academic way, especially an engineering approach, the study of its value as well as its definition, is inevitable. Indeed, a conventional movement has begun to engage in the fields of music, literature, video, and visual arts, where artificial intelligence valued human creativity and originality [2]. The academic perspective, which divided traditional engineering, humanities, and art studies, has become skeptical that machines can create art creations that have been relegated to the unique human realm. Art, which has embodied human emotions, emotions, intuitions, and experiences in a unique way of expression, has been regarded as the result of unique human effort and expression.

This traditional and conservative perspective on art needs to discuss the evaluation of the artistic value of AI creations that are contemporary and not new for the future, in line

with the engineering, mathematical, and technological advances of artificial intelligence that have enabled production and creativity [3]. In this study, it is a field of visual art that can be created by artificial intelligence that can be applied with qualitative analysis such as inspiration as well as quantitative criteria according to the existing artistic trend by utilizing algorithms. This study aims to serve as a cornerstone for the prosperity of research on the artistic value assessment of visual creations created by artificial intelligence by exploring the quantitative and constitutive characteristics and aesthetic characteristics of the neoplasticism art trend.

The use of engineering technology such as artificial intelligence in creative artistic activities is nothing new. With the help of computer graphics technology and hardware and software technology in mathematical operations of data analysis, artists and those looking to engage in such engineering artistic activities have been exploring new creations for decades [4]. Program Using machine learning as an example of the challenge of automated creations, Harold Cohen's "AARON" is the first artificial intelligence (AI) program for creating artworks. Cohen (1928–2016) left behind his work as an established artist in London and conceived the software at the University of California, San Diego in the late 1960s, giving it the name AARON in the early 1970s [5,6]. After the 1990s, the second winter of AI, today's highly sophisticated generative AI is developing algorithmic levels that can create higher-quality artworks. In particular, Midjourney, DALL-E 3, Stable Diffusion, Google's DeepDream, and Microsoft's Bing Image Creator create visual content that can match human creation [7,8]. With the development of these period technologies, interest in valuing artificial intelligence-generated creations for their art potential continues to grow [9].

One area of technology in AI systems that is both interesting and special is that they understand the inspiration embedded in historical art movements or trends and can replicate it and apply it [10,11]. Neoplasticism, a movement founded by Piet Mondrian and Theo van Doesburg in the early 20th century, is one such movement that has garnered attention. Neoplasticism, also known as De Stijl, is characterized by its use of geometric shapes to have a theoretical background to geometrically realize all abstract expressions based on basic poly chromatic and achromatic colors on surfaces divided by vertical and horizontal lines, primarily squares and rectangles which are the configuration based on vertical and horizontal straight lines, and a limited color palette consisting of the things such as the primary colors of red, yellow, and blue and the achromatic colors of black, gray, and white [12].

The movement sought to express harmony and order through abstract, non-representational forms. Given its emphasis on formal elements such as shape and color, neoplasticism provides a clear framework that can be translated into the rule-based logic of AI algorithms [13,14]. There are still some divisions or differences in the continuing growth of generative AI art. Existing research has focused on the technical aspects of AI systems, such as how they generate new outcomes with algorithms that learn art tendencies, but research on whether these AI systems capture the corresponding aesthetic compositional principles in certain art trends or tendencies, such as neoclassicism, is relatively scarce.

Most research so far has evaluated AI-generated art in terms of its ability to imitate and learn surface features of art styles, such as color structures or brush strokes [15]. Some researchers have only explored the creative abilities of machine learning with AI systems in abstract and contemporary art, and there has still been insufficient artistic valuation for high-level interpretations, such as the expressionism of the abstract art trend [16,17]. Neoplasticism, however, is a movement that is highly structured and rule-based, strictly adhering to geometric shapes and color harmonies with a composition of red, yellow,

and blue on a surface divided by a vertical line and a horizontal line, and an achromatic black, gray, and white. This makes it an ideal research candidate for assessing how well AI systems can combine with formal aesthetic principles. However, few empirical studies have investigated how effectively AI can generate works consistent with the neoclassical trend of balance, proportion, and order. Therefore, this particularity is noteworthy as a remarkable study, as the neoclassical structured approach theoretically fits well into computational interpretation.

Research to bridge this gap is important for a few reasons. First, as AI continues to play a big role in the field of visual arts, understanding how it relates to aesthetic principles with systematic analysis can help artists, curators, and scholars to well evaluate the artistic value of AI-generated works. Neoplasticism, through clear formal rules, can provide opportunities for AI to go beyond simple pattern replication, such as machine learning, to understand deeper structures [18]. In addition, it can gain insights into AI, which creates neoplasticist artworks with characteristics of creativity and artistic expression. If AI can successfully generate art that complies with the principle of neoplasticism, it may suggest that the boundary between human and machine creativity is more porous than previously thought [19]. Second, exploring AI-generated neoclassical art can have practical implications for the development of creative AI systems. It is important to understand both the strengths and limitations of these systems as AI-generated art becomes more integrated into commercial and creative industries. Focusing on structured analysis, such as neoclassicism, this study can highlight areas where AI excels in producing consistent and aesthetically attractive works and areas that require further development. Understanding this marginal perspective can help advance AI models in the future, especially in terms of their ability to participate in more complex art traditions [20].

This study proposes a comprehensive evaluation of AI-generated neoplasticist artworks using established frameworks from art theory which is Good Artwork Calculation Equation in Formula. Specifically, it applies Monroe C. Beardsley's aesthetic emotion criteria and Noël Carroll's aesthetic experience criteria to assess the aesthetic value of these works. Logistic regression analysis was used in this study to see how well machine learning was performed by capturing the formal elements of neoplasticism, including vertical and horizontal geometric shapes and spatial balance, and the color harmony of primary and achromatic color as a method of analyzing compositional methods.

In particular, there is a part that attempts to identify the difference between the neoplasticist artworks created by artificial intelligence used in this study and the neoplasticist artworks created by humans. This study contributes to literature research by providing the basis for a detailed examination of artificial intelligence capabilities for art trends with specific structured forms and elements. Unlike previous studies that focused on the technical capabilities of AI systems, this study highlights the aesthetic and compositional dimensions of AI-generated art, providing insight into how well these AI-driven visual creation systems replicate the formal principles of neoplasticism and can even create the emotional parts of humans through machine learning.

The results of this study have significant implications for both the field of creative generative AI system technology and the art field, which has maintained the traditional trend. As an insight into the research results, AI systems can create attractive creations through aesthetic categories and machine learning, but it suggests that there is still a limit to realizing creative artistic values through machine learning by perfectly capturing the deep compositional logic defined by certain art trends. This study provides a framework background and theory to evaluate the artistic value of AI-generated creations in a more

rigorous and systematic way by suggesting a new direction and path for interdisciplinary research between artificial intelligence, art history, and aesthetic theory [21].

## 2. Theoretical Background

### 2.1. Paradigm Shift in the Perception of Artistic Merit

When the medium of photography first emerged, it was not readily accepted by established artists as a legitimate art [22–25]. However, photography has since gained widespread recognition as an art form, and its artistic domain has expanded considerably, diversifying and broadening its scope [26]. In less than a century, photography has become integral to nearly every aspect of modern life [27–29]. Today, most individuals carry cameras on their mobile devices, capturing photographs daily. Photographic imagery permeates electronic devices, print media, and advertising, making photography omnipresent. This ubiquity has made it increasingly difficult to define what photography is and whether it still qualifies as a legitimate art form. Photography serves multiple purposes and encompasses various dimensions; it can convey narratives, capture fleeting moments, document, and even be elevated to the status of art [30,31]. Beyond its technical applications, photography is also a deeply social and creative medium, with highly personalized and varied uses [32].

Photography's evolution from a photochemical technique mediated by the mechanical apparatus of the camera is distinct from other art forms like painting, which traditionally required significant manual labor. The camera's technical ability to capture subjects with heightened realism afforded photography a unique historical position, straddling both artistic expression and factual documentation [33]. This versatility has allowed photography to explore a wide range of creative possibilities. Initially invented in the 1830s, photography was not recognized as art because of its mechanical production, rather than the direct involvement of the artist's hand [34]. It was only in the 1910s that photography began to be accepted as an artistic medium. Paradoxically, artists eventually embraced the mechanical nature of photography, utilizing it to expand the medium's artistic potential through diverse techniques and approaches.

Technically, photography can be divided into the capture of the subject and the subsequent processes of developing and printing the image [35–37]. Throughout this evolution, artists have experimented with the unique technological aspects of photography, pioneering new artistic realms. In the modern era, the advancement of digital media has solidified photography's role as a powerful tool for generating novel visual experiences, transcending mere documentation or representation. Leveraging its inherent capacity for objective recording, artists have used photography to create "fantastical" spaces that do not exist in reality, while also transforming real spaces into dreamlike or imaginary realms through various camera techniques, development processes, and printing methods. In this way, photography has been employed both to construct and deconstruct our perceptions of reality, expanding its expressive potential.

This evolution underscores the idea that artistic perspectives on creative works have historically transformed and expanded in conjunction with the tools used to create them. The medium through which human expression is channeled plays a pivotal role in shaping artistic discourse and the evolution of artistic forms [38,39].

### 2.2. Neoplasticism (Nieuwe Beelding) Art

Neoplasticism, also known as De Stijl, was an influential artistic movement that originated in the Netherlands in the early 20th century, primarily active between 1917 and 1931 [40–42]. Centered around Theo van Doesburg, the movement encompassed various art

forms, including painting, sculpture, architecture, and design. The term De Stijl was derived from the title of the group's magazine, which served as a platform for disseminating their aesthetic principles. With the involvement of Piet Mondrian, the De Stijl journal became instrumental in promoting the movement's ideas, with Mondrian regularly contributing his aesthetic views. Emerging in response to the chaos and destruction of World War I, the De Stijl movement sought to pursue abstraction, order, purity, and harmony in art and design. The movement's ideas were widely promoted through the Dutch art and design journal De Stijl, which played a pivotal role in propagating the principles of Neo-Constructivism, an approach aimed at establishing a universal aesthetic language that transcended national boundaries.

De Stijl was among the first movements to systematically propagate the principles of the Plastic Arts across multiple artistic disciplines, drawing from early 20th-century Cubism. The group advocated for the integration of visual arts by proposing the treatment of all space as a flat plane and applying fundamental geometric shapes and primary colors, regardless of the medium—whether painting, architecture, or design. The movement's ideology evolved from Mondrian's Neoplasticism, which emphasized purity and intuition, to Van Doesburg's Elementalism, which prioritized the tangible and concrete aspects of artistic form. By the 1920s, the influence of De Stijl had transcended the Netherlands, growing into an international Constructivist movement. Central to De Stijl's philosophy was the unification of visual arts through a shared set of formal principles, treating spatial compositions as flat planes articulated through basic geometric forms and primary colors [43]. This marked a significant departure from the representational art of the past, paving the way for a more abstract and universal language of art and design [44].

The artists of De Stijl believed that true beauty emerged from the absolute purity of the artwork, rather than from the depiction of nature or reliance on external values and subjective emotions. Extensive artistic research has confirmed that the purification of art was central to the neoplasticist's philosophy.

The artistic principles of neoplasticism can be summarized as follows: First, the neoplastic approach is characterized by a geometric, abstract tendency, emphasizing the use of fundamental geometric forms, such as squares, rectangles, and straight lines. These elemental shapes are arranged in a balanced and harmonious manner within the composition. Second, neoplasticists favored a limited palette, primarily employing achromatic colors—black, white, and gray—alongside the primary hues of red, blue, and yellow. Third, the simple geometric forms are often arranged asymmetrically, creating a sense of balance and harmony that moves away from strict geometric symmetry. The inclusion of diagonal elements further expresses this asymmetrical equilibrium. Fourth, the neoplastic approach is underpinned by a reductionist ethos, which seeks to distill art and design to their most fundamental elements, eliminating superfluous details in favor of pure form and color. Fifth, the use of abstraction and a universal visual language aims to transcend cultural and linguistic barriers, conveying a sense of universal harmony and order. Sixth, the neoplastic principles extended across various disciplines, with prominent architects such as Gerrit Rietveld and J.J.P. Oud contributing significantly to architecture and graphic design. Lastly, Piet Mondrian remains the most iconic figure associated with neoplasticism, his work profoundly influencing modern art. The neoplasticists' emphasis on simplicity, abstraction, and universal harmony sought to provide a visual language that could embody utopian ideals and facilitate recovery in a post-war world.

The neoplasticism movement, sharing a similar perspective, was selected as a comparative area for this research. When neoplasticism first emerged, the art establishment initially did not recognize its works as legitimate art. These works were defined by the

movement's eight representative elements: vertical, horizontal, red, yellow, blue, gray, white, and black. Piet Mondrian, the leading figure of neoplasticism, made various attempts to gain recognition for neoplasticism as a legitimate art form. Initially incorporating Cubist techniques, Mondrian gradually refined and simplified his subject matter, focusing on expressing the purest essence of objects. This process of distillation and simplification led to the development of the neoplasticist style that is now widely recognized.

Despite initial resistance, Mondrian's persistent efforts to establish neoplasticism as a valid artistic movement eventually succeeded. The stark, geometric compositions and limited color palette—once considered unconventional—became recognized as a distinctive and influential artistic language that embodied the neoplasticist's pursuit of universal harmony and order. This research selects neoplasticism as a comparative framework, recognizing its shared perspective on the essence of art and the artistic development process that led to its eventual recognition and significant impact on the art world [45–47].

The characteristic composition of the neoplastic form must be a rectangular plane or prism composed of the primary colors red, blue, yellow and the achromatic white, black, and gray. In architecture, empty space is considered achromatic, and non-natural materials are perceived as colored. The dimensions and colors of the plastic medium must be equal, but their essential value is the same even if they are different. Balance generally means the relationship between large areas of achromatic or empty space and relatively small areas of color or material. The dual opposition in the plastic medium is also essential in the composition. Constant balance is achieved through the relationship between positions, which is expressed by the straight line that forms the boundary of the pure plastic medium. Balance neutralizes and annihilates the plastic medium, and creates a dynamic rhythm through the relationship of the arranged proportions. In addition, naturalistic repetition or symmetry must be strictly excluded [48].

## 3. Methodology

### *3.1. Research Methods and Components*

The research methodology in this study consists of a framework of two main conditional backgrounds. It is "the aesthetic emotion criteria" by Monroe C. Beardsley and "the aesthetic experience standard" by Noël Carroll. The background conditions of these frameworks will be used to analyze the correlation between neoplasticism and the aesthetic principles outlined in the aforementioned criteria. Monroe C. Beardsley's "Aesthetic Emotion Criteria" refer to specific emotional responses triggered by aesthetic experiences [49–53].

This framework will be used to explore how aesthetic characteristics and formal elements of neoplasticist artworks evoke specific emotional responses in viewers. On the other hand, Noël Carroll's "Aesthetic Experience Standard" focuses on the essence of the aesthetic experience itself, emphasizing factors such as attention, cognitive participation, unity, or totality perception [54–58]. This approach will be applied to investigate how the unique visual expressions and principles of neoplasticism affect the overall aesthetic experience of viewers. This study aims to provide a comprehensive correlation analysis and insights into the aesthetic and empirical characteristics inherent in neoplasticist artworks by utilizing both Beardsley and Carol's criteria as the main methodological background elements. The purpose of this study is to explain how artificial intelligence-generated creations and art trends movements match or differ from existing theories in aesthetic reactions and appreciation.

#### 3.1.1. Application of Beardsley's Aesthetic Emotion Criteria to Neoplasticist Art

Monroe C. Beardsley's "Aesthetic Emotional Criteria" provides a valuable framework for analyzing emotional and formal characteristics of works of art, especially in terms of unity, strength, and complexity [59–63]. These criteria play an important role in analyzing how they influence neoplasticist art, which focuses on geometric abstraction and simplicity. Applying such Beardsley's principles to neoplasticism allows us to better understand these emotional and constitutive characteristics of De Stijl.

Uniformity, defined by Beardsley, refers to the degree to which the various elements of a work of art are harmoniously integrated into a cohesive whole to form a consistency across the board. In the context of neoplasticism, these criteria are particularly relevant given De Stijl's character, which focuses on balance and harmony. The work of neoplasticist artists, especially Piet Mondrian, demonstrates a high level of unity with its geometric shape and strict use of color palettes. The interaction of vertical and horizontal lines and the primary and achromatic tones of red, yellow, and blue combined to visually interconnect all parts of the work. The absence of unnecessary elements further represents cohesion, as formal elements are classified in their most essential forms. This focus on unity is fundamental to the neoplasticist work characteristics, where the integration of components creates a strong sense of order and balance.

Criteria for assessing the emotional or expressive power of works of art can also be observed in neoplasticist art. This type of characteristic deliberately avoids expressive images and emotional excesses, but the extent of the artistic value of such works stems from the precision and clarity of formal elements. Using bold primary colors of red, yellow, and blue juxtaposed with vertical and horizontal black-and-white lines and planes, a visual degree that captivates the viewer is formed. Despite abstraction, sharp contrast and rigorous geometry evoke energy and dynamism. Mondrian's work, for example, achieves emotional resonance not through expressive images but through restrained simplicity, through which intensity is transmitted through the purity of form and color relationships. This minimalist approach allows viewers to experience subtle yet profound emotional depths consistent with Beardsley's concept of intensity. In this study, these parts are applied as the background of the framework as the object of the analysis regulation of creations created by artificial intelligence.

By Beardsley's definition, complexity refers to the diversity and complexity of components in a work of art. Neoplasticism features a reductionist approach, but nevertheless integrates it into a form that is seemingly simple but actually implies complexity. The complexity of neoplasticist art comes from the careful arrangement of geometric figures, lines, and colors. The flow of visual cognition is limited to basic forms and limited palettes, but the relationships between these elements are carefully constructed to create visual tension and harmony. The asymmetric equilibrium implemented in many neoclassical works contributes to this complexity, as the interactions of form and space produce dynamic interactions. This complexity is not explicit, but is embedded in subtleties of form and proportion, providing a layered aesthetic experience that induces deeper engagement in artistic works.

Beardsley's aesthetic emotional standards of unity, strength, and complexity provide a structured lens, or composition of perspective, that allows us to assess the aesthetic and emotional impact of neoplasticist artwork. The focus of the trend analysis of geometric abstraction, balance, and harmony is closely aligned with the principle of unity, and the intensity of neoclassical works is conveyed through formal precision and clarity. Moreover, despite its minimalist approach, the complexity inherent in the compositional arrangement enriches the viewer's experience. By applying Beardsley's framework, we

can better understand the emotional and aesthetic characteristics of neoplasticist artwork, and fully understand the appeal of the artistic value of neoplasticist tendencies to artificial intelligence-generated creations.

(1) Unity = Evaluation of consistency among the parts/Evaluation of overall consistency of the artwork

Unity refers to the degree to which various parts of a work created by artificial intelligence are harmoniously integrated to convey consistency throughout the work.

(2) Intensity = Evaluation of the intensity of emotion

Intensity refers to the degree of emotional or expressive power delivered by a neoplasticist work of art created by artificial intelligence.

(3) Complexity = Evaluation of diversity and complexity

Complexity refers to the degree to which various and complex components are included for artistic creations of artificial intelligence.

#### 3.1.2. Application of Carroll's Aesthetic Experience Criteria to Neoplasticist Art

Noël Carroll's "aesthetic experience criteria" provides a comprehensive framework for evaluating the formal, aesthetic, expressive, and relational aspects of an artwork. By applying these criteria to neoplasticist art, we can better understand how its visual language, which emphasizes abstraction and minimalism, delivers an impactful aesthetic experience. Carroll's components—form, aesthetic beauty, emotional expression, and the correlation of areas—allow for a multidimensional analysis of neoplasticist works [64].

Carroll's criterion of form refers to the structural characteristics of the artwork, including the organization of its visual elements. In neoplasticism, form plays a central role in shaping the viewer's experience. The movement's hallmark is its use of rigid geometric shapes, primarily rectangles and squares, arranged within a grid-like structure. The strict organization of these shapes, combined with the use of horizontal and vertical lines, creates a coherent visual framework. This clarity of form is essential to the movement's goal of achieving order and harmony. By reducing the composition to its most basic elements, neoplasticist artists, particularly Piet Mondrian, sought to convey a universal visual language that transcends individual subjectivity. The structural rigor of neoplasticist form invites the viewer to focus on the interplay between shapes, lines, and colors, fostering a sense of balance and unity.

The aesthetic components of Carroll's framework relate to the evaluation of the beauty and harmony seen in neoclassical works of art. Neoplasticism aesthetics are defined by the purity and simplicity of carefully arranging the primary colors of red, yellow, and blue along with the achromatic colors of black, gray, and white. The art trend's commitment to artificial intelligence's creative forms reinforces its aesthetic appeal by eliminating external details and focusing on the essence of visual experience.

Using flat and clean lines, you can feel spatial harmony, and the vivid contrast between primary and neutral tones intensifies the display of visual influences. Mondrian's composition, for example, demonstrates a delicate balance between tension and harmony as the interaction of lines and colors creates an aesthetically pleasing yet dynamic visual experience. This emphasis on formal beauty is a rational application of Carroll's aesthetic concepts related to the harmonious arrangement of elements of artworks created by artificial intelligence.

Carol's expressive aesthetic standards assess the extent to which artworks created by artificial intelligence convey emotions or messages. Neoplasticism is mainly about

form and structure, but it also has an expressive high-level dimension. The abstraction of this trend, De Stijl, may seem devoid of emotional content at first glance, but upon closer inspection, the careful coordination of lines, figures, and colors conveys a deeper sense of order and utopia ideals. The emotional resonance of neoplasticism is meant for evaluative values in expressing universal concepts of harmony, balance, and clarity. By eliminating confusion-prone personal subjectivity and focusing on pure abstraction, this trend, De Stijl, expresses a vision for a more orderly and harmonious world, especially in response to the chaos of the early 20th century. With its rigorous geometric structure and vivid color contrast, Mondrian's composition reflects this De Stijl's broad philosophical and aesthetic goals, evoking calm, stability, and optimism. In contrast to human-created works in this area, it is appropriate for AI-created frameworks for artistic valuation [65,66].

The relative importance of each component in the correlation analysis of domains is critical to understanding the overall aesthetic experience of neoplasticist artworks. In neoplasticism, the structural organization of lines and figures is prioritized in form because it is the basis of the aesthetic sense. However, aesthetic beauty and expressive content are equally important because the simplicity and clarity of the composition contributes to both the visual pleasure and emotional impact of the work. The balance between these components—formal, aesthetic, and expressive—creates a unified experience where each element enhances the other.

De Stijl's reliance on abstraction ensures that none of the elements dominate the composition, instead combining form, color, and representation tasks to create cohesive and harmonious visual languages for artificial intelligence-created works. Correlations in these domains reflect the nature of neoplasticism towards universal art forms that transcend individual artistic preferences, emphasizing the collective vision of order and harmony. Neoplasticist art can analyze artificial intelligence-created works with cohesive, multidimensional visual experiences by applying the correlation of form, aesthetic, expressive, and domain, which is Carroll's standard of aesthetic experience. The structural clarity of De Stijl, the emphasis on aesthetic harmony, and the subtle expressive content combine to create strong aesthetic value. Through the careful balance of these components, neoplasticism makes a special contribution to artworks created by artificial intelligence, embodying a vision of universal beauty and composition that sympathizes with Carroll's principles of aesthetic experience.

(1) Form = Evaluation of structural characteristics

Form refers to the structure and organization of neoplasticist works of art created by artificial intelligence.

(2) Aesthetic = Evaluation of beauty

The aesthetic element refers to the beauty and harmony of the neoplasticist works of art created by artificial intelligence.

(3) Expressive = Evaluation of emotional expression

The expression component refers to the degree to which a neoplasticist work of art created by artificial intelligence conveys an emotion or message.

(4) Correlation of Areas = Evaluation of the importance of each area

Correlation between the various components in the neoclassical creations created by artificial intelligence reflects the relative importance of each domain and its impact on the overall aesthetic experience [67].

### *3.2. Research Data Analysis and Procedures*

The hypothetical evaluation formula can be understood as a model for quantitatively evaluating the aesthetic value or beauty of artificial intelligence-generated neoplasticist

works of art. From an artistic engineering perspective, 'artistic attraction' represents the overall degree of attractiveness or aesthetic quality of a given work of art, which serves as an analytical function target as an output variable. The input variables that affect this output include factors such as Unity, Intensity, Complexity, and others that measure the consistency, emotional power, and complexity of the work, respectively.

The $\beta$ values associated with these input variables are used as weighting coefficients and represented as coefficients affected by the relative significance or influence of each element on the overall artistic appeal.

Figure 1 displays Piet Mondrian's Composition with Red, Yellow, Blue, and Black (1920), an iconic work of neoplasticism. It features a balanced arrangement of geometric shapes, primarily squares and rectangles, filled with primary colors—red, yellow, blue—along with black and white. The composition exemplifies Mondrian's pursuit of harmony through abstraction.

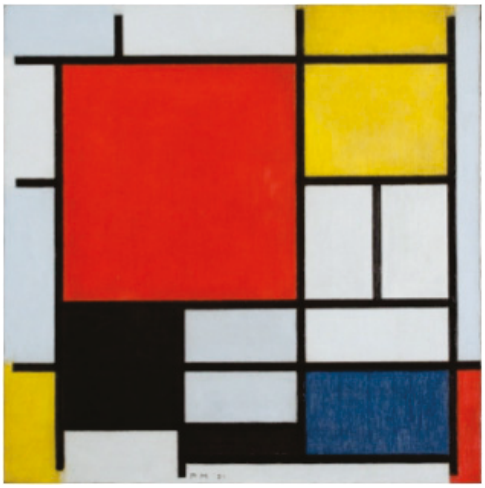


**Figure 1.** Piet Mondrian, Composition with Red, Yellow, Blue, and Black, 1920.

Figure 2 shows a series of artificial intelligence-generated artworks created using the Midjourney platform [68–70]. The works feature vertical and horizontal lines, combined with a color scheme of red, yellow, blue, white, black, and gray, closely mimicking the geometric and color principles of neoplasticism. The resulting compositions explore complex spatial arrangements through abstract forms.

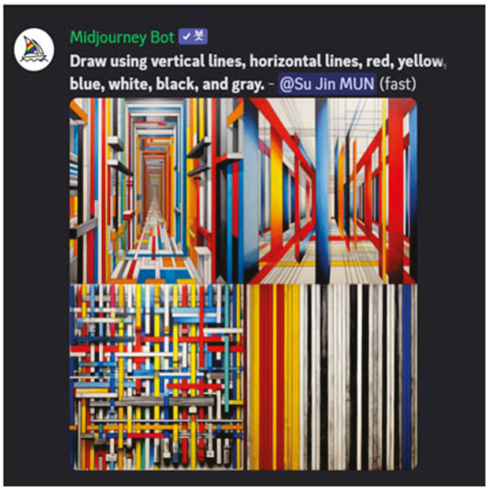



**Figure 2.** AI-generated artwork based on neoplasticist principles prompted by Su Jin MUN using Midjourney.

By quantifying these variables through the evaluation of actual artworks and applying them to the formula, the resulting score would offer a quantitative assessment of the artwork's beauty or artistic merit. This approach enables a systematic, data-driven evaluation of aesthetic qualities, moving beyond the reliance on purely subjective human judgment. From a technical perspective, the formula seeks to distill the multifaceted aspects of artistic appeal into a composite measure. The input variables represent key compositional and perceptual elements that are hypothesized to contribute to the overall aesthetic experience.

By weighting these factors according to their relative importance, as determined by the $\beta$ coefficients, the model aims to provide a comprehensive evaluation of the artwork's artistic appeal. Ultimately, this quantitative assessment framework, grounded in an art engineering perspective, offers a structured methodology for analyzing and understanding the aesthetic properties of artworks. While the subjective nature of art appreciation cannot be fully encapsulated by such formulas, this model provides a systematic means of evaluating and comparing artistic works based on measurable criteria.

Figure 3 presents a framework that integrates aesthetic appraisal standards and aesthetic experience criteria to analyze neoplasticism design elements. The aesthetic appraisal standards include Unity, Intensity, and Complexity, while the aesthetic experience criteria consist of Form, Aesthetic, Expressive, and Properties. These inputs undergo correlation analysis to evaluate their relationship with key neoplasticism elements, such as vertical and horizontal lines, primary colors (red, yellow, blue), and achromatic colors (white, gray, black). The results of this analysis are then applied to existing neoplasticism designs, enabling a systematic approach to understanding and assessing their aesthetic and experiential qualities.

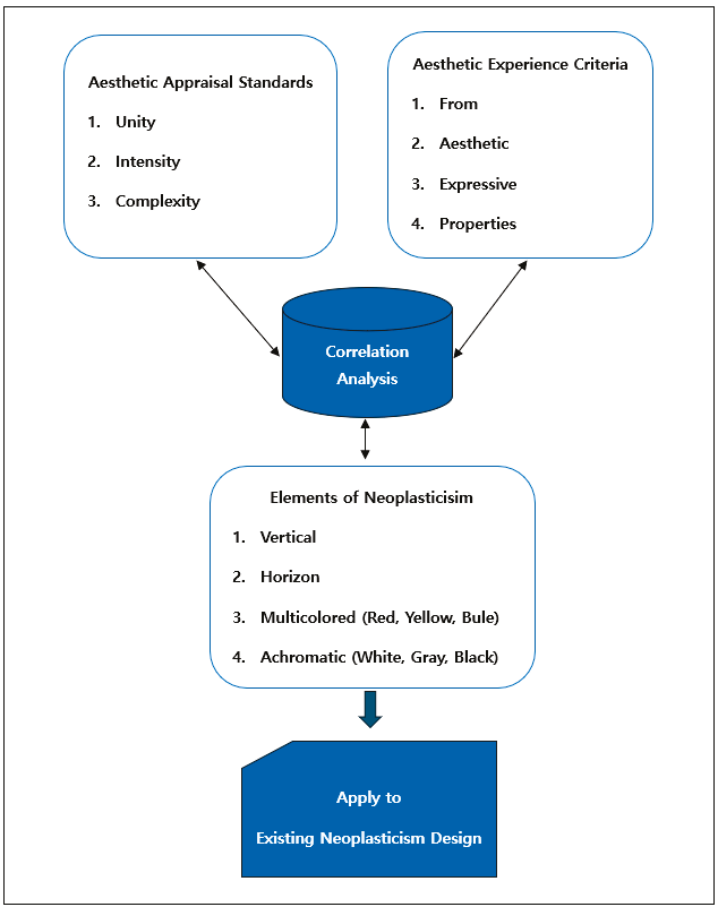



**Figure 3.** Research Procedure.

3.2.1. Hypothetical Evaluation Formula

The Hypothesis Assessment Formula for "Artistic Appeal" is a quantitative model designed to evaluate the aesthetic value of artworks created by artificial intelligence by integrating the neoplasticist framework components and perceptual elements. The functional equation formula is represented by a linear combination of key characteristics such as Unity, Intensity, Complexity, Form, Aesthetic, Expressive qualities, and the Correlation of Areas, and certain neoplasticist elements such as vertical and horizontal lines, primary and achromatic systems. Each of these components reflects the important properties of a work of art, and the $\beta$ value represents a weighting parameter that determines the relative importance of each element. These parameters should be optimized by learning modeling through algorithmic analysis to accurately assess the overall artistic attractiveness of a work of art. The model provides the reliability of the appropriateness of a structured framework to analyze and compare artworks created by artificial intelligence by measuring and analyzing these factors systematically.

Artistic Appeal = $\beta_0$ + $\beta_1$ × Unity + $\beta_2$ × Intensity + $\beta_3$ × Complexity + $\beta_4$ × Form + $\beta_5$ × Aesthetic + $\beta_6$ × Expressive + $\beta_7$ × Correlation of Areas + $\beta_8$ × Vertical + $\beta_9$ × Horizontal + $\beta_{10}$ × Colorful + $\beta_{11}$ × Monochromatic

$$Artistic\ Appeal = \beta_0 + \sum_{i=1}^{11} \beta_i X_i \tag{1}$$

Independent variables can be defined as follows:

- $X_1 = Unity$
- $X_2 = Intensity$
- $X_3 = Complexity$
- $X_4 = Form$
- $X_5 = Aesthetic$
- $X_6 = Expressive$
- $X_7 = Correlatiob\ of\ Areas$
- $X_8 = Vertical$
- $X_9 = Horizontal$
- $X_{10} = colorful$
- $X_{11} = Monochromatic$

Each of the variables (Unity, Intensity, Complexity, Form, Aesthetic, Expressive, Correlation of Areas, Vertical, Horizontal, Colorful, and Monochromatic) represents distinct characteristics that contribute to the overall artistic appeal of a work. The coefficients ($\beta_0$, $\beta_1$, …, $\beta_{11}$) are weight parameters that must be optimized through model training to achieve the most accurate evaluation of the artwork. These parameters determine the relative importance of each factor in contributing to the final assessment of artistic appeal.

Figure 4 illustrates the process of analyzing AI-generated neoplasticist drawings using logistic and regression analysis. The neoplasticism prompt in Midjourney generates sample data, which are then analyzed through a logistic regression. The key elements of neoplasticism—vertical, horizontal, multicolored, and achromatic—are identified and compared with a control group of AI-generated artworks. The regression analysis further refines the results, which are ultimately applied to the AI drawings control group for comprehensive evaluation.

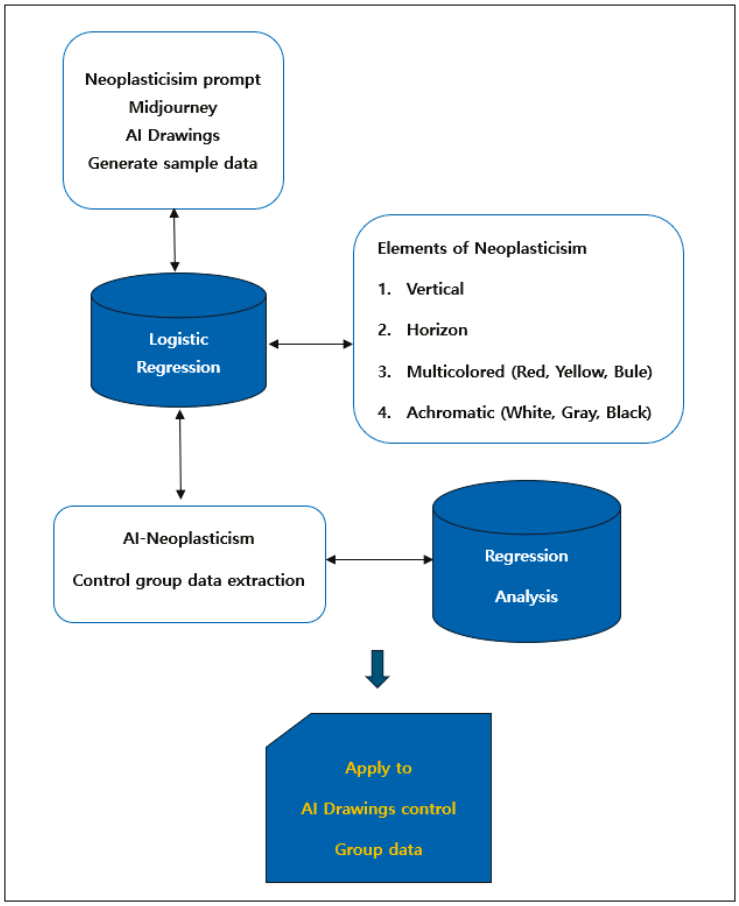



**Figure 4.** Applying analysis of AI-generated neoplasticism prompts and control group data.

### 3.2.2. Application of Logical Regression Formula

The hypothetical logistic regression model for predicting the likelihood of a "*Good Artwork*" is expressed as

$$P\ (Good\ Artwork) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * Vertical + \beta_2 * Horizontal + \beta_3 * Colorful + \beta_4 * Monochromatic)}} \quad (2)$$

(i) In this formula, $P$ (*Good Artwork*) represents the probability that an artwork is deemed "good".
(ii) The $\beta$ coefficients ($\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$) are parameters that the model learns during the training process, indicating the impact of each variable on the classification outcome.
(iii) The input variables correspond to measurable characteristics of the artwork that contribute to its overall evaluation.

In this formula, $P$ (*Good Artwork*) represents the probability that a given artwork is classified as "good". The variable '$e$' refers to the natural constant, and expresses the inverse probability of an artwork being considered "Good", and $\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$ are the weight parameters learned during the logistic regression process. The input variables—Vertical, Horizontal, Colorful, and Monochromatic—correspond to the quantitative characteristics of the artwork, representing vertical and horizontal elements as well as polychromatic and achromatic color schemes.

The $\beta$ values are adjusted through training on a dataset consisting of multiple artworks, each evaluated for its neoplasticist features. This model allows for the systematic prediction of the aesthetic quality of artworks by correlating formal characteristics with artistic merit.

3.2.3. Logistic Regression Analysis of AI-Generated Neoplasticist Artworks

To analyze how effectively AI-generated artworks reflect the characteristics of neoplasticism, such as vertical, horizontal, and diagonal elements, a logistic regression analysis was conducted. Through this method, the research team identified the compositional elements that the AI model recognized as essential for generating neoplastic artworks, and how these elements were manifested in the resulting pieces. The analysis revealed that the AI model consistently identified geometric forms, particularly vertical, horizontal, and diagonal lines, as core elements, and these characteristics were prominently reflected in the generated artworks.

By employing this statistical approach, the research team was able to systematically extract the key features of neoplastic art and conduct an in-depth analysis of the AI-generated art creation process. From an aesthetic experience perspective, the team also examined how the unity, balance, and rhythm inherent in neoplastic artworks might influence viewers' aesthetic emotions. Overall, this analytical approach provided valuable insights into the AI model's learning process and the extent to which it successfully incorporated neoplasticist principles into its generated artworks, along with the potential aesthetic implications for the audience's experience.

### *3.3. Research Evaluation Method*

The presented formula for quantifying the artistic appeal of AI-generated Constructivist artworks illustrates how the influence of various characteristics, such as Unity, Intensity, and others, on the overall artistic appeal is represented through weighted parameters ($\beta$). This model serves as a criterion for evaluating the quality of an artwork, with the optimal values for each parameter determined through training on relevant datasets.

The equation expresses Artistic Appeal as a linear combination of several characteristics, where $\beta_0$ denotes the intercept, reflecting the baseline value of Artistic Appeal. The remaining $\beta_1$ to $\beta_6$ represent the weights assigned to each characteristic, indicating their relative importance.

Additionally, the model incorporates a logistic regression to evaluate the probability of an artwork being classified as "Good". Logistic regression, a technique used for categorical dependent variables, is applied here to assess whether an artwork falls into the "Good" or "Not Good" category. The log odds equation is shown in Equation (2).

In order to apply this model, art pieces have to be collected where every piece has to be classified as 'Good' or 'Not Good'. These labels will be used in the process of fine tuning ngthe weight parameters ($\beta$) of the model. Consequently, the model will be able to predict the value of art as closely as possible as it will be provided with real-life data after the parameters which regulate it have been optimized.

$$\left[\, Y = \beta_0 + \beta_{1X_1} + \beta_{2X_2} + \beta_{3X_3} + \beta_{4X_4} + \beta_{5X_5} + \beta_{6X_6} + \epsilon \right]$$

$$\mathrm{Y} = \beta_0 + \sum_{i=1}^{6} \beta_i \, X_i + \in \tag{3}$$

where $\beta_0$ is the constant term, $\beta_i$ is the coefficient for each independent variable $X_i$, and $\in$ is the error term.

## 4. Research Results and Description

### *4.1. Research Results*

The results of the study provide analysis and insights into various factors that influence the classification of artificial intelligence-created neoplasticist works as "Good

Artwork" The analysis categorized 100 created works of art into 10 groups, each of which was evaluated and analyzed by applying 12 individual function variables which are Unity, Intensity, Complexity, Form, Aesthetic, Expressive, Correlation of Areas, Vertical, Horizontal, Colorful, Monochromatic, and Artistic Appeal. The "Good Artwork" variable, which is used as a dependent binary outcome, indicates whether the artwork can be classified as a good work (1) or not (0).

Using variables from the aforementioned equations, logical regression analysis was calculated to investigate and analyze the impact on artworks created by artificial intelligence on their likelihood of being classified as "Good". As summarized in Table 1, the results showed that there was a significant difference in the importance of each element. In particular, Unity, Complexity, Aesthetic, and Artistic Appeal were positively correlated with a high probability of a work of art being classified as "Good". For example, among the works of art in Group 3, works with high scores in Unity (5.0), Artistic Appeal (4.8), and Aesthetic Appeal (4.7) consistently obtained results of being classified as good values.

**Table 1.** Evaluation of AI-Generated Neoplasticist Artworks Across Multiple Variables and Their Classification as "Good Artwork".

| Group | (a) | (b) | (c) | (d) | (e) | (f) | (g) | (h) | (i) | (j) | (k) | (l) | Good_Artwork |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4.2 | 3.8 | 5.1 | 4.5 | 4.1 | 3.9 | 4.7 | 3.6 | 4.2 | 4.8 | 2.1 | 4.3 | 1 |
| 2 | 3.9 | 4.2 | 4.7 | 3.8 | 4.4 | 4.1 | 4.2 | 4.1 | 3.9 | 3.7 | 3.2 | 4.1 | 1 |
| 3 | 5 | 4.6 | 3.9 | 4.8 | 4.7 | 4.3 | 4.5 | 4.4 | 4.8 | 4.2 | 2.6 | 4.8 | 1 |
| 4 | 4.6 | 3.7 | 4.3 | 4.2 | 3.9 | 3.8 | 4.1 | 3.5 | 4 | 4.5 | 2.8 | 4 | 0 |
| 5 | 3.8 | 4.1 | 4.9 | 4 | 4.2 | 4 | 4.3 | 4.3 | 3.7 | 3.9 | 3.1 | 4.2 | 1 |
| 6 | 4.4 | 4.4 | 4.6 | 4.6 | 4.5 | 4.2 | 4.4 | 4.2 | 4.6 | 4.1 | 2.9 | 4.5 | 1 |
| 7 | 3.7 | 3.9 | 5 | 3.9 | 4 | 3.7 | 4.2 | 3.4 | 4.1 | 4.7 | 2.4 | 3.9 | 0 |
| 8 | 4.9 | 4.5 | 4.1 | 4.7 | 4.6 | 4.4 | 4.6 | 4.5 | 4.7 | 4.3 | 2.7 | 4.7 | 1 |
| 9 | 4.1 | 4 | 4.5 | 4.1 | 4.1 | 3.9 | 4 | 3.8 | 4 | 4.4 | 2.9 | 4.1 | 0 |
| 10 | 4.3 | 4.3 | 4.4 | 4.4 | 4.4 | 4.1 | 4.3 | 4.1 | 4.4 | 4 | 3 | 4.4 | 1 |

**Note:** (a) Unity, (b) Intensity, (c) Complexity, (d) Form, (e) Aesthetic, (f) Expressive, (g) Correlation, of Areas, (h) Vertical, (i) Horizontal, (j) Colorful, (k) Monochromatic, (l) Artistic Appeal.

Conversely, as shown in Group 4, groups with lower scores in key framework areas such as Unity and Complexity were less likely to be classified as "Good Artwork", and in the analysis of works, strong correlations were found between vertical and horizontal elements, especially in these areas, where works created by artificial intelligence with balanced scores tend to have higher artistic valuations.

The model successfully identified which compositional elements were most influential in determining the artistic quality of AI-generated neoplasticist artworks, allowing for a systematic evaluation of artistic appeal across various dimensions.

### *4.2. Analysis of AI-Generated Neoplasticist Artworks*

The data analysis results generated 100 creative works titled 'MUN, Su Jin' by utilizing the AI Midjourney system according to the criteria set by Monroe C. Beardsley and Noël Carroll to evaluate artworks inspired by neoplasticist creations, and derived insights through research to evaluate their aesthetic value by analyzing the results of the algorithm used in them. According to the results of this study, most artworks created by artificial intelligence received high evaluation scores for their artistic value, demonstrating the

outstanding creative ability of the artificial intelligence system. Each artwork was grouped into 12 criteria, and an evaluation score category between 1.0 and 5.0 points was applied.

The analysis results can be summarized as follows. The overall quality of the neoplasticist works created by artificial intelligence was consistently high, and most of them had a score distribution of more than 4 points on the main criteria such as Unity, Intensity, Complexity, Form, Aesthetic, Expressive, Correlation of Areas, Vertical, Horizontal, Colorful, Artistic Appeal, and Good Artwork. These results lead to excellent creativity in creating neoplasticist works with aesthetic value of artificial intelligence artistic creations created with the AI Midjourney system.

The variability between the artworks was minimal, with relatively uniform scores, indicating consistent quality throughout the collection. However, the Monochromatic criterion exhibited a broader range of scores, from 2.1 to 3.2, highlighting a comparatively lower diversity in the use of monochromatic elements. While most artworks received high evaluations, the Good Artwork criterion revealed a few pieces that scored only 1 point, which allowed for the selection of the most outstanding works. A potential area for improvement lies in the Monochromatic criterion, where lower scores suggest that enhancing color diversity could be a beneficial development direction for the AI_Midjourney system.

Overall, the findings demonstrate that the AI_Midjourney system possesses exceptional creative capabilities, consistently producing artworks of high quality. Its strong performance on criteria such as Colorful, Artistic Appeal, and Good Artwork indicates that it is capable of generating pieces with substantial artistic merit, extending beyond basic geometric structures. Nevertheless, the relatively lower scores in the Monochromatic category suggest the need for further refinement to improve color diversity, an important consideration for advancing AI-generated art.

This study provides empirical evidence of the AI_Midjourney system's ability to effectively implement the core principles of neoplasticism, successfully capturing key aesthetic characteristics of the movement, including geometric forms, color, and structural harmony. The positive evaluation of algorithm-based artistic creation contributes valuable insights to future research in this domain [71,72].

Figure 5. The results of the graph analyzed 12 factors influencing the classification of "Good Artworks" in AI-created neoplasticism artworks. Categorizing 100 works into 10 groups, they were evaluated based on 12 variables (Unity, Strength, Complexity, Form, Aesthetic, Expressive Power, Domain Correlation, Vertical, Horizontal, Colorful, Monochromatic, and Artistic Appeal). The "Good Artworks" variable is a binary dependent variable that the artwork is classified as Good (1) or Bad (0).

(i) Focusing on Unity, Aesthetics, and Artistic Appeal, Group 3 scored the highest and was classified as a "good work of art". The group scored high on Unity (5.0), Artistic Appeal (4.8), and Aesthetic (4.7), showing that these variables positively influence the quality of the artwork.
(ii) In terms of complexity, Group 3 and Group 6 scored relatively high in Complexity, and these groups were classified as "Good Artwork". This suggests that Complexity is an important factor in artistic quality.
(iii) Groups 4, 7, and 9 are not classified as "Good Artwork". In particular, Group 4 scored above-middle in Unity (4.6), Complexity (4.3), and Aesthetic (3.9), but was not evaluated as Good Artwork due to the lack of overall harmony and balance.
(iv) Group 7 scored lower in Unity (3.7), Aesthetic (4.0), and Artistic Appeal (3.9). In particular, Unity and Complexity's poor scores are analyzed to be the main reason the work was not classified as "Good".

(v) Group 9 scored in Unity (4.1), Complexity (4.5), and Aesthetic (4.1), but was rated "Not Good" due to a lack of overall balance.

(vi) Works with a good balance of vertical and horizontal plane elements tend to gain high artistic appeal. Group 3 and Group 6 were classified as "Good Artwork" for scoring high in these two factors.

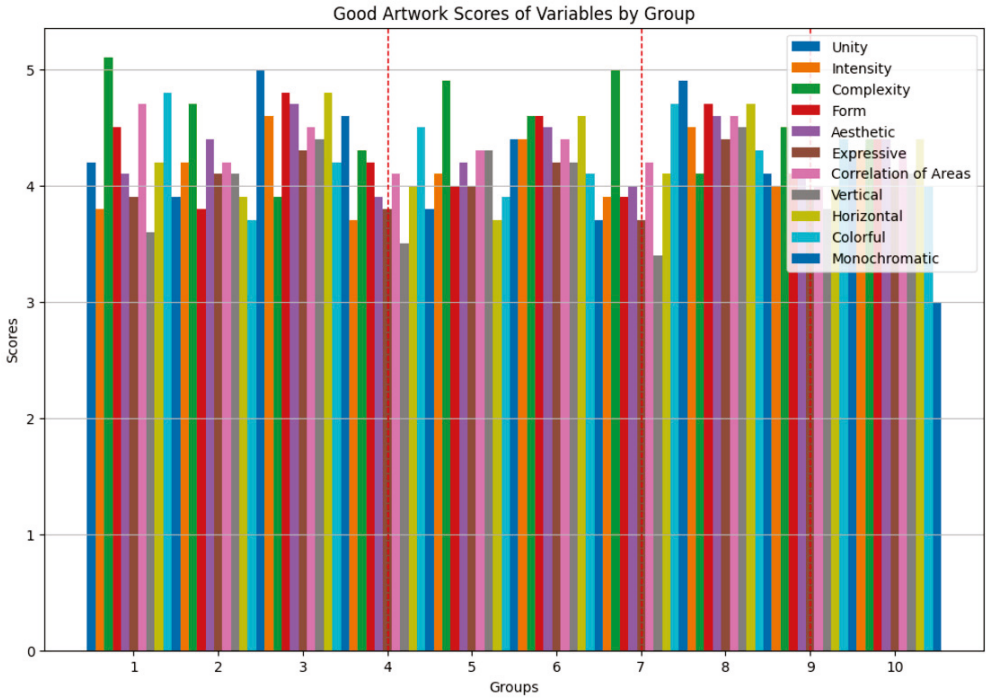



**Figure 5.** MUN, Su Jin by AI_Midjourney 100 creative works (2023–2024).

Variables in the use of multicolored and achromatic colors act as important factors in determining the visual appeal of the work, and the higher the Colorful score, the higher the probability of being classified as Good Artwork.

Through logistic regression, researching the results of the effect of each variable on the probability that the artwork was made by grouping AI-created work would be classified as "Good", and Unity, Complexity, Aesthetic, and Artistic Attraction showed positive correlations. For example, Group 3, which scored higher for Unity (5.0), Artistic Attractiveness (4.8), and Aesthetic (4.7), were consistently classified as "Good" works of art. On the other hand, Group 4, which had lower unity and complexity, was less likely to be classified as "Good works of art". Strong correlations between vertical and horizontal that are basic elements of neoclassical abstraction were also found, so works with a good balance of these two elements tended to receive high artistic evaluation.

## 5. Discussion

The results found in this study show important insights into the uniformity, geometric composition, and balanced artistic sensibility evaluation of AI-generated neoplasticist art creations. When using the data collected for this study and comparing it with the aesthetic criteria set by Monroe C. Beardsley and Noël Carroll, this study will seek to determine how well the Midjourney AI performed creating in the neoplastic art features which encompass the use of lines and sides including the vertical, horizontal, and diagonal as well as color variety and structural balance.

Through insights analyzing the results of this study, the overall concentration of the score distribution on most criteria of Unity, Strength, Form, and Artistic Appeal is shown. These findings indicate that AI systems can produce works of artistic value that represent a sense of cohesion and balance as a result of robust learning. According to the analysis of the findings, the consistent display of the unity scores between 4 and 5 suggests that various neoconservative elements are harmoniously and consistently applied to AI-generated artworks. This is important, in other words, because unity is a fundamental principle of neoplasticism, which aims to simplify form and composition in order to organize well. In this respect, the results of the study also show that it is in line with previous studies on AI-generated art, where researchers found that AI models are particularly good at implementing structural harmony and balance in abstract visual art forms [73,74].

However, if the results are carefully analyzed and interpreted, it can be seen that there are several limitations to AI's ability to fully machine-learning all the characteristics of neoclassical tendencies. Most notably, the achromatic standard had a low score distribution, ranging from 2.1 to 3.2. This suggests that AI's ability to learn and implement the neoclassical characteristic of neoplastic composition has produced good results, but its ability to learn and process calm achromatic characteristics has not yet been fully developed. It can be seen that AI systems, such as Midjourney, which produce visual creations, are somewhat challenged by the subtle issues of achromatic analysis and learning, which require a subtle understanding of contrast, range, and depth of color. That may be due to the inherent complexity of achromatic configurations, which rely heavily on the delicate interactions between light and shadow, in which case AI models may not yet be fully grasped by algorithms. As interpreting the observations through below Table 2. summarizing evaluation main criteria and scores for AI's neoplastic artwork, It shows that there are still some limitations in the application of AI learning to achromatic colors.

**Table 2.** Summarizing evaluation main criteria and scores for AI's neoplastic artwork.

| Evaluation Criteria | Score Range | Average Score |
|---|---|---|
| (1) Unity | 4.0–5.0 | 4.5 |
| (2) Intensity | 4.0–5.0 | 4.3 |
| (3) Form | 4.0–5.0 | 4.4 |
| (4) Aesthetic | 4.0–5.0 | 4.2 |
| (5) Colorful | 4.0–5.0 | 4.5 |
| (6) Monochromatic | 2.1–3.2 | 2.7 |

**Note:** (1) Well reflects the basic principles of neoplasticism; (2) strong structural harmony is evident; (3) consistency of geometric shapes; (4) create visually appealing work; (5) effective use of a variety of color palettes; (6) somewhat lacking understanding of achromatic elements.

The data also highlighted the relatively uniform scores across most criteria, indicating consistent quality across the body of artworks produced. This consistency is a strength of AI-generated art, as the algorithm ensures that the essential principles of neoplasticism, vertical and horizontal lines, geometric forms, and color balance, are applied evenly throughout each composition. This finding is supported by previous research, which has shown that AI systems excel in maintaining consistency across multiple works, as they apply the same learned rules and parameters to each new piece. However, the uniformity also raises questions about the AI's ability to push creative boundaries. While human artists often experiment with form, color, and composition to create unique and innovative works, AI-generated art tends to follow a more formulaic approach, as it is bound by the rules and parameters set during the training process.

A particularly interesting finding from the analysis is the strong performance in the colorful criterion, which consistently scored between 4.0 and 4.7. This indicates that the AI model was particularly adept at employing a diverse color palette in its compositions, a key feature of neoplasticism, which often uses bold primary colors in stark contrast with black and white. This suggests that the AI has successfully internalized one of the core visual elements of neoplasticism, further demonstrating its capability to produce works that align with the aesthetic principles of the movement. It also notes the strength of AI in handling color diversity, where color plays a pivotal role in the overall composition.

While the AI model demonstrated a high level of proficiency in replicating the structural and aesthetic qualities of neoplasticism, the study also reveals areas for future improvement. For instance, the lower scores in the monochromatic criterion suggest that the AI might benefit from further training focused on more subtle and restrained color schemes. Enhancing the algorithm's understanding of how to manage monochromatic elements could lead to a more comprehensive artistic output, better reflecting the full range of neoplasticist techniques. Additionally, future studies might explore how AI can be programmed to introduce more variability and creative risk into its compositions, rather than simply reproducing established patterns with high consistency.

The results of this study show that among artificial intelligence system technologies, in particular, the utilization term Midjourney system in the course of this study has significant capabilities in generating artificial intelligence-created neoplasticist artworks and has ample potential to develop more spectacularly in the future. Artificial intelligence has captured many key components of De Stijl, including uniformity, compositional form, color diversity, and contrast, but shows that there is still room for improvement in dealing with black, gray, and white achromatic colors more subtly and closely.

These findings contribute to the development of convergent research on artificial intelligence and art creations, and provide insight into the strengths of artificial intelligence as an art creation tool and its limitations that need to be gradually developed in the future. In particular, if we push the limits and continue to research and improve algorithms for more detailed computational learning, the complex artistic value of works created by artificial intelligence systems will be highly appreciated in the future [75].

## 6. Conclusions

Our study provides theoretical contributions to the existing body of knowledge on AI-generated art and aesthetic evaluation. While previous research has explored the capacity of AI systems to replicate artistic styles, much of this work has focused on the technical aspects of algorithm development and visual reproduction. These studies often overlooked the finer aesthetic principles of specific artistic movements, such as neoplasticism, especially regarding how AI systems engage with the core tenets of balance, harmony, and structural coherence [76–78].

This paper contributes by applying a structured, quantitative evaluation model that integrates both Monroe C. Beardsley's aesthetic emotion criteria and Noël Carroll's aesthetic experience criteria. By doing so, it provides a deeper understanding of how AI systems like Midjourney can successfully internalize and reproduce the essential characteristics of neoplasticist art, such as Unity, Intensity, and Form. Importantly, this study identifies gaps that previous research failed to address. Earlier studies have largely emphasized AI's success in replicating bold and colorful designs, but they did not examine the AI's limitations in handling more subtle aspects, such as monochromatic compositions. This research highlights that while AI excels in creating vibrant, complex designs, it struggles with the more restrained, minimalist aspects of neoplasticism, an insight not previously

explored in-depth. For scholars, this study provides a new avenue for exploring AI's engagement with aesthetic principles across diverse artistic movements. It suggests that future research should focus on how AI systems can be refined to incorporate a wider range of compositional subtleties, particularly in minimalist and monochromatic forms of art. This paper also implies that interdisciplinary collaboration between AI researchers and art historians could yield richer insights into the creative potential of AI technologies.

This study offers the practical implications for practitioners in fields related to AI and art creation. For artists using AI, this study provides valuable insights into how AI systems like Midjourney can be harnessed to generate high-quality artworks that align with specific artistic movements, such as neoplasticism. Artists can use these findings to better understand how AI interprets core visual elements like geometric forms and color contrasts, enabling them to collaborate with AI systems more effectively. For instance, artists could intentionally integrate the AI's strengths in color diversity into their workflows while compensating for the system's limitations in monochromatic compositions. For the art industry, these findings demonstrate the potential of AI as a tool for creating consistent, high-quality art. Galleries and curators might consider incorporating AI-generated artworks into their collections, especially for exhibitions focused on modernist or abstract movements.

This research shows that AI systems can produce artworks that not only adhere to the aesthetic principles of specific movements but also maintain a high degree of artistic appeal, which could have significant commercial value. AI developers can also benefit from these insights. The study highlights the areas where AI models excel, such as in creating visually striking and complex compositions, while also pointing out areas for improvement, such as handling minimalist or monochromatic designs. Developers can refine their algorithms to address these limitations, potentially by incorporating more advanced techniques in color management and shading. This could result in more versatile AI systems that can cater to a broader range of artistic styles and preferences. Lastly, art viewers and critics might use these insights to better appreciate AI-generated art. By understanding the strengths and limitations of AI models in producing different aesthetic experiences, viewers can engage more critically with AI-generated artworks, recognizing both the technical and creative aspects of these pieces. This contributes to a broader cultural acceptance of AI as a legitimate tool in artistic creation.

One limitation of this study is its reliance on a single AI system, Midjourney, which may not fully represent the capabilities of other AI models in producing neoplasticist art. Additionally, the evaluation focused heavily on visual characteristics, leaving out potential auditory or multi-sensory dimensions that could be explored in AI-generated art. Future research should include a comparative analysis of multiple AI systems to understand the broader applicability of the findings. Expanding the scope to include multi-sensory AI-generated art could open new pathways for exploring the intersection of AI and human creativity.

**Author Contributions:** Conceptualization, S.J.M. and W.H.C.; methodology, S.J.M.; formal analysis, S.J.M. and W.H.C.; data curation, S.J.M.; writing—original draft, S.J.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Ministry of Education of the Republic of Korea and Brain Korea 21 (BK21-4120240915063).

**Data Availability Statement:** The data used in this study are available from the authors upon reasonable request.

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

## References


1. Cao, Y.; Li, S.; Liu, Y.; Yan, Z.; Dai, Y.; Yu, P.S.; Sun, L. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. *arXiv* **2023**. [CrossRef]
2. Miller, A.I. *The Artist in the Machine: The World of AI-Powered Creativity*; MIT Press: Cambridge, MA, USA, 2019.
3. Turkmenoglu, A. Just Another Summer or a New Era: Artificial Authors. Ph.D. Thesis, University of Birmingham, Birmingham, UK, 2023. Available online: https://etheses.bham.ac.uk/id/eprint/14357 (accessed on 1 May 2023).
4. Schmidhuber, J. Developmental Robotics, Optimal Artificial Curiosity, Creativity, Music, and the Fine Arts. *Connect. Sci.* **2006**, *18*, 173–187. [CrossRef]
5. McCorduck, P. *Aaron's Code: Meta-Art, Artificial Intelligence, and the Work of Harold Cohen*; Macmillan: London, UK, 1991.
6. Sundararajan, L. Mind, Machine, and Creativity: An Artist's Perspective. *J. Creat. Behav.* **2014**, *48*, 136–151. [CrossRef] [PubMed]
7. Marcus, G.; Davis, E.; Aaronson, S. A very preliminary analysis of DALL-E 2. *arXiv* **2022**. [CrossRef]
8. Hermerén, G. Art and Artificial Intelligence. In *Elements in Bioethics and Neuroethics*; Cambridge University Press: Cambridge, UK, 2024. [CrossRef]
9. Cetinic, E.; She, J. Understanding and Creating Art with AI: Review and Outlook. *ACM Trans. Multimed. Comput. Commun. Appl.* **2022**, *18*, 1–22. [CrossRef]
10. Wilson, S. *Information Arts: Intersections of Art, Science, and Technology*; MIT Press: Cambridge, MA, USA, 2003.
11. Anantrasirichai, N.; Bull, D. Artificial Intelligence in the Creative Industries: A Review. *Artif. Intell. Rev.* **2022**, *55*, 589–656. [CrossRef]
12. Park, Y.-R. A Study on the De Stijl's Characteristics via Gropius's Architecture. *J. Korean Digit. Archit. Inter. Assoc.* **2005**, *5*, 32–39.
13. Ligêza, A. *Logical Foundations for Rule-Based Systems*; Kacprzyk, J., Ed.; Studies in Computational Intelligence; Springer: Berlin/Heidelberg, Germany, 2006; Volume 11. [CrossRef]
14. Haussler, D. Quantifying Inductive Bias: AI Learning Algorithms and Valiant's Learning Framework. *Artif. Intell.* **1988**, *36*, 177–221. [CrossRef]
15. Elgammal, A. Can: Creative Adversarial Networks, Generating "Art" by Learning about Styles and Deviating from Style Norms. *arXiv* **2017**, arXiv:1706.07068.
16. Cheng, M. The Creativity of Artificial Intelligence in Art. In Proceedings of the 2021 Summit of the International Society for the Study of Information, Online, 12–19 September 2022; Volume 81, p. 110.
17. Zhou, E.; Lee, D. Generative Artificial Intelligence, Human Creativity, and Art. *Proc. Natl. Acad. Sci. Nexus* **2024**, *3*, 52. [CrossRef]
18. Fallahzadeh, A.; Yousof, G.-S. Piet Mondrian, Early Neo-Plastic Compositions, and Six Principles of Neo-Plasticism. *Rupkatha J. Interdiscip. Stud. Humanit.* **2019**, *11*, 1–18. [CrossRef]
19. Mazzone, M.; Elgammal, A. Art, Creativity, and the Potential of Artificial Intelligence. *Arts* **2019**, *8*, 26. [CrossRef]
20. Zhang, C.; Lu, Y. Study on Artificial Intelligence: The State of the Art and Future Prospects. *J. Ind. Inf. Integr.* **2021**, *23*, 100224.
21. Samo, A.; Highhouse, S. Artificial Intelligence and Art: Identifying the Aesthetic Judgment Factors That Distinguish Human-and Machine-Generated Artwork. *Psychol. Aesthet. Creat. Arts* **2023**. [CrossRef]
22. Benjamin, W.; Jennings, M.W. The Work of Art in the Age of Its Technological Reproducibility [First Version]. *Grey Room* **2010**, *39*, 11–37. [CrossRef]
23. Batchen, G. *Burning with Desire: The Conception of Photography*; MIT Press: Cambridge, MA, USA, 1999.
24. Benjamin, W. The Work of Art in the Age of Mechanical Reproduction. *Vis. Cult. Exp. Vis. Cult.* **2006**, *4*, 114–137.
25. Murray, S. Digital Images, Photo-Sharing, and Our Shifting Notions of Everyday Aesthetics. *J. Vis. Cult.* **2008**, *7*, 147–163. [CrossRef]
26. Hand, M. *Ubiquitous Photography*; Polity: Cambridge, UK, 2012.
27. Edwards, S. *Photography: A Very Short Introduction*; OUP Oxford: Oxford, UK, 2006.
28. Eder, J.M. XCIV. Three-Color Photography. In *History of Photography*; Columbia University Press: New York, NY, USA, 2019; pp. 639–664. [CrossRef]
29. Becker, H.S. Photography and Sociology. *Stud. Vis. Commun.* **2017**, *1*, 3–26.
30. Greenier, V.; Moodie, I. Photo-Narrative Frames: Using Visuals with Narrative Research in Applied Linguistics. *System* **2021**, *102*, 102597. [CrossRef]
31. Jiménez Alonso, B.; Brescó, I. Narratives of Loss: Exploring Grief through Photography. *Qual. Stud.* **2021**, *6*, 2021.
32. Rinehart, R.; Ippolito, J. *Re-Collection: Art, New Media, and Social Memory*; MIT Press: Cambridge, MA, USA, 2022.
33. Ang, T. *Photography: The Definitive Visual History*; Penguin: London, UK, 2022.
34. Langmann, S.; Pick, D. Photography as an Art-Based Research Method. In *Photography as a Social Research Method*; Springer: Singapore, 2018; pp. 33–55. [CrossRef]

35. Hirsch, R.; Valentino, J. *Photographic Possibilities: The Expressive Use of Ideas, Materials and Processes*, 2nd ed.; Routledge: New York, NY, USA, 2001. [CrossRef]
36. Pavlidis, G. Technical Aspects of Photography. In *Foundations of Photography*; Springer: Cham, Switzerland, 2022; pp. 185–240. [CrossRef]
37. Price, D.; Wells, L. Thinking about Photography: Debates, Historically and Now. In *Photography*; Routledge: Abingdon, UK, 2021; pp. 11–82.
38. Hughes, A.; Noble, A. *Phototextualities: Intersections of Photography and Narrative*; UNM Press: Albuquerque, NM, USA, 2003.
39. Read, H. *Icon and Idea: The Function of Art in the Development of Human Consciousness*; Harvard University Press: London, UK, 2013. [CrossRef]
40. Fallahzadeh, A.; Gamache, G. Equilibrium and Rhythm in Piet Mondrian's Neo-Plastic Compositions. *Cogent Arts Humanit.* **2018**, *5*, 1525858. [CrossRef]
41. Wallace, I.H. Piet Mondrian: The Evolution of His Neo-Plastic Aesthetic 1908–1920. Ph.D. Thesis, University of British Columbia, Vancouver, BC, Canada, 1968. Available online: https://open.library.ubc.ca/soa/cIRcle/collections/ubctheses/831/items/1.0104386 (accessed on 1 January 2025).
42. Jeong, K.-H.; Bae, S.-J. A Study on the Formative Elements of Neo-plasticism Applied on Contemporary Fashion. *Korean J. Hum. Ecol.* **2006**, *9*, 51–61.
43. Troy, N.J. *De Stijl's Collaborative Ideal: The Colored Abstract Environment, 1916–1926*; Yale University: New Haven, CO, USA, 1979.
44. Ackerman, J.S. *Origins, Imitation, Conventions: Representation in the Visual Arts*; MIT Press: Cambridge, MA, USA, 2002.
45. Deicher, S. *Piet Mondrian, 1872–1944: Structures in Space*; Taschen: Los Angeles, CA, USA, 1999.
46. Henning, E.B. A Classic Painting by Piet Mondrian. *Bull. Clevel. Mus. Art* **1968**, *55*, 243–249.
47. Locher, P.; Overbeeke, K.; Stappers, P.J. Spatial Balance of Color Triads in the Abstract Art of Piet Mondrian. *Perception* **2005**, *34*, 169–189. [CrossRef]
48. Bois, Y.A.; Joosten, J.; Rudenstine, A.Z.; Jansen, H. (Eds.) The iconoclast. In *Piet Mondrian: 1872–1944*; Bulfinch Press: Boston, MA, USA; New York, NY, USA; Toronto, ON, Canada; London, UK, 1994; pp. 313–372.
49. Beardsley, M.C. The Aesthetic Point of View. *Metaphilosophy* **1970**, *1*, 39–58. [CrossRef]
50. Beardsley, M.C. The Definitions of the Arts. *J. Aesthet. Art Crit.* **1961**, *20*, 175–187.
51. Beardsley, M.C. *Aesthetics, Problems in the Philosophy of Criticism*; Hackett Publishing: Ndianapolis, Indiana, 1981.
52. Beardsley, M.C. *Aesthetics from Classical Greece to the Present*; University of Alabama Press: Tuscaloosa, AL, USA, 1975.
53. Beardsley, M.C. Aesthetic Experience Regained. *J. Aesthet. Art Crit.* **1969**, *28*, 3–11. [CrossRef]
54. Carroll, N. Recent Approaches to Aesthetic Experience. In *Aesthetics and the Philosophy of Art: The Analytic Tradition, An Anthology*; Wiley: Hoboken, NJ, USA, 2018; p. 170.
55. Carroll, N. *Beyond Aesthetics: Philosophical Essays*; Cambridge University Press: Cambridge, UK, 2001.
56. Carroll, N. The Philosophy of Motion Pictures-PhilPapers. Available online: https://philpapers.org/rec/CARTPO-112 (accessed on 1 January 2025).
57. Carroll, N. *On Criticism*; Routledge: New York, NY, USA, 2008. [CrossRef]
58. Durà-Vilà, V. Attending to Works of Art for Their Own Sake in Art Evaluation and Analysis: Carroll and Stecker on Aesthetic Experience. *Br. J. Aesthet.* **2016**, *56*, 83–99.
59. Carroll, N. On the Historical Significance and Structure of Monroe Beardsley's Aesthetics: An Appreciation. *J. Aesthetic Educ.* **2010**, *44*, 2–10. [CrossRef]
60. Leder, H.; Nadal, M. Ten Years of a Model of Aesthetic Appreciation and Aesthetic Judgments: The Aesthetic Episode–Developments and Challenges in Empirical Aesthetics. *Br. J. Psychol.* **2014**, *105*, 443–464. [CrossRef]
61. Hitsuwari, J.; Ueda, Y.; Yun, W.; Nomura, M. Does Human–AI Collaboration Lead to More Creative Art? Aesthetic Evaluation of Human-Made and AI-Generated Haiku Poetry. *Comput. Hum. Behav.* **2023**, *139*, 107502. [CrossRef]
62. Shusterman, R. Aesthetic Experience: From Analysis to Eros. *J. Aesthet. Art Crit.* **2006**, *64*, 217–229. [CrossRef]
63. Wreen, M. *Beardsley's Aesthetics*; Hackett Publishing Company: Indianapolis, IN, USA, 2005.
64. LEÓN, M.J.A. Art in Three Dimensions by Carroll, Noël. *J. Aesthet. Art Crit.* **2012**, *70*, 231–233. [CrossRef]
65. Shusterman, R. *Pragmatist Aesthetics: Living Beauty, Rethinking Art*; Rowman & Littlefield: Lanham, MD, USA, 2000.
66. Berleant, A. *Sensibility and Sense: The Aesthetic Transformation of the Human World*; Andrews UK Limited: Luton, UK, 2011.
67. Lin, F.; Xu, W.; Li, Y.; Song, W. Exploring the Influence of Object, Subject, and Context on Aesthetic Evaluation through Computational Aesthetics and Neuroaesthetics. *Appl. Sci.* **2024**, *14*, 7384. [CrossRef]
68. Jaruga-Rozdolska, A. Artificial Intelligence as Part of Future Practices in the Architect's Work: MidJourney Generative Tool as Part of a Process of Creating an Architectural Form. *Architectus* **2022**, *3*, 95–104. [CrossRef]

69. Wen, Y.; Jain, N.; Kirchenbauer, J.; Goldblum, M.; Geiping, J.; Goldstein, T. Hard Prompts Made Easy: Gradient-Based Discrete Optimization for Prompt Tuning and Discovery. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 51008–51025.
70. Xie, Y.; Pan, Z.; Ma, J.; Jie, L.; Mei, Q. A Prompt Log Analysis of Text-to-Image Generation Systems. In Proceedings of the ACM Web Conference 2023, Austin, TX, USA, 1–5 May 2023; pp. 3892–3902. [CrossRef]
71. Fenfen, L.; Zimin, Z. Research on Deep Learning-Based Image Semantic Segmentation and Scene Understanding. *Acad. J. Comput. Inf. Sci.* **2024**, *7*, 43–48. [CrossRef]
72. Fenfen, L.; Zimin, Z. Research on Image Style Transfer and Artistic Creation Algorithm Based on Generative Adversarial Networks. *Int. J. New Dev. Eng. Soc.* **2024**, *8*, 1–6. [CrossRef]
73. Zhang, Z.; Fort, J.M.; Mateu, L.G. Exploringthe Potential of Artificial Intelligence as a Tool for Architectural Design: A Perception Study Using Gaudí'sworks. *Buildings* **2023**, *13*, 1863. [CrossRef]
74. Cohen, M.W.; Cherchiglia, L.; Costa, R. Evolving Mondrian-Style Artworks. In *Computational Intelligence in Music, Sound, Art and Design*; Correia, J., Ciesielski, V., Liapis, A., Eds.; Springer: Cham, Switzerland, 2017; pp. 338–353. [CrossRef]
75. Yusa, I.M.M.; Yu, Y.; Sovhyra, T. Reflections on the Use of Artificial Intelligence in Works of Art. *J. Aesthet. Des. Art Manag.* **2022**, *2*, 152–167. [CrossRef]
76. Pavlidis, T. *Algorithms for Graphics and Image Processing*; Springer: Berlin/Heidelberg, Germany, 2012.
77. Roddy, S. Creative Machine-Human Collaboration: Toward a Cybernetic Approach to Artificial Intelligence and Machine Learning Techniques in the Creative Arts. In *AI and the Future of Creative Work*; Routledge: Oxfordshire, UK, 2023.
78. Grba, D. Deep Else: A Critical Framework for Ai Art. *Digital* **2022**, *2*, 1–32. [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.

computers

MDPI

Article

# "Optimizing the Optimization": A Hybrid Evolutionary-Based AI Scheme for Optimal Performance

**Agathoklis A. Krimpenis [1,*] and Loukas Athanasakos [2]**

[1] Department of Mechanical Engineering, Hellenic Mediterranean University, Estavromenos, 71410 Heraklion, Greece
[2] Core Department, National and Kapodistrian University of Athens, 34400 Psachna, Greece; lathanasakos@core.uoa.gr
* Correspondence: akrimpenis@hmu.gr

**Abstract:** Optimization algorithms for solving technological and scientific problems often face long convergence times and high computational costs due to numerous input/output parameters and complex calculations. This study focuses on proposing a method for minimizing response times for such algorithms across various scientific fields, including the design and manufacturing of high-performance, high-quality components. It introduces an innovative mixed-scheme optimization algorithm aimed at effective optimization with minimal objective function evaluations. Indicative key optimization algorithms—namely, the Genetic Algorithm, Firefly Algorithm, Harmony Search Algorithm, and Black Hole Algorithm—were analyzed as paradigms to standardize parameters for integration into the mixed scheme. The proposed scheme designates one algorithm as a "leader" to initiate optimization, guiding others in iterative evaluations and enforcing intermediate solution exchanges. This collaborative process seeks to achieve optimal solutions at reduced convergence costs. This mixed scheme was tested on challenging benchmark functions, demonstrating convergence speeds that were at least three times faster than the best-performing standalone algorithms while maintaining solution quality. These results highlight its potential as an efficient optimization approach for computationally intensive problems, regardless of the included algorithms and their standalone performance.

**Keywords:** metaheuristic algorithms; normalization; stochastic approach; distributed optimization; exploration; exploitation

## 1. Introduction

Optimization is defined as the selection of the best possible solution based on predetermined criteria out of a set of available alternative solutions [1]. Optimization problems arise in all fields that require quantitative data management, predominantly in computer science, engineering [2], operational research, and economics. The development of optimization methods has been of interest in mathematics for centuries [3], and research interest in this area remains undiminished. In the simplest case, an optimization problem involves maximizing or minimizing a real-valued function by systematically selecting input values from a permissible set and calculating the output values of the function, attempting to approach the ideal solution. Generally, optimization entails finding the "best available" values of some objective function given a specified range of values (or inputs) [4]. The best values of a function may consist of a multitude of local optima, with the goal of the optimization process being to find the global optimum extremum. To find ideal minima or maxima in an optimization problem, heuristic methods

such as Genetic Algorithms, the Firefly Algorithm, and many other related evolutionary-based optimization algorithms are used. In most cases with optimization problems, the objective is to minimize or maximize scoring functions.

Many automated software tools for optimization generally apply either a maximization or a minimization task, not both [5]. Therefore, a maximization problem can be converted into a minimization one (and vice versa) by adding a negative sign in the objective function's formula. Since evolutionary-based algorithms operate stochastically, especially in multi-objective optimization schemes, finding optimum objective function values is expressed using the Pareto front theory. The Pareto principle (also known as the 80/20 rule) states that approximately 80% of the effects come from 20% of the causes [6]. Based on this, the Pareto front is a set of non-dominated solutions, chosen as optimal if no goal can be improved without sacrificing at least one other goal [7]. In an optimization problem, the entire solution space is simultaneously the output value domain of the function. In constrained optimization, only sub-sets of the output value space constitute the solution space. Additionally, the imposed constraints could be either expressed as equalities or inequalities.

Recently, significant interest has focused on finding solutions for a wide range of challenging optimization problems by scientists and researchers from various fields. This interest extends beyond academic and research purposes to address the expanding needs in real-life applications. The most widespread practice involves recommending algorithms based on elements observed in nature. This is due to the remarkable similarity between the way natural laws (such as natural phenomena) interact and the behavior of living entities (such as genes or a swarm of insects) in problem-solving. A comparison is made with how the human mind tackles similar problems.

Algorithms for global optimization are usually divided into two major categories: deterministic and stochastic [8]. In deterministic global optimization, the solution search space is the complete domain of all independent or input variables. At each step, the value range of one of the variables is chosen. Subsequently, the lower and upper bounds of the objective function values are calculated with regard to the current region. If these bounds are within a given distance from each other, the upper limit is accepted as the best optimum in the current region. If it is better than the previous overall best optimum, the current replaces the previous; otherwise, the region is discarded. If the bounds are farther apart, a variable for branching and a branching point are selected, and the region is divided into smaller sub-regions. The algorithm terminates when there are no more regions to examine. In the stochastic approach to the global optimization of a function, two phases are considered: global and local. In the 'global phase', the function is evaluated at a number of random points generated from a uniform distribution. In the 'local phase', the corresponding points are used as starting points for a local minimum search. The effectiveness of methods utilizing multiple stages depends on both the performance of the overall stochastic phase and the corresponding local minimization phase.

As the state-of-the-art progresses, and with the current popularity of Artificial Intelligence, the main scientific application fields that employ optimization algorithms are engineering [9] (e.g., optimizing aerodynamic shapes, designing lightweight yet robust structures); finance [10] (e.g., balancing risk and return or calibrating trading algorithms); business development [11] (e.g., optimizing customer clusters for personalized marketing); molecular modeling [12] (e.g., optimizing materials for catalytic reactions or simulating protein structures); biology [13] (e.g., designing synthetic gene circuits or accurately assembling fragmented DNA sequences into complete genomes); and machine learning [14] (e.g., optimizing communication and model aggregation for federated learning or optimizing neural network architecture).

*Motivation*

In metaheuristic algorithms, the term 'meta' characterizes algorithms that operate at a higher level than heuristics, generally performing better than simple heuristic schemes. All metaheuristic algorithms compensate for the focus on local searches of solutions by exploring the entire solution space [15]. Solution diversity is often achieved through randomization. Randomization is an effective way to limit local search compared to searching the entire solution space. This minimizes the chances of the algorithm getting trapped in local minima. Almost all metaheuristic algorithms are usually suitable for nonlinear modeling and global optimization. The primary components of any metaheuristic algorithm are intensification (exploitation) and diversification (exploration). Diversification means searching for different solutions to the extent that allows for exploration of as much as possible of the available value range, while intensification means focusing on nearby available values, utilizing the information that a good solution is in that region. Selection of the best ensures that the solutions eventually converge to the optimal one. These metaheuristic algorithms can also be used in hybrid schemes [16], commonly referred to as hyper-heuristics [17], where a combination of heuristic and metaheuristic algorithms can be used in tandem. These hybrid schemes, although generally more flexible and adaptable to problem-solving, still rely on selecting one algorithm that ultimately solves the optimization problem [18]. In this article, a thorough build and evaluation of a complex algorithmic scheme for the optimization of problems will be presented. The developed system constitutes a general tool for "Optimizing the Optimization", following an approach that is different from hyper-heuristics, but still applicable to a plethora of problems that require obtaining the global optimum for an analytically expressed objective function. In the literature, there appear to be numerous metaheuristic optimization algorithms that could be employed to that end [19] for design and parameter-tuning purposes. However, as this study encompasses introducing a new way of utilizing them, the purpose is to create a tool that optimizes not only the problem at hand but also the optimization process itself. Within this manuscript, Section 2 ("Materials and Methods") describes the methodology we followed, the quantization of different parameters that we performed, the structure of the individual algorithms and the hybrid scheme. This section also explains the benchmark objective functions that we selected. Section 3 ("Results") presents the outcomes of the conducted research in detailed tables and figures, while Section 4 ("Discussion") thoroughly explains the findings we made by conducting this research, and the breakthrough we identified. Finally, Section 5 ("Conclusions") provides the main points of the conducted research, as well as highlighting areas for improvement and the potential for future work.

## 2. Materials and Methods

Taking into consideration the review article by Stützle and López-Ibáñez [20], this study aims to create a novel hybrid evolution-based algorithmic scheme to optimize the optimization process, by employing metaheuristic algorithms in an advanced complex scheme. To achieve this goal, a set of optimization algorithms with different characteristics was selected to solve the problem. Note that these algorithms were selected indicatively, as they are well established and utilized in a vast field of applications; they were not selected because of their "better" performance over other more specialized modern ones. Subsequently, through the proposed innovative scheme, these algorithms collaborate to solve the optimization problem, aiming for the optimal solution with the fewest possible evaluations of the objective function. Simultaneously, the approach maximizes the exploration of the solution space and intensifies the search for optima. The architecture of the developed scheme is dynamic and provides the capability to use any combination of

algorithms. However, for proof of concept, we selected the following four algorithms, as in the literature these have been proven very efficient in coping with technological problems that involve large numbers of input and output parameters: the Genetic Algorithm [21], Firefly Algorithm [22], Harmony Search Algorithm [23], and Black Hole Algorithm [19].

### *2.1. Grouping and Normalization of Unique Initial Values*

Optimization algorithms consist of a set of input parameters, which vary for each, depending on their principle and operational philosophies. To create a unified method that can call them simultaneously and cooperatively, these values must be recognized and normalized based on their correlation. As has been mentioned in the literature, all optimization algorithms are governed by two fundamental characteristics: the degree of exploration of the solution space (exploration) and the degree of intensification of specific local regions of the solution (exploitation). Based on this, the normalization of all individual algorithm parameters for the set of algorithms was done by focusing on which of these two fundamental characteristics they influence.

The process begins with the identification of parameters common to all of the selected algorithms:

- **Objective Function**: The function to be solved is common to all algorithms, as this is the base that all optimization algorithms utilize to compare solutions. The algorithms evaluate the quality of solutions based on their ability to minimize or maximize the objective function. In each iteration, all the algorithms' processes are assessed through the objective function.
- **Problem Dimensions**: The dimensions of the objective function are common to all algorithms. Any problem can have dimensions ranging from 1 to N, with each dimension representing an independent variable based on the function's form and the solving objective.
- **Value Range**: The solution space, according to the objective function, has a predetermined range within which algorithms should search for possible solutions. This range is determined by constraints imposed by the objective function itself or the additional constraints set and can vary from a very narrow value field to infinity. The range is characterized by a set of minimum and maximum values, with each dimension having its declared minimum and maximum. Depending on the problem type and constraints, the sets of minima or maxima can be characterized by different values for each independent variable.
- **Number of Iterations**: One of the essential optimization algorithm variables is the total number of iterations it will perform before stopping and returning the best values it has identified up to that point. Since these algorithms are stochastic, and the ideal solution is often unknown, the number of iterations is one of the most crucial factors for the success of an algorithm to converge. For this study, all algorithms must be applied for the same number of iterations. Due to the different architecture of each algorithm, to achieve this, the iterations are measured by the number of objective function evaluations.
- **Population of Solutions**: Each algorithm initializes and maintains a predetermined number of potential solutions. This number is predetermined, and despite changes made to facilitate convergence to an optimal solution, the size remains constant. The number of maintained solutions is equally important for an optimization algorithm because there is a correlation between the population and the speed and quality of convergence, especially as the difficulty level, which is imposed by the objective function, increases.

In addition to the initial input values, which are the same for all of the optimization algorithms used, each algorithm has a unique set of parameters shaped by its particular characteristics. To enable these algorithms to be called under the same initial conditions, these specific parameters for each algorithm should be mapped based on their influence on either the degree of exploration or the degree of exploitation. Furthermore, for better uniformity, specific algorithm variables that characterize both the degree and the rate of parameterization have been grouped for their influence on the "mutation rate". These variables can take real values from 0 to 1, and for each algorithm, the mapping of these values is performed as follows:

**Genetic Algorithm (GA):**

- Beta: This value determines the rate of selecting individuals from the solution set for mutation. Due to its influence on expanding the search in the solution space, this value corresponds to "Exploration".
- PC: The term "PC" comes from "Population to Children", which translates to the correlation of the total population with the number of "offspring" or new values generated for each iteration. This variable determines the percentage of differentiation in the set of values, thus corresponding to "Exploitation".
- Gamma: The Gamma variable determines the intensity of crossover between the parameterized values. As this parameterization affects the way the solution space is explored, this term corresponds to "Exploration".
- MU: MU determines the percentage of features parameterized in each iteration for each chromosome. In other words, this value determines how much each chromosome to be parameterized changes. This value corresponds to the "Mutation Rate".
- Sigma: The last parameter that characterizes only GAs is Sigma. This parameter determines the size of parameterization for each chromosome to be parameterized, depending on MU. This variable also corresponds to the "Mutation Rate".

**Firefly Algorithm (FA):**

- Alpha: This variable determines the degree of randomization with which fireflies move that do not have the highest "intensity" or suitability as per the objective function. Thus, their degree of change is partly determined by this parameter, and for this reason, it corresponds to the "Mutation Rate".
- Beta: This term determines the degree of attraction with which fireflies are attracted to the "radiance" of the "brightest", or in other words, the better solution. The larger this term, the greater the attraction exhibited by the optimal fireflies toward the others, creating a tendency to gather the others towards them, thereby increasing the "Exploitation" characteristic of the algorithm.
- Gamma: This term collaborates with the Beta value and determines the degree to which the remaining fireflies react to this attraction, or how much light they allow to be absorbed and influence them. Conversely, this term determines the "Exploration".

**Harmony Search Algorithm (HSA):**

- Harmony Memory Considering Rate (HMRC): With this term, the algorithm decides how many of the stored values (melodies) in the form of a solution population it will choose to use per iteration compared to new ones that will be generated. Thus, the size of the solution space in which the algorithm will search for new solutions is determined, and it corresponds to "Exploration".
- Pitch Adjusting Rate (PAR): This variable characterizes the maximum allowable differentiation that an existing solution can have when it is deemed necessary to parameterize. In short, with the term PAR, it is determined how close a new solution will appear to the existing solution, thus determining the "Mutation Rate".

- Maximum Pitch Adjustment Proportion/Index (MPAP/MPAI): These two parameters essentially have the exact influence on a harmony search algorithm. One is used when the function to be solved is continuous, while the other is used when we have a function with discrete values. In practice, these variables define the range of conversion that all parameterized values-melodies will undergo for each iteration. Consequently, it directly affects "Exploitation".

For the **Black Hole Algorithm (BHA),** there are no specific predefined variables beyond the common initial input values described earlier.

In summary, to enable the four selected algorithms to "run" in a mixed scheme with the same initial values, the individual unique variables governing the operation of each algorithm should be grouped based on some common features characterizing all of the optimization algorithms. More specifically:

- Exploration Degree:
  - Beta, Gamma (Genetic Algorithm)
  - Gamma (Firefly Algorithm)
  - HMRC (Harmony Search Algorithm)
- Exploitation Degree:
  - PC (Genetic Algorithm)
  - Beta (Firefly Algorithm)
  - MBAP/MBAI (Harmony Search Algorithm)
- Mutation Rate:
  - MU, Sigma (Genetic Algorithm)
  - Alpha (Firefly Algorithm)
  - PAR (Harmony Search Algorithm)

By grouping the above variables, in combination with the common parameters introduced in each algorithm, a mixed scheme can be designed with which they will be invoked.

### 2.2. *Algorithmic Scheme*

The goal is to create a hybrid optimization algorithm scheme that calls upon these algorithms with the aim of solving a common objective function. To achieve this goal, a novel algorithm will be developed, acting as the leader for this hybrid scheme. If all the algorithms are considered players in a cooperative problem-solving game, the leader algorithm will determine their initial values, input variables for each iteration, as well as when and how they should stop. To start, the selection of an objective function is crucial. Reference functions from the literature will be used for evaluating the performance of the scheme and for measuring the time required for the scheme to reach predefined accuracy levels for known solutions. By utilizing these functions, the search range of solutions, the function's elasticity for use with multiple independent variable values, the number of local optima, and the global optimum, along with the value extracted from this reference objective function, are known from the beginning. With the entire value domain known, it becomes possible to evaluate each optimization algorithm separately and the overall performance of the scheme, both individually and in comparison, to find a solution with each algorithm separately if the scheme is not utilized.

#### 2.2.1. Objective Functions

For proof of concept, three different reference functions from the literature [24] will be utilized. Each one has its own value range and optimal value at its specific point, and all were selected to provide solutions for multiple choices of independent variables, depending

on the preselection made during the initialization of values. This allows the same function to be tested within various dimensions, thus evaluating the solution quality as complexity increases. The functions used for this particular study appear in Table 1.

**Table 1.** Reference functions selected, their value range and optimized result.

| Function | Objective Function | Value Range | Optimized Result |
|---|---|---|---|
| Sphere | $f(x) = \sum_{i=1}^{D} x_i^2$ | $-\infty \leq x \leq +\infty$ | $x^* = f(0, \ldots, 0)$, $f(x^*) = 0$ |
| Schwefel | $f(x) = 418.9829d - \sum_{i=1}^{d} x_i sin(\sqrt{\lvert x_i \rvert}$ | $-50 \leq x_i \leq 50$ | $f(420.9687, \ldots, 420.9687) = 0$ |
| Alpine N.2 | $f(x) = \prod_{i=1}^{D} \sqrt{x_i} sin(x_i)$ | $0 \leq x_i \leq 10$ | $x^* = f(7.917, \ldots, 7.917)$, $f(x^*) = 2.808^D$ |

2.2.2. Architecture of the Algorithmic Scheme

Having defined the objective function for optimization and prepared the algorithms regarding their input variables and their association with their respective influencing factors (Exploration, Exploitation, and Mutation Rate), the leader algorithm provides initial values to feed them. After the initial feeding, all algorithms attempt to solve the problem with equal treatment from the leader. For optimization of the optimization, the algorithms need to interact with each other from iteration to iteration. This is achieved by setting all algorithms to simultaneously optimize the same problem for a finite number of iterations, evaluating their results after each iteration. The algorithm that exhibits the best performance will feed all of the others with the optimal solution, while in the case of the worst-performing algorithm, it will be fed with the entire population of solutions from the best one, thus aiding through the elitism method to increase the chances of finding a better point in the next iteration (see Figure 1). For the others, apart from being fed with the optimal solution, the rest of the population remains constant from their previous iteration. In this way, by setting a small number of iterations for each individual algorithm at each step of the mixed scheme, convergence to the global optimum is sought with as few evaluations of the objective function as possible overall.

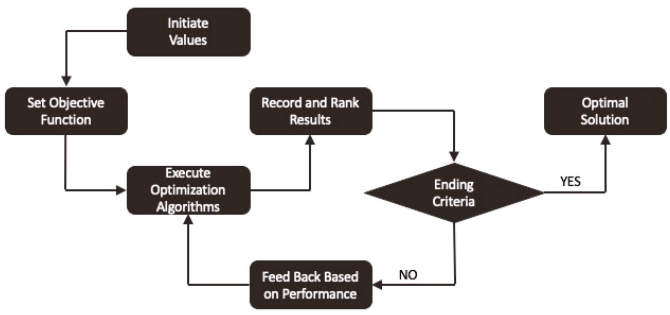



**Figure 1.** Overview flow chart of the hybrid algorithmic scheme.

By keeping the values for the exploration degree, exploitation degree, and mutation rate constant and in the middle of their range, it is ensured that the exploration and exploitation levels of the algorithms will remain at their average and equal for all players. The leader, by feeding all algorithms with information from the algorithm that has the best performance, indirectly attempts to maximize both the exploration and exploitation degrees simultaneously for the scheme. This is achieved without sacrificing time to evaluate the objective function more times, while at the same time minimizing the risk of the mixed scheme getting trapped in a local optimum.

The mixed optimization scheme guided by the leader will repeat the above process until a termination criterion is satisfied. This criterion may either concern the overall convergence degree (e.g., the order of magnitude of convergence relative to an ideal termination value) or a predefined, finite number of iterations for a problem with an unknown ideal value.

#### 2.2.3. Initial Values

These values, as defined by the optimization algorithms' literature, are the following:

> Problem dimensions (independent variables) (e.g., d = 2);
> Minimum value in the domain (where: Min = [value] * d, for creating a minimum for each dimension);
> Maximum value in the domain (where: Max = [value] * d, for creating a maximum for each dimension);
> Number of iterations (e.g., iterations = 20);
> Population size (e.g., population = 1000);
> Exploration rate (e.g., exploration = 0.5);
> Exploitation rate (e.g., exploitation = 0.5);
> Mutation rate (e.g., mutation_rate = 0.5).

These initial values are manually defined intuitively, compressed into a single variable with slave values (variable type: Structure), and fed as a 'package' to all of the optimization algorithms, which, in turn, distribute them to their corresponding specific parameters.

#### 2.2.4. Special Values

To achieve the smooth operation of the scheme, some additional initial values must be determined, characterized as special for the leader. These values assist the leader algorithm in making decisions regarding the distribution of results per iteration step of the mixed scheme and determining the criterion by which it will terminate the process and extract the optimal result. These variables are being created through the results of each iteration within the leader algorithm, and are automatically allocated accordingly. These values are defined as follows:

> Optimal solution: The optimal solution is the best-performing value presented by the best algorithm compared to the others, e.g., for a problem with two independent variables, the point (1,1).
> Fitness of the optimal solution (Cost): Fitness is defined as the value given at the output of the objective function at the point of the optimal solution, e.g., Cost = 2, related to fitness for the point (1,1) and the function $f(x, y) = x + y$.
> Selection index: To allow the leader to determine which player performs better or worse, there must be a logical index (true/false) that, in each iteration of the mixed scheme, takes the appropriate value concerning the performance of the players. This selection index adjusts to the compressed single variable fed to the algorithms. Thus, the algorithms know whether they should consider either a better solution or an entire population for their next optimization attempt.
> Distributed performance list: For feedback to the user during the execution of the mixed scheme and for returning the optimal solution at all levels (optimal cost, optimal position), a distributed performance list is created. This list ranks the algorithms from best to worst and returns the corresponding index for each.

#### 2.2.5. Objective Function Evaluations

In addition to the above parameters, it is crucial to align how each algorithm will attempt to improve the objective function. The aim of the overall study is the optimization

of optimization; therefore, for the results to be reliable, there must be homogeneity among the players of the mixed scheme. Each player-algorithm, for a set number of optimization iterations, examines a possible optimal solution a predetermined number of times. This means that, when assigning initial values to individual optimization algorithms, it must be considered that for one iteration of the mixed scheme, all algorithms will have evaluated the objective function an equal number of times, thus avoiding bias in favor of any algorithm performing more evaluations than the others in an iteration.

### *2.3. Programming Environment*

For the implementation of the above, a programming language should be selected to transfer these relationships into a digital computational environment. In theory, any programming language can fulfill this need. For the purposes of this study, Python has been chosen. Especially in the field of optimization, Python is an indispensable tool due to its low complexity and flexibility in quickly transforming models and mathematical relationships into code.

For this study, all the code was implemented using Python version 3.7.10 within the Anaconda programming environment. Sublime Text software (Version 4126) was used for code processing, while the implementation and execution of the code were carried out using version 5.0.0 of the Spyder software. All development and testing were performed on an Apple MacBook Pro laptop with an M1 Pro processor and 16 GB of RAM, purchased in Greece.

### *2.4. Hybrid Scheme Rules and Programming*

After structuring the player-algorithms and the mixed scheme at a theoretical level, the next step is to translate all this knowledge into a computational environment using Python. Starting with the optimization algorithms, all of them need to be implemented in code. As optimization algorithms have been a topic in the literature for several years, and Python is an open-source language, implementations are readily available for use by the programming community. These implementations are widely accessible but require extensive parameterization to adapt and integrate into the mixed scheme of this research.

For each algorithm, an implementation was chosen that aligns with the goals of the mixed scheme and was parameterized according to the following rules:

1. All algorithms should "accept" and "perceive" the values of the domain in the same way. This means that for each function $f(x_i)$ fed into the algorithms, where $x_i$ represents the value from the domain for each time instance, it should have a common variable format for both the leader algorithm and the players. This format was chosen as "NumPy Array". In practice, NumPy Arrays are numeric lists created by the "NumPy" library, which is part of Python's source code libraries. This choice was made to ensure input uniformity based on an independent, standalone library from the programming language.
2. All algorithms should output the optimal value from the domain and the set of solution populations as a list, decoded by NumPy. This choice serves the leader algorithm to manage, compare, and present data independently of how they are generated in the individual player algorithms.
3. All algorithms should be able to assign initial and special values appropriately, according to the correspondences made in their theoretical study.
4. All algorithms should be able to extract the same parameters in the same format. Specifically, the output values for all of the algorithms are as follows: (i) optimal input value, (ii) optimal value of the objective function, and (iii) total population at the last iteration.

5. All algorithms should parameterize the value of the maximum iterations given to them, according to the number of evaluations of the objective function they perform in total, so that all of the algorithms evaluate the objective function the same number of times in each "call".
6. All algorithms should be able to accept the objective function encoded in the same way.
7. The algorithms should be in the form of a library, which will be imported by the leader, who, in turn, can call on the execution method, providing the compressed initial and special values and the objective function as input variables.

Having parameterized the algorithms accordingly, the next step is to structure the leader algorithm. The structure of the leader is described in pseudocode as in Figure 2.

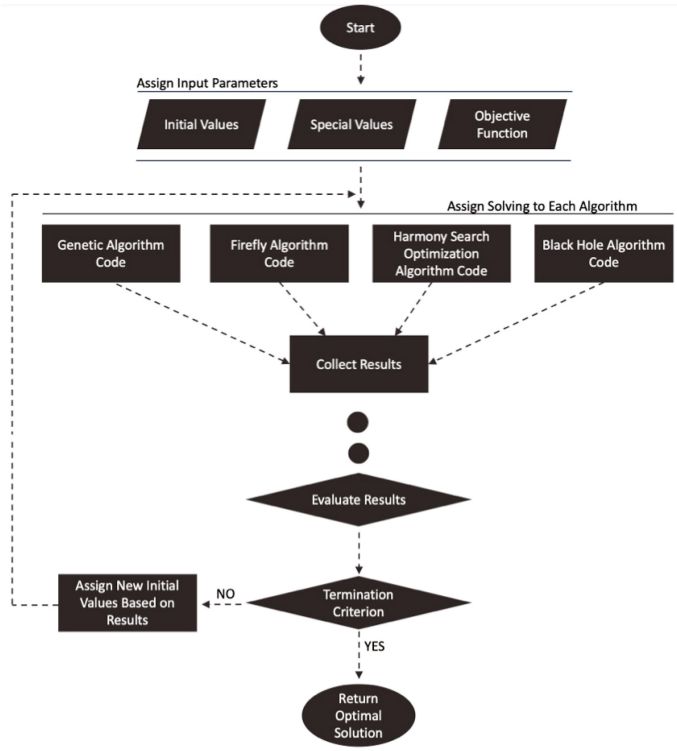



**Figure 2.** Pseudocode sequence of the hybrid scheme leader.

The form of the leader can either act as a coordinator of the mixed scheme or simply as a single controller of the individual algorithms without creating interaction between them for their isolated evaluation. Also, it must be clarified that in order to maximize the convergence speed, after each iteration the optimal results obtained by any of the algorithms, but not whole populations, are seeded to all the algorithms of the next iteration. The first stage of this study is to test the performance of the algorithms for each reference function separately, in order to draw conclusions regarding their optimization capability. Assessing the quality of solutions for each optimization algorithm separately, based on the aforementioned benchmark functions, is then applied to the mixed scheme with the same problem. The goal is to achieve comparable or better results with overall fewer evaluations of the objective function.

## 3. Results

### *3.1. Testing Optimization Algorithms Individually*

For the evaluation of the algorithms, a problem-solving approach was applied to the three reference functions, where the optimal value is previously known from the literature. This

knowledge, however, was not introduced to the algorithms; thus, an objective assessment of the algorithms' ability to optimize can be performed. By varying the total number of iterations, problem dimensions, and solution population for each trial, the results of Tables 2–4 were obtained. For each evaluation, the algorithms executed a fixed number of iterations with a predefined population of solutions, and the solution values retrieved were the ones that achieved the best fitness at the end of each complete cycle of iterations. For each different set of parameters, a completely new trial was performed, generating new initial values and running independently of the others. Finally, for each reference function, a final trial with the maximum referenced population of solutions was perceived as indicative of the effectiveness of each algorithm in providing a solution that was within $10^{-2}$ of the desired one.

**Table 2.** Sphere reference function optimization results. Legend: ITE = Number of iterations with a complete population, POP = Population number, GA = Genetic Algorithm, DEV = Absolute deviation of returned solution to the ideal one, FA = Firefly Algorithm, HAS= Harmony Search Algorithm, BHA = Black Hole Algorithm.

| **Dimensions—Independent Values = 2** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **ITE** | **POP** | **GA** | **DEV-GA** | **FA** | **DEV-FA** | **HSA** | **DEV-HSA** | **BHA** | **DEV-BHA** |
| 50 | 200 | $2.249 \times 10^{-7}$ | $2.249 \times 10^{-7}$ | 0.415 | 0.415 | 0.171 | 0.171 | 0.020 | 0.020 |
| | 400 | $2.990 \times 10^{-7}$ | $2.990 \times 10^{-7}$ | 0.340 | 0.340 | 0.018 | 0.018 | 0.343 | 0.343 |
| | 1000 | $2.599 \times 10^{-8}$ | $2.599 \times 10^{-8}$ | 0.318 | 0.318 | 0.096 | 0.096 | 0.082 | 0.082 |
| 240 | 200 | $1.868 \times 10^{-21}$ | $1.868 \times 10^{-21}$ | 0.010 | 0.010 | 0.407 | 0.407 | 0.247 | 0.247 |
| | 400 | $1.762 \times 10^{-21}$ | $1.762 \times 10^{-21}$ | 0.456 | 0.456 | 0.082 | 0.082 | 0.075 | 0.075 |
| | 1000 | $3.057 \times 10^{-23}$ | $3.057 \times 10^{-23}$ | 0.095 | 0.095 | 0.013 | 0.013 | 0.057 | 0.057 |
| Comparison of each algorithm reaching an accuracy of $10^{-2}$ | | | | | | | | | |
| Algorithm | ITE | POP | Best Value Returned | | | | | | |
| GA | 20 | 1000 | $1.880 \times 10^{-5}$ | | | | | | |
| FA | 1200 | 1000 | $2.464 \times 10^{-4}$ | | | | | | |
| HSA | 1000 | 1000 | $4.654 \times 10^{-3}$ | | | | | | |
| BHA | 700 | 1000 | $1.455 \times 10^{-3}$ | | | | | | |
| **Dimensions—Independent Values = 3** | | | | | | | | | |
| **ITE** | **POP** | **GA** | **DEV-GA** | **FA** | **DEV-FA** | **HSA** | **DEV-HSA** | **BHA** | **DEV-BHA** |
| 50 | 200 | $7.241 \times 10^{-5}$ | $7.241 \times 10^{-5}$ | 2.209 | 2.209 | 1.702 | 1.702 | 0.488 | 0.488 |
| | 400 | $4.668 \times 10^{-5}$ | $4.668 \times 10^{-5}$ | 2.061 | 2.061 | 7.337 | 7.337 | 2.300 | 2.300 |
| | 1000 | $1.557 \times 10^{-5}$ | $1.557 \times 10^{-5}$ | 1.572 | 1.572 | 0.318 | 0.318 | 3.049 | 3.049 |
| 240 | 200 | $2.727 \times 10^{-10}$ | $2.727 \times 10^{-10}$ | 0.469 | 0.469 | 0.354 | 0.354 | 3.053 | 3.053 |
| | 400 | $1.009 \times 10^{-10}$ | $1.009 \times 10^{-10}$ | 1.600 | 1.600 | 0.635 | 0.635 | 3.828 | 3.828 |
| | 1000 | $5.415 \times 10^{-11}$ | $5.415 \times 10^{-11}$ | 0.512 | 0.512 | 3.337 | 3.337 | 3.256 | 3.256 |
| Comparison of each algorithm reaching an accuracy of $10^{-2}$ | | | | | | | | | |
| Algorithm | ITE | POP | Best Value Returned | | | | | | |
| GA | 20 | 1000 | $1.694 \times 10^{-3}$ | | | | | | |
| FA | 2400 | 1000 | $1.130 \times 10^{-3}$ | | | | | | |
| HSA | 30,000 | 1000 | $7.109 \times 10^{-3}$ | | | | | | |
| BHA | - | - | No convergence for more than $10^{-1}$ | | | | | | |

**Table 3.** Schwefel reference function optimization results. Legend: ITE = Number of iterations with a complete population, POP = Population number, GA = Genetic Algorithm, DEV = Absolute deviation of returned solution to the ideal one, FA = Firefly Algorithm, HSA = Harmony Search Algorithm, BHA = Black Hole Algorithm.

| **Dimensions—Independent Values = 2** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **ITE** | **POP** | **GA** | **DEV-GA** | **FA** | **DEV-FA** | **HSA** | **DEV-HSA** | **BHA** | **DEV-BHA** |
| 50 | 200 | $3.198 \times 10^{-5}$ | $3.198 \times 10^{-5}$ | 44.615 | 44.615 | 131.036 | 131.036 | 101.910 | 101.910 |
| | 400 | $2.634 \times 10^{-5}$ | $2.634 \times 10^{-5}$ | 7.410 | 7.410 | 53.005 | 53.005 | 44.155 | 44.155 |
| | 1000 | $2.589 \times 10^{-5}$ | $2.589 \times 10^{-5}$ | 41.289 | 41.289 | 8.540 | 8.540 | 118.362 | 118.362 |
| 240 | 200 | $2.545 \times 10^{-5}$ | $2.545 \times 10^{-5}$ | 4.721 | 4.721 | 0.720 | 0.720 | 3.762 | 3.762 |
| | 400 | $2.545 \times 10^{-5}$ | $2.545 \times 10^{-5}$ | 48.525 | 48.525 | 31.229 | 31.229 | 196.613 | 196.613 |
| | 1000 | $2.545 \times 10^{-5}$ | $2.545 \times 10^{-5}$ | 43.090 | 43.090 | 97.473 | 97.473 | 60.672 | 60.672 |
| Comparison of each algorithm reaching an accuracy of $10^{-2}$ | | | | | | | | | |
| Algorithm | | ITE | POP | | Best Value Returned | | | | |
| GA | | 30 | 1000 | | $1.120 \times 10^{-4}$ | | | | |
| FA | | 10,000 | 1000 | | $7.456 \times 10^{-2}$ | | | | |
| HSA | | 10,000 | 1000 | | $3.748 \times 10^{-2}$ | | | | |
| BHA | | - | - | | No convergence | | | | |
| **Dimensions—Independent Values = 3** | | | | | | | | | |
| **ITE** | **POP** | **GA** | **DEV-GA** | **FA** | **DEV-FA** | **HSA** | **DEV-HSA** | **BHA** | **DEV-BHA** |
| 50 | 200 | 0.430 | 0.430 | 334.301 | 334.301 | 445.006 | 445.006 | 370.047 | 370.047 |
| | 400 | 0.013 | 0.013 | 101.121 | 101.121 | 114.760 | 114.760 | 270.679 | 270.679 |
| | 1000 | 0.001 | 0.001 | 338.982 | 338.982 | 153.056 | 153.056 | 152.497 | 152.497 |
| 240 | 200 | $3.818 \times 10^{-5}$ | $3.818 \times 10^{-5}$ | 128.789 | 128.789 | 136.208 | 136.208 | 306.345 | 306.345 |
| | 400 | $3.818 \times 10^{-5}$ | $3.818 \times 10^{-5}$ | 347.940 | 347.940 | 223.536 | 223.536 | 272.299 | 272.299 |
| | 1000 | $3.818 \times 10^{-5}$ | $3.818 \times 10^{-5}$ | 218.610 | 218.610 | 255.159 | 255.159 | 189.642 | 189.642 |
| Comparison of each algorithm reaching an accuracy of $10^{-2}$ | | | | | | | | | |
| Algorithm | | ITE | POP | | Best Value Returned | | | | |
| GA | | 50 | 1000 | | $3.546 \times 10^{-3}$ | | | | |
| FA | | - | - | | No convergence | | | | |
| HSA | | 150,000 | 1000 | | $9.954 \times 10^{-2}$ | | | | |
| BHA | | - | - | | No convergence | | | | |

**Table 4.** Alpine N.2 reference function optimization results. Legend: ITE = Number of iterations with a complete population, POP = Population number, GA = Genetic Algorithm, DEV = Absolute deviation of returned solution to the ideal one, FA = Firefly Algorithm, HSA = Harmony Search Algorithm, BHA = Black Hole Algorithm.

| **Dimensions—Independent Values = 2** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **ITE** | **POP** | **GA** | **DEV-GA** | **FA** | **DEV-FA** | **HSA** | **DEV-HSA** | **BHA** | **DEV-BHA** |
| 50 | 200 | 7.885 | 0.001 | 7.836 | 0.048 | 7.343 | 0.541 | 6.681 | 1.203 |
| | 400 | 7.885 | 0.001 | 7.795 | 0.089 | 7.872 | 0.012 | 7.759 | 0.125 |
| | 1000 | 7.885 | 0.001 | 7.710 | 0.174 | 7.846 | 0.038 | 7.865 | 0.019 |
| 240 | 200 | 7.885 | 0.001 | 7.871 | 0.013 | 7.753 | 0.131 | 7.678 | 0.206 |
| | 400 | 7.885 | 0.001 | 7.861 | 0.023 | 7.617 | 0.267 | 7.592 | 0.292 |
| | 1000 | 7.885 | 0.001 | 7.878 | 0.006 | 7.295 | 0.589 | 7.671 | 0.213 |
| Comparison of each algorithm reaching an accuracy of $10^{-2}$ | | | | | | | | | |
| Algorithm | | ITE | | POP | | Best Value Returned | | | |
| GA | | 3 | | 1000 | | 7.821 | | | |
| FA | | 300 | | 1000 | | 7.847 | | | |
| HSA | | 300 | | 1000 | | 7.463 | | | |
| BHA | | 200 | | 1000 | | 7.803 | | | |
| **Dimensions—Independent Values = 3** | | | | | | | | | |
| **ITE** | **POP** | **GA** | **DEV-GA** | **FA** | **DEV-FA** | **HSA** | **DEV-HSA** | **BHA** | **DEV-BHA** |
| 50 | 200 | 22.130 | 0.013 | 14.552 | 7.588 | 16.963 | 5.171 | 10.118 | 12.022 |
| | 400 | 22.141 | 0.001 | 16.225 | 5.915 | 19.221 | 2.919 | 20.047 | 2.093 |
| | 1000 | 22.143 | 0.003 | 18.846 | 3.294 | 20.776 | 1.364 | 21.504 | 0.636 |
| 240 | 200 | 22.143 | 0.003 | 21.433 | 0.707 | 19.410 | 2.730 | 9.617 | 12.523 |
| | 400 | 22.143 | 0.003 | 20.100 | 2.040 | 20.087 | 2.053 | 20.170 | 1.970 |
| | 1000 | 22.143 | 0.003 | 16.885 | 5.255 | 18.975 | 3.165 | 20.260 | 1.880 |
| Comparison of each algorithm reaching an accuracy of $10^{-2}$ | | | | | | | | | |
| Algorithm | | ITE | | POP | | Best Value Returned | | | |
| GA | | 30 | | 1000 | | 22.042 | | | |
| FA | | 3000 | | 1000 | | 22.075 | | | |
| HSA | | 13,000 | | 1000 | | 22.055 | | | |
| BHA | | - | | - | | No convergence | | | |

#### 3.1.1. Sphere Reference Function

The optimal solution of the function is 0, at the point $f(0,0)$. We set a fixed range of values with a minimum of $[-10]$ and a maximum of $[10]$ for each dimension, and we obtained the solutions as seen in Table 2, with a graphical representation in Figure 3.

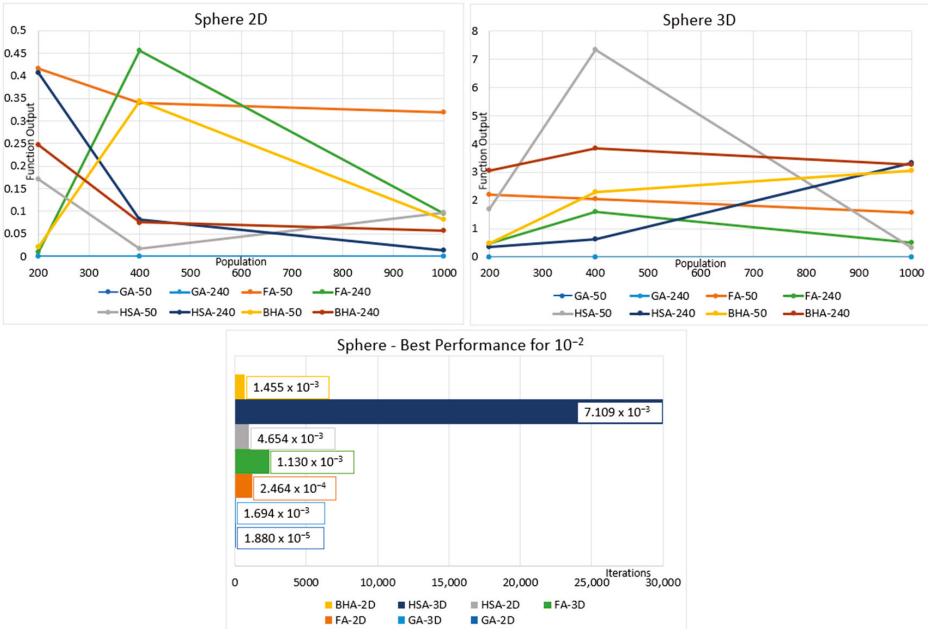



**Figure 3.** Representation of the results for the Sphere concerning 2D (**upper left**), 3D (**upper right**) and accuracy (**bottom**).

### 3.1.2. Schwefel Reference Function

The optimal solution of the function is 0, at the point $f(420.9687, \ldots, 420.9687)$. We set a fixed range of values with a minimum of $[-500]$ and a maximum of $[500]$ for each dimension, and we obtained the solutions as seen in Table 3, with a graphical representation in Figure 4.

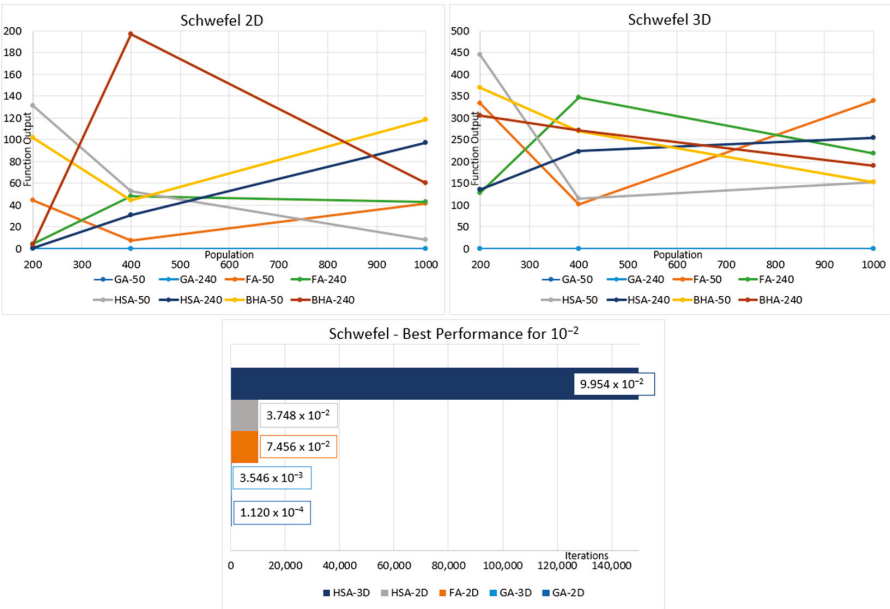



**Figure 4.** Representation of the results for the Schwefel concerning 2D (**upper left**), 3D (**upper right**) and accuracy (**bottom**).

### 3.1.3. Alpine n.2 Reference Function

The optimal solution of the function is $2.808^D$, where D is the number of dimensions, at the point $f(7.917, \ldots, 7.917)$. For two dimensions, the optimal value is 7.884764, while

for three dimensions, it is 22.140698. We set a fixed range of values with a minimum of [0] and a maximum of [10] for each dimension, and we obtained the solutions as seen in Table 4, with a graphical representation in Figure 5.

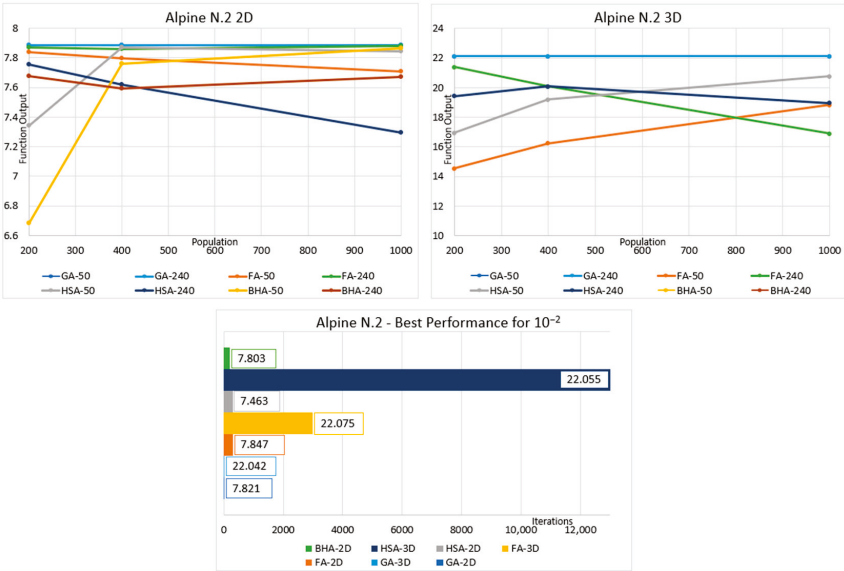



**Figure 5.** Representation of the results for the Alpine N.2 concerning 2D (**upper left**), 3D (**upper right**) and accuracy (**bottom**).

### *3.2. Testing the Hybrid Scheme*

Having examined how each algorithm individually solves the optimization problem, it is necessary to test the solution of the same problem by the hybrid scheme described in Section 2.2. With the same initial values defined for all of the algorithms and having configured the way values will be fed, we obtained the solution arrays of Table 5 for each reference function, which are graphically represented in Figure 6.

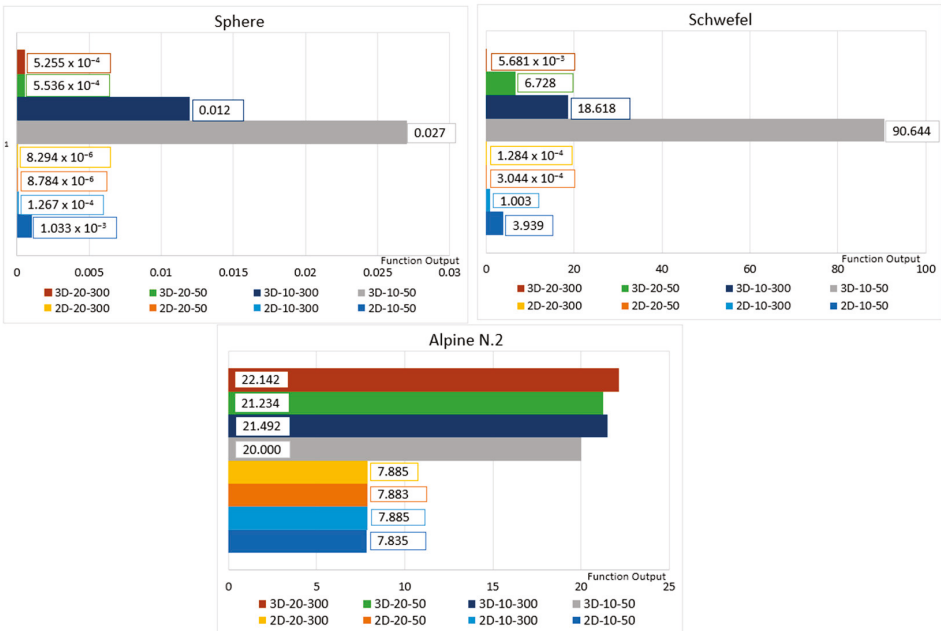



**Figure 6.** Representation of the results regarding the performance of the Hybrid scheme for Sphere (**left**), Schwefel (**right**), and Alpine N.2 (**center**).

**Table 5.** Hybrid scheme optimization results for all reference functions.

| Number of iterations with a complete population for each algorithm | Population for each algorithm | Total runs of the hybrid scheme | Returned output value of objective function | Total hybrid scheme deviation |
|---|---|---|---|---|
| **Sphere Reference Function** | | | | |
| Dimensions—Independent Values = 2 | | | | |
| 10 | 50 | 2 | $1.033 \times 10^{-3}$ | 0.001 |
| 10 | 300 | 2 | $1.267 \times 10^{-4}$ | 0.0001 |
| 20 | 50 | 2 | $8.784 \times 10^{-6}$ | $10^{-6}$ |
| 20 | 300 | 2 | $8.294 \times 10^{-6}$ | $10^{-6}$ |
| Total effort to reach an accuracy of $10^{-2}$ | | | | |
| 7 | 50 | 2 | $2.658 \times 10^{-2}$ | |
| Dimensions—Independent Values = 3 | | | | |
| 10 | 50 | 2 | 0.027 | 0.027 |
| 10 | 300 | 2 | 0.012 | 0.012 |
| 20 | 50 | 2 | $5.536 \times 10^{-4}$ | $10^{-4}$ |
| 20 | 300 | 2 | $5.255 \times 10^{-4}$ | $10^{-4}$ |
| Total effort to reach an accuracy of $10^{-2}$ | | | | |
| 10 | 50 | 2 | $2.700 \times 10^{-2}$ | |
| **Schwefel reference function** | | | | |
| Dimensions—Independent Values = 2 | | | | |
| 10 | 50 | 2 | 3.939 | 3.939 |
| 10 | 300 | 2 | 1.003 | 1.003 |
| 20 | 50 | 2 | $3.044 \times 10^{-4}$ | $10^{-4}$ |
| 20 | 300 | 2 | $1.284 \times 10^{-4}$ | $10^{-4}$ |
| Total effort to reach an accuracy of $10^{-2}$ | | | | |
| 16 | 50 | 2 | $5.532 \times 10^{-2}$ | |
| Dimensions—Independent Values = 3 | | | | |
| 10 | 50 | 2 | 90.644 | 90.644 |
| 10 | 300 | 2 | 18.618 | 18.618 |
| 20 | 50 | 2 | 6.728 | 6.728 |
| 20 | 300 | 2 | $5.681 \times 10^{-3}$ | $10^{-3}$ |
| Total effort to reach an accuracy of $10^{-2}$ | | | | |
| 20 | 300 | 2 | $5.681 \times 10^{-3}$ | |
| **Alpine N.2 reference function** | | | | |
| Dimensions—Independent Values = 2 | | | | |
| 10 | 50 | 2 | 7.835 | 0.049 |
| 10 | 300 | 2 | 7.885 | 0.001 |
| 20 | 50 | 2 | 7.883 | −0.001 |
| 20 | 300 | 2 | 7.885 | 0.001 |
| Total effort to reach an accuracy of $10^{-2}$ | | | | |
| 10 | 50 | 2 | 7.835 | |
| Dimensions—Independent Values = 3 | | | | |
| 10 | 50 | 2 | 20.000 | 2.140 |
| 10 | 300 | 2 | 21.492 | 0.647 |
| 20 | 50 | 2 | 21.234 | 0.906 |
| 20 | 300 | 2 | 22.142 | 0.002 |
| Total effort to reach an accuracy of $10^{-2}$ | | | | |
| 20 | 300 | 2 | 22.137 | |

## 4. Discussion

### *4.1. Evaluation Summary for Individual Algorithms*

All of the individual trials were performed for the same fixed number of iterations, number of population, and dimensions. This does not reflect the way each algorithm is "preferably" configured, but this has been selected on purpose. For example, since the Genetic Algorithms are inherently ideal for solving similar problems, they were expected to behave much better, across all of the trials. At the same time, a population of 1000 solutions per iteration is extreme, but was chosen to provide the rest of the algorithms an equalizing factor compared to the GA.

All algorithms behave differently in solving the same optimization problem. It was observed that the number of individuals in the population is extremely important compared to the number of objective function evaluations. In cases where the number is extremely low, the performance of the algorithms, except for the Genetic Algorithm, is very low and unstable, as there is not enough time to create a satisfactory convergence level. Similarly, as complexity increases in the dimensions and value ranges, things become even more challenging for the algorithms. Solutions have a very wide range, while their distance from the ideal solution fluctuates significantly depending on the initial parameter settings. In addition, since the nature of the selected optimization algorithms is stochastic, the initial values are generated at random. This randomness explains the large variability that all the algorithms have, apart from the better-suited GA. This variability was expected and desired, to highlight the different approach that each algorithm has, since this can become the strengthening factor when using these algorithms in a hybrid scheme. Finally, it was observed that, except for the Genetic Algorithms, when convergence concerns the optimal value at 0, the other algorithms significantly struggle to converge, and sometimes do not converge at all. This reinforces the argument that different algorithms are more suitable for different optimization problems. Some specific observations include:

- Genetic Algorithm: Exceptional ability to solve problems in a very short time for the majority of optimization problems.
- Particle Swarm Algorithm: Good problem-solving ability, but exhibits high instability depending on the combinations of initial values. Solution accuracy rarely exceeds $10^{-2}$, and it struggles to converge for optimal values close to 0.
- Harmony Search Algorithm: The agnostic way, meaning the way it globally evaluates the entire domain value it maintains as a population, requires an exceptionally large number of iterations to solve a problem. Nevertheless, the solutions converge with very high accuracy to the desired target.
- Black Hole Algorithm: The inability to adjust its reaction method makes it extremely unstable, to the point of not being able to converge in some cases. Its results have a high degree of randomness, regardless of the initial parameter values provided. As the domain of definition increases, so does the instability it exhibits.

### *4.2. Comparing the Hybrid Scheme to Individual Algorithms*

From the above process, we observed that the mixed-solution approach to optimization problems demonstrates several benefits. When comparing the individual performance of algorithms, our main observation is that algorithms require a certain number of evaluations of the objective function to achieve convergences of reasonable accuracy. The best performance was observed from the Genetic Algorithm, which consistently behaved well in solving all reference problems. Other algorithms required significantly more time to solve the same problem, especially when the dimensions (independent variables of the problem) were more than two. In such cases, many algorithms struggled to converge to

a global optimum, and in some cases, as in the Black Hole Algorithm, convergence was not achievable.

For minimization of the run time of the mixed optimization scheme, each algorithm was assigned to a different thread of a multi-thread CPU. This is easily applied, as each algorithm in each iteration is independent. The mixed optimization system that was introduced and implemented in this study seems to solve all the individual problems presented by the algorithms without sacrificing performance in any area. In all cases, the system achieves convergence with significantly fewer evaluations of the objective function, as shown in Figure 7. Based on individual evaluations for each algorithm, the difference in performance is at least three times better, meaning the mixed approach needs to evaluate the objective function by only one-third of the number needed by the best-performing standalone algorithm. This can be translated by multiplying the number of iterations to the number of populations selected to retrieve a total "effort" made by the processor. For example, for the best performing GA, in solving the Schwefel function in 3D, the processor running 50 ITE * 1000 POP returned with a best value within $10^{-3}$ from the target, while with the hybrid scheme, the same calculation only cost the processor 20 ITE * 300 POP, for very similar accuracy. This translates to roughly eight times less computational cost for the same accuracy. If we compare this to HSA, for example, on the same reference function it required some significant 150,000 ITE * 1000 POP to achieve a solution with an accuracy of $10^{-2}$, thus requiring a 250 times higher computational cost! Additionally, during the use of the mixed approach, an index was used to indicate the algorithm that achieved the optimal solution. This was done to evaluate the system itself, as a scheme that always takes the optimal solution from the same algorithm has limitations in its use. According to this logic, while individual evaluation showed that the performance of the Genetic Algorithm was stable and better than the others, during the use of the mixed approach, it was observed that the Genetic Algorithm did not always provide the optimal solution, making the scheme flexible against the diversity of optimization problems.

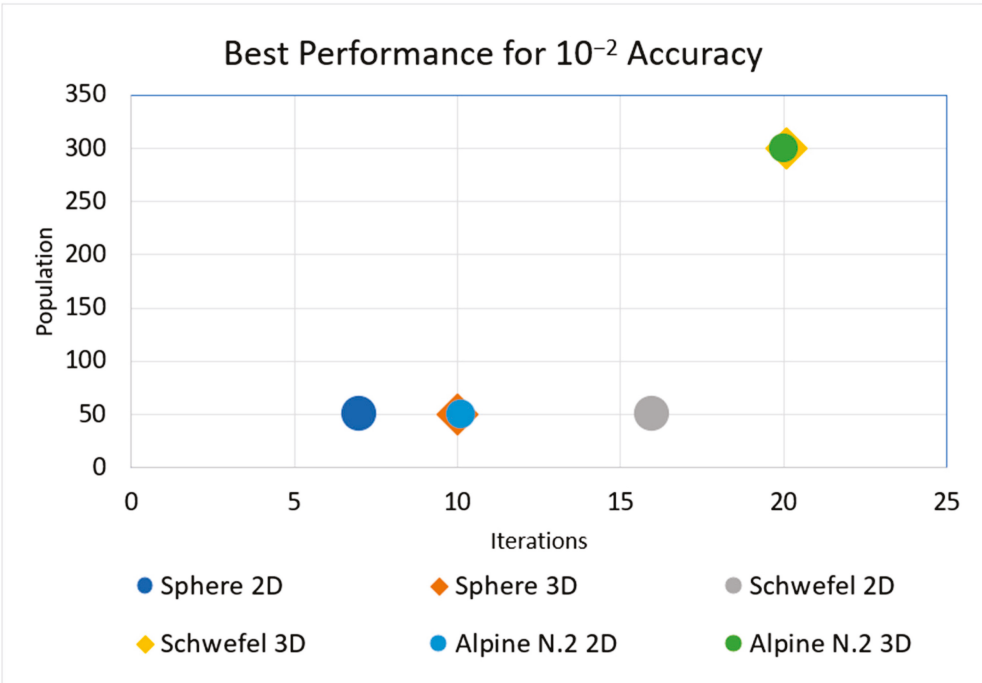



**Figure 7.** Performance graph for solution accuracy of the hybrid scheme. Lower is better.

### *4.3. Identified Breakthrough*

Although only few other similar approaches have been published, such as in [25] and [26], the significant advantage observed in the proposed mixed approach is that, with very few iterations during its initial call, the algorithm with the best performance dramatically accelerates the others, giving them a significant boost towards the global optimum without trapping them in a local optimum. Thus, with a form of elitism, the best-performing algorithm in the scheme collaboratively shares its findings with the team, which often surpasses it in performance in subsequent steps. Therefore, by leveraging the capabilities of different algorithms, the global optimum in an optimization problem is found much more quickly, without the scheme being affected by the poor performance of a specific algorithm-player. For this to happen, all algorithms must perform poorly on the same optimization problem. The scheme's diversity, capitalizing on each algorithm's performance quality for a different problem, helps the entire optimization process converge to a global optimum in a better time frame at a significantly lower calculation cost, regardless of the specific form of the problem's objective function. Practically, this means that any optimization algorithm can be utilized, and the scheme has dynamic length, meaning it can be functional with two algorithms up to N, where N is the total number of optimization algorithms in the literature.

## 5. Conclusions

Through this research, an attempt was made to create an innovative mixed solving scheme for optimization problems. The goal was for this scheme to be able to solve optimization problems by performing the fewest possible evaluations of the objective function. Observing the scheme, we conclude that such a method could evolve into a powerful solving tool applicable to all areas facing optimization problems. Through calculation trials using these algorithms to optimize predefined reference functions with known best solutions, we realized that by using a hybrid scheme with multiple optimization algorithms, the entire process of optimization gets optimized, and the best solution can be found at least three times faster than by using the standalone best performing algorithm and in some cases at two or three magnitudes faster. Even though the selected reference functions are relatively easy to solve using GA, they pose as a powerful tool to demonstrate the potential of the hybrid scheme by merging more modern algorithms that are not optimized for solving such functions, but perform better in other applications. Following this logic, by using a hybrid tool like the one proposed in this manuscript, complex problems that require a lot of time for selecting and fine-tuning the most appropriate optimization algorithm can be solved just by employing a group of algorithms with uniform input values, competing with each other and sharing their best performers after each iteration.

While the present study focuses on demonstrating the feasibility of applying such a scheme by solving problems with known optima, the prospects for the future development of the scheme are excellent. A tremendous advantage is the dynamic use of algorithms. This scheme can exploit as few as two algorithms in its basic form and up to as many as the user wishes to introduce. Thus, in theory, a user could solve all possible problems for all possible objective functions. As long as any of the involved algorithms converges for a specific problem with a well-formulated objective function in a well-formulated solution space, the proposed method converges. Moreover, the proposed mixed algorithm will converge at least as fast, but in reality, multiple times faster, than the fastest of the involved algorithms. The scheme would automatically promote the best algorithm for solving, obtaining results in the shortest possible time, at a fraction of the calculation time needed

for any individual optimization algorithm, and requiring the fewest parameter adjustments for each different problem.

There are a few identified areas in which the scheme can be improved and the research can be progressed. Firstly, for optimization algorithms themselves, the scheme can dynamically evolve along with the development of optimization algorithms by systematically evaluating the literature and considering various optimization schemes. Furthermore, the scheme and how it feeds algorithms in each iteration can be improved, with researchers looking to improve the scheme by experimenting with different feeding criteria for algorithms between each iteration of the mixed scheme. The scheme itself could even take the form of an optimization algorithm, making the solving process more efficient. Another point of improvement could be the selected programming language. Even with the use of the same programming language (Python), an experienced software engineer could dramatically optimize the code, from writing style to resource allocation for solving, significantly improving the time required to complete the process. Additionally, lower-level programming languages, generally much faster than Python and with greater accuracy in resource distribution for optimization, could be explored, such as C++. Finally, the specialization of the scheme regarding the field of interest could be an improvement. Depending on the field to which an optimization problem belongs (e.g., mechanical engineering, physics, or biology), the scheme could be parameterized by domain experts to better fit the problems associated with each research field.

**Author Contributions:** A.A.K.: Conceptualization, Methodology, Writing—original draft preparation, Writing—review and editing, Supervision. L.A.: Conceptualization, Formal analysis and investigation, Resources, Writing—original draft preparation, Software creation, Algorithm integration. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** No data were used for this research, only mathematical modeling.

**Conflicts of Interest:** The authors declare no conflicts of interest. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

## References

1. Dantzig, G. *The Nature of Mathematical Programming*; Mathematical Programming Glossary; INFORMS Computing Society: Catonsville, MD, USA, 2014.
2. Martins, J.R.R.A.; Ning, A. Engineering Design Optimization. Available online: https://www.cambridge.org/highereducation/books/engineering-design-optimization/B1B23D00AF79E45502C4649A0E43135B#overview (accessed on 16 September 2024).
3. Du, D.-Z.; Pardalos, P.M.; Wu, W. History of Optimization. In *Encyclopedia of Optimization*; Floudas, C.A., Pardalos, P.M., Eds.; Springer US: Boston, MA, USA, 2008; pp. 1538–1542. ISBN 978-0-387-74758-3.
4. El-Omari, N.K.T. Sea Lion Optimization Algorithm for Solving the Maximum Flow Problem. *Int. J. Comput. Sci. Netw. Secur.* **2020**, *20*, 30–68. [CrossRef]
5. Weir, M.D.; Hass, J.; Giordano, F.R. *Thomas' Calculus*; Pearson Addison Wesley: Boston, MA, USA, 2005; ISBN 978-0-321-18558-7.
6. Bunkley, N. Joseph Juran, 103, Pioneer in Quality Control, Dies. *The New York Times*. Available online: https://www.nytimes.com/2008/03/03/business/03juran.html (accessed on 16 September 2024).
7. Mohanty, R.; Suman, S.; Das, S.K. Modeling the Axial Capacity of Bored Piles Using Multi-Objective Feature Selection, Functional Network and Multivariate Adaptive Regression Spline. In *Handbook of Neural Computation*; Elsevier: Amsterdam, The Netherlands, 2017; pp. 295–309. ISBN 978-0-12-811318-9.
8. Liberti, L.; Kucherenko, S. Comparison of Deterministic and Stochastic Approaches to Global Optimization. *Int. Trans. Oper. Res.* **2005**, *12*, 263–285. [CrossRef]
9. Haggag, S.; Desokey, F.; Ramadan, M. A Cosmological Inflationary Model Using Optimal Control. *Gravit. Cosmol.* **2017**, *23*, 236–239. [CrossRef]

10. Index. In *An Introduction to Algorithmic Finance, Algorithmic Trading and Blockchain*; Emerald Publishing Limited: Leeds, UK, 2020; pp. 185–191. ISBN 978-1-78973-894-0.
11. Sieja, M.; Wach, K. The Use of Evolutionary Algorithms for Optimization in the Modern Entrepreneurial Economy: Interdisciplinary Perspective. *EBER* **2019**, *7*, 117–130. [CrossRef]
12. Guimaraes, G.L.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P.L.C.; Aspuru-Guzik, A. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. *arXiv* **2017**, arXiv:1705.10843.
13. He, J.; Mattsson, F.; Forsberg, M.; Bjerrum, E.J.; Engkvist, O.; Nittinger, E.; Tyrchan, C.; Czechtizky, W. Transformer Neural Network for Structure Constrained Molecular Optimization. *ChemRxiv* **2021**. [CrossRef]
14. Brownlee, J. Why Optimization Is Important in Machine Learning. 2021. Available online: https://MachineLearningMastery.com (accessed on 16 September 2024).
15. Gandomi, A.H.; Yang, X.-S.; Talatahari, S.; Alavi, A.H. Metaheuristic Algorithms in Modeling and Optimization. In *Metaheuristic Applications in Structures and Infrastructures*; Elsevier: Amsterdam, The Netherlands, 2013; pp. 1–24. ISBN 978-0-12-398364-0.
16. Blum, C.; Puchinger, J.; Raidl, G.; Roli, A. Hybrid Metaheuristics. In *Hybrid Optimization. Springer Optimization and Its Applications*; van Hentenryck, P., Milano, M., Eds.; Springer: New York, NY, USA, 2011; Volume 45. [CrossRef]
17. Burke, E.; Gendreau, M.; Hyde, M.; Kendall, G.; Ochoa, G.; Özcan, E.; Qu, R. Hyper-heuristics: A survey of the state of the art. *J. Oper. Res. Soc.* **2013**, *64*, 1695–1724. [CrossRef]
18. Farag, A.A.; Ali, Z.M.; Zaki, A.M.; Rizk, F.H.; Eid, M.M.; El-Kenawy, E.M. Exploring Optimization Algorithms: A Review of Methods and Applications. *J. Artif. Intell. Metaheuristics* **2024**, *7*, 8–17. [CrossRef]
19. Kumar, S.; Datta, D.; Singh, S.K. Black Hole Algorithm and Its Applications. In *Computational Intelligence Applications in Modeling and Control*; Azar, A.T., Vaidyanathan, S., Eds.; Studies in Computational Intelligence; Springer International Publishing: Cham, Switzerland, 2015; Volume 575, pp. 147–170. ISBN 978-3-319-11016-5.
20. Stützle, T.; López-Ibáñez, M. Automated Design of Metaheuristic Algorithms. In *Handbook of Metaheuristics. International Series in Operations Research & Management Science*; Gendreau, M., Potvin, J.Y., Eds.; Springer: Cham, Switzerland, 2019; Volume 272. [CrossRef]
21. Lambora, A.; Gupta, K.; Chopra, K. Genetic Algorithm—A Literature Review. In Proceedings of the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, Indi, 14–16 February 2019; IEEE: Faridabad, India; pp. 380–384.
22. Yang, X.-S. Firefly Algorithms for Multimodal Optimization. In *Stochastic Algorithms: Foundations and Applications*; Watanabe, O., Zeugmann, T., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2009; Volume 5792, pp. 169–178. ISBN 978-3-642-04943-9.
23. Gao, X.Z.; Govindasamy, V.; Xu, H.; Wang, X.; Zenger, K. Harmony Search Method: Theory and Applications. *Comput. Intell. Neurosci.* **2015**. [CrossRef] [PubMed]
24. Jamil, M.; Yang, X.-S. A Literature Survey of Benchmark Functions For Global Optimization Problems. *arXiv* **2013**, arXiv:1308.4008.
25. Kerschke, P.; Trautmann, H. Automated Algorithm Selection on Continuous Black-Box Problems by Combining Exploratory Landscape Analysis and Machine Learning. *Evol. Comput.* **2019**, *27*, 99–127. [CrossRef] [PubMed]
26. Liu, J.; Moreau, A.; Preuss, M.; Rapin, J.; Roziere, B.; Teytaud, F.; Teytaud, O. Versatile black-box optimization. In Proceedings of the 2020 Genetic and Evolutionary Computation Conference (GECCO'20), Cancún, Mexico, 8–12 July 2020; Association for Computing Machinery: New York, NY, USA; pp. 620–628. [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.

computers

MDPI

Article

# Beyond Snippet Assistance: A Workflow-Centric Framework for End-to-End AI-Driven Code Generation

**Vladimir Sonkin [1] and Cătălin Tudose [2,3,*]**

[1] Luxoft Serbia, 11079 Beograd, Serbia; vladimirsonkin455@gmail.com
[2] Luxoft Romania, 020335 Bucharest, Romania
[3] Faculty of Automatic Control and Computers, National University of Science and Technology Politehnica Bucharest, 060042 Bucharest, Romania
* Correspondence: catalin.tudose@gmail.com

**Abstract:** Recent AI-assisted coding tools, such as GitHub Copilot and Cursor, have enhanced developer productivity through real-time snippet suggestions. However, these tools primarily assist with isolated coding tasks and lack a structured approach to automating complex, multi-step software development workflows. This paper introduces a workflow-centric AI framework for end-to-end automation, from requirements gathering to code generation, validation, and integration, while maintaining developer oversight. Key innovations include automatic context discovery, which selects relevant codebase elements to improve LLM accuracy; a structured execution pipeline using Prompt Pipeline Language (PPL) for iterative code refinement; self-healing mechanisms that generate tests, detect errors, trigger rollbacks, and regenerate faulty code; and AI-assisted code merging, which preserves manual modifications while integrating AI-generated updates. These capabilities enable efficient automation of repetitive tasks, enforcement of coding standards, and streamlined development workflows. This approach lays the groundwork for AI-driven development that remains adaptable as LLM models advance, progressively reducing the need for human intervention while ensuring code reliability.

**Keywords:** artificial intelligence; LLM; AI code review; prompt engineering; routine tasks; software development automation; developer productivity improvement; JAIG

## 1. Introduction

The rapid evolution of Artificial Intelligence (AI) in software development is reshaping how code is written, tested, and deployed. Tools like GitHub Copilot [1] and Cursor [2] provide real-time context-aware code suggestions, enhancing developer productivity. However, these tools primarily function as assistive mechanisms rather than end-to-end automation solutions. They focus on individual coding tasks, such as auto-completing code snippets or refactoring small code blocks, but lack the ability to orchestrate and automate entire development workflows.

As Large Language Models (LLMs) [3,4] continue to advance, they offer the potential to move beyond isolated snippet assistance and enable full automation of multi-step software development processes. To address this gap, we introduce workflow-centric AI code generation. This is an approach that structures AI-driven development as a chain of prompts, each representing a step in the software engineering process. This framework extends beyond simple code generation by incorporating automated validation using LLMs, test generation, test execution, and iterative refinement, ensuring that AI-generated code

not only meets functional requirements but also adheres to correctness and best practices. By making the workflow iterative, the system continuously refines AI-generated code until it aligns with project specifications and passes validation checks.

### 1.1. Originality and Novelty

While existing AI-assisted coding tools improve efficiency at the snippet level, they do not automate the entire software development lifecycle (SDLC). Unlike prior solutions that enhance developer productivity through isolated suggestions, our framework offers a structured execution model capable of handling end-to-end development workflows.

The novelty of this approach lies in the following:

- Automating complete software development processes, including code generation, validation, and testing.
- A structured execution pipeline based on the concept of a Chain Of Thought that dynamically refines AI-generated code based on feedback loops, validation mechanisms, and rollback strategies.
- Automatic Context Discovery, dynamically identifying the minimal relevant code files for a given request by analyzing project structure and dependencies. This allows the framework to work efficiently with large projects without feeding the entire codebase, selecting only the necessary code to fulfill the task. It improves LLM accuracy and simplifies the selection of relevant context from extensive codebases.
- Self-healing automation enables the framework to detect and refine incorrectly generated code that does not fulfill the requirements by providing feedback to LLM and automatically regenerating the code.
- Automatic Code Merging, preserving manual code modifications while seamlessly integrating AI-generated updates. The framework compares AI output with manually altered code, allowing developers to modify code alongside AI automation rather than relying solely on automatic code generation.
- Separation of Concerns, ensuring that requirements (what should be done) remain independent from workflows (how it should be done). Developers can modify requirements without changing workflows or update workflows and regenerate the code based on the same requirements.

### 1.2. Original Contribution

The primary contribution of this work is the development of a workflow-centric framework for AI-driven software automation, which does the following:

- Moves beyond snippet-level assistance to provide comprehensive AI-driven automation.
- Introduces the Prompt Pipeline Language (PPL), a declarative scripting format that allows for the definition of multi-step AI execution workflows.
- Implements automated code validation and rollback mechanisms, ensuring that generated code meets quality standards before integration.
- Demonstrates a reference implementation, JAIG (Java AI-powered Generator), showcasing how structured workflows enhance AI-assisted development.

This approach enables developers to transition from AI-assisted coding to AI-driven automation, significantly reducing manual intervention, improving efficiency, and ensuring software quality, particularly for repetitive tasks.

The remainder of this paper is structured as follows:

- Section 2 (Related Work) reviews existing AI-assisted coding tools and their limitations in automating software development workflows.

- Section 3 (Materials and Methods) introduces the workflow-centric framework, describes the Prompt Pipeline Language (PPL), and presents the reference implementation, JAIG.
- Section 4 (Automating Development with Workflow-Oriented Prompt Pipelines) details the structured execution model, separation of concerns, and AI-assisted code merging.
- Section 5 (Enhancing the Reliability of LLM-Generated Code) explores mechanisms for self-healing automation, automated validation, and error correction.
- Section 6 (Prompt Pipeline Language) explains the iterative execution approach used for workflow automation.
- Section 7 (Practical Applications) demonstrates real-world use cases, including feature implementation, refactoring, debugging, and test generation.
- Section 8 (Results) presents empirical findings on efficiency, accuracy, and reliability.
- Section 9 (Discussion) interprets the results, compares the framework with existing AI-assisted approaches, and outlines limitations.
- Section 10 (Conclusion) summarizes key findings and discusses future research directions.
- Supplementary Materials (Example: Workflow written in Prompt Pipeline Language) provides a detailed example of a PPL-based workflow.

## 2. Related Work

Previous research and commercial tools, such as GitHub Copilot [1] and Cursor [2], primarily focus on providing immediate assistance by predicting and suggesting code snippets.

GitHub Copilot is an instrument for code completion and automatic programming created by GitHub and OpenAI that integrates with various IDEs, supporting the developer with suggestions for code snippets and auto-completion.

Cursor understands the code, suggests improvements, and can also write pieces of code. It offers AI code completion, error correction, and natural language commands. While it extends its functionality to include features such as in-file navigation and editing, its scope still centers on enhancing individual coding tasks rather than automating comprehensive workflows.

While beneficial, these approaches can overlook larger tasks like the following:

- Refactoring large, complex codebases [5];
- Automating repetitive tasks such as boilerplate code generation for CRUD operations [6–8] or DTO (Data Transfer Object) classes [9];
- Implementing multi-step workflows like creating database models, updating service layers, and generating API endpoints;
- Conducting automated code reviews to ensure adherence to coding standards and best practices;
- Detecting and fixing code smells, such as duplicated code or overly complex methods [5];
- Generating comprehensive unit and integration tests to ensure high code coverage [10];
- Replacing deprecated APIs with modern alternatives;
- Addressing security vulnerabilities by identifying and fixing common issues like SQL injection [11,12] or hardcoded secrets.

GitHub Copilot has already gained popularity among developers, and it may be regarded as a substitute for human pair programming [13]. Sometimes it is referred to as the "AI Pair Programmer". An empirical study that collected and analyzed the data from Stack Overflow and GitHub discussions revealed that the major programming languages used with it are JavaScript and Python, while the main IDE with which it integrates is Visual Studio Code 1.97 [14].

A comprehensive analysis of the results of using GitHub Copilot concludes that the code snippets it suggested have the highest correctness score for Java (57%), while JavaScript is the lowest (27%). The complexity of the code suggestions is low, no matter the programming language for which they were provided. Also, the generated code proved to need simplifications in many cases, or it relied on undefined helper methods [15].

Another metric of large interest is the quality of the code that is generated. According to the assessment from [16], GitHub Copilot was able to generate valid code with a 91.5% success rate. Examining the code correctness, 164 problems were provided, resulting in 47 (28.7%) correctly solved, 84 (51.2%) partially correctly solved, and 33 (20.1%) incorrectly solved.

Also, primarily known for its text generation capabilities, LLM can be adapted for NER (Named-Entity Recognition) tasks and be used for entity recognition in border security [17]. Nature-inspired and bio-intelligent computing paradigms using LLMs have revealed effective and flexible solutions to practical and complex problems [18].

Generating unit tests is also a crucial task in software development, but many developers tend to avoid it, being tedious and time-consuming. It is, therefore, an ideal type of task to be delegated to an AI assistant. An empirical study conducted to generate Python tests with the help of GitHub Copilot [19] assessed the usability of 290 tests generated by GitHub Copilot for 53 sampled tests from open-source projects. The results showed that within an existing test suite, 45.28% of the generated tests were passing, while 54.72% of generated tests were failing, broken, or unusable. If the GitHub Copilot tests are created without an existing test suite in place, 92.45% of the tests are failing, broken, or unusable.

Without a unifying workflow, even with the aid of AI tools, developers must still control every step of the process and perform the integration, testing, refactoring, and merging of generated snippets manually. This reliance on manual oversight and intervention often leads to inconsistencies, errors, and inefficiencies in the development process, limiting the potential productivity gains of AI assistance.

Several attempts were made to automate the entire process of software development, which includes writing, building, testing, executing code, and pushing it to the server. AutoDev [20] is an AI-driven software development framework for autonomous planning and execution of intricate software engineering tasks. It starts with the definition of complex software engineering objectives assigned to AI agents that can perform regular operations on the codebase: file editing, retrieval, building, execution, testing, and pushing.

AlphaCodium [21] follows a test-based, multi-stage, code-oriented iterative process organized as a flow, using LLMs to address the problems of developing code. Its applicability is mainly for regular code generation tasks that cover a large part of the time of a developer.

The influence of AI-driven software development automation redefines traditional practices, offering innovative solutions to long-standing challenges. AI may be applied in different phases of the process, with various results: automated code generation, debugging, maintenance, and decision-making processes [22].

AI is effectively integrated into the business process, pushing various industries, such as software engineering, automation, education, accounting, mining, legal services, and media [23]. Automation was applied to remove roadblocks during the last decades, and AI-based tools may replicate the same mindset nowadays [24].

By adopting standardized workflows, developers can leverage AI more effectively to automate these tasks while ensuring that the generated code aligns with project requirements, coding standards, and best practices, reducing the need for constant manual oversight.

## 3. Materials and Methods

The methodology consists of several key components as follows:

- Basic Workflow Building Blocks: These are the basic capabilities that enable other, more comprehensive features.
- Prompt Pipeline Language (PPL): A declarative scripting format for defining multi-step automation workflows.
- Reference Implementation (JAIG): A Java-based AI-powered code generation tool that integrates these concepts.
- Evaluation Metrics: Assessing accuracy, efficiency, and reliability of AI-generated code through empirical testing.
- Algorithmic Approach: Automatic Context Discovery, Automatic Rollbacks, and Multi-Step Execution methods.

Each subsection describes the corresponding aspect of the methodology in detail.

### *3.1. Basic Workflow Building Blocks*

To overcome the limitations of existing AI-assisted coding tools, we propose a workflow-centric framework that automates the entire software development lifecycle, from requirements gathering to code generation, testing, and deployment. The framework is designed to address key challenges in AI-driven development workflows and ensure seamless integration of AI-generated outputs into existing systems.

The framework introduces the following foundational components to enhance reliability and automation:

1. Files/Folders Inclusion in Prompts: Enables developers to specify the path of files or folders in a prompt, ensuring that these files are included in the input context.
2. Automated Response Parsing and File Organization: After AI-generated code is produced, it must be structured correctly within the project's directory hierarchy.
3. Automatic Rollbacks: If the generated code fails to meet expectations, the framework reverts changes to maintain a consistent project state.
4. Automatic Context Discovery: Dynamically analyzes project structure and dependencies to determine the minimum set of relevant files required for a given request.

### *3.2. Implementation of the Prompt Pipeline Language*

To facilitate structured execution, we introduce the Prompt Pipeline Language (PPL), a declarative scripting format that allows developers to define multi-step automation workflows. This language enables the execution of chained prompts, where the output of one prompt becomes the input for the next, thereby reducing the need for manual intervention.

The key components of PPL include the following:

- Directives: Commands that control execution flow, such as #repeat-if for retry conditions and #save-to for output file destinations.
- Reusability Mechanisms: Support for placeholders and reusable templates that allow developers to define standardized workflows across multiple projects.
- Self-Healing Mechanisms: Integration of automated test generation and validation to detect errors and trigger corrective actions before merging AI-generated code into production.

PPL is discussed in detail in Section 5.

### *3.3. Reference Implementation: JAIG (Java AI-Powered Generator)*

As a proof of concept, we implemented the proposed framework in JAIG, a command-line tool that automates Java development tasks using AI-driven workflows. JAIG leverages OpenAI's GPT models for code generation and integrates with IntelliJ IDEA for seamless adoption by developers. JAIG features include the following:

1. Automated Code Generation: Developers provide structured prompts containing project context and requirements, which JAIG processes to generate Java classes, methods, and APIs.
2. Workflow Execution: Uses PPL to execute multi-step automation, ensuring that generated code is structured, compiled, and tested before integration.
3. Code Validation and Refinement: Implements automatic rollback mechanisms to revert faulty AI-generated outputs and enables iterative improvement through prompt modifications.
4. Test Generation and Execution: Automatically generates and executes unit tests to ensure the correctness of generated code.
5. Code Merging: Prevents AI-generated code from overwriting manually edited files by performing intelligent diffs and merges.

However, some capabilities (such as Automatic Context Discovery and Automatic Workflow Generation) are still under development.

### *3.4. Evaluation Methodology*

To assess the framework's effectiveness, we conducted a comparative study, tracking developer performance with and without AI-driven automation. The evaluation focused on the following:

- Automation Efficiency: Categorized tasks as fully automated, partially automated, or requiring manual intervention to measure the framework's automation impact.
- Productivity Gains: Analyzed task completion times across code generation, testing, refactoring, and documentation updates.
- Code Reliability: Evaluated compilation success, automated rollbacks, and validation accuracy to ensure AI-generated code correctness.
- Workflow Adaptability: Examined modifications to workflows as an alternative to manual refactoring.

By structuring AI-assisted development into iterative, self-correcting workflows, the study aimed to validate the framework's impact on efficiency, reliability, and scalability in real-world software projects. The results of the measurements following this methodology are presented in Section 8 (Results).

### *3.5. Algorithmic Approach for Workflow Execution*

The core algorithms powering the framework play a critical role in ensuring the accuracy, reliability, and efficiency of AI-driven development workflows. This section outlines the primary algorithms used for Automatic Context Discovery, Automatic Rollbacks, and Multi-Step Prompt Execution.

#### 3.5.1. Automatic Context Discovery Algorithm

As projects grow in complexity, providing the necessary context to LLMs becomes increasingly challenging. In small projects, developers can explicitly specify relevant files and documentation, but in large-scale systems, manually selecting the right context for each AI request is inefficient and error-prone. Without a structured approach, excessive or missing context can lead to irrelevant responses or incomplete code generation.

To address this, the framework introduces an automated context discovery mechanism that dynamically determines the minimum necessary context to resolve a given issue. This is achieved through a two-step process:

Step 1: Generate Project Index

Before an LLM can efficiently assist with development, it must understand the overall structure of the codebase. To enable this, the framework automatically generates a table of contents (ToC) that summarizes the key components of the project, including the following:

- Main classes, interfaces, and their relationships.;
- API endpoints and exposed services;
- Core business logic modules;
- Database schema and key entities;
- Configuration files and external dependencies.

This project index acts as a navigational map, allowing the framework to locate and retrieve relevant components without analyzing the entire codebase.

Step 2: Select Relevant Context

Once the project index is created, the framework dynamically identifies the minimal context required to address the developer's query. This process involves the following:

1. Understanding the Developer's Request: The framework parses the prompt to identify the affected modules, classes, or APIs.
2. Dependency Analysis: The framework traces dependencies between classes and functions to determine what additional files must be included.
3. Automatic Context Assembly: Based on the identified dependencies, the framework dynamically constructs the optimal context before passing it to the LLM.

For example, if a developer requests:

"Modify the user authentication service to support multi-factor authentication".

The framework will:

- Identify the UserService class as the main modification target;
- Retrieve dependencies like UserRepository, AuthController, and SecurityConfig;
- Include relevant API documentation and configuration files;
- Exclude unrelated parts of the project, preventing unnecessary context overload;

#### 3.5.2. Automatic Rollback Handling Algorithm

To maintain stability in automated workflows, the system implements an automatic rollback mechanism whenever AI-generated code needs to be regenerated. Rollback is triggered not only when validation fails but also whenever code regeneration is required, such as updating requirements, modifying workflows, compilation failures, or any other event requiring code regeneration.

The rollback algorithm operates as follows:

1. Snapshot Creation: Before applying AI-generated modifications, the system stores the original code state to ensure recovery if needed.
2. Code Regeneration: When code needs to be updated or re-executed, the system generates new output based on the latest prompt and workflow state.
3. Validation Check: The newly generated code undergoes compilation, static analysis, and automated tests to verify correctness.
4. Rollback Trigger: If validation fails or if regeneration was triggered due to prompt updates, the system automatically restores the last working version and logs the failure for developer review.
5. Loop Prevention: To avoid infinite rollback-regeneration loops, workflows must define a limited number of refinement attempts before escalating the issue for manual intervention.
6. Successful Execution and Code Integration: If validation passes, the AI-generated code is parsed, validated, and merged with the existing codebase.

Automatic rollbacks are essential for maintaining an agile and error-resistant development process. Developers can safely experiment with different prompts, refining them iteratively until the desired output is achieved while ensuring that faulty or incomplete code does not disrupt the stability of the existing codebase, as in Figure 1.

**Figure 1.** Applying automatic rollback when the generated code is unsatisfactory.

3.5.3. Multi-Step Prompt Pipeline Execution

The Prompt Pipeline Language enables the structured execution of AI-generated workflows. The pipeline execution follows these steps:

1. Step Definition: The workflow is defined using PPL, specifying input dependencies and execution order.
2. Sequential Execution: Each step is processed in order, feeding the output of one prompt into the next.
3. Error Handling and Retries: If a step fails validation, it is retried up to a predefined threshold before triggering a rollback.
4. Final Integration: Once all steps are completed, the final AI-generated output is merged into the project.

By leveraging these algorithms, our framework ensures that AI-driven automation remains robust, scalable, and adaptable to complex software engineering tasks.

### *3.6. Prompt Directives*

To ensure flexibility and configurability in AI-driven automation workflows, the framework provides mechanisms for prompt configuration. These capabilities allow developers to configure how the framework processes the LLM input and post-processes the output.

Prompts can be configured using directives, which define execution behavior, override default settings, and specify storage locations for generated artifacts. Directives override global configuration settings when applied within a prompt; otherwise, the framework defaults to the global configuration.

For example, while the default LLM model might be GPT-4o [25], users may opt for a more powerful GPT-o1 model for specific prompts requiring advanced reasoning capabilities. The GPT-o1 series is designed for complex problem-solving tasks in research, strategy, coding, math, and science domains [26]. By specifying a directive within a prompt, developers can dynamically adjust which model is used.

Some examples of directives include the following:

- #model: <model_name>: selects the AI model used for processing a prompt. Developers can switch between models for different levels of complexity (e.g., GPT-4o for general tasks, GPT-o1 for high-reasoning requirements).
- #temperature: <value>: adjusts the creativity level of responses. A higher value produces more varied outputs, while a lower value ensures deterministic results.
- #save-to: <path_to_file>: saves a generated response to the specified path.

By default, the framework parses LLM responses as Java code and organizes them under /src/main/java following the Maven folder structure [27]. The #save-to directive de-

fines where the generated content should be stored for no-code artifacts, like documentation or configuration files.

By leveraging directives, developers gain fine-grained control over workflow execution, ensuring that AI-driven automation remains adaptable, efficient, and tailored to project requirements.

### *3.7. AI-Assisted Code Merging*

In software engineering, requirements serve as the foundation for product development [28]. In Agile environments [29], the codebase is continuously evolving due to the following:

1. Manual code updates: Developers modify the codebase directly.
2. Automatic code updates: Changes in requirements (prompt modifications) triggering AI-driven code regeneration.

A common challenge in AI-assisted development is managing these concurrent modifications. AI-generated code may overwrite manual updates, leading to lost developer contributions and introducing inconsistencies.

To solve this, the framework introduces AI-assisted code merging, ensuring the following:

- Manual code updates are preserved while incorporating AI-generated updates.
- AI-generated code is adjusted based on recently modified project files.
- Developers can independently modify the codebase or the requirements without introducing conflicts.

Figure 2 illustrates how AI-assisted code merging resolves conflicts between manually updated code and AI-regenerated code due to changes in prompts. The process follows these key steps:

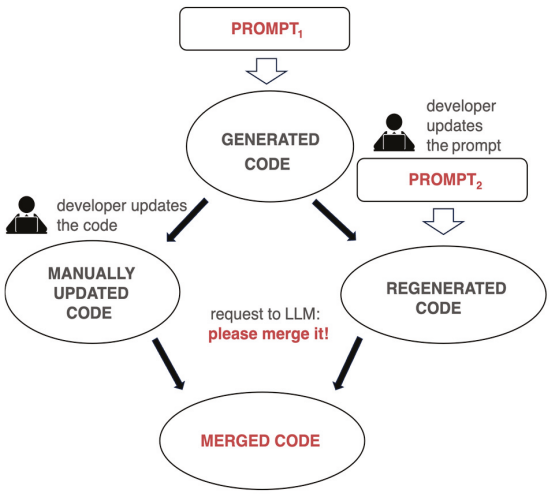



**Figure 2.** Merging generated output with manually updated code.

1. Initial Code Generation: AI generates code based on an initial prompt (Prompt$_1$).
2. Manual Updates: Developers modify the generated code as needed.
3. Prompt Update and Regeneration: If requirements change, an updated prompt (Prompt$_2$) triggers AI to regenerate the code.
4. Merging Process: Instead of replacing manual updates, the framework requests the LLM to merge the regenerated code with the manually updated version.
5. Final Output: The merged code incorporates both AI-generated updates and developer modifications, ensuring consistency.

This approach prevents code loss, maintains developer control, and allows seamless AI-driven automation with human oversight.

### *3.8. Reusable Prompt Templates*

To enhance efficiency, the framework supports prompt templates, making it possible for prompts to dynamically adapt to project-specific requirements using placeholders, as shown in Figure 3.

- Instead of hardcoding values in every prompt, placeholders such as [[Entity]] and [[requirements]] are used.
- At runtime, these placeholders are automatically substituted with actual values from structured data files (e.g., course.yaml).
- LLM processes the dynamically assembled prompt, ensuring that responses are tailored to specific tasks.

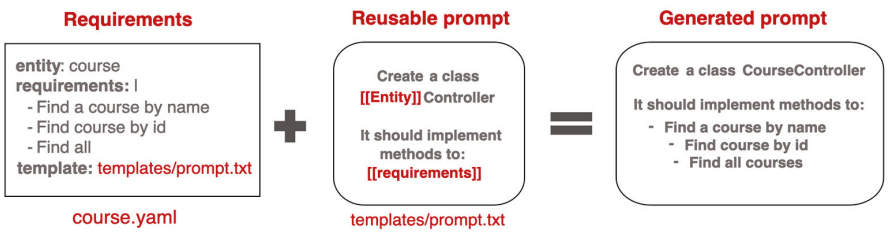



**Figure 3.** The process of substituting placeholders at runtime.

In Figure 3, placeholders such as [[Entity]] or [[requirements]] are automatically substituted based on provided input, such as course.yaml:

With the foundational components of the framework established, including LLM response parsing, prompt directives, AI-assisted code merging, and reusable templates, we are now ready to create full-scale workflows that implement complex tasks with AI assistance.

## 4. Automating Development with Workflow-Oriented Prompt Pipelines

Traditional software development follows a structured workflow involving problem analysis, design, coding, testing, and deployment. These phases are typically broken down into smaller steps executed sequentially.

A key principle in LLM-driven automation is the Chain of Thought [30,31], where prompts are processed in sequence, with each output feeding into the next. This approach mirrors the Divide and Conquer strategy [32,33], breaking down complex tasks into manageable, AI-assisted steps.

For example, developing a new feature typically involves the following:

- Updating the domain model;
- Modifying the service layer;
- Creating API endpoints;
- Writing unit tests.

While AI tools assist with individual coding tasks, a prompt pipeline structures the entire workflow, automating multi-step execution, refining outputs iteratively, and minimizing manual intervention. By leveraging structured automation, developers can ensure that generated code remains aligned with requirements, reducing effort while improving software quality.

This approach enables developers to automate the entire feature development process, ensuring that the generated code is consistent with the requirements.

### *4.1. Prompt Pipeline for Workflow Automation*

A prompt pipeline enables structured automation by defining a sequence of prompts, each focusing on a specific development task. This approach ensures consistency and standardization across multiple business domains, simplifying code generation at scale.

Each prompt targets a specific aspect of the development process. The following are examples:

- One prompt generates the domain model;
- Another handles the service layer;
- A separate prompt defines API interactions;
- Additional prompts cover testing and documentation.

By breaking tasks into specialized prompts, the framework ensures higher accuracy, improved modularity, and easier debugging. Additionally, LLMs perform more efficiently and reliably when focused on smaller, well-defined tasks rather than generating large, complex code blocks in a single step. This targeted approach minimizes errors and enhances predictability, as discussed in more detail in Section 5: Enhancing the Reliability of LLM-Generated Code.

Figure 4 illustrates how a prompt pipeline processes requirements, dynamically inserting them into prompts. The key benefit is that pipelines and requirements remain independent to ensure the following:

- Developers can modify the pipeline to adjust how code is generated;
- Or they can update requirements without altering the workflow.

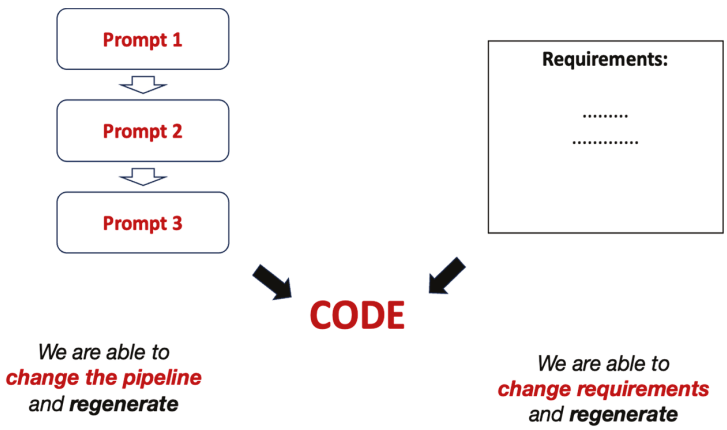



**Figure 4.** A structured prompt pipeline enables automated code generation, ensuring that requirements and workflow logic remain independent.

This separation of concerns provides greater flexibility and maintainability, making AI-driven development more adaptable to evolving project needs.

### *4.2. Separation of Concerns in Workflow Automation*

Separation of concerns [34] is a key principle of the workflow-oriented approach, ensuring that requirements (what should be done) are decoupled from workflows (how it should be done). This decoupling provides flexibility: the same requirements can be reused across multiple workflows, or workflows can be updated independently without altering the requirements. This modularity enhances maintainability, allows faster iterations, and prevents unintended side effects when making changes.

Additionally, the framework leverages a built-in template engine with support for Velocity Templates, enabling dynamic prompt generation based on requirements. This allows for conditional logic, loops, and reusable placeholders, ensuring that textual prompts can be dynamically structured depending on specific project needs. By combining this templating approach with automated workflows, the system can adapt prompt definitions dynamically, improving both flexibility and efficiency.

Figure 5 illustrates this principle, showing how development workflows and business requirements can be defined independently yet work together to generate consistent, reliable code.

The following two scenarios of independent changes are:

- Changing the workflow: When a workflow is updated (e.g., to generate additional documentation), the same requirements can be reapplied to the updated workflow. Steps that have already been executed (e.g., domain model generation) are not re-executed unless affected by the changes. However, if an early step in the workflow is modified, such as domain model generation, all subsequent dependent steps must be reprocessed to ensure alignment across the entire workflow.
- Changing the requirements: When requirements change, the framework automatically rolls back changes from previous workflows and regenerates only the necessary parts of the codebase, ensuring minimal manual intervention while keeping the implementation aligned with updated specifications. This allows developers to focus on refining the requirements while the system ensures consistency. Ideally, once a workflow is established, developers only need to update the requirements or introduce new ones, while the workflow handles the rest automatically, reducing manual overhead.

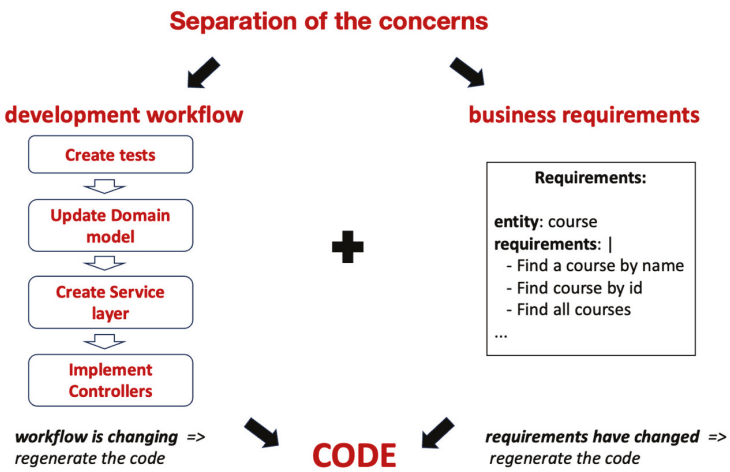



**Figure 5.** Separation of concerns between development workflow and business requirements.

The principle of separation of concerns is widely applied in software development. For example, the Spring Framework [35] separates configuration from business logic, allowing updates to settings without altering the underlying code. Similarly, in the workflow-oriented AI automation approach, workflows and requirements remain independent, allowing seamless updates to one without disrupting the other.

## 5. Enhancing the Reliability of LLM-Generated Code

AI-powered code generation offers the potential to streamline development workflows, reducing manual effort in repetitive tasks. However, its effectiveness is heavily dependent on reliability. While LLMs have made significant progress in code generation, they remain prone to misinterpretations, inconsistencies, and errors. Mistakes in one step can propagate

through subsequent steps, potentially disrupting the entire workflow. For instance, an incorrectly generated domain model may lead to errors in the service layer and API endpoints, requiring manual intervention to resolve.

To mitigate these risks, breaking down development into smaller, well-defined steps is crucial. LLMs perform best when generating small, focused code snippets rather than entire systems in a single attempt. However, even when using small steps, minor errors can compound over multiple iterations. For example, in a 10-step workflow with a 99% reliability rate per step, the overall success rate drops to 90%, making full automation unreliable without safeguards.

Previously, we introduced the Chain of Thought approach, which structures code generation into sequential steps, ensuring that each step produces validated output before proceeding to the next. This structured execution isolates errors early, preventing flawed code from affecting downstream processes. Instead of regenerating an entire feature due to a single incorrect component, the pipeline identifies and corrects only the affected step.

However, even with structured workflows, errors and inconsistencies can still occur. To address these challenges, this section explores key mechanisms to improve the reliability of LLM-generated code, focusing on automated validation and self-healing techniques that allow workflows to detect and correct errors efficiently.

### *5.1. Self-Healing of the Generated Code*

Self-healing mechanisms [36] enable the system to recover from errors autonomously, minimizing the need for manual intervention. These mechanisms employ a multi-tiered approach, applying different recovery strategies depending on whether the issue arises from syntax errors, logical inconsistencies, or missing dependencies. Without an automated correction system, developers must manually inspect, debug, and rerun the workflow, defeating the purpose of automation.

To prevent this, self-healing can be implemented through the two following mechanisms:

- Code Validation with Automated Tests: Generates and runs tests to verify that the code aligns with the requirements, identifying errors automatically.
- Code Validation Using LLM Self-Assessment: Uses an LLM to analyze the generated code against predefined requirements, ensuring that even when tests pass, logical errors are caught before they introduce downstream failures.

These mechanisms help maintain code accuracy while reducing the likelihood of manual intervention.

### *5.2. Code Validation with Automated Tests*

Ensuring the reliability of each step in the workflow requires automated validation, which verifies the correctness of the generated code before moving to the next step. Tests should be dynamically generated based on the same requirements, ensuring validation is always aligned with the intended functionality. Figure 6 illustrates this test-based validation workflow.

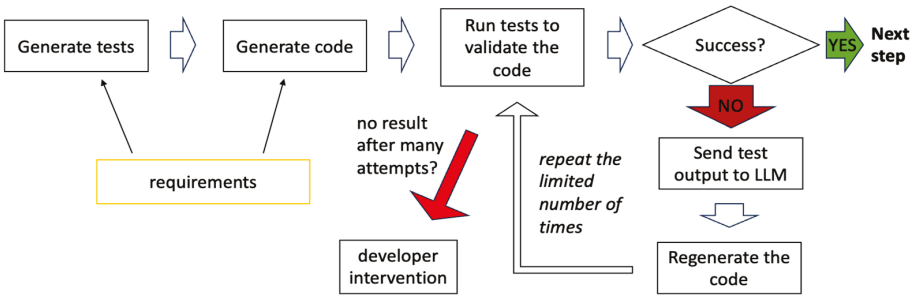



**Figure 6.** Validating the code with automatically generated tests.

To prevent infinite loops when correcting errors, it is essential to define a retry limit. If the generated code fails validation, it should be retried a predetermined number of times before escalating the issue. This ensures that trivial fixable issues can be resolved automatically while preventing excessive computational overhead.

If the maximum retry attempts are reached without success, the workflow triggers developer intervention, allowing a human to do the following:

- Refine the requirements to improve clarity for LLM processing,
- Select an alternative model for validation or code regeneration,
- Adjust the prompt pipeline to enhance the quality of generated output

Setting an appropriate retry threshold balances automation efficiency with error-handling robustness.

### *5.3. Code Validation Using LLM Self-Assessment*

While automated tests verify syntactic and functional correctness, LLM-based self-assessment ensures that the generated code aligns with higher-level requirements and intended business logic. A test may pass even if the generated code does not fully align with the requirements. To address this, LLM-based self-assessment introduces an additional layer of validation. The LLM compares the generated code against the original requirements and provides feedback if discrepancies are detected.

This approach ensures that even if all tests pass, the generated code is still reviewed for logical correctness before integration. Figure 7 illustrates this validation process.

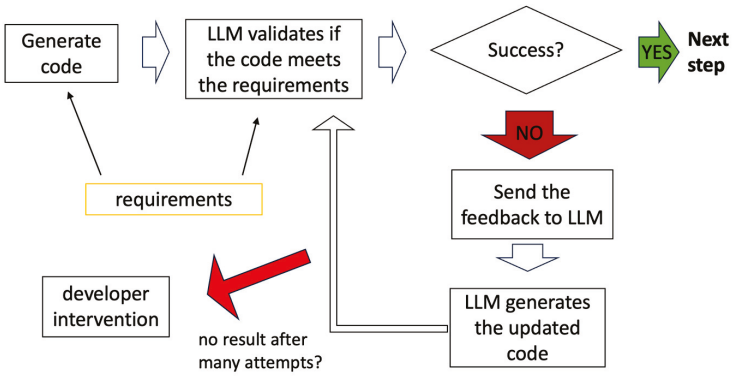



**Figure 7.** Comparing the code with the requirements and providing feedback.

### *5.4. Combining Tests and Self-Assessment*

Combining automated tests and LLM-based self-assessment offers a comprehensive validation strategy. Tests verify the technical correctness of the code, while self-assessment ensures it meets business requirements, addressing scenarios where tests alone may fall short.

The validating LLM can be different from the generating LLM, adding another layer of reliability as in the following:

- Generating Model: GPT-4o generates the code.
- Validating Model: Claude Sonnet [37,38] validates the code by comparing it to the requirements.

By assigning different models for generation and validation, the system reduces the risk of bias in self-assessment, ensuring that inconsistencies in one model's output can be caught before they introduce downstream failures. This division of responsibilities ensures that even complex, multi-step workflows remain accurate and aligned with expectations.

By integrating automated tests and self-assessment mechanisms, workflows can effectively catch both technical errors and requirement misinterpretations. Combining these

strategies with multiple LLMs for generation and validation significantly enhances reliability, reducing the need for manual corrections and improving the feasibility of fully automated code generation.

## 6. Prompt Pipeline Language

The Prompt Pipeline Language (PPL) provides a structured method for defining and executing AI-driven workflows. It enables prompt chaining, validation, self-healing, and automation, ensuring that generated outputs are reliable, adaptable, and maintainable.

PPL ensures that AI-generated code evolves dynamically, reducing manual intervention while preserving quality. Workflows defined in PPL can be reused, modified, and extended without disrupting previous steps, making automation scalable and efficient.

### *6.1. Key Features of PPL*

PPL enhances AI-driven development by providing the following:

- Prompt Chaining: The output of one prompt automatically becomes the input for the next, structuring complex tasks into smaller, manageable steps.
- Execution Control: Conditions (e.g., validation failures) can trigger retries, alternative strategies, or human intervention.
- Self-Healing and Feedback Loops: If a step fails, the workflow refines itself, using feedback to improve generated results.
- Automated Workflow Updates: Workflows and requirements remain independent, allowing incremental modifications without affecting validated outputs.
- Templated Prompt Execution: PPL supports placeholders and conditional logic, dynamically generating prompts based on structured templates (detailed in Supplementary Materials).

The execution behavior of PPL workflows is governed by directives, which define dependencies, error handling, and automation behavior (see Supplementary Materials for details).

### *6.2. Iterative Execution Algorithm and Workflow Adaptation*

PPL follows an iterative execution model to ensure AI-generated code meets validation criteria before integration. The workflow adapts dynamically, refining outputs through structured validation and self-healing mechanisms.

The process consists of the following steps:

1. Prompt Execution: The prompt is combined with the requirements and sent to the LLM for processing.
2. Validation and Self-Assessment: The generated output is validated using automated tests and self-assessment mechanisms.
3. Self-Healing and Refinement: If validation fails (e.g., due to a compilation error or unmet requirements), targeted feedback is provided, and the prompt is re-executed to improve the response.
4. Human Oversight and Adjustments: If automatic corrections fail after a predefined number of attempts, the system escalates the issue to a developer for manual intervention.
5. Incremental Workflow Updates: When new steps are introduced into the workflow, only the affected steps are re-executed, preserving validated outputs and avoiding unnecessary regeneration.
6. Final Integration and Execution: AI-generated outputs are merged into the existing codebase while maintaining manual modifications. The final code is verified and ready for execution.

This adaptive approach ensures that AI-driven automation remains efficient, scalable, and resilient to changes. A detailed example of a PPL-generated workflow is provided in the Supplementary Materials.

### 6.3. Automated Workflow Generation

Given that AI is capable of generating code, it is reasonable to explore its potential in generating workflows as well. Manually designing AI-driven workflows can be time-consuming and error-prone, especially for large-scale projects. Automating this process simplifies pipeline creation while ensuring best practices are applied across various use cases.

Workflow Generation Process:

1. Analyze Task Description: The system extracts key requirements from natural language specifications.
2. Define Optimal Prompt Sequence: The generator constructs a logical order of prompts using best practices for validation, execution, and error handling.
3. Apply Execution Directives: Execution conditions, validation rules, and feedback loops are automatically assigned for optimal efficiency.
4. Optimize for Performance: The generator eliminates redundant prompts and ensures efficient dependency resolution.
5. Generate and Validate Workflow: A trial run is performed to verify correctness before deployment.

By automating workflow creation, developers eliminate the need for manual pipeline engineering while improving scalability, adaptability, and maintainability. This approach ensures that AI-driven workflows are not only efficient but also reusable and easily extensible for evolving project requirements.

## 7. Practical Applications of Workflow-Oriented AI Code Generation

The workflow-oriented approach to AI-driven code generation provides numerous practical applications by automating repetitive and routine tasks that developers frequently encounter. By integrating this framework, developers can optimize various stages of the software development lifecycle, improving efficiency, accuracy, and maintainability. The following applications demonstrate its impact:

### 7.1. Feature Implementation and Code Generation

- Automated Feature Development: The framework can generate boilerplate code, including REST API endpoints [39], service classes, and database models using reusable prompt pipelines.
- Efficient CRUD Operations: Repetitive tasks such as CRUD generation [6–8], standard controllers, and DTO classes [9] can be fully automated, accelerating development while ensuring consistency.
- Improved Development Workflows: This approach is particularly beneficial in established coding practices where standardized templates allow for seamless automation.

### 7.2. Code Quality and Optimization

- Automated Code Review: The framework analyzes codebases, detects adherence to coding standards, and flags potential bugs [40]. It generates detailed reports with suggestions for improvement, enabling a more structured and thorough review process.
- Refactoring Large Codebases: The framework enhances maintainability by breaking down large, complex methods into smaller, modular functions, improving readability and efficiency.

- Fixing Code Smells: The system identifies issues such as duplicated code, deep nesting, and long methods [5] and suggests structured refactoring improvements.

*7.3. Error Resolution and Refactoring*

- Automated Bug Fixing: The framework can detect and resolve simple bugs that do not require extensive debugging. Future iterations may incorporate advanced debugging capabilities.
- Replacing Deprecated APIs: AI-powered automation scans for outdated API usage and replaces them with modern alternatives, provided it has the necessary mappings.
- Security Vulnerability Analysis: The system proactively scans for common security issues, such as SQL injection, and recommends best practices for mitigation.

*7.4. Test Generation and Validation*

- Automated Unit Test Creation: The framework generates unit tests based on the existing class structure and method signatures, improving test coverage and reducing manual effort.
- Enhanced Test-Driven Development (TDD): Developers can seamlessly integrate AI-generated test cases into their projects, ensuring higher code reliability.
- Automated Test Execution: The system runs tests on generated code to validate correctness, refining and regenerating code if necessary.

*7.5. Documentation Generation*

- Automated Code Documentation: The framework automatically documents REST API endpoints, class hierarchies, and method details in a structured format, ensuring alignment with the latest code updates. Documentation generation can be integrated into CI/CD pipelines, ensuring documentation remains consistent with ongoing code changes.
- Customizable Output Formats: Documentation can be generated in various formats (e.g., Markdown, HTML, PDF) to accommodate different project needs and improve maintainability.

## 8. Results

To evaluate the impact of the framework, we conducted a comparative study involving two developers of similar qualification levels. One developer performed tasks manually, while the other used JAIG, the reference implementation of our framework. Both developers logged their time in JIRA, tracking the duration of different tasks, including code review, test generation, API creation, and documentation updates.

The results reveal significant efficiency improvements, especially in repetitive and structured tasks.

*8.1. Automation Efficiency*

We classified tasks into the following three categories:

- Fully Automated: The task is completed entirely without human intervention. The system generates, validates, and integrates the output based on predefined workflows. The generated results are of sufficient quality to eliminate the need for manual review.
- Partially Automated: The workflow automates most of the process, but a human must review or adjust the output before committing the results.
- Requiring Manual Intervention: The task is too complex or context-dependent to be fully automated. AI may provide recommendations, but a developer must manually implement or finalize the solution.

Structured and repetitive tasks, such as code reviews, test generation, and API creation, were highly automated. Conversely, business logic implementation and complex bug fixes require more human involvement due to contextual decision-making, as illustrated in Table 1.

**Table 1.** Task automation by category.

| Task | Fully Automated | Partially Automated | Manual Intervention |
|---|---|---|---|
| Code Review | 80% | 20% | 0% |
| Documentation Generation | 100% | 0% | 0% |
| Unit Test Generation | 95% | 5% | 0% |
| API Endpoint Creation | 90% | 10% | 0% |
| Business Logic Implementation | 30% | 50% | 20% |
| Simple Bug Fixes | 20% | 50% | 30% |
| Complex Bug Fixes | 0% | 40% | 60% |

The framework focuses on automating repetitive development tasks. However, for highly complex, non-repetitive scenarios, conventional AI chat interactions may be more suitable. As LLM models evolve, the percentage of tasks that require manual intervention is expected to decrease, further expanding automation capabilities.

### *8.2. Measured Time Savings*

Compared to manual workflows, JAIG-enabled automation delivered the following efficiency gains:

- A 92% reduction in code review time, as AI automatically generated structured review reports.
- A 96% decrease in unit test creation time enables rapid test coverage expansion.
- A 100% acceleration in documentation updates, ensuring API and system documentation remained consistent with generated code. This process was fully automated, eliminating the need for manual documentation maintenance.

Although initial workflow creation required time investment, it was compensated by long-term automation benefits, leading to a faster and more structured development process.

### *8.3. Self-Healing Mechanisms and Reliability*

Self-healing mechanisms increase reliability by iteratively refining AI-generated code in response to compilation failures, test failures, or incorrect logic. Instead of requiring manual debugging, the system performs the following:

1. Detects errors and provides feedback to AI.
2. Regenerates code based on the failure context.
3. Repeats until the issue is resolved or escalates to human intervention.

This approach ensures AI-generated code is validated before integration, reducing compilation errors from 8% to 2% and increasing correct implementation rates from 87% to 97%.

Consistency, a key reliability factor, refers to the system's ability to generate stable, reproducible solutions that consistently meet the same functional requirements across multiple executions. JAIG achieved 96% consistency, meaning that while the exact code output may differ, the generated solutions remain functionally equivalent and aligned with the given requirements. This stability minimizes variability in AI-driven development, ensuring predictable and reliable automation.

### *8.4. Workflow Modification: An Alternative to Code Refactoring*

A key advantage of a workflow-driven approach is that modifying workflows replaces the need for manual code refactoring.

Instead of editing existing code, developers can perform either of the following:

1. Adjust workflow rules to change how code is generated.
2. Add new workflow steps to extend functionality, such as automatically updating API documentation whenever a new endpoint is created.

Let us discuss these two opportunities separately.

#### 8.4.1. Modifying Workflow Rules: REST to gRPC Migration

We measured the efficiency of migrating an API from REST to gRPC. In a traditional refactoring approach, developers must manually modify API controllers, rewrite service definitions, update unit tests, and regenerate documentation.

With JAIG and workflow automation, the business requirements remain unchanged, but the workflow is switched from REST API generation to gRPC service generation. The steps related to unit test generation and documentation updates remain intact, ensuring consistency.

As a result, we observed the following:

- Automatic code regeneration without modifying individual files.
- Consistent updates to service definitions, method signatures, and client stubs.
- An 80% faster migration compared to manual refactoring.

#### 8.4.2. Adding New Workflow Steps: OpenAPI Documentation Generation

Another advantage of workflow-driven automation is the ability to seamlessly introduce new steps. We evaluated a case where an API was initially generated without OpenAPI documentation, requiring manual updates afterward.

With workflow automation, we added a single step to automatically generate OpenAPI documentation alongside API creation. This allows the following:

- Eliminated manual effort to keep documentation up to date.
- Ensured consistency between generated API specifications and actual implementations.
- Reduced documentation effort by 90%, as updates were now automated.

By allowing workflows to be adjusted without modifying existing code, this approach provides a more scalable and maintainable alternative to traditional refactoring.

The empirical findings demonstrate that structured workflow automation significantly improves efficiency, reduces manual effort, and enhances code reliability. In the following discussion, we analyze these results in depth, comparing our framework to existing AI-assisted coding approaches, outlining its advantages, and identifying areas for future research.

## 9. Discussion

### *9.1. Interpretation of Results*

The evaluation confirms that the workflow-centric AI-driven framework boosts efficiency, accuracy, and reliability. Manual coding effort was reduced by 85%, with up to 95% of time savings in testing and refactoring. The 99% compilation success rate after refinements ensures stability, while automated rollbacks and test-driven validation minimize semantic errors and enhance overall code quality.

Based on these results, we identify several key takeaways that highlight the advantages of a workflow-centric AI-driven approach.

The key takeaways are as follows:

1. Highly structured tasks (code reviews, test generation, API creation) are nearly fully automated, leading to significant time savings.
2. Workflows are reusable. The initial setup effort is compensated by long-term efficiency gains.

3. Workflow modifications replace traditional refactoring, making large-scale changes faster and more structured.
4. Self-healing mechanisms improve reliability, reducing error rates and ensuring higher-quality AI-generated code.
5. Incremental automation is possible: new steps (e.g., documentation updates) can be added seamlessly to existing workflows.

### *9.2. Comparison with Existing AI-Assisted Development Approaches*

While AI-powered code assistants such as GitHub Copilot and Cursor provide valuable snippet-based suggestions, they primarily operate as reactive tools, assisting developers in real-time coding. In contrast, our proposed framework automates entire development workflows, covering code validation, refactoring, automated testing, and integration. The inclusion of automatic context, rollback mechanisms, and iterative refinements ensures greater consistency and correctness, distinguishing it from existing AI-assisted development approaches. Unlike Copilot, which may produce contextually incorrect or incomplete code, the framework continuously evaluates AI-generated solutions, refining outputs based on structured feedback mechanisms, as illustrated in Table 2.

Tools like AutoDev [20] and AlphaCodium [21] represent significant advancements in AI-driven software development but take fundamentally different approaches from our framework. AutoDev focuses on fully autonomous execution of software development tasks within a containerized environment, while AlphaCodium enhances test-based iterative code refinement, particularly for competitive programming. By contrast, our workflow-centric framework integrates AI into developer-guided automation workflows, balancing automation with structured validation, iterative refinement, and human oversight. While each approach has its strengths, a direct comparison is challenging due to differing objectives. A more detailed evaluation of these methodologies is left for future research.

**Table 2.** Comparison of JAIG vs. GitHub Copilot.

| Feature | GitHub Copilot | Workflow-Centric AI Framework (JAIG) |
|---|---|---|
| Code Generation | Suggests snippets based on local context | Collects relevant project context, generates structured code, validates outputs, and refines through iterative feedback |
| Workflow Automation | None, focused on single-task assistance | Full automation of multi-step workflows |
| Validation and Testing | Requires manual review and testing | Built-in validation, test generation, and self-healing mechanisms |
| Error Handling | Manual | Automatic rollbacks and iterative refinement |
| Integration with Human Developers | Complements manual coding | Automates structured workflows while allowing human intervention in complex cases |

### *9.3. Limitations and Challenges*

While the workflow-centric AI-driven framework demonstrates significant efficiency gains, some limitations and challenges need to be addressed for broader adoption and further improvement.

#### 9.3.1. Scalability Challenges for Large, Enterprise-Level Codebases

As software projects scale, challenges arise in efficiently handling large, multi-repository codebases with complex interdependencies. The framework's ability to select relevant context dynamically is key, but performance overhead may increase when processing large-scale enterprise systems.

Potential solutions are as follows:

- Optimizing context discovery algorithms to efficiently extract the minimal required code sections for LLMs.
- Implementing incremental code generation to reduce processing overhead by modifying only affected components rather than regenerating entire workflows.
- Exploring the capability of handling large-scale workflows in parallel.

#### 9.3.2. Dependence on LLM-Generated Code and Potential Subtle Errors

LLMs, while powerful, may introduce subtle logic errors, security vulnerabilities, or inconsistencies in generated code, even when passing syntactic and semantic validation. Over-reliance on AI-generated outputs without deeper validation poses risks, particularly in safety-critical applications.

Potential solutions are as follows:

- Enhancing multi-step validation with external static analysis tools and AI-assisted formal verification techniques.
- Leveraging ensemble AI models combining different LLMs or rule-based systems to improve robustness.
- Ensuring a human-in-the-loop review process for critical components before deployment.

#### 9.3.3. Handling Domain-Specific Languages (DSLs)

The framework is primarily designed for general-purpose programming languages, but AI-driven automation faces challenges when handling domain-specific languages (DSLs) with strict syntax and semantics. While the framework already supports generating non-code artifacts (e.g., documentation, configuration files, Cucumber tests, API integration tests) and organizing them via the #save-to directive, DSL-specific constraints may impact the accuracy and usability of AI-generated outputs.

Potential solutions are as follows:

- Extending the framework's prompt pipeline language (PPL) to support custom DSLs and structured metadata.
- Integrating specialized LLMs trained on domain-specific corpora to improve accuracy in DSL-based projects.
- Implementing custom transformation layers that automatically parse LLM-generated responses into valid DSL syntax.

#### 9.3.4. Initial Learning Curve and Adoption Complexity

While the framework provides automation benefits, adopting workflow-based AI development requires the following:

- Familiarity with prompt engineering and structured pipeline-based execution.
- Understanding how to modify workflows to achieve the desired automation outcome.
- Balancing human intervention with AI-driven processes to avoid over-reliance on automation.

Potential solutions:

- Providing predefined workflow templates for common development tasks (e.g., API generation, test automation, refactoring).
- Developing intuitive UI tools to assist in workflow creation and modification, reducing reliance on manual prompt configuration.
- Offering gradual onboarding strategies, allowing teams to transition from semi-automated to fully automated workflows in phases.

9.3.5. Handling Edge Cases and Complex Business Logic

AI-driven workflows excel in structured, repetitive tasks, but challenges arise when dealing with the following:

- Highly domain-specific business logic that lacks clear formalization.
- Scenarios requiring nuanced human decision-making, such as architectural trade-offs or strategic code restructuring.
- Edge cases that involve implicit dependencies across multiple components.

Potential solutions:

- Hybrid AI + Human review approach: Let AI handle structured parts while human developers oversee complex decisions.
- Enriching AI-generated prompts with historical project context to improve decision accuracy for complex scenarios.
- Introducing adaptive self-learning mechanisms, where the framework continuously refines AI behavior based on developer feedback.

### *9.4. Future Research Directions*

To further enhance automation reliability and extend its applicability, future research should explore the following:

1. Dialog-Based Workflow Management: Enabling interactive AI that asks clarifying questions during requirement formulation to ensure all necessary details are provided upfront. This reduces ambiguities and improves the accuracy of AI-generated code.
2. Dynamic Workflow Adaptation: Allowing workflows to adjust in real-time based on project changes, reducing the need for manual intervention and making AI-driven development more responsive.
3. Human-in-the-Loop Automation: Investigating methods for incremental developer feedback throughout the workflow, allowing AI to refine outputs dynamically based on real-time user input rather than only escalating after failure.
4. Self-Optimizing Workflows: Using reinforcement learning and historical data to continuously refine AI-generated workflows, improving efficiency and adaptability over time.
5. Hybrid AI Models for Task Optimization: Combining lightweight AI models for routine tasks with advanced LLMs for complex decision-making, improving efficiency, performance, and cost-effectiveness.
6. Scalability and Deployment in Production: Investigating how workflow-driven automation scales across large projects and multi-developer teams, ensuring maintainability, security, and efficiency in real-world applications.
7. Testing Workflow Reliability: Developing systematic methods to verify AI-generated workflows across varying requirements, ensuring correctness, stability, and consistency.
8. Interactive Workflow Optimization: Studying real-time feedback mechanisms that help developers refine AI-generated workflows dynamically, adapting to evolving requirements and constraints.
9. AI-Assisted Code Explainability and Justification: Research methods to generate human-readable explanations of AI-generated code, providing insights into design choices and ensuring better trust and adoption among developers.
10. AI-Driven Security Analysis and Compliance Enforcement: Investigating how AI can automatically detect security vulnerabilities and ensure compliance with industry regulations (e.g., OWASP, GDPR, HIPAA) before deployment.

11. Cross-Project Knowledge Transfer: Developing AI techniques that learn from past projects to improve workflow recommendations, optimize context selection, and enable intelligent code reuse across multiple repositories.

## 10. Conclusions

The workflow-centric AI-driven framework advances beyond existing AI-assisted coding tools by automating the entire software development lifecycle. Structured execution, self-healing mechanisms, and automated validation enhance consistency, minimize human intervention, and accelerate software delivery.

As AI technology evolves, this approach has the potential to redefine software engineering by shifting developer focus from routine coding to higher-level architecture and strategic decision-making. Future research will address open challenges such as self-optimizing workflows, adaptive dialog-based workflow management, and enterprise-scale automation. In parallel, near-term improvements will enhance usability, scalability, and real-world adoption.

### *10.1. Real-World Applications and Industry Adoption*

With increasing AI adoption, this framework has the potential to transform enterprise software development by streamlining the following key engineering processes:

- Automated API Development: Generating REST/gRPC services, API documentation, unit, and integration tests with minimal manual effort.
- Continuous Code Review: Integrating into CI/CD pipelines to analyze code changes, enforce coding standards, detect potential issues, and generate structured review reports, ensuring high-quality code before integration.
- AI-Assisted Refactoring: Standardizing modernization efforts by replacing legacy patterns, outdated tools, deprecated libraries, and inefficient code structures with optimized implementations, reducing technical debt and improving maintainability.
- Code Compliance and Security Enforcement: Ensuring generated code adheres to industry standards, security policies, and regulatory requirements (e.g., OWASP, GDPR, HIPAA), reducing security risks and improving auditability.

These capabilities position the framework as a scalable AI-powered solution for software automation, driving efficiency in finance, healthcare, e-commerce, and enterprise SaaS development.

### *10.2. Future Framework Enhancements*

To further refine the framework and expand its applicability, the following areas of improvement are proposed:

- Broader Language Support: Expanding beyond Java to support Python, JavaScript, and C++, enabling adoption across diverse tech stacks.
- Embedding Code Debugging in the Workflow: Integrating AI-driven debugging tools directly into the workflow. After code generation, the debugger will automatically execute the generated code, capturing runtime insights such as variable states, execution flow, and errors. This feedback loop will iteratively refine the code before validation and integration, improving reliability and reducing the need for manual debugging.
- CI/CD Pipeline Integration: Enabling automated deployment of AI-generated code. Enhancements will focus on seamless integration with DevOps tools (e.g., Jenkins, GitHub Actions, GitLab CI/CD) to automate builds, testing, and deployment workflows. Additionally, post-deployment monitoring and rollback mechanisms will enhance reliability and performance.

- Workflow Visualization and Debugging: Introducing graphical representations of workflows for better execution tracking, issue analysis, and optimization. This may include a graphical debugger that maps workflow steps, highlights errors, and enables interactive debugging.
- Integration with Industry-Standard Tools: Strengthening compatibility with IDEs (e.g., IntelliJ, VS Code), testing frameworks (JUnit, Mocha), and project management platforms (Jira, GitLab) to ensure seamless adoption in development environments.

By implementing these improvements, the workflow-centric AI framework can evolve into a fully autonomous software development assistant, enhancing productivity while maintaining software quality and developer control.

**Supplementary Materials:** The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/computers14030094/s1, Workflow written in Prompt Pipeline Language Example.

**Author Contributions:** Conceptualization, V.S.; formal analysis, V.S.; investigation, V.S. and C.T.; methodology, V.S.; software, V.S.; supervision, V.S. and C.T.; validation, V.S. and C.T.; writing—original draft, V.S.; writing—review and editing, C.T. 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 available in a publicly accessible repository. https://github.com/sonkin/JAIG (accessed on 1 March 2025) and https://jaig.pro/ (accessed on 1 March 2025).

**Conflicts of Interest:** Author Vladimir Sonkin is employed by the company Luxoft Serbia. Author Cătălin Tudose is employed by the company Luxoft Romania. The 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. GitHub Copilot. Available online: https://github.com/features/copilot (accessed on 1 March 2025).
2. Cursor, the AI Code Editor. Available online: https://www.cursor.com/ (accessed on 1 March 2025).
3. Iusztin, P.; Labonne, M. *LLM Engineer's Handbook: Master the Art of Engineering Large Language Models from Concept to Production*; Packt Publishing: Birmingham, UK, 2024.
4. Raschka, S. *Build a Large Language Model*; Manning: New York, NY, USA, 2024.
5. Fowler, M. *Refactoring: Improving the Design of Existing Code*, 2nd ed.; Addison-Wesley Professional: Boston, MA, USA, 2018.
6. Bonteanu, A.M.; Tudose, C.; Anghel, A.M. Multi-Platform Performance Analysis for CRUD Operations in Relational Databases from Java Programs using Spring Data JPA. In Proceedings of the 13th International Symposium on Advanced Topics in Electrical Engineering (ATEE), Bucharest, Romania, 23–25 March 2023.
7. Bonteanu, A.M.; Tudose, C.; Anghel, A.M. Performance Analysis for CRUD Operations in Relational Databases from Java Programs Using Hibernate. In Proceedings of the 2023 24th International Conference on Control Systems and Computer Science (CSCS), Bucharest, Romania, 24 May 2023.
8. Bonteanu, A.M.; Tudose, C. Performance Analysis and Improvement for CRUD Operations in Relational Databases from Java Programs Using JPA, Hibernate, Spring Data JPA. *Appl. Sci.* **2024**, *14*, 2743. [CrossRef]
9. Tudose, C. *Java Persistence with Spring Data and Hibernate*; Manning: New York, NY, USA, 2023.
10. Tudose, C. *JUnit in Action*; Manning: New York, NY, USA, 2020.
11. Martin, E. *Mastering SQL Injection: A Comprehensive Guide to Exploiting and Defending Databases*; Independently Published; 2023. Available online: https://www.amazon.co.jp/-/en/Evelyn-Martin/dp/B0CR8V1TKH (accessed on 1 March 2025).
12. Caselli, E.; Galluccio, E.; Lombari, G. *SQL Injection Strategies: Practical Techniques to Secure Old Vulnerabilities Against Modern Attacks*; Packt Publishing: Birmingham, UK, 2020.
13. Imai, S. Is GitHub Copilot a Substitute for Human Pair-programming? In An Empirical Study. In Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Companion Proceedings, Pittsburgh, PA, USA, 22–24 May 2022; pp. 319–321.

14. Nguyen, N.; Nadi, S. An Empirical Evaluation of GitHub Copilot's Code Suggestions. In Proceedings of the 2022 Mining Software Repositories Conference, Pittsburgh, PA, USA, 22–24 May 2022; pp. 1–5.
15. Zhang, B.Q.; Liang, P.; Zhou, X.Y.; Ahmad, A.; Waseem, M. Demystifying Practices, Challenges and Expected Features of Using GitHub Copilot. *Int. J. Softw. Eng. Knowl. Eng.* **2023**, *33*, 1653–1672. [CrossRef]
16. Yetistiren, B.; Ozsoy, I.; Tuzun, E. Assessing the Quality of GitHub Copilot's Code Generation. In Proceedings of the 18th International Conference on Predictive Models and Data Analytics in Software Engineering, Singapore, 17 November 2022; pp. 62–71.
17. Suciu, G.; Sachian, M.A.; Bratulescu, R.; Koci, K.; Parangoni, G. Entity Recognition on Border Security. In Proceedings of the 19th International Conference on Availability, Reliability and Security, Vienna, Austria, 30 July–2 August 2024; pp. 1–6.
18. Jiao, L.; Zhao, J.; Wang, C.; Liu, X.; Liu, F.; Li, L.; Shang, R.; Li, Y.; Ma, W.; Yang, S. Nature-Inspired Intelligent Computing: A Comprehensive Survey. *Research* **2024**, *7*, 442. [CrossRef] [PubMed]
19. El Haji, K.; Brandt, C.; Zaidman, A. Using GitHub Copilot for Test Generation in Python: An Empirical Study. In Proceedings of the 2024 IEEE/ACM International Conference on Automation of Software Test, Lisbon, Portugal, 15–16 April 2024; pp. 45–55.
20. Tufano, M.; Agarwal, A.; Jang, J.; Moghaddam, R.Z.; Sundaresan, N. AutoDev: Automated AI-Driven Development. *arXiv* **2024**, arXiv:2403.08299.
21. Ridnik, T.; Kredo, D.; Friedman, I. Code Generation with AlphaCodium: From Prompt Engineering to Flow Engineering. *arXiv* **2024**, arXiv:2401.08500.
22. Alenezi, M.; Akour, M. AI-Driven Innovations in Software Engineering: A Review of Current Practices and Future Directions. *Appl. Sci.* **2025**, *15*, 1344. [CrossRef]
23. Babashahi, L.; Barbosa, C.E.; Lima, Y.; Lyra, A.; Salazar, H.; Argôlo, M.; de Almeida, M.A.; de Souza, J.M. AI in the Workplace: A Systematic Review of Skill Transformation in the Industry. *Adm. Sci.* **2024**, *14*, 127. [CrossRef]
24. Ozkaya, I. The Next Frontier in Software Development: AI-Augmented Software Development Processes. *IEEE Softw.* **2023**, *40*, 4–9. [CrossRef]
25. Chatbot App. Available online: https://chatbotapp.ai (accessed on 1 March 2025).
26. Using OpenAI o1 Models and GPT-4o Models on ChatGPT. Available online: https://help.openai.com/en/articles/9824965-using-openai-o1-models-and-gpt-4o-models-on-chatgpt (accessed on 1 March 2025).
27. Varanasi, B. *Introducing Maven: A Build Tool for Today's Java Developers*; Apress: New York, NY, USA, 2019.
28. Sommerville, I. *Software Engineering*, 10th ed.; Pearson: London, UK, 2015.
29. Anghel, I.I.; Calin, R.S.; Nedelea, M.L.; Stanica, I.C.; Tudose, C.; Boiangiu, C.A. Software Development Methodologies: A Comparative Analysis. *UPB Sci. Bull* **2022**, *83*, 45–58.
30. Ling, Z.; Fang, Y.H.; Li, X.L.; Huang, Z.; Lee, M.; Memisevic, R.; Su, H. Deductive Verification of Chain-of-Thought Reasoning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 36407–36433.
31. Li, L.H.; Hessel, J.; Yu, Y.; Ren, X.; Chang, K.W.; Choi, Y. Symbolic Chain-of-Thought Distillation: Small Models Can Also "Think" Step-by-Step. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; Volume 1, pp. 2665–2679.
32. Cormen, T.H.; Leiserson, C.; Rivest, R.; Stein, C. *Introduction to Algorithms*; MIT Press: Cambridge, MA, USA, 2009.
33. Smith, D.R. Top-Down Synthesis of Divide-and-Conquer Algorithms. *Artif. Intell.* **1985**, *27*, 43–96. [CrossRef]
34. Daga, A.; de Cesare, S.; Lycett, M. Separation of Concerns: Techniques, Issues and Implications. *J. Intell. Syst.* **2006**, *15*, 153–175. [CrossRef]
35. Walls, C. *Spring in Action*; Manning: New York, NY, USA, 2022.
36. Ghosh, D.; Sharman, R.; Rao, H.R.; Upadhyaya, S. Self-healing systems—Survey and synthesis. *Decis. Support Syst.* **2007**, *42*, 2164–2185. [CrossRef]
37. Claude Sonnet Official Website. Available online: https://claude.ai/ (accessed on 1 March 2025).
38. Claude 3.5 Sonnet Announcement. Available online: https://www.anthropic.com/news/claude-3-5-sonnet (accessed on 1 March 2025).
39. Fielding, R.T. Architectural Styles and the Design of Network-Based Software Architectures. Ph.D. Thesis, University of California, Irvine, CA, USA, 2000.
40. Martin, R.C. *Clean Code: A Handbook of Agile Software Craftsmanship*; Pearson: London, UK, 2008.


**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.

computers

MDPI

Article

# Improving Road Safety with AI: Automated Detection of Signs and Surface Damage

**Davide Merolla [1], Vittorio Latorre [1], Antonio Salis [2,*] and Gianluca Boanelli [2]**

[1] Dipartimento di Bioscienze e Territorio, Università degli Studi del Molise, 86090 Pesche, Italy; merolla.davide96@gmail.com (D.M.); vittorio.latorre@unimol.it (V.L.)
[2] Research & Innovation Division, Tiscali Italia S.p.A., 09122 Cagliari, Italy; gianluca.boanelli@tiscali.com
* Correspondence: antonio.salis@tiscali.com

**Abstract:** Public transportation plays a crucial role in our lives, and the road network is a vital component in the implementation of smart cities. Recent advancements in AI have enabled the development of advanced monitoring systems capable of detecting anomalies in road surfaces and road signs, which can lead to serious accidents. This paper presents an innovative approach to enhance road safety through the detection and classification of traffic signs and road surface damage using advanced deep learning techniques (CNN), achieving over 90% precision and accuracy in both detection and classification of traffic signs and road surface damage. This integrated approach supports proactive maintenance strategies, improving road safety and resource allocation for the Molise region and the city of Campobasso. The resulting system, developed as part of the CTE Molise research project funded by the Italian Minister of Economic Growth (MIMIT), leverages cutting-edge technologies such as cloud computing and High-Performance Computing with GPU utilization. It serves as a valuable tool for municipalities, for the quick detection of anomalies and the prompt organization of maintenance operations.

**Keywords:** road safety; traffic sign detection; road surface damage detection; smart cities; YOLO; convolutional neural network (CNN); predictive maintenance; High-Performance Computing (HPC); cloud computing

## 1. Introduction

According to the latest World Health Organization (WHO) report, public road networks are the lifeblood of modern societies, playing a crucial role in the transportation of goods and people, which is fundamental for trade, commerce, and tourism. They also enable easy access to jobs, education, healthcare, and social activities, both in urban and rural areas. However, ensuring road safety remains a critical challenge.

Harsh weather conditions can accelerate road degradation, and as traffic volume grows—including heavy traffic—frequent repairs become necessary. Any lapse in maintenance can lead to severe incidents, resulting in fatalities worldwide. The 2023 WHO report states that approximately 1.19 million people die each year due to road traffic crashes [1], highlighting the urgency of proactive and efficient road monitoring systems.

This paper presents a novel deep learning-based approach for real-time traffic sign detection and road damage assessment, designed to support a Road Management System for public municipalities. While previous studies have focused either on road condition monitoring or traffic sign detection separately, our method integrates both tasks into a unified AI-driven framework.

By leveraging state-of-the-art computer vision techniques, our approach aims to provide higher accuracy, real-time processing capabilities, and improved adaptability to various environmental conditions. Unlike traditional rule-based or sensor-heavy solutions, our system operates efficiently using standard cameras and cloud-based AI models, reducing costs and making large-scale deployment feasible.

This research is conducted as part of the Molise CTE research project, funded by the Italian Ministry of Economic Growth (MIMIT). The project aims to harness emerging technologies such as cloud computing, High-Performance Computing, Artificial Intelligence, and AR/VR to develop and demonstrate next-generation Smart City solutions.

The structure of this paper is as follows. Section 2 presents a comprehensive Literature Review, positioning our approach within existing research and highlighting key differentiators. Section 3 describes the proposed AI-powered road monitoring system, detailing its architecture, data pipeline, and cloud-based infrastructure. Section 4 outlines the computational experiments, providing a rigorous performance evaluation with relevant benchmarks. Section 5 discusses Challenges, Solutions, and Integration with Municipal Maintenance Applications, addressing potential limitations, their mitigations, and demonstrating the practical applicability of our solution. Section 6 presents ongoing developments, future improvements, and potential exploitation opportunities, before concluding the paper.

## 2. Literature Review

### *2.1. Introduction*

The detection and classification of traffic signs and road damage are vital components of intelligent transportation systems. Recent advancements in deep learning—particularly using convolutional neural networks (CNNs) and YOLO (You Only Look Once) architecture—have markedly enhanced the accuracy and efficiency of these systems. Moreover, the integration of GPS data has opened new avenues for comprehensive road monitoring and predictive maintenance. This section reviews seminal and recent literature in traffic sign detection and classification, road damage detection, and GPS-based road condition monitoring.

### *2.2. Traffic Sign Detection Approaches*

#### 2.2.1. Early and Seminal Work

Early research laid the groundwork for automated traffic sign detection using traditional machine learning methods. A key paper by Maldonado-Bascon et al. employs color segmentation and Support Vector Machines (SVMs) to detect and recognize road signs, demonstrating robustness against transformations and occlusions [2]. Another seminal approach by Fang, Chen, and Fuh integrates neural networks for feature extraction and Kalman filters for tracking, enabling reliable performance under diverse environmental conditions [3].

#### 2.2.2. YOLO Architecture and Enhancements

In recent years, the YOLO family of algorithms has become a popular choice for real-time object detection, including traffic signs. YOLOv4 was shown by Yang and Zhang (2020) to significantly improve detection accuracy over YOLOv3 for Chinese traffic signs [4]. Zhang (2023) similarly demonstrated that YOLOv3 outperformed R-CNN algorithms in both speed and accuracy [5].

Enhancements to YOLO have been introduced to address specific challenges:

- Lightweight Models. Sign-YOLO integrates the Coordinate Attention (CA) module and High-BiFPN to improve multi-scale semantic fusion, achieving significant gains in precision, recall, and speed on the CCTSDB2021 dataset [6].

- Advanced Feature Extraction. PVF-YOLO employs Omni-Dimensional Convolution (ODconv) and Large Kernel Attention (LKA) to further boost detection accuracy and speed [7].

### *2.3. Traffic Sign Classification*

#### 2.3.1. CNN-Based Classification

Once detected, traffic signs need to be accurately classified into specific categories. Convolutional neural networks (CNNs) have shown exceptional performance in this area. Ciresan et al. (2012) achieved state-of-the-art results on German traffic signs using CNNs [8]. More recently, models like TSR-YOLO embed advanced modules tailored to complex traffic scenarios, further improving classification metrics [9].

#### 2.3.2. Traffic Sign Damage Classification

Relatively few papers address traffic sign damage classification. Trpkovic, Selmic, and Jevremovic (2021) employed a CNN to identify and classify damaged and vandalized traffic signs [10]. Acilo et al. (2018) used transfer learning with ResNet-50 to detect signs' compliance status and physical degradation, achieving high accuracy [11].

#### 2.3.3. Generative AI for Synthetic Data

Generative Adversarial Networks (GANs), particularly DCGAN, have been used to address the imbalance between damaged and undamaged traffic signs in training datasets. By generating realistic synthetic images of damaged traffic signs, researchers can enhance model robustness and accuracy [12]. This approach not only balances the dataset but also improves performance in recognizing signs under diverse and challenging conditions.

### *2.4. Road Damage Detection*

#### 2.4.1. Public Datasets

Datasets such as Mapillary and the Road Damage Detection (RDD) dataset are widely used for training and evaluating models aimed at identifying potholes, cracks, and surface wear. Their comprehensive annotations make them valuable resources for developing robust road damage detection systems.

#### 2.4.2. Deep Learning Methods

Deep CNNs and YOLO architectures have been particularly successful in road damage detection. Zhang et al. (2018) utilized a deep CNN to detect road cracks, achieving high precision and recall [13]. Maeda et al. (2018) applied YOLO to detect multiple types of road damage, demonstrating effectiveness in real-world settings [14].

### *2.5. Integration of GPS Data*

Using GPS data enriches time series analysis with spatial context, enhancing road damage detection and monitoring.

- Mobile Sensor Networks. Strutu et al. (2013) proposed a mobile sensor network-based system incorporating 3D accelerometers, GPS, and video modules for road surface monitoring [15]. Perttunen et al. (2011) similarly demonstrated that accelerometers and GPS on mobile phones could effectively detect road surface anomalies [16].
- Low-Cost Systems. Tarun and Esther (2023) introduced a Raspberry Pi and GPS-based road sign detection system, showcasing high detection precision and efficient real-time operation [17].

By combining sensor data with GPS coordinates, researchers can identify patterns in road degradation related to specific routes or environmental conditions, enabling proactive maintenance strategies.

### 2.6. Additional Insights from Recent Advances

Lim et al. (2023) provide a comprehensive overview of state-of-the-art traffic sign recognition, categorizing research into conventional machine learning and deep learning approaches [18]. Key developments include the following:

- Preprocessing and Feature Extraction. Sophisticated techniques for data augmentation, color normalization, and feature selection help address variations in lighting, weather, and camera angles.
- Model Generalization. The use of diverse training datasets is crucial to ensuring that models remain robust under varying backgrounds and environmental conditions.
- Predictive Maintenance. Integrating sensor data, GPS, and advanced neural network models allows for the implementation of predictive maintenance strategies, preventing road and sign failures before they occur [19].

### 2.7. Summary

Significant advancements have been made in traffic sign detection, classification, and road damage detection using deep learning—particularly YOLO and CNN-based methods. Seminal works laid the foundation using machine learning and traditional feature extraction, while modern approaches leverage real-time detection and classification via YOLO variants. Damage-specific research, although still limited, suggests promising avenues for applying CNNs and Generative Adversarial Networks to handle imbalanced datasets. Furthermore, the integration of GPS data offers spatial context essential for predictive maintenance. Overall, these trends highlight the potential for intelligent transportation systems to substantially enhance road safety, maintenance, and driver assistance.

## 3. Methodology

### 3.1. Experimental Workflow and Datasets

Figure 1 illustrates the overall workflow for both road sign and road damage detection/classification tasks. We begin by gathering images (from the Mapillary Vistas and RDD 2022 datasets), then apply data quality checks and data augmentation to expand or balance these datasets. For road signs, we perform two steps: (1) detection using YOLOv8, and (2) classification into damaged or not damaged using a convolutional neural network (CNN) with attention mechanisms. For road damage (treated purely as anomalies), a single detection step via YOLOv8 is sufficient.

Mapillary Vistas

The Mapillary Vistas dataset is a large-scale, street-level imagery dataset designed for semantic segmentation and object detection in various conditions [20].

- Classes: 401.
- Images: 41,906.
- Size: 32.8 GB.
- Train/Validation Split: 80%/20%.

RDD 2022

The Road Damage Detection (RDD) 2022 dataset focuses on identifying and classifying road-surface damages such as cracks and potholes [21].

- Classes: 4.
- Images: 34,007.
- Size: 9.6 GB.
- Train/Validation Split: 80%/20%.

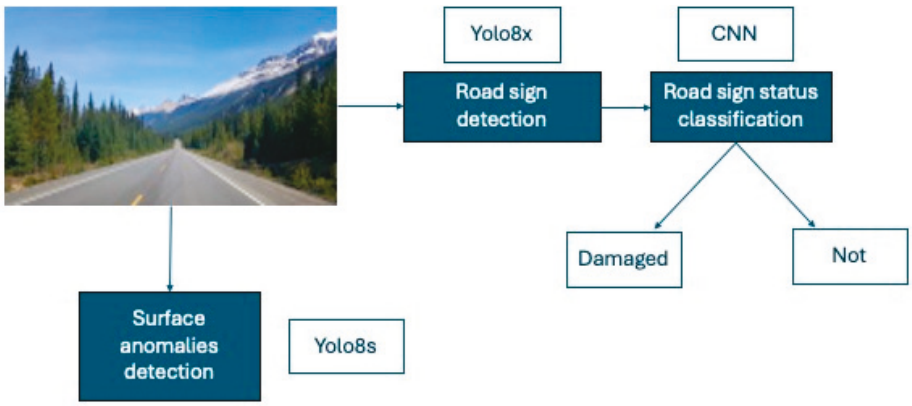



**Figure 1.** Experimental workflow. The diagram shows how images flow from collection through dataset expansion, data augmentation, and detection/classification.

### 3.2. YOLOv8 Architecture

Figure 2 illustrates the YOLOv8 network structure, which consists of three main components: Backbone, Neck, and Head. The Backbone (shown on the left side of the figure) is responsible for extracting hierarchical features from the input image through a series of convolutional layers and specialized blocks (e.g., C2f and Bottleneck). These blocks enhance the network's learning capacity while maintaining low computational overhead.

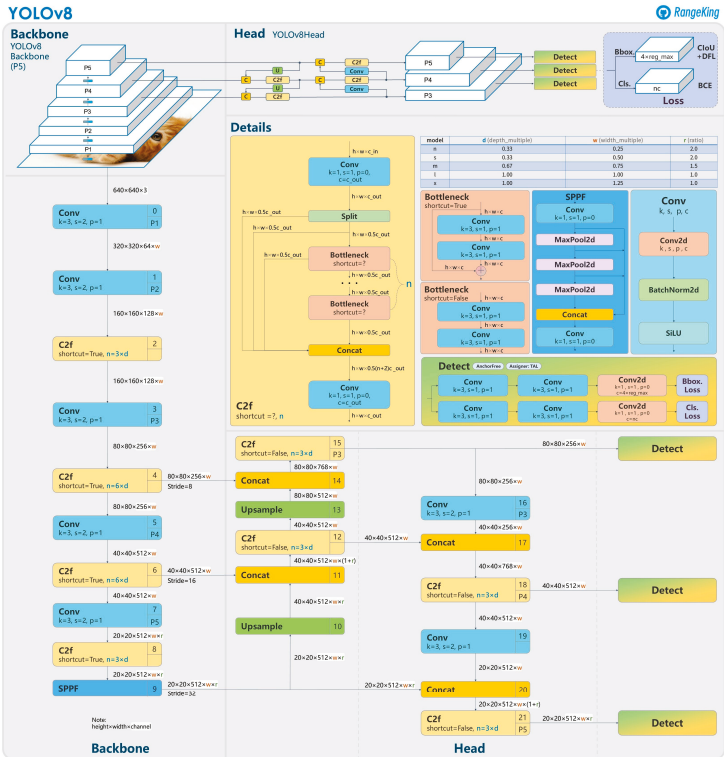



**Figure 2.** YOLOv8 architecture.

The middle section highlights additional key details of the architecture, including the following:

Split and Concat: Operations that split and concatenate feature maps, allowing multi-scale information to be merged efficiently.

C2f (Cross Stage Partial Networks v2) and Bottleneck: Blocks featuring shortcut connections that facilitate gradient flow and improve feature representation.

- SPPF (Spatial Pyramid Pooling-Fast): A module that partitions feature maps into regions of different sizes, enabling the network to capture multi-scale context.

The Neck (center-right) includes layers such as Upsample and Concat to fuse feature maps at various resolutions, improving multi-scale feature representation. Finally, on the right side is the Head, which processes detection outputs (Detect) across three different scales (P3, P4, P5). These convolutional layers further refine the fused features to produce final predictions for classes, bounding box coordinates, and confidence scores. The Loss function (top-right) combines classification, bounding box regression, and objectness terms to ensure balanced network training.

In our implementation, we used YOLOv8 "out of the box", i.e., with no architectural modifications to the Ultralytics-released source code. We only customized certain training parameters—such as the number of epochs, batch size, and learning rate—to suit our dataset and experimental objectives. By adopting the native YOLOv8 framework, we leveraged its official optimizations, ensuring fast inference and competitive performance for both traffic sign detection and road damage assessment tasks.

### *3.3. First Phase: Detection*

#### 3.3.1. Data Preparation

Before training, data manipulation is performed to ensure optimal performance and accuracy:

- Augmentation: rotations, scaling, flips, color adjustments.
- Normalization: scaling pixel values (0–1).
- Label Smoothing: reducing overfitting by softening hard labels.
- Anchor Box Calculation: custom anchors to improve detection of varying object sizes.

#### 3.3.2. YOLOv8s for Road Surface Damage

We use a smaller YOLOv8 variant (YOLOv8s) to balance speed and accuracy for road damage detection. Using a lightweight model is sufficient and efficient.

| **Parameter** | **Value** |
|---|---|
| Pretrained | Yes (Ultralytics checkpoint) |
| Epochs | 160 |
| Image Size | 640 |
| Patience | 100 |
| Cache | RAM |
| Device | GPU |
| Batch Size | 64 |

An example of road damage detection is shown in Figure 3.

YOLOv8s is used to identify anomalies (e.g., cracks, potholes) in the road surface.

**Figure 3.** Road damage detection.

3.3.3. YOLOv8x for Road Signs

For road sign detection, we employ the larger YOLOv8x model to achieve higher precision—this is critical for correctly identifying smaller signs with varied shapes. We prioritized accuracy over speed in this task since sign recognition demands finer resolution and often more complex feature extraction.

| Parameter | Value |
|---|---|
| Pretrained | Yes (Ultralytics checkpoint) |
| Epochs | 100 |
| Image Size | 640 |
| Patience | 100 |
| Cache | RAM |
| Device | GPU |
| Batch Size | Auto |

An example of traffic sign detection is shown in Figure 4.

**Figure 4.** Traffic sign detection.

YOLOv8x locates and classifies various signs in real time with high accuracy.

*3.4. Second Phase: Road Sign Classification*

After YOLOv8 detects road signs, we crop each sign from the frame and classify it as damaged or not damaged. Damaged signs include those with graffiti, stickers, rust, or physical deformations. The initial dataset is unbalanced (6025 damaged vs. 34,315 undamaged). We address this with the following:

- Focal Loss: weighs hard-to-classify examples more, mitigating class imbalance.

- Cutout Regularization: randomly removes sections of the image during training, improving robustness.

### *3.5. Enhanced CNN with Attention*

In a subsequent step, we integrate attention mechanisms into a CNN to further improve classification accuracy:

1. Input Layer: $128 \times 128 \times 3$ images.
2. Convolutional Blocks: Each block has Convolution $\rightarrow$ BatchNorm $\rightarrow$ ReLU, plus Channel and Spatial Attention modules, followed by Max-Pooling.
3. Fully Connected Layers: Flatten $\rightarrow$ Dense (256, ReLU) $\rightarrow$ Dropout (0.5).
4. Output Layer: Dense (1, Sigmoid).
5. Optimizer and Loss: Adam with Focal Loss to handle class imbalance.

Attention Mechanisms

- Channel Attention: emphasizes relevant feature channels (e.g., small damages).
- Spatial Attention: focuses on crucial areas of the sign where anomalies might appear.

Data Augmentation and Regularization

- Augmentations: flips, rotations, shifts, shear, zoom.
- Cutout Regularization: random masking to handle occlusions.

Training and Evaluation

We train for 10 epochs (batch size 32, 80–20 train–validation split) and use ReduceLROnPlateau if validation accuracy stalls. The best model is saved via ModelCheckpoint, yielding 90% accuracy on validation data.

### *3.6. Generative AI for Class Balancing (Stable Diffusion)*

We further address class imbalance using Stable Diffusion v2.1 (fine-tuned) to generate 18,000 synthetic images of damaged traffic signs:

| **Parameter** | **Value** |
|---|---|
| Pretrained | Yes (Stable Diffusion v2.1) |
| Epochs | 50 |
| Image Size | $512 \times 512$ |
| Patience | 10 |
| Cache | RAM |
| Device | GPU |
| Batch Size | 32 |
| Conditioning Method | Image + Text Prompt Encoding |
| Optimization | AdamW (LR 5e-5, Cosine Scheduler) |
| Loss Functions | Contrastive + Perceptual (LPIPS) |

By conditioning both on existing images and textual damage descriptions (e.g., "rust", "graffiti"), we generate realistic variations of damaged signs, effectively tripling the size of the damaged class. This improves the CNN's ability to recognize real-world damaged signs. Contrastive Loss ensures generated images differ significantly between damaged vs. undamaged categories, further aiding discrimination, while Perceptual Loss (LPIPS) helps maintain visual realism in synthetic images.

## 4. Computational Experiments

This section presents the experimental setup and results for our deep learning models applied to road sign detection (YOLOv8x), road surface damage detection (YOLOv8s), and road sign classification (CNN). We first describe the computational resources utilized (Section 4.1), then detail the performance metrics and results for each model (Sections 4.2 and 4.3).

### *4.1. Hardware and Training Environment*

#### 4.1.1. YOLO Models on Google Colab

- GPU: NVIDIA Tesla T4.
- CUDA Cores: 2560.
- Tensor Cores: 320.
- GPU Memory: 16 GB GDDR6.
- Memory Bandwidth: 320 GB/s.
- Theoretical FP32 Performance: Up to 8.1 TFLOPS.
- CPU: Intel(R) Xeon(R) CPU, 2 vCPUs @ 2.3 GHz.
- RAM: 12.7 GB available in Colab.
- Disk: 100 GB available storage.

Google Colab's NVIDIA Tesla T4 GPU provided ample computational capacity for training both YOLOv8x (used for detecting road signs) and YOLOv8s (used for detecting road surface damages). The hardware configuration allowed for efficient handling of large datasets and the complex operations required by the YOLO architecture.

#### 4.1.2. CNN on Reevo Servers

- CPU: 24 vCPUs @ 2.5 GHz.
- RAM: 32 GB.

The CNN for road sign classification was trained on Reevo servers. This setup featured more CPU cores and higher RAM, which supported the parallel data preprocessing and training steps necessary for our classification task.

### *4.2. YOLO Models: Performance Metrics and Results*

We evaluated YOLOv8x (for road sign detection) and YOLOv8s (for road surface damage detection) using a set of common metrics:

- mAP50: Mean Average Precision at 50% IoU threshold.
- mAP50-95: Mean Average Precision averaged over IoU thresholds from 50% to 95%.
- Precision: The ratio of true positive detections to the total positive detections.
- Recall: The ratio of true positive detections to the total number of actual positives.
- Box Loss: Measures the error in bounding box predictions.
- Object Loss: Assesses the error in distinguishing objects from the background.

Each subsection below provides a closer look at how these metrics evolved during training, alongside commentary on the associated diagrams.

#### 4.2.1. YOLOv8x for Road Sign Detection

1. Accuracy Metrics

   Refer to Figure 5 for the training curves of mAP50, mAP50-95, precision, and recall:

   - mAP50: Progressively increases and stabilizes around 0.9, reflecting high accuracy in detecting road signs at a 50% IoU threshold.

- mAP50-95: Improves more gradually, stabilizing near 0.7, demonstrating robustness across varying IoU thresholds.
- Precision: Fluctuates initially but trends upward, indicating fewer false positives over time.
- Recall: Increases steadily to about 0.8, reflecting the model's ability to detect actual positive instances effectively.

Interpretation: The consistently high mAP50 confirms the model's strong detection capabilities, while the gradual rise in mAP50-95 indicates good performance even with more stringent IoU thresholds.

2. Box Loss

Refer to Figure 6 for the box loss curve:

- The box loss decreases sharply during the initial epochs before stabilizing.
- Lower values imply more precise bounding box predictions.

Interpretation: As training progresses, YOLOv8x refines its bounding box coordinates, leading to improved localization of road signs.

3. Object Loss

Refer to Figure 7 for the object loss curve.

- Object loss drops rapidly in early epochs and then plateaus.
- Lower values indicate improved capability in distinguishing objects from the background.

Interpretation: YOLOv8x becomes increasingly effective at differentiating road signs from the surrounding environment, boosting overall detection performance.

#### 4.2.2. YOLOv8s for Road Surface Damage Detection

1. Accuracy Metrics

Refer to Figure 8 for the training curves of mAP50, mAP50-95, precision, and recall.

- All metrics show steady improvement, indicating the model's growing proficiency in detecting road surface damages.

Interpretation: As with YOLOv8x, YOLOv8s demonstrates robust and consistent accuracy improvements, verifying its suitability for identifying damaged road surfaces.

2. Box Loss

Refer to Figure 9 for the box loss curve.

- The box loss decreases over time, highlighting enhanced precision in bounding box predictions for damaged regions.

Interpretation: The model learns to localize damaged sections more accurately with each epoch.

3. Object Loss

Refer to Figure 10 for the object loss curve.

- Like YOLOv8x, the object loss shows a downward trend, reflecting improved discrimination between damaged and undamaged surfaces.

Interpretation: YOLOv8s becomes more reliable at identifying genuine damage, reducing false positives on intact road surfaces.

### *4.3. CNN for Road Sign Classification*

After detecting road signs, we employed a CNN to classify them into specific categories. Refer to Figure 11 for the evolution of training and validation metrics (accuracy, precision, and recall):

- Training Accuracy (blue line): Rises rapidly in the early epochs and converges at around 90%, indicating effective feature learning on the training set.
- Precision (green line): Remains consistently high, demonstrating the model's ability to correctly identify positive instances.
- Validation Metrics (orange, red, pink lines): Closely follow the training metrics, reflecting stable model performance and strong generalization.

Interpretation:

The CNN attains approximately 90% classification accuracy on the validation set, underscoring its robust performance in identifying various road sign types.

### *4.4. Figures and Their Descriptions*

Below are placeholders for each figure referenced in the above subsections. Insert the actual images and descriptions once they are ready.

Figure 5 presents the accuracy metrics of the YOLOv8x model during training, evaluated on the validation set. The key metrics displayed include the following:

- mAP50 (dark blue line): This metric steadily increases, stabilizing around 0.9, indicating high accuracy in detecting road signs at a 50% IoU threshold.
- mAP50-95 (orange line): Shows a gradual increase over epochs, stabilizing near 0.7, demonstrating the model's robustness across varying IoU thresholds.
- Precision (cyan line): Initially fluctuates but trends upwards, suggesting improved true positive detections while reducing false positives.
- Recall (pink line): Shows a steady upward trend, reaching around 0.8, confirming the model's effectiveness in capturing actual positive instances.

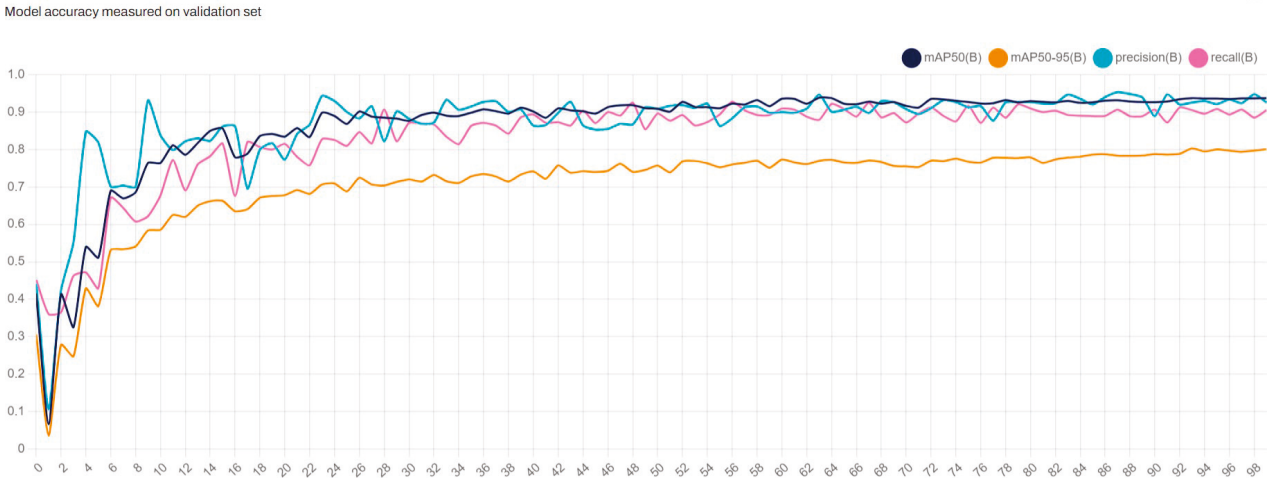



**Figure 5.** YOLOv8x accuracy.

The graph highlights how the YOLOv8x model progressively improves its accuracy across different IoU thresholds, making it a reliable solution for road sign detection.

Figure 6 illustrates the box loss values for the YOLOv8x model during training, showing the difference between predicted and actual bounding boxes for detected objects. The two curves represent the following:

- Training Box Loss (dark blue line): The loss decreases significantly within the initial epochs and continues to decline gradually, indicating that the model is learning to predict bounding box coordinates more accurately.
- Validation Box Loss (orange line): Initially higher, this value also decreases over time, though it stabilizes at a slightly higher level than the training loss.

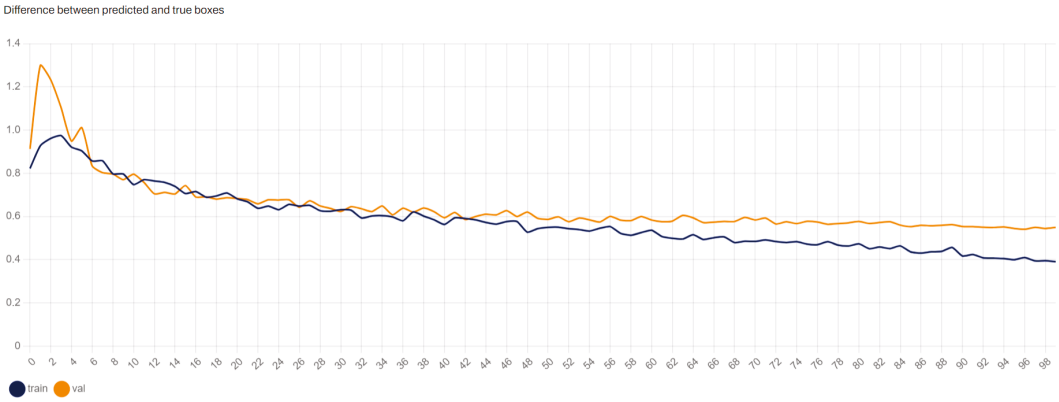



**Figure 6.** YOLOv8x box loss.

The declining trend in both training and validation box loss suggests that the model is effectively refining its bounding box predictions, leading to improved localization of road signs over successive training epochs.

Figure 7 presents the object loss values for the YOLOv8x model during training, which measures the model's ability to correctly distinguish between background and objects of interest. The two curves indicate the following:

- Training Object Loss (dark blue line): The loss decreases significantly in the early epochs and continues to decline steadily, demonstrating improved model confidence in object detection.
- Validation Object Loss (orange line): Initially higher, this loss follows a similar decreasing trend, though it stabilizes slightly above the training loss.

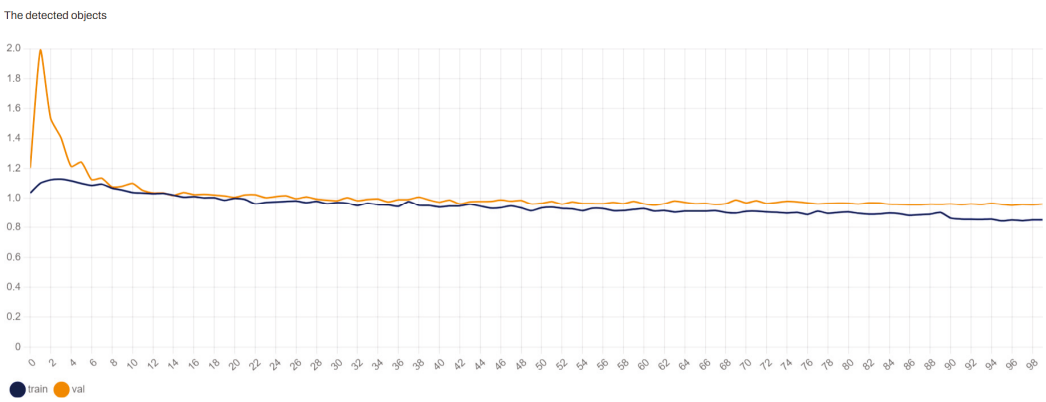



**Figure 7.** YOLOv8x object loss.

The overall downward trend in object loss confirms that YOLOv8x is learning to better differentiate road signs from the background, leading to more reliable and accurate detections over successive training epochs.

Figure 8 presents the accuracy metrics of the YOLOv8s model during training, evaluated on the validation set. The key metrics displayed include the following:

- mAP50 (dark blue line): This metric gradually increases and stabilizes around 0.65, indicating a reasonable level of accuracy in detecting road surface damages at a 50% IoU threshold.
- mAP50-95 (orange line): Exhibits a slower but steady improvement, stabilizing around 0.35, reflecting the model's performance across a broader range of IoU thresholds.
- Precision (cyan line): Shows fluctuations but follows an overall increasing trend, indicating progressive improvements in reducing false positives.
- Recall (pink line): Consistently trends upwards, stabilizing around 0.55, confirming the model's ability to detect most instances of road damage.

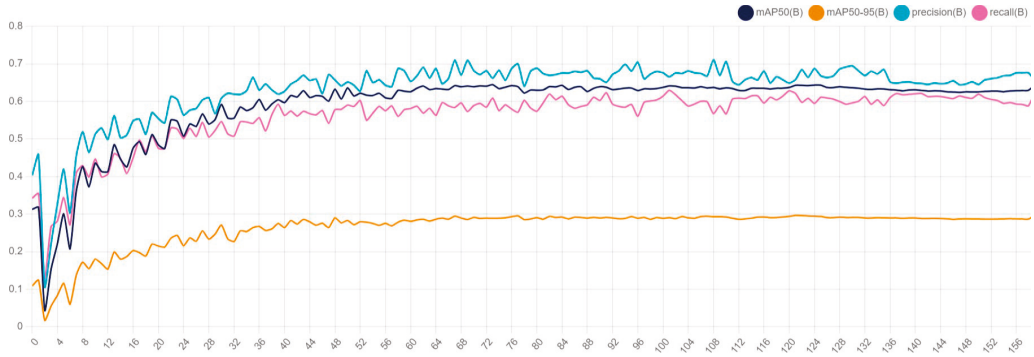



**Figure 8.** YOLOv8s accuracy.

The graph demonstrates that YOLOv8s effectively improves its accuracy over training epochs, successfully detecting road surface damages while adapting to different IoU thresholds.

Figure 9 illustrates the box loss evolution during training for the YOLOv8s model, which quantifies the accuracy of the model's bounding box predictions for detected road surface damages. The trends observed indicate a clear and steady improvement in localization precision:

- Training Box Loss (dark blue line): Shows a consistent downward trend, demonstrating that the model is progressively learning to refine its bounding box predictions and more accurately delineate damaged road areas.
- Validation Box Loss (orange line): Initially fluctuates but gradually stabilizes, remaining slightly higher than the training loss. This behavior suggests that while the model continues to improve, it maintains a balance between training and real-world generalization.

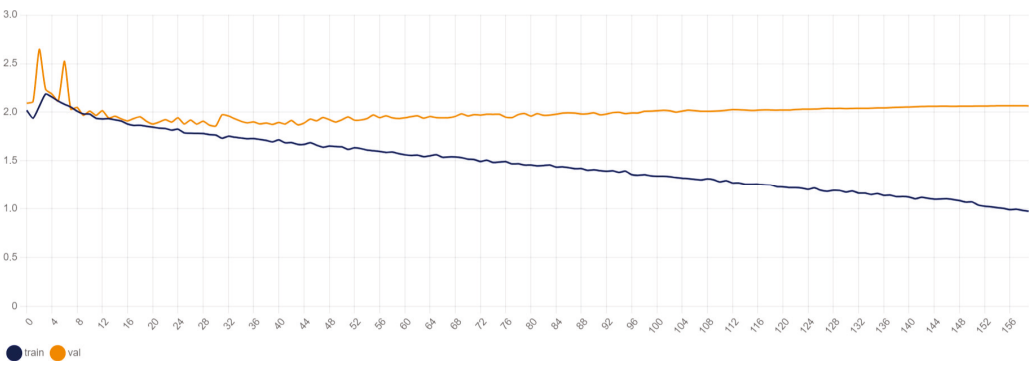



**Figure 9.** YOLOv8s box loss.

The overall reduction in box loss confirms that YOLOv8s effectively optimizes its bounding box localization, ensuring precise and reliable detection of road surface damages across different conditions.

Figure 10 illustrates the object loss progression during training for the YOLOv8s model, which measures how well the model differentiates between damaged and undamaged road surfaces. The trends observed indicate a significant improvement in the model's ability to correctly classify areas of interest:

- Training Object Loss (dark blue line): Displays a steady decline, indicating that the model is continuously refining its ability to distinguish road surface damages from the background.
- Validation Object Loss (orange line): While initially higher, it gradually stabilizes over time, suggesting that the model maintains its capability to generalize to unseen validation data.

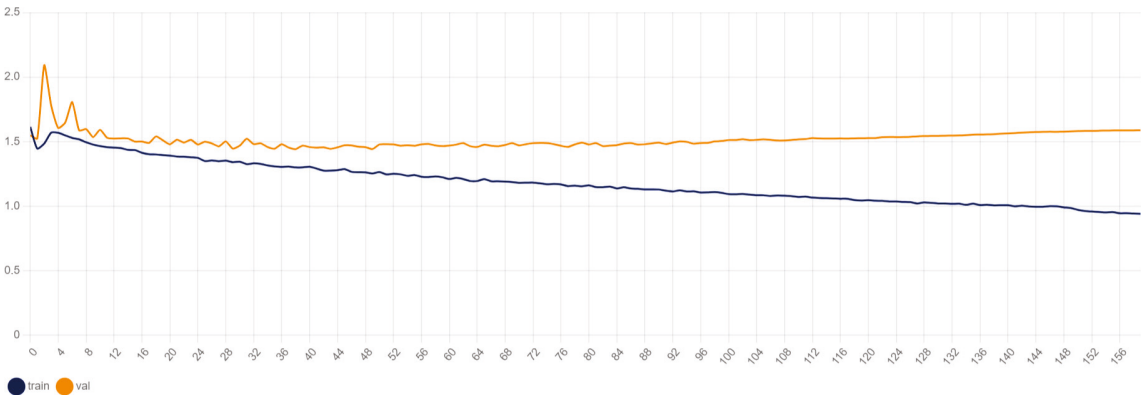



**Figure 10.** YOLOv8s object loss.

The overall decreasing trend in object loss confirms that YOLOv8s is becoming increasingly effective at detecting road damage, reducing classification errors, and enhancing detection reliability.

Figure 11 presents the evolution of key performance metrics for the CNN model during training, showcasing both training and validation results. The trends indicate high classification accuracy and stability across epochs:

- Training Accuracy (blue line): Rapidly increases and stabilizes around 90%, confirming the model's strong learning capability.
- Validation Accuracy (yellow line): Closely follows the training accuracy, indicating robust generalization on unseen data.
- Training Precision (green line): Maintains a high and stable value, demonstrating the model's effectiveness in correctly classifying positive instances.
- Validation Precision (red line): Shows some fluctuations but remains consistently high, reinforcing the model's reliability in classification.
- Training Recall (dark purple line): Approaches 1.0, ensuring minimal false negatives.
- Validation Recall (light purple line): Also remains high, further confirming the model's ability to correctly identify road sign classes.

**Figure 11.** CNN training and validation metrics.

The figure highlights the CNN's strong classification capabilities, achieving approximately 90% accuracy while maintaining reliable precision and recall scores. This performance confirms the model's effectiveness in correctly classifying detected road signs.

### *4.5. Summary of Findings*

YOLOv8x (Road Sign Detection): Achieves high mAP50 (~0.9), reflecting strong detection capabilities. Box loss and object loss trends confirm efficient learning for both localization and foreground/background separation.

YOLOv8s (Road Surface Damage Detection): Shows steady improvement in accuracy metrics, with reduced box and object loss over time, confirming its effectiveness in identifying damaged road surfaces.

CNN (Road Sign Classification): Reaches approximately 90% accuracy, indicating reliable classification of road signs once they are detected by the YOLO models.

Overall, the proposed models demonstrate promising performance for detecting and classifying road signs, as well as identifying road surface damage. Future work will focus on expanding the dataset, refining hyperparameters, and exploring advanced regularization techniques to further enhance performance.

Tables 1–3 clearly summarize the achieved results:

**Table 1.** Final performance metrics for YOLOv8x (road sign detection).

| Metric | Final Value (Approx.) |
|---|---|
| mAP (50) | 0.92 |
| mAP (50–95) | 0.78 |
| Precision | 0.88 |
| Recall | 0.85 |
| Box Loss (Training) | 0.30 |
| Box Loss (Validation) | 0.45 |
| Object Loss (Training) | 0.70 |
| Object Loss (Validation) | 0.90 |

**Table 2.** Final performance metrics for YOLOv8s (road surface damage detection).

| Metric | Final Value (Approx.) |
|---|---|
| mAP (50) | 0.75 |
| mAP (50–95) | 0.35 |
| Precision | 0.90 |
| Recall | 0.80 |
| Box Loss (Training) | 0.70 |
| Box Loss (Validation) | 1.80 |
| Object Loss (Training) | 0.80 |
| Object Loss (Validation) | 1.60 |

**Table 3.** Final performance metrics for CNN (road sign classification).

| Metric | Training (Approx.) | Validation (Approx.) |
|---|---|---|
| Accuracy | 0.95 | 0.88 |
| Precision | 0.85 | 0.80 |
| Recall | 1.00 | 0.90 |

## 5. Challenges, Solutions, and Integration with Municipal Maintenance Applications

In this section, we report the practical issues that were encountered during the design and implementation of our system and the implemented solutions to such issues. Then, we show how the trained model has been integrated with the tools available to the municipality.

The first issue in the application is the imbalance in the dataset, as most images are classified as "not damaged". This was managed by implementing the Focal Loss function, which helps to down-weight the loss assigned to well-classified examples, thus focusing more on the difficult, minority class. We also applied CutOut data augmentation to enhance the robustness of the model.

Another issue with the classification of a sign to be damaged or not is that often it is required to identify subtle details such as small stickers or scratches. To address this, we incorporated attention mechanisms (Spatial and Channel Attention) within the CNN architecture, enabling the model to focus on regions of interest within each image.

Furthermore, we had to consider that the images in the dataset varied significantly in terms of lighting conditions, angles, and weather. To counteract this, we applied extensive data augmentation techniques such as rotations, shifts, and brightness variations. Additionally, we included a diverse set of images in the dataset to improve model generalization.

We also observed signs of overfitting in the model during training. To mitigate this, we employed Dropout layers and BatchNormalization, and implemented early stopping based on validation performance to prevent the model from memorizing the training data.

Finally, the training of such deep learning models on large datasets requires significant computational power. We leveraged High-Performance Computing (HPC) and cloud resources to accelerate the training process, making use of powerful GPUs and distributed computing.

*Model Integration with the Municipal Maintenance Applications*

The software has been developed so that it can be seamlessly integrated into a mobile application tailored for municipal maintenance operators. Through this application, maintenance teams can access real-time information on road and traffic sign anomalies detected by the system. The key feature of the application is the use of georeferencing to display layers of defective road surfaces and traffic signs on a GIS map, allowing for

immediate visualization and action planning. Our approach combines top-down and bottom-up strategies. In addition to centralized monitoring, we encourage citizen reports, which contribute to the dataset of damaged traffic signs through a mobile application. These reports help balance the dataset classes and improve the accuracy and precision metrics of the convolutional neural network (CNN) through a continuous training process.

The application is designed with an intuitive user interface, enabling users to filter anomalies by type (e.g., potholes, damaged signs). As shown in Figure 12, the dashboard displays icons for different types of anomalies:

- Road Damage: Potholes, cracks, and other surface issues.
- Traffic Sign Damage: Defaced, rusty, or obstructed signs.

Each icon on the map provides detailed information about the anomaly, including its exact location, description, and images captured by the detection system.

Furthermore, the solution is designed to be easily replicable across municipalities of different sizes and can be implemented on smartphones. By leveraging cloud computing and scalable data storage, the system can manage large volumes of data and provide real-time updates to maintenance teams. This adaptability makes it suitable for application in any urban area, allowing municipalities to efficiently monitor and maintain their road infrastructure. Users can submit reports of damaged traffic signs, including GPS location and an image of the sign. This information is used to update the dataset in real time and prioritize maintenance interventions, ensuring a timely and efficient response.

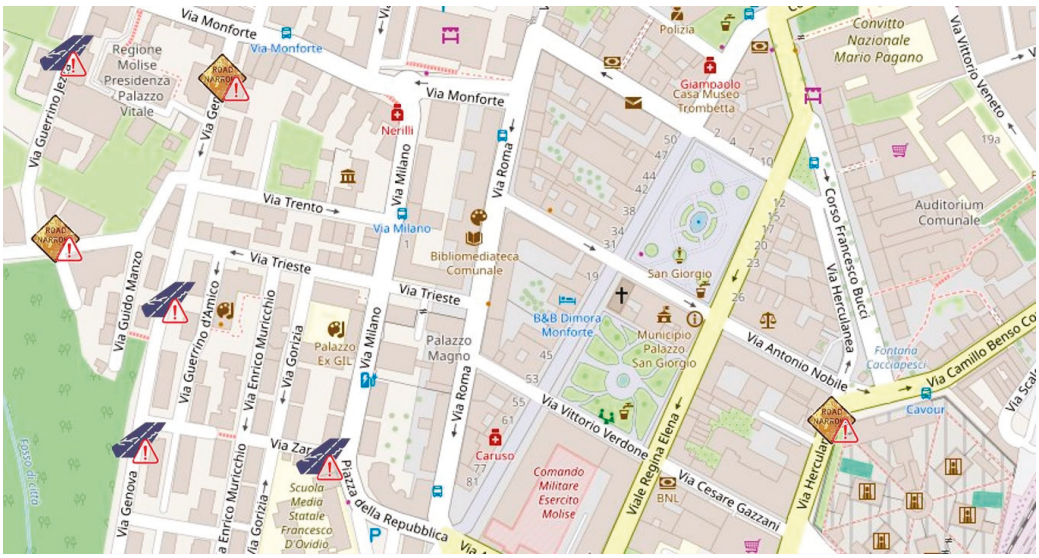



**Figure 12.** Dashboard showing georeferenced layers of road and traffic sign anomalies in the urban area of Campobasso, as explained by warning symbols.

The application is also designed to include additional features in the future, such as predictive maintenance algorithms, to analyze historical data to forecast potential road and traffic sign issues before they occur. This will enable municipalities to transition from reactive to proactive maintenance strategies, reducing costs and improving road safety.

## 6. Conclusions

In this study, we successfully developed and trained YOLO models for road sign detection and CNN models for classifying road signs as damaged or not damaged. Our approach utilized data augmentation and cutout regularization techniques to enhance the robustness and generalization of our models. Computational experiments conducted

on Google Colab and Reevo servers demonstrated the effectiveness of our methods in handling large datasets and complex computations.

For future work, we propose the following extensions to enhance the capabilities and applications of our models:

- Incorporating Retroreflectivity Factors: To further refine the classification of road signs, we plan to include retroreflectivity factors in our analysis. This involves detecting and classifying faded or discolored signs, which can significantly impact road safety. Developing models that can identify such signs will be crucial for timely maintenance and replacement.
- Leveraging Generative AI for Data Labeling: The process of manually labeling large datasets is time-consuming and prone to human error. By employing generative AI techniques, we can automate the labeling process, thereby reducing the time and effort required. This will also enable us to handle larger datasets more efficiently.

By implementing these extensions, we aim to improve the accuracy and reliability of road sign detection and classification systems. This will contribute to better road safety and maintenance practices, ultimately benefiting road users and maintenance authorities.

**Author Contributions:** Conceptualization, D.M. and A.S.; methodology, D.M., A.S. and V.L.; validation, D.M. and V.L.; formal analysis, D.M.; software, D.M.; writing—original draft preparation, D.M., A.S. and G.B.; writing—review and editing, D.M., A.S., V.L. and G.B.; supervision, A.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Molise CTE Project, funded by MIMIT (FSC 2014–2020), grant #D33B22000060001.

**Data Availability Statement:** All relevant data are included in the paper. Further inquiries can be directed to the corresponding author.

**Acknowledgments:** We would like to extend our gratitude to the project coordinators and funding bodies for their support and resources, which were instrumental in the successful completion of this research.

**Conflicts of Interest:** Antonio Salis and Gianluca Boanelli are employees of Tiscali Italia. The other authors declare no conflicts of interest.

## References

1. World Health Organization. Road Safety. 2023. Available online: https://www.who.int/health-topics/road-safety (accessed on 15 December 2024).
2. Maldonado-Bascón, S.; Lafuente-Arroyo, S.; Gil-Jiménez, P.; Gómez-Moreno, H.; López-Ferreras, F. Road-sign detection and recognition based on support vector machines. *IEEE Trans. Intell. Transp. Syst.* **2007**, *8*, 264–278. [CrossRef]
3. Fang, C.; Chen, S.-W.; Fuh, C. Road-sign detection and tracking. *IEEE Trans. Veh. Technol.* **2003**, *52*, 1329–1341. [CrossRef]
4. Yang, W.; Zhang, W. Real-time traffic signs detection based on YOLO network model. In Proceedings of the IEEE 2020 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Chongqing, China, 29–30 October 2020; pp. 354–357.
5. Zhang, X. Traffic sign detection based on YOLO v3. In Proceedings of the 2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA), Shenyang, China, 29–31 January 2023; pp. 1044–1048.
6. Song, W.; Suandi, S.A. Sign-YOLO: A novel lightweight detection model for Chinese traffic sign. *IEEE Access* **2023**, *11*, 113941–113951. [CrossRef]
7. Xu, T.; Ren, L.; Shi, T.; Gao, Y.; Ding, J.-B.; Jin, R.-C. Traffic sign detection algorithm based on improved YOLOX. *Inf. Technol. Control* **2023**, *52*, 966–983. [CrossRef]
8. Ciresan, D.C.; Meier, U.; Masci, J.; Schmidhuber, J. Multi-column deep neural networks for traffic sign classification. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3642–3649.

9. Song, W.; Suandi, S.A. TSR-YOLO: A Chinese traffic sign recognition algorithm for intelligent vehicles in complex scenes. *Sensors* **2023**, *23*, 749. [CrossRef] [PubMed]
10. Trpkovic, A.; Selmic, M.; Jevremovic, S. Model for the identification and classification of partially damaged and vandalized traffic signs. *KSCE J. Civ. Eng.* **2021**, *25*, 3953–3965. [CrossRef]
11. Acilo, N.; Cruz, A.G.S.D.; Kaw, M.K.L.; Mabanta, M.D.; Pineda, V.G.G.; Roxas, E.A. Traffic sign integrity analysis using deep learning. In Proceedings of the 2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA), Penang, Malaysia, 9–10 March 2018; pp. 107–112.
12. Dewi, C.; Chen, R.-C.; Liu, Y.-T.; Tai, S.-K. Synthetic data generation using DCGAN for improved traffic sign recognition. *Neural Comput. Appl.* **2021**, *34*, 21465–21480. [CrossRef]
13. Zhang, A.; Wang, K.C.; Li, B.; Yang, E.; Dai, X.; Peng, Y.; Fei, Y.; Liu, Y.; Li, J.Q.; Chen, C. Automated pixel-level pavement crack detection on 3D asphalt surfaces using a deep-learning network. *Comput. -Aided Civ. Infrastruct. Eng.* **2017**, *32*, 805–819. [CrossRef]
14. Maeda, H.; Sekimoto, Y.; Seto, T.; Kashiyama, T.; Omata, H. Road damage detection using deep neural networks with images captured through a smartphone. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 5207–5212.
15. Strutu, M.; Stamatescu, G.; Popescu, D. A mobile sensor network based road surface monitoring system. In Proceedings of the 2013 17th International Conference on System Theory, Control and Computing (ICSTCC), Sinaia, Romania, 11–13 October 2013; pp. 630–634.
16. Perttunen, M.; Mazhelis, O.; Cong, F.; Kauppila, M.; Leppänen, T.; Kantola, J.; Collin, J.; Pirttikangas, S.; Haverinen, J.; Ristaniemi, T.; et al. Distributed road surface condition monitoring using mobile phones. *J. Ambient Intell. Smart Environ.* **2011**, *3*, 64–78.
17. Tarun, R.; Esther, B.P. Real-time regional road sign detection and identification using Raspberry Pi. In Proceedings of the 2023 International Conference on Networking and Communications (ICNWC), Chennai, India, 5–6 April 2023; pp. 1–5.
18. Lim, X.R.; Lee, C.P.; Lim, K.M.; Ong, T.S.; Alqahtani, A.; Ali, M. Recent advances in traffic sign recognition: Approaches and datasets. *Sensors* **2023**, *23*, 4674. [CrossRef] [PubMed]
19. Pensa, D. Integration of GPS Data into Predictive Models for Tyre Maintenance. Master's Thesis, Politecnico di Milano, Milan, Italy, 2017. Available online: https://www.politesi.polimi.it/retrieve/a81cb05c-7f41-616b-e053-1605fe0a889a/tesi.pdf (accessed on 15 December 2024).
20. Neuhold, G.; Ollmann, T.; Rota Bulò, S.; Kontschieder, P. The Mapillary Vistas Dataset for semantic understanding of street scenes. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4990–4999. [CrossRef]
21. Arya, D.; Maeda, H.; Ghosh, S.K. RDD2022: A multi-national image dataset for automatic road damage detection. *Geosci. Data J.* **2022**, arXiv:2209.08538. [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.

computers

MDPI

Article

# Detection of Surgical Instruments Based on Synthetic Training Data †

**Leon Wiese [1,*,‡], Lennart Hinz [1,‡], Eduard Reithmeier [1], Philippe Korn [2] and Michael Neuhaus [2]**

[1] Institute of Measurement and Automatic Control, Faculty of Mechanical Engineering at the Leibniz University, An der Universität 1, 30832 Garbsen, Germany
[2] Department of Oral and Maxillofacial Surgery, Hannover Medical School, Carl-Neuberg-Straße 1, 30625 Hannover, Germany
* Correspondence: leon.wiese@imr.uni-hannover.de; Tel.: +49-511-762-13354
† This paper is an extended version of our paper published in SPIE MI107, https://doi.org/10.1117/12.3005798.
‡ Current address: Insitute of Measurement and Automatic Control, Leibniz University Hannover, An der Universität 1, 30832 Garbsen, Germany.

**Abstract:** Due to a significant shortage of healthcare staff, medical facilities are increasingly challenged by the need to deploy current staff more intensively, which can lead to significant complications for patients and staff. Digital surgical assistance systems that track all instruments used in procedures can make a significant contribution to relieving the load on staff, increasing efficiency, avoiding errors and improving hygiene. Due to data safety concerns, laborious data annotation and the complexity of the scenes, as well as to increase prediction accuracy, the provision of synthetic data is key to enabling the wide use of artificial intelligence for object recognition and tracking in OR settings. In this study, a synthetic data generation pipeline is introduced for the detection of eight surgical instruments during open surgery. Using 3D models of the instruments, synthetic datasets consisting of color images and annotations were created. These datasets were used to train common object detection networks (YOLOv8) and compared against networks solely trained on real data. The comparison, conducted on two real image datasets with varying complexity, revealed that networks trained on synthetic data demonstrated better generalization capabilities. A sensitivity analysis showed that synthetic data-trained networks could detect surgical instruments even at higher occlusion levels than real data-trained networks. Additionally, 1920 datasets were generated using different parameter combinations to evaluate the impact of various settings on detection performance. Key findings include the importance of object visibility, occlusion, and the inclusion of occlusion objects in improving detection accuracy. The results highlight the potential of synthetic datasets to simulate real-world conditions, enhance network generalization, and address data shortages in specialized domains like surgical instrument detection.

**Keywords:** surgical instrument detection; synthetic data generation; object detection; YOLOv8

## 1. Introduction

Hospitals are increasingly facing challenges with a diminishing pool of qualified personnel, which not only impacts the workload of the remaining staff but also significantly affects the quality of the procedures performed [1]. To reduce the workload, intelligent assistance systems can be integrated into the operating room. For instance, detection

algorithms can be utilized to identify the instruments used during surgery, which can, in turn, facilitate a variety of statistical analyses and documentation tasks.

The accuracy of learning-based recognition algorithms depends significantly on the available data. The larger and more diverse the datasets are, the better the training results of neural networks [2]. However, it is often difficult to have a sufficiently large amount of training data. This is especially relevant in areas which are hard to access. Additionally, the annotation process is often labor-intensive and error-prone. Furthermore, an automated annotation process can lead to significant cost savings through reducing the workload of the annotation process. In addition, data protection concerns are a major obstacle to the efficient annotation and generation of datasets, particularly in the medical context, due to critical and often unresolvable issues regarding data security and data sovereignty [3]. To overcome these problems, synthetic data generation can be utilized. This has the advantage of allowing any desired amount of training data to be produced. Additionally, such an approach enables the creation of a more diverse training dataset through domain randomization, thanks to the multitude of adjustable parameters, which leads to higher generalization in recognition algorithms [4]. However, automated approaches require the availability of powerful computing systems capable of rendering the images. The required computational power increases with the resolution of the images being rendered.

Synthetic data can be generated in different ways; according to Schieber et al. [5], the approach can be divided into two methods. The first method generates a multitude of new synthetic images from a small set of real images. This involves simple image-based manipulations such as the crop-and-paste method. By relying on existing images, this approach comes with various advantages and disadvantages. Assuming that all real images are already in the target domain, there is no or only a very small domain gap in this approach [6]. Furthermore, no additional information (3D models, materials, etc.) needs to be available. A disadvantage of this approach is that, due to the absence of a domain gap, the recognition algorithms exhibit low generalization. Additionally, with the help of the crop-and-paste method, objects can only be scaled and positioned, which can lead to complications with objects that can feature different geometric configurations (i.e., scissors). The domain gap reflects the differences between two domains, for example, in terms of the properties of their datasets or their distributions with regard to the training data and the actual network input. The larger the domain gap, the more different the domains and the data generated or collected in them are. A large domain gap can be reduced, among other methods, through domain adaptation. In this process, either unsupervised domain adaptation [7] or supervised domain adaptation [8] can be utilized.

The second method for creating a synthetic training dataset is based on the use of synthetically generated images, rendered from virtual scenes. Here, it is necessary to provide additional information in advance, particularly the three-dimensional geometries of all relevant objects. In addition, the material properties and textures of these objects must be known in order to enable realistic renderings. This information forms the basis for a render-based generation process in which the 3D models are rendered in different poses and in varying scenarios. By manipulating a variety of parameters concerning lighting, background and camera perspective, the models can be rendered under different conditions. This approach allows precise control over the simulation conditions and enables the systematic generation of extensive and varied datasets. The large number of adjustable parameters significantly affect the domain gap between real and synthetic images. This results in a large number of adjustable system parameters for synthetic scenes, which need to be precisely adapted to the application environment of the respective algorithms. However, it is important to ensure that a proper generalization of the networks is achieved.

This trade-off is crucial for training models that are robust to fluctuations and uncertainties in complex scenes while providing high predictive accuracies.

*Related Works*

The increasing acceptance of object detection algorithms leads to a rising demand for datasets beyond established research datasets like COCO [9] or ImageNet [10]. The application areas are becoming more specialized and harder to access, so that common data acquisition approaches are not able to generate sufficient datasets. For instance, it is nearly impossible to train a neural network to detect non-rigid linear objects using the datasets [9,10]. However, this becomes feasible when using a synthetic dataset [11]. Based on these challenges, various working groups are engaging in the synthetic generation of datasets for object detection applications.

Sagues-Tanco et al. [12] describe in their work a method by which a synthetic dataset can be generated based on the crop-and-paste method; this dataset was used to detect kitchen objects using a Mask-RCNN. A comparison between synthetic and real data showed that the synthetic data provided similarly good detection results. The creation of the data, due to the automatic generation, is significantly faster. However, this work also shows that the crop-and-paste method is highly dependent on manually labeled initial data.

To address this issue, it is possible to resort to fully synthetic generation methods using computer generated images. Dirr et al. have developed a method in [11] which, starting from a 3D model of a cable, develops a purely synthetic dataset. This dataset forms the basis for recognition and segmentation of cables in various positions. It has been demonstrated that a purely synthetic dataset is capable of providing sufficient information to solve the detection and segmentation task. These results are attributed to the deformation adaptation within the generation process, which allows for the simulation of any cable configurations, while adhering to given constraints.

A more detailed examination of the individual generation parameters and their impact on the generalization of detection algorithms is described in the paper [13] by Tremblay et al. It was shown that datasets with unrealistic parameter combinations produce a higher generalization of detection algorithms. Moreover, it was demonstrated that synthetic networks fine-tuned with real data perform better than those trained solely on real data.

Through the completely synthetic generation of data, there are additional advantages besides automatic annotation. For instance, ref. [14] have adapted the generation process such that, in addition to RGB and annotations, depth maps of the scenes are produced. With the help of these, a 6DoF pose estimation of the objects can be performed. A similar approach was used in [15] to produce a synthetic dataset for the recognition and pose estimation of 3D objects.

Given that creating sufficiently large datasets is particularly challenging in the medical field, this area is especially suitable for the integration of synthetic data generation processes. Wang et al. demonstrated in [16] that the crop-and-paste method is suitable for improving the detection accuracy of laparoscopic instruments. Similar to the findings in [13], it was shown that networks trained on synthetic datasets exhibit higher generalization than networks trained using conventionally labeled data.

The research group led by Ireland et al. presented the SORT dataset, which was generated by using a synthetic generation process [17]. This dataset was utilized to train various detection algorithms for counting surgical instruments. A comparison between conventional datasets and the synthetic dataset revealed that networks trained on synthetic data produced inferior results.

In a subsequent study, it was demonstrated that the use of synthetic datasets composed of rendered images can improve detection accuracy [18]. Using this dataset, improved results were achieved for four out of six medical objects. The slightly lower detection accuracy for the remaining two objects was attributed to the absence of adjustable 3D models in the dataset.

By adapting the models, it was demonstrated that networks trained exclusively on synthetic data outperformed conventionally trained networks across all classes [19]. Additionally, it was shown that synthetic data-trained networks exhibit higher generalization capabilities.

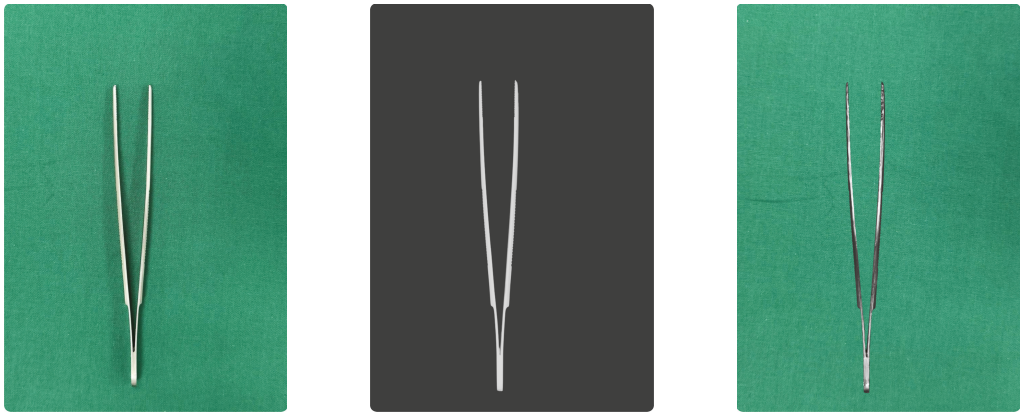
Besides the categorization from Schieber et al. there are different approaches for generating synthetic datasets. Lin et al. utilize text-to-image generators to enhance few-shot object detection algorithms [20].

## 2. Generation Pipeline for Synthetic Dataset

Since no dataset is available for the crop-and-paste method, this approach cannot be utilized. Therefore, this study adopts a fully synthetic approach. Furthermore, to the best of our knowledge, no previous work has been published that generates a synthetic dataset based on virtual scenarios and multiple parameter combinations to train neural networks for detecting surgical instruments during an operation.

Based on the advantages of render-based synthetic data generation as outlined above, a pipeline for the creation of a dataset for the recognition of surgical instruments is being developed using Blender (Blender Foundation, Amsterdam, Netherlands). Blender enables the creation and rendering of 3D scenes, and BlenderProc [21] is used to apply labels to the objects of the rendered images. This pipeline mainly follows the work of [19], but extends the adjustable generation parameters. As mentioned in Section 1, for render-based generation approaches, it is necessary to provide the geometric 3D information of the objects of interest. To provide high-resolution 3D geometries of all respective surgical instruments (here: 8), a 3D scanner GOM ATOS (Carl Zeiss GOM Metrology GmbH, Brunswick, Germany) based on fringe projection profilometry is used to reconstruct the 3D geometry of different surgical instruments from four different superior groups: needle holder, tweezers, scraper and scissors.

Figure 1a shows a real image of tweezers and Figure 1b depicts the 3D model. Since the fringe projection system is incapable of capturing textures, these are added manually. Therefore, Blender is used to generate nine different metallic textures which are assigned to the respective objects. The nine examined textures each exhibit varying degrees of mattness and reflectivity. Additionally, two of these textures feature blood splatters on the metallic surface. In order to fully utilize the possibilities of the synthetic generation approach, uniform colors are used in addition. This aims to achieve higher variance in the generated datasets. Figure 1c shows a rendering of the tweezers with a metallic texture on the tweezers surface. Since the applied 3D scanner reconstructs objects only in a fixed configuration, additional special multi-configuration approaches (e.g., for scissors) or non-rigid manipulation approaches (e.g., for tweezers) are necessary. Therefore, to be able to represent all possible configurations of the instruments, the 3D models of the four scissors and the needle holder are adapted, which results in further degrees of freedom, in addition to the object pose and texture.



(**a**) Real image (**b**) 3D Model (**c**) Rendered image

**Figure 1.** Surgical instruments.

In addition to the models and textures of the instruments, the environments in which the synthetic images are generated need to be defined. To create a the dataset, four environmental scenes with different complexities were created. The first is a square area on which the instruments can be placed. In this scene, only the color of the surface can be varied. In a second scene, the instruments are positioned in medical trays using Blender's physics simulation. Because the possible positions of the instruments are confined to the surface of the trays, a higher instrument density is expected in the generated images. This scene is intended to simulate the trays of instruments found in an operating room. A richer geometric structure arises from a plane shaped by the combined effects of sine and cosine waves. In each new render scene, the surface is generated with different parameters. Unlike this scene, the instruments are not positioned arbitrarily on the surface but are placed using Blender's physics simulation. In addition to the structure variability of the surface, the color of the surface is also adjustable, which leads to a more complex parameter space. The fourth scene features a virtually reconstruction of an operating theater, which is used to provide a realistic environment. Based on the instrument models and scenes, the synthetic dataset was created. Figure 2 displays the whole generation process for one set of images. This set is defined by a set of seven parameters:

- Scene ($\epsilon$)
- Camera object distance ($\chi$)
- Scene texture ($\gamma_{scene}$)
- Light intensity ($\lambda$)
- Instrument texture ($\gamma_{inst}$)
- Occlusion objects in scene ($\theta$)
- Number of instruments per scene ($\rho_{inst}$)

$\rho_{inst}$ denotes the number of instruments, $\gamma_{inst}$ corresponds to the texture of the instruments, the scene corresponds with $\epsilon$, the texture of the scene corresponds with $\gamma_{scene}$ (not for the virtual operating model), $\lambda$ denotes the light intensity, the position of the camera is defined by the distance between camera and object, which is denoted by $\chi$, and the use of occluding objects is denoted by $\theta$.

Based on the scene, the instruments are either placed on the surface or positioned by simulating a drop of the instrument. When the *plane surface* is selected, a non-overlapping positioning strategy is used to place the instruments on the surface. This results in low-complexity images, as no instruments overlap in the generated scenes. A more complex positioning method is employed when the scenes *random surface*, *medical box* and *medical lab* are selected. In these cases, all instruments are randomly positioned within the given scene, using Blender's physics simulation to simulate a drop onto the surface. The images

are generated by using $m$ camera positions, which ultimately result in $m$ rendered images. For experimental purposes, six Azure Kinect cameras (Microsoft, Redmond, Washington, DC, USA) were employed in a simulated operating room, each positioned to capture the operating table. These cameras offer a resolution of 4096 × 3076 pixels. To optimize computational efficiency, the captured images were segmented into patches of 512 × 512 pixels, which serve as the input for detecting surgical instruments. To ensure the transferability of results from simulated to real-world images, the dimensions of the rendered images were also set to 512 × 512 pixels. This standardization not only facilitates transferability but also preserves the 1:1 aspect ratio required for the employed network topology.

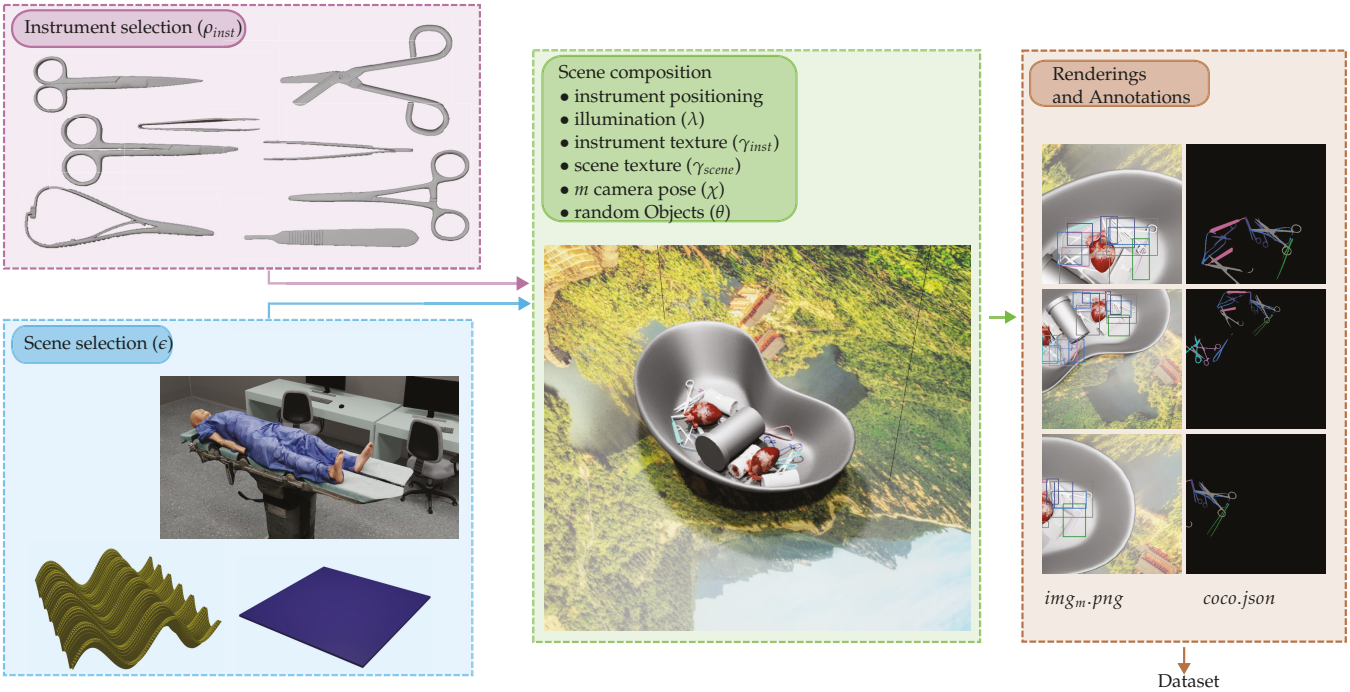
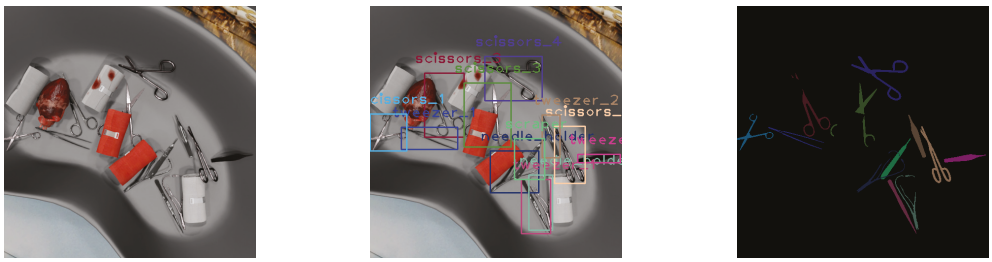



**Figure 2.** Dateset generation process.

BlenderProc is used to generate not only the rendered images but also the corresponding annotations and segmentation masks. This procedure is used to generate a database of 300k images, which are provided with the corresponding annotations. To create a highly diverse dataset, care was taken to ensure that the selected scene parameters were as evenly distributed as possible.

Figure 3 shows a rendering example (Figure 3a, the corresponding bounding box annotations in Figure 3b and the segmentation mask in Figure 3c, where the color indicates the object instance. As can be seen in Figure 3a, in addition to the instruments, occluding objects are integrated into the simulation.



(**a**) Rendered image (**b**) Bounding Box annotations (**c**) Segmentation mask

**Figure 3.** Example rendering.

## 3. Training for Detection Algorithms

The synthetic dataset created in Section 2 is used in the training of an object detection algorithm that localizes and classifies surgical instruments. Based on the good object detection results shown in [22], the YOLOv8 architecture with pretrained weights and the well-maintained implementation from Ultralytics [23] is used.

### *3.1. Real and Synthetically Trained Networks*

To ensure comparability between a conventional and synthetic dataset, the same model architecture is trained. The real dataset includes 1468 manually annotated images. The validation of the two trained networks is carried out using two real datasets. The first is a subset of the dataset on which the real network is trained. The second dataset is unknown to both networks and is in a slightly different environmental domain than the dataset on which the real network is based on. This dataset is used to assess the generalization of the networks and contains 100 manually annotated images. Based on the unknown real dataset, the networks are fine-tuned on it and also compared against each other. To continue evaluating the generalization, another dataset unknown to the networks is created for comparison. As shown in [19], the eight considered classes feature a considerable misclassification rate with respect to the parent classes. To reduce these misclassifications, similar instruments are grouped into classes. As a result, four classes are formed, which are composed as follows:

- needle holder: {needle holder};
- tweezers: {tweezers_1, tweezers_2};
- scraper: {scraper};
- scissors: {scissors_1, scissors_2, scissors_3, scissors_4}.

To utilize an optimal combination of the hyperparameters 'model size', 'batch size', 'learning rate', 'weight decay', 'optimizer' and 'pretrained' for the training process of the networks, we refer to the parameters identified in [19]. This leads to four trained networks: a network trained purely on synthetic data $net_{syn}$, a network trained purly on real data $net_{real}$, a synthetically trained network fine-tuned on real images $net_fine_{syn}$ and a real network fine-tuned on real images $net_fine_{real}$.

Figure 4 shows examples from the used datasets. Figure 4a is part of the dataset that is used to train the $net_{real}$. Figure 4b displays a synthetically generated image used for training the $net_{syn}$ in the virtual surgical environment with metallic textures of the surgical instruments' surfaces. Figure 4c,d shows examples from the two test datasets, where v1 is also the dataset used for fine-tuning. Dataset v2, based on the generated settings (occlusion, distance and occlusion objects), exhibits a higher level of complexity compared to dataset v1.

(**a**) (**b**) (**c**) (**d**)

**Figure 4.** Examples of the different datasets. (**a**) Example of a real training dataset. (**b**) Example of a synthetic dataset. (**c**) Example of a test dataset v1. (**d**) Example of a test dataset v2.

### *3.2. Impact of Generation Parameter*

In addition to examining the generalization of the networks, this study investigates the impact of various generation parameters with regard to the training results of the networks. For this purpose, different datasets were generated based on the parameters shown in Table 1.

**Table 1.** Generation parameters for synthetic dataset generation.

| Parameter | Options |
|---|---|
| Scene ($\epsilon$) | plane surface, random surface, medical box, virtual environment |
| Camera object distance ($\chi$) | 0.5, 1, 1.5, 2.0 |
| Scene texture ($\gamma_{scene}$) | Image, Color |
| Light intensity ($\lambda$) | $[10, 60]$, $[60, 110]$, $[110, 160]$, $[160, 210]$, $[600, 700]$ |
| Instrument texture ($\gamma_{inst}$) | Color, Metallic Texture |
| Occlusion objects in scene ($\theta$) | Yes, No |
| Number of instruments per scene ($\rho_{inst}$) | $[5, 10]$, $[10, 30]$, $[30, 50]$ |

The parameters used are derived both from the selected scenes and environmental influences. Additionally, unrealistic values were intentionally introduced to create high diversity within the datasets, particularly in aspects such as textures. For the parameters of camera distance, light intensity, and the number of instruments per scene, random values were employed. Care was taken to ensure that objects remain recognizable in the resulting images.

For the scalar parameter of camera distance, a normal distribution was applied, with the mean set as the selected value and a variance of 0.2. For the parameters of light intensity and the number of instruments per scene, random scalar values are selected within predefined intervals. This approach ensures variability while maintaining the integrity of the rendered images.

To be able to map any interactions between the individual parameters, all 1920 configurations of the parameters are generated. For each set, 500 images are generated. To assess the correlation between a tweaked parameter and the net results, the YOLOv8 architecture is trained for 20 epochs. After training, the accuracy of the recognition algorithm was checked using a real dataset v1.

## 4. Results

The results of the training processes carried out in Section 3 are shown in the Tables 2 and 3. The various metrics are calculated according to the specifications given in [24]. In addition to the average precision (AP), the tables also lists the average recall (AR). Both metrics are from the COCO Challenge [9]. In addition to the tables, Figure 5 depicts the AP of the models for the two assessed datasets. The AP indicates the average proportion of objects classified as positive by the model that are actually recognized correctly. Similarly, the AR describes the average proportion of relevant objects that are correctly recognized by the model. Both values enable a comprehensive evaluation of the model performance in the context of the recognition task. Based on the the metrics AP50, AP75, AR1, AR10, and AR100, a detailed performance analysis is carried out. Average precision (AP) is calculated at different intersection-over-union (IoU) thresholds: AP50 at 50% IoU and AP75 at 75% IoU. Average recall (AR) evaluates detection performance based on a fixed number of predicted bounding boxes per object. AR1 considers only one

bounding box per object, while AR10 and AR100 use the top 10 and 100 bounding boxes, respectively, for recall calculation.

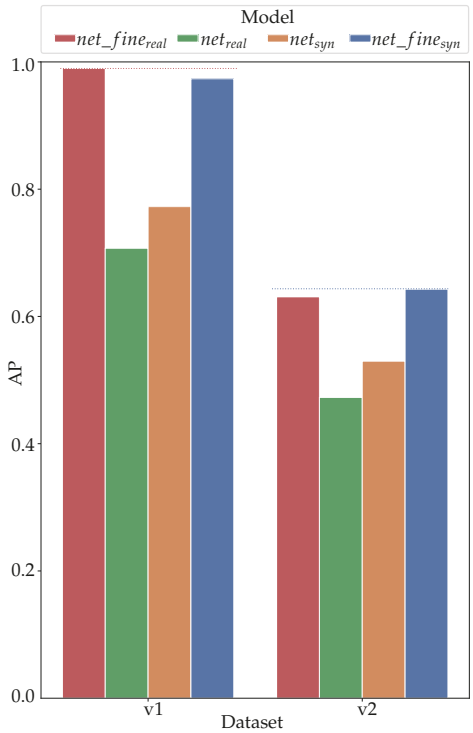



**Figure 5.** Training results.

**Table 2.** Results of nets trained only on real data.

| Dataset | Model | AP<br>AR1 | AP50<br>AR10 | AP75<br>AR100 | $AP_s$<br>$AR_s$ | $AP_m$<br>$AR_m$ | $AP_l$<br>$AR_l$ |
|---|---|---|---|---|---|---|---|
| v1 | $net_{real}$ | 0.707<br>0.594 | 0.706<br>0.658 | 0.681<br>0.658 | 0.058<br>0.056 | 0.626<br>0.665 | 0.690<br>0.702 |
| | $net_fine_{real}$ | **0.990**<br>**0.814** | **0.987**<br>**0.887** | **0.950**<br>**0.887** | **0.642**<br>**0.650** | **0.859**<br>**0.882** | **0.930**<br>**0.936** |
| v2 | $net_{real}$ | 0.473<br>0.423 | 0.473<br>0.458 | 0.424<br>0.458 | 0.042<br>0.050 | 0.348<br>0.419 | 0.875<br>0.927 |
| | $net_fine_{real}$ | 0.631<br>0.517 | 0.632<br>0.541 | 0.556<br>0.541 | 0.207<br>0.205 | 0.468<br>0.517 | 0.862<br>0.877 |

It can be concluded that both datasets with real and synthetic data provide sufficient information to assess the performance of the networks for the recognition and classification of surgical instruments. It can be observed that the less complex dataset v1 achieves better results in all criteria relevant for the composition of the training data. Furthermore, the results indicate that the network trained on real data in combination with fine-tuning on dataset v1 produces the best results. Only for the AP and the AR of the medium-sized (m) and large (l) objects does the network trained using the synthetic dataset with fine-tuning on the v1 dataset show a better AP value for $AP_m$ and a better AP value for $AP_l$.

**Table 3.** Results of nets trained on synthetic data.

| Dataset | Model | AP<br>AR1 | AP50<br>AR10 | AP75<br>AR100 | $AP_s$<br>$AR_s$ | $AP_m$<br>$AR_m$ | $AP_l$<br>$AR_l$ |
|---|---|---|---|---|---|---|---|
| v1 | $net_{syn}$ | 0.773<br>0.643 | 0.769<br>0.691 | 0.754<br>0.691 | 0.000<br>0.000 | 0.680<br>0.700 | 0.670<br>0.679 |
| | $net_fine_{syn}$ | **0.974**<br>**0.794** | **0.969**<br>**0.865** | **0.951**<br>**0.865** | **0.515**<br>**0.544** | **0.844**<br>**0.875** | **0.850**<br>**0.858** |
| v2 | $net_{syn}$ | 0.530<br>0.454 | 0.527<br>0.481 | 0.509<br>0.481 | 0.039<br>0.117 | 0.429<br>0.461 | 0.849<br>0.852 |
| | $net_fine_{syn}$ | 0.643<br>0.534 | 0.643<br>0.556 | 0.576<br>0.556 | 0.134<br>0.150 | 0.482<br>0.542 | 0.863<br>0.876 |

When considering networks that have not been fine-tuned on dataset v1, it becomes evident that networks trained on synthetic data produce better results. The average precision (AP) of $net_{real}$ decreases from 0.990 to 0.707, corresponding to a degradation of 38.67%. Similarly, the average recall at an intersection-over-union (IoU) threshold of 1 (AR1) drops from 0.814 to 0.594, representing a reduction of 33.58%.

A similar trend can be observed with the network $net_{syn}$. Here, the AP decreases from 0.974 to 0.773, which corresponds to a relative change of 24.84%. The AR1 decreases from 0.794 to 0.643 which is a reduction of 25.51%.

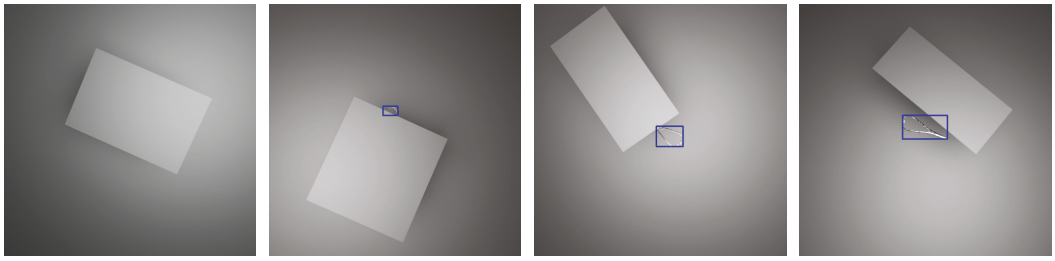
The evaluation results on the more complex dataset v2 show that both networks achieved lower performance compared to dataset v1. It is evident that the network trained on synthetic data consistently outperforms the network trained on real data across nearly all areas. When considering the fine-tuned networks, a similar trend is observed, with both networks achieving improved results. However, the performance gap between the two has narrowed.

Tables 2 and 3 provide an overview of the performance of the individual networks in detecting medical instruments. A specific analysis of the detection results based on object sizes (s, m, l) offers a more detailed insight into the networks responses to varying object sizes. The classification of objects is based on the area of their bounding boxes, without considering whether an object is small or merely partially occluded. Since the primary application involves the detection of surgical instruments during an surgery—where the instruments are often partially occluded due to being held in hands—it is essential to examine how the degree of occlusion affects the detection performance of different networks.

For this purpose, an additional dataset is created based on the procedure introduced in Section 2. This dataset contains 3800 images in which objects exhibit visibility levels ranging from 0% to 100%. Figure 6 illustrates four exemplary visibility levels of the objects.

Figure 7 illustrates the detection rates of the networks in relation to object visibility. A similar trend to the results observed on dataset v2 (see Tables 2 and 3) is evident. The network trained on synthetic data ($net_{syn}$) achieves the best detection performance. However, a notable difference is that the fine-tuned network exhibits a lower detection rate compared to $net_{syn}$.

For $net_{syn}$, objects with a visibility below 15% are not detected with sufficient accuracy, whereas objects with a visibility above 15% achieve good detection rates. In contrast, $net_fine_{syn}$ only reaches acceptable detection rates at visibility levels exceeding approximately 30%.



(**a**) 0% Visibility (**b**) 16% Visibility (**c**) 56% Visibility (**d**) 100% Visibility

**Figure 6.** Different object visibility states of a scissor (framed by the blue bounding box).

For networks trained on real datasets, the trends differ. The $net_{real}$ requires a significantly higher visibility threshold (<75%) to reliably detect objects. Similarly, the fine-tuned real network ($net_fine_{real}$) achieves good detection rates only when object visibility exceeds 60%.

**Figure 7.** Correlation between detection rates and object visibility.

*Generation Parameter Impact*

The preceding synthetic dataset from Section 3 is based on random parameter combinations. As shown in Figure 5, the synthetic dataset proved to be a suitable foundation for the detection of surgical instruments. Given the wide range of adjustable generation parameters, the impact of parameter sampling with regard to the detection performance is further investigated.

For this purpose, datasets are generated for all combinations of the parameters listed in Table 1. A total of 1920 datasets are created, each of which is used to train a detection algorithm. The trained models are evaluated using the real test dataset v1. The results of this investigation are depicted in Figure 8.

The six graphs illustrate the relationship between the selected parameter (x-axis) and the achieved mean average precision (mAP) on the test dataset (y-axis). The mAP averages the AP over all classes. The results are presented in relation of the chosen generation scenes. A focused analysis of the parameter *scene* reveals that the scene *med_box* achieves the highest median results for nearly all parameter configurations. Exceptions are observed in scenarios where the light intensity ($\lambda$) is in the range of $[160, 210]$ or where the number of instruments in the scene is the range of $[10, 30]$, in which case the scene *med_lab* performs

better. The scenes *plane_surface* and *random_surface* do not yield the highest median results for any of the analyzed parameter configurations.

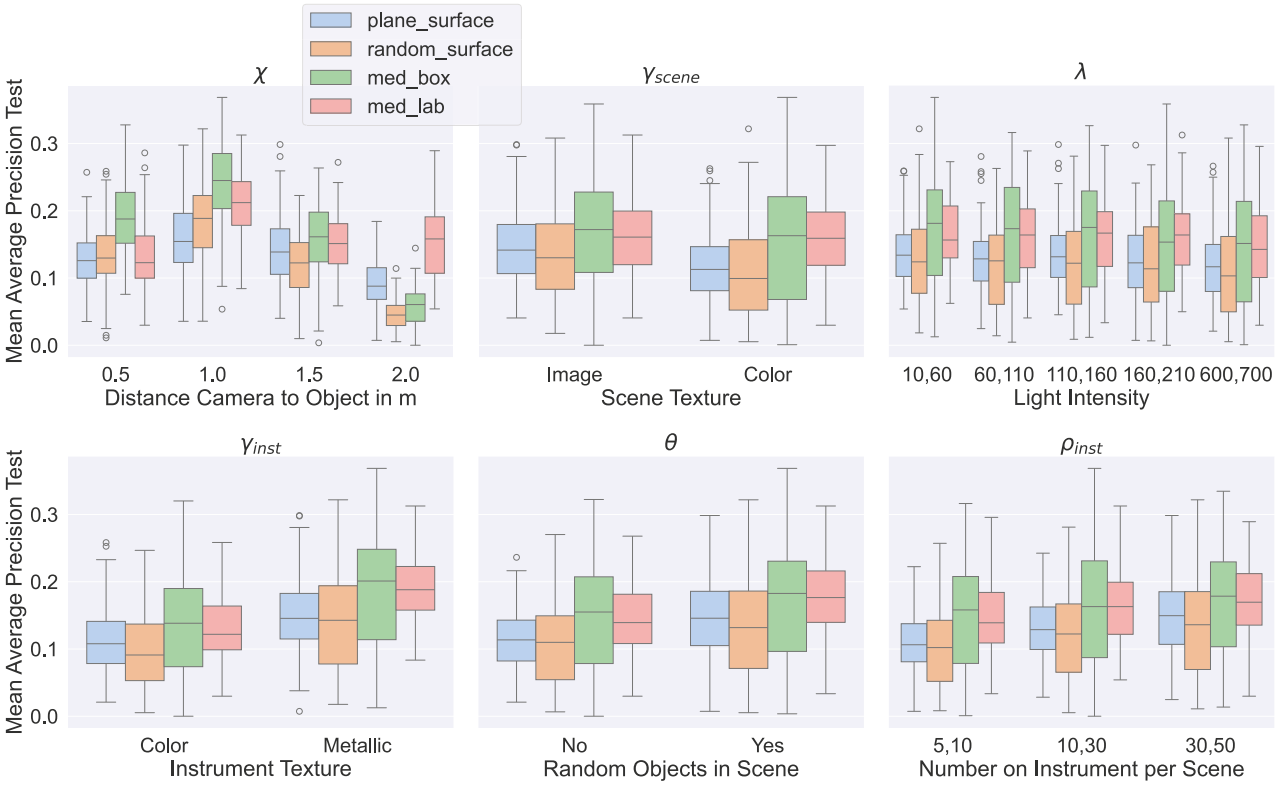



**Figure 8.** Impact of parameter sampling with regard to the detection performance.

When analyzing the impact of the distance between the render camera and the object ($\chi$), it is observed that the highest median mAP values across all four generation scenes occur at a distance of 1 m, with an average mAP of 0.196. As the distance increases, detection performance declines. An exception is observed in the *med_lab* scene, where a distance of 2 m produces slightly better results in direct comparison. Regarding the choice between an image-based background and a solid color background for the scene ($\gamma_{scene}$), it is evident that using an image as texture yields slightly better results, with an average improvement of 0.019 %. From the results of the varying light intensities ($\lambda$), no correlation between light intensity and mAP can be identified. Therefore, it is not possible to determine an optimal range of intensity. For the parameter related to instrument textures ($\gamma_{inst}$), metallic textures consistently achieve higher detection performances across all four scenes. On average, the mAP with metallic textures is 0.046 points higher compared to other textures. Examining the parameter for adding occlusion objects to the scene shows that including additional objects improves detection performance. The mAP increases from 0.128 (no additional objects) to 0.156 when occlusion objects are included. Finally, regarding the number of instruments per scene ($\rho_{inst}$), the results indicate that scenes with a higher number of instruments yield better detection outcomes.

## 5. Discussion

Based on the observations from Section 4, it can be concluded that a fully synthetic dataset can be utilized to train a detection and classification algorithm for surgical instruments. The comparison presented in the Tables 2 and 3 between networks trained on real and synthetic data shows that networks trained on synthetic data achieve better results on

previously unseen data (dataset v1). This behavior can be attributed to the higher diversity of the synthetic dataset, which is supposed to enable better generalization of the network. According to [25], such generalization is achieved when the training dataset exhibits a high degree of diversity. After performing fine-tuning on the networks using dataset v1, it can be observed that the network trained on real data achieves better results. This can be attributed to the similarities between dataset v1 and the real training dataset, which allow the newly adjusted network parameters to be better conditioned for detecting objects in this dataset. When analyzing the results of the four networks on the more complex dataset v2, it becomes evident that all networks achieved lower detection performance. However, it is also clear that the networks trained on synthetic data, benefiting from a larger and more diverse dataset and the resulting higher generalization capability, produced better results compared to the others. Furthermore, it can be observed that the performance could be improved through fine-tuning on dataset v1. This improvement can be attributed to the similarity between datasets v1 and v2, as both were captured in the same laboratory using the same cameras. The main differences include in the visibility of the instruments and variations in lighting conditions.

The trend shown in Figure 7 reflects the previous observations: networks trained on synthetic data require significantly lower object visibility compared to networks trained on real data. This behavior can primarily be attributed to the better generalization capability of the networks trained on synthetic datasets. Additionally, it may be influenced by the characteristics of the synthetic dataset itself. As illustrated in the example images in Figure 6, the dataset consists entirely of synthetic images, where the selected parameters do not fully reflect real-world conditions. Only the instrument texture ($\gamma_{inst}$) and lighting ($\lambda$) are adjusted to approximate real scenarios. It is hypothesized that the domain gap between the synthetic dataset and real training datasets is too large. A similar trend is observed for the fine-tuned synthetic network, which performs worse at lower object visibility compared to the non-fine-tuned network. In contrast, the fine-tuned real network demonstrates better performance, which can be attributed to the underlying dataset v1, characterized by a higher degree of instrument occlusion.

To minimize the domain gap, the influence of generation parameters on detection performance is to be emphasized. The results demonstrate that the optimal scene is *med_box*. This behavior can be attributed to the selected test dataset, as datasets generated with $\epsilon = med_box$ produce object positions similar to those in the real test dataset.

The optimal distance between the camera and the object can be attributed to the fact that all objects in the real test dataset are captured using multiple fixed cameras, each positioned at approximately the same distance from the surgical instruments. As a result, the bounding boxes in the test dataset exhibit a similar size. Comparing these bounding box sizes to those generated with a parameter combination involving a distance of $\chi = 1m$ reveals a similar size distribution.

Furthermore, it can be observed that using images as scene textures is preferable. This may be due to the fact that images create a better contrast between the background and the object, making the objects stand out more prominently and thus improving their detectability.

An analysis of the instrument texture ($\gamma_{inst}$) indicates that the use of metallic textures is likely to yield better detection results. This can be attributed to the fact that the instruments in the test dataset consistently exhibit a metallic texture. Variations in appearance arise only due to reflections caused by different lighting conditions, which were also simulated during the generation of the synthetic data. The anticipated increase in diversity through the use of surreal textures could not be achieved. This suggests that the domain gap should not become excessively large, as is the case when using surreal textures.

It can also be observed that the inclusion of occlusion objects led to improved detection results. This improvement can be attributed to the increased level of occlusion within the datasets caused by the presence of these additional objects. Furthermore, there is a possibility that the surgical instruments, during the simulation, may interact with the occlusion objects (e.g., by resting on them), allowing the instruments to be viewed from different perspectives.

For a similar reason, it is hypothesized that increasing the number of surgical instruments during the generation process improves detection performance. A higher number of instruments in the same space leads to increased occlusion, which may encourage the learning of more relevant object features. Furthermore, as the number of instruments per generation process increases, the number of object instances per image also rises. This results in a higher information density, enabling the networks to train more efficiently.

## 6. Conclusions

A comparison between a network trained on real data and one trained solely on synthetic data revealed that the synthetic dataset achieved better results on any unseen dataset. This finding led to the conclusion that networks trained on synthetic datasets exhibit better generalization capabilities, making them less sensitive to unknown data.

This presumed generalization advantage was further investigated using another synthetic dataset. During the generation of this dataset, the visibility of objects was predefined, enabling an analysis of how sensitive the networks are to varying levels of object visibility. The results show that networks trained on synthetic data can correctly detect surgical instruments even at lower visibility levels.

An analysis of the different parameters of the generation process revealed their varying influences on detection performance. The results demonstrate that a large domain gap significantly reduces detection accuracy, as observed with the given test dataset. Furthermore, it was shown that more realistic parameter combinations resulted in improved detection outcomes. Notably, the detection performance exhibited a strong dependence on the distance between the camera and the object, which directly impacts the size of the bounding boxes.

All results are discussed based on a single dataset, which limits their generalizability and makes them valid only for this specific dataset. Therefore, it is essential to conduct a comprehensive study to identify which dataset properties are particularly relevant for ensuring that the trained networks exhibit high generalization capabilities. Additionally, it is necessary to expand the number of surgical instruments used, as the set employed in this study represents only a fraction of the instruments used in an actual surgical procedure. A larger instrument dataset would not only enhance applicability but also enable a more comprehensive assessment of the generalizability achieved by the proposed approach. Nevertheless, the preliminary results indicate that, if the application domain of the networks is known, an optimal dataset can be generated using the synthetic generation process.

The findings of this study demonstrate that networks trained on synthetic datasets achieve higher accuracy compared to those trained on traditional manually created datasets. This approach is particularly advantageous in domains where access to large-scale datasets is restricted. Furthermore, the proposed methodology facilitates the generation of extensive datasets while ensuring compliance with privacy regulations and promoting diversity within the data. These attributes make the approach particularly valuable in medical applications, where it has the potential to improve recognition accuracy and significantly alleviate the workload of clinical staff.

## 7. Future Work

Based on this study, several aspects should be addressed in future investigations. First, further investigations are needed to explore the applicability of synthetic data, as it has been shown that such data can be used to train better-performing networks. Given the ability to generate unlimited synthetic data, it is important to evaluate the extent to which an increased dataset size continues to provide additional information and performance improvements.

Moreover, to derive more generalizable conclusions about the influence of various generation parameters, additional real datasets should be considered. It is also necessary to expand the examined parameter space to identify further dependencies and their effects on detection performance.

In addition to expanding the generation parameters, further investigation should focus on determining the extent to which an optimal parameter combination can be identified. This combination should aim to achieve the smallest possible domain gap while maximizing generalization.

Furthermore, the capabilities of synthetic data generation should be expanded, as the annotation effort for segmentation and detection tasks is notoriously high. This effort increases significantly when, in addition to detection and classification, keypoints on objects also need to be annotated. Keypoints are unique points on the object. Keypoints not only raise the annotation workload but also make the process more challenging, as they must consistently be placed at the same locations, which is particularly error-prone in manual annotation.

For these reasons, it would be valuable to investigate how the generation process can be adapted to utilize the existing information for automatically annotating defined keypoints. Such annotations could facilitate more robust pose estimation of objects [26,27].

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

**Funding:** This research was funded by Leibniz Young Investigator Grants program by the Leibniz University Hannover grant number 11-76251-114/2022.

**Data Availability Statement:** Dataset available on request from the authors.

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

## References

1. Ball, J.; Murrells, T.; Rafferty, A.; Morrow, E.; Griffiths, P. 'Care left undone' during nursing shifts: Associations with workload and perceived quality of care. *BMJ Qual. Saf.* **2013**, *23*, 116–125. [CrossRef] [PubMed]
2. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 843–852.
3. El Mestari, S.; Lenzini, G.; Demirci, H. Preserving data privacy in machine learning systems. *Comput. Secur.* **2024**, *137*, 103605. [CrossRef]
4. Zhou, X.; Koltun, V.; Krähenbühl, P. Simple Multi-dataset Detection. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 7561–7570.
5. Schieber, H.; Demir, K.; Kleinbeck, C.; Yang, S.; Roth, D. Indoor synthetic data generation: A systematic review. *Comput. Vis. Image Underst.* **2024**, *240*, 103907. [CrossRef]
6. Figueira, A.; Vaz, B. Survey on synthetic data generation, evaluation methods and GANs. *Mathematics* **2022**, *10*, 2733. [CrossRef]
7. Mattolin, G.; Zanella, L.; Ricci, E.; Wang, Y. ConfMix: Unsupervised Domain Adaptation for Object Detection via Confidence-based Mixing. *arXiv* **2022**, arXiv:2210.11539. Available online: https://arxiv.org/abs/2210.11539 (accessed on 6 January 2025).

8. Nugraha, A.; Novanto, Y.; Rahayudi, B. Supervised Virtual-to-Real Domain Adaptation for Object Detection Task using YOLO. *arXiv* **2023**, arXiv:2302.13891. Available online: https://arxiv.org/abs/2302.13891 (accessed on 6 January 2025).
9. Lin, T.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13, pp. 740–755.
10. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference On Computer Vision And Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
11. Dirr, J.; Gebauer, D.; Yao, J.; Daub, R. Automatic image generation pipeline for instance segmentation of deformable linear objects. *Sensors* **2023**, *23*, 3013. [CrossRef] [PubMed]
12. Sagues-Tanco, R.; Benages-Pardo, L.; López-Nicolás, G.; Llorente, S. Fast synthetic dataset for kitchen object segmentation in deep learning. *IEEE Access* **2020**, *8*, 220496–220506. [CrossRef]
13. Tremblay, J.; Prakash, A.; Acuna, D.; Brophy, M.; Jampani, V.; Anil, C.; To, T.; Cameracci, E.; Boochoon, S.; Birchfield, S. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 969–977.
14. Thalhammer, S.; Patten, T.; Vincze, M. SyDPose: Object Detection and Pose Estimation in Cluttered Real-World Depth Images Trained using Only Synthetic Data. In Proceedings of the 2019 International Conference On 3D Vision (3DV), Quebec City, QC, Canada, 16–19 September 2019; pp. 106–115.
15. Tremblay, J.; To, T.; Birchfield, S. Falling things: A synthetic dataset for 3d object detection and pose estimation. In Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2038–2041.
16. Wang, A.; Islam, M.; Xu, M.; Ren, H. Generalizing Surgical Instruments Segmentation to Unseen Domains with One-to-Many Synthesis. In Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, USA, 1–5 October 2023; pp. 4608–4614.
17. Irel, J.; Radwan, I.; Herath, D.; Goecke, R. Can Synthetic Data Improve Multi-Class Counting of Surgical Instruments? In Proceedings of the 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Sydney, Australia, 30 November–2 December 2022; pp. 1–8.
18. Hydock, K.; Elliott, A.; Busch, M.; Lipchak, L.; Blair, D.; Chapman, J. Generation of Synthetic Data for Medical Decision Support Applications. In Proceedings of the 2023 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), St. Louis, MO, USA, 27–29 September 2023; pp. 1–7.
19. Wiese, L.; Hinz, L.; Reithmeier, E. Medical instrument detection with synthetically generated data. In *Medical Imaging 2024: Imaging Informatics for Healthcare, Research, and Applications*; SPIE: Bellingham, WA, USA, 2024; Volume 12931, pp. 45–53.
20. Lin, S.; Wang, K.; Zeng, X.; Zhao, R. Explore the power of synthetic data on few-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 638–647.
21. Denninger, M.; Sundermeyer, M.; Winkelbauer, D.; Olefir, D.; Hodan, T.; Zidan, Y.; Elbadrawy, M.; Knauer, M.; Katam, H.; Lodhi, A. Blenderproc: Reducing the reality gap with photorealistic rendering. In Proceedings of the 16th Robotics: Science And Systems, RSS 2020, Workshops, Virtual, 12–16 July 2020.
22. Hussain, M. YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. *Machines* **2023**, *11*, 677. [CrossRef]
23. Ultralytics YOLOv5: A State-of-the-Art Real-Time Object Detection System. 2021. Available online: https://docs.ultralytics.com (accessed on 27 November 2024).
24. Padilla, R.; Netto, S.; Da Silva, E. A Survey on Performance Metrics for Object-Detection Algorithms. In Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), Niteroi, Brazil, 1–3 July 2020; pp. 237–242.
25. Zhang, W.; Jiang, J.; Shao, Y.; Cui, B. Efficient Diversity-Driven Ensemble for Deep Neural Networks. In Proceedings of the 2020 IEEE 36th International Conference On Data Engineering (ICDE), Dallas, TX, USA, 20–24 April 2020; pp. 73–84.
26. Li, Y.; Zhang, S.; Wang, Z.; Yang, S.; Yang, W.; Xia, S.; Zhou, E. Tokenpose: Learning keypoint tokens for human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11313–11322.
27. Hein, J.; Seibold, M.; Bogo, F.; Farshad, M.; Pollefeys, M.; Fürnstahl, P.; Navab, N. Towards markerless surgical tool and hand pose estimation. *Int. J. Comput. Assist. Radiol. Surg.* **2021**, *16*, 799–808. [CrossRef] [PubMed]


**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.

computers

MDPI

Article

# Leveraging Azure Automated Machine Learning and CatBoost Gradient Boosting Algorithm for Service Quality Prediction in Hospitality

**Avisek Kundu [1,2], Seeboli Ghosh Kundu [3], Santosh Kumar Sahu [4,*] and Nitesh Dhar Badgayan [5,*]**

1 Technology Consulting (Data Science, ML & AI), Ernst & Young LLP, Gurgaon 122002, India
2 Department of Operations and IT, IBS, Hyderabad (A Constituent of ICFAI Foundation for Higher Education), Hyderabad 501203, India
3 Symbiosis Centre for Management Studies, Bengaluru Campus, Symbiosis International (Deemed University), Pune 560100, India
4 School of Mechanical Engineering, VIT-AP University, Besides A.P. Secretariat, Amaravati 522237, India
5 KPMG, Mumbai 400011, India
* Correspondence: sksahumech@gmail.com (S.K.S.); nitesh.badgayan@gmail.com (N.D.B.)

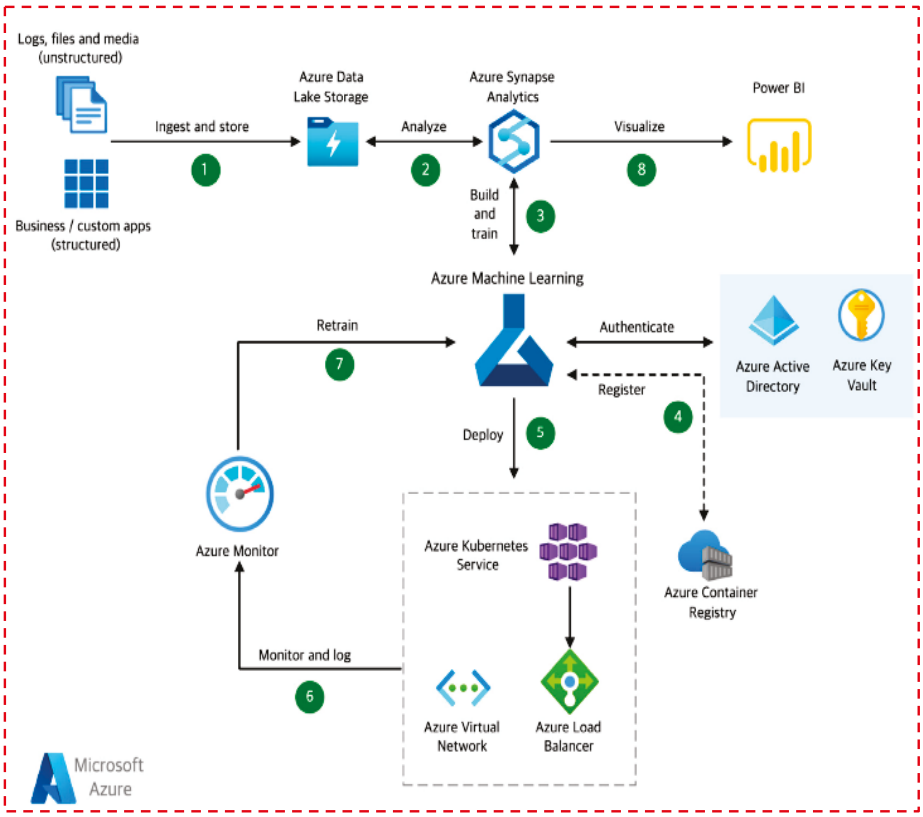
**Abstract:** The importance of measuring service quality for business performance has been widely recognized in service marketing literature due to its pivotal influence on customer satisfaction and its long-term impact on customer loyalty. The SERVQUAL model, comprising five dimensions—reliability, assurance, tangibility, empathy, and responsiveness—provides a measurable framework for evaluating the overall customer satisfaction. This study endeavors to ascertain whether all SERVQUAL dimensions carry equal weight in their effect on the overall service quality and to estimate the service quality based on various input features. To achieve this, questions were framed to assess the impact of variables such as gender, age, marital status, highest level of education, and frequency of hotel stays. The importance of each feature relative to the five SERVQUAL dimensions was investigated using machine learning models, specifically, CatBoost and Microsoft Azure Automated Machine Learning (AutoML) studio. This study revealed that both CatBoost and Azure AutoML identified the frequency of hotel stays and age group as the dominant predictors of service quality. Additionally, Azure AutoML highlighted the marital status as a more significant factor, suggesting its potential influence on customer preferences. The comparative modeling results demonstrated a strong alignment between the feature importance derived from CatBoost and Azure AutoML, enabling decision-makers to identify which dimensions are influenced by specific predictors and focus on targeted improvements.

**Keywords:** service quality; SERVQUAL model; CatBoost; Azure Automated Machine Learning

## 1. Introduction

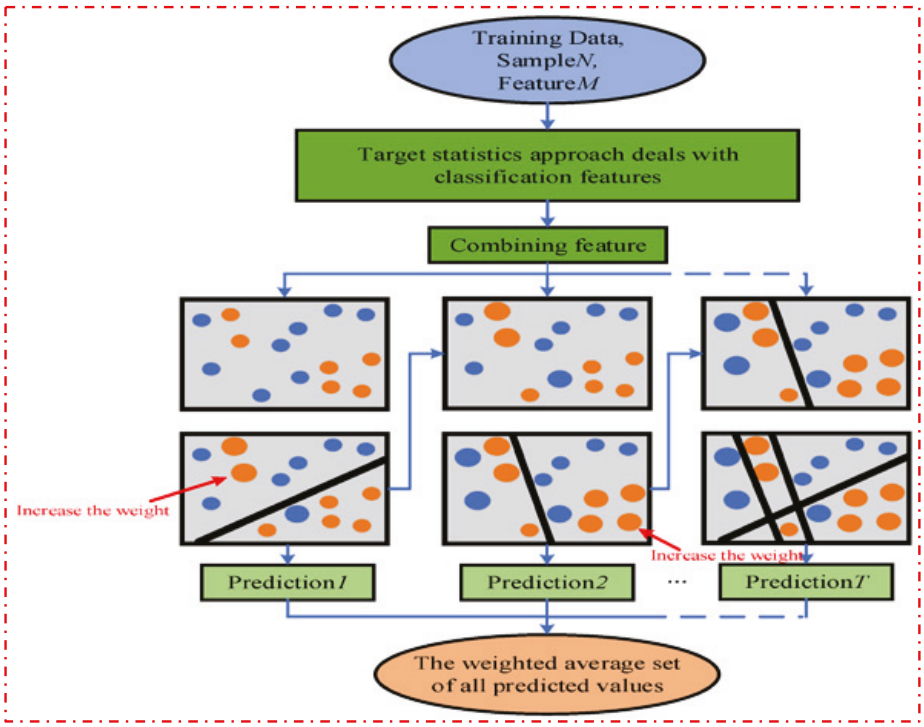
SERVQUAL is a well-established framework for assessing service quality across five dimensions: reliability, assurance, tangibility, empathy, and responsiveness. These dimensions are critical in understanding and improving customer satisfaction within service industries. Traditional methods of measuring service quality often rely on subjective assessments or manual data collection, which can be time-consuming and prone to biases. Machine learning (ML) models, however, offer a more objective and data-driven approach to predict and assess service quality. To enhance the effectiveness of SERVQUAL, machine

learning techniques offer a more objective, scalable, and efficient approach for modeling and predicting service quality across its dimensions. Cloud computing platforms have gained popularity recently due to their scalability, flexibility, accessibility, and collaboration. One cloud platform is Microsoft Azure, which offers various cloud services, including automated machine learning (AutoML). AutoML simplifies and accelerates the process of building, training, and deploying ML models. It also automatically selects the features and carries out hyperparameter tuning without extensive manual tuning of hyperparameters. This ability enhances its applicability across domains, from engineering to business applications [1]. It generally does not require coding as it enables users to explore multiple algorithms and configurations efficiently and effectively. It often results in high-performing models that align with best practices in machine learning [2]. Complementing the automated modeling process, gradient-boosting algorithms, such as CatBoost, can handle categorical data without extensive programming. One advantage of this algorithm is that it prevents overfitting through techniques like ordered boosting and oblivious trees [3]. It can handle categorical data directly without any encoding. Besides this, it can capture complex nonlinear relationships in data, enhancing accuracy. It also supports GPU acceleration, which makes it practical for handling big data [4]. Figure 1a,b show the architecture of Azure AutoML and CatBoost.




(**a**)

**Figure 1.** *Cont.*




(**b**)

**Figure 1.** Architectures of (**a**) Azure AutoML [5] and (**b**) CatBoost [6].

The comparative modeling of SERVQUAL dimensions comprises five distinct factors, as shown in Figure 2. It pertains to an organization's capacity to address customer requisites [7–9]. Reliability relates to an organization's proficiency in providing persistent and precise services [10]. Assurance relates to an organization's positive declaration intended to give confidence, or, in other words, a promise [11,12]. Tangibility refers to the physical aspects of the service, which can include the appearance of the facilities, equipment, and personnel [13]. Empathy relates to the ability to understand and share the feelings of another [14,15]. Responsiveness pertains to reacting quickly and positively [12,15].

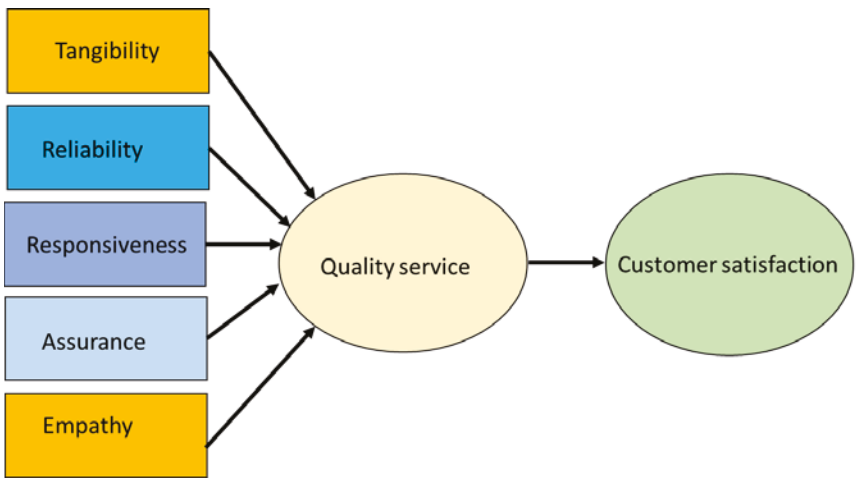



**Figure 2.** SERVQUAL Model: factors influencing service quality and customer satisfaction [9].

The following section summarizes the literature on the application of SERVQUAL in different service-based industries and the application of varying ML algorithms for

predictive modeling. Park et al. [16] propose a decision-support framework to evaluate hotel service quality by analyzing online reviews. The framework identifies critical service attributes, conducts sentiment analysis to gauge guest satisfaction, benchmarks service quality across hotels, and assesses attribute-specific quality associations. German et al. [17] examined factors influencing consumers in the Philippines to select package delivery services during the COVID-19 pandemic, utilizing the pro-environmental planned behavior theory and SERVQUAL. Lizarelli et al. [18] propose an integrative framework combining SERVQUAL, Analytical Kano (A-Kano), and Quality Function Deployment (QFD) with fuzzy techniques to address imprecision and uncertainty in customer perception data. The framework involves four phases: identifying quality attributes, integrating SERVQUAL and A-Kano with a Fuzzy Inference System (FIS), linking outputs with QFD using a 2-tuple fuzzy linguistic representation, and prioritizing improvement projects. It was tested in an entrepreneurial education firm; the approach enhances service quality assessment and prioritizes technical requirements, offering a novel methodology for integrating SERVQUAL, A-Kano, and QFD to advance service quality improvements. Adler et al. [19] examined a fundamental limitation in Gradient Boosting Machines (GBMs), where base learners (usually decision trees) tend to bias feature importance (FI) toward high-cardinality categorical variables. Despite their predictive solid performance, this bias impacts FI measures. The study introduced a cross-validated, unbiased base learner approach that mitigated this issue with minimal computational overhead, improving the FI accuracy while maintaining the GBM's predictive power across synthetic and real-world datasets. Rivero et al. [20] examined how customer satisfaction (CS) is influenced during disruptive events, using the SERVQUAL model to mediate the effects of service innovation (SI) and service recovery (SR) on CS. In the context of Typhoon Odette in the Philippines, the study analyzed responses from 584 participants, testing seven hypotheses with Partial Least Squares Structural Equation Modeling. The findings revealed that SERVQUAL mediates the relationship between SI and CS partially, and fully mediates SR to CS, but the direct SR-CS relationship is unsupported. The results indicated that during massive disruptions, CS relies more on human-centric SERVQUAL dimensions than product restoration alone, emphasizing that SR efforts should focus on SERVQUAL aspects to enhance CS effectively. Stefano et al. [21] discussed the importance of quality in determining both product or service performance and customer satisfaction, emphasizing how consumers' perceptions shape their overall assessment of a service. Given the abstract nature of service quality—driven by intangibility, heterogeneity, and inseparability—this study defined it as perceived by the customer and highlighted the gap between expectations and actual perceptions. Customer expectations, rooted in perceived needs, can differ widely from real needs, thus shaping satisfaction. To address quality in service provision, this paper evaluated a large hotel's service quality using fuzzy SERVQUAL and fuzzy AHP, revealing significant improvement areas in customer satisfaction. Rosário et al. [22] reviewed 74 studies on AutoML, highlighting its benefits for businesses. AutoML reduces time and resources for developing models, accelerates decision-making, and enables the creation of accurate predictive models. It also improves model performance, enhances accessibility, and democratizes innovation. As businesses grow, AutoML scales to handle larger datasets, driving efficiency, accuracy, and innovation. Paladino et al. [23] assessed the performance of three AutoML tools—PyCaret, AutoGluon, and AutoKeras—on heart disease datasets. The study compared these tools to traditional machine learning models and found that AutoML tools outperformed the conventional models, with AutoGluon achieving the highest accuracy (78–86%). While AutoML simplified model creation, the study highlighted the need to address its limitations. The findings suggested that AutoML could significantly

improve heart disease diagnosis and prevention. Abdulrab et al. [24] emphasized the significance of each SERVQUAL dimension and discussed the challenges and technological integrations faced by the hospitality industry. Table 1 explains the major contributions from references.

**Table 1.** Major contributions from references.

| Major Contribution | Description | Reference |
|---|---|---|
| Innovative Framework | The study introduces a novel approach to modeling SERVQUAL dimensions by combining Azure AutoML and CatBoost for enhanced service quality prediction. | [1–3] |
| Automation of Service Quality Modeling | It highlights how AutoML automates processes such as feature selection, hyperparameter tuning, and model evaluation, making service quality assessments more efficient. | [1,4] |
| Handling Categorical Data | CatBoost is utilized to handle categorical data directly without the need for encoding, preventing overfitting and improving model accuracy. | [3,4] |
| Scalability and Efficiency | The integration of Azure AutoML with CatBoost provides a scalable and efficient framework for service quality modeling that can be applied across different industries. | [1,2,4] |
| Cross-Validation for Reliability | The approach emphasizes the use of cross-validation to ensure that findings are robust and reliable, improving the trustworthiness of results. | [4,5] |
| Improved Accuracy in Prediction | By combining AutoML with gradient-boosting algorithms, the study improves the accuracy of service quality predictions in comparison to traditional methods. | [2,4,6] |
| Categorical Data Handling in Service Quality | CatBoost's ability to handle categorical data without the need for encoding and its prevention of overfitting are key features that benefit service quality modeling. | [3,6] |
| Suitability for Cross-Domain Applications | The use of Azure AutoML makes the framework applicable across various service industries, from hospitality to engineering, due to its flexibility and automation. | [2,5] |
| Modeling SERVQUAL Dimensions | The study successfully applies the SERVQUAL model to assess customer satisfaction, using a combination of modern ML techniques to predict and analyze service quality dimensions. | [7–9] |

The literature review underscores the extensive research conducted on assessing SERVQUAL dimensions using various machine learning (ML) algorithms. However, a significant gap exists in the comparative modeling of SERVQUAL dimensions using cloud platforms like Azure AutoML in conjunction with advanced boosted algorithms such as CatBoost. Existing studies have primarily relied on traditional ML approaches or fuzzy-based techniques, leaving unexplored the potential of integrating these advanced tools for a more streamlined and accurate analysis. This study addresses the identified gap by employing a novel approach to model SERVQUAL parameters using Azure AutoML and CatBoost with hotel customer data. Azure AutoML automates feature selection, hyperparameter tuning, and model evaluation, offering a coding-free and efficient solution for building high-performing models. CatBoost complements this by excelling in handling categorical data without the need for encoding and mitigating overfitting through techniques such as ordered boosting and oblivious trees. Together, these tools ensure robust

and reliable results by enabling the cross-validation of findings and accurately identifying key predictors of service quality. The novelty of this study lies in its unique combination of Azure AutoML and CatBoost, an approach not previously explored in the literature. While prior research has evaluated SERVQUAL parameters through traditional ML algorithms or fuzzy-based techniques, this integration offers a new perspective by combining the automation and scalability of cloud platforms with the precision of advanced gradient-boosting algorithms. This approach not only streamlines the modeling process but also enhances accuracy, effectively addressing categorical data challenges and preventing overfitting. By bridging this gap, the study establishes a comparative dimension modeling framework using cutting-edge tools, paving the way for scalable, efficient, and accurate service quality assessments in the hospitality industry.

## 2. Methods

### 2.1. Survey Methodology

Using an online survey with 617 complete responses from respondents, this study leverages Azure AutoML and CatBoost regressor ensemble models to capture the complex patterns of different dimensions and demographic features. It also uses different predictors to estimate customer satisfaction scores. The final response rate was calculated as the percentage of completed responses relative to the total sample contacted. A response rate of 28.04% (617 complete responses out of 2200 targeted respondents contacted via online survey) was achieved for this study. The response rate of similar studies ranges between 25% and 40%. To improve response rates and minimize non-response bias, a structured follow-up process was implemented. Participants who did not respond received up to 3 reminders at intervals of 72 h (3 days) post-initiation of the survey. A reminder was sent each week for one month (4 additional reminders) to respondents who did not respond to the initial invitation. The reminders were sent to encourage participation while respecting respondent burden, aiming to maximize complete and reliable data collection. Partial responses were also tracked, where respondents answered only a portion of the survey's questions. The partial response rate was 25.21% (208 out of 825), reflecting the proportion of respondents who initiated but did not complete the entire survey. There were 825 total respondents (partial responses and complete responses). A total of 208 respondents provided partial responses (post the total 7 reminders). For analysis, techniques such as imputation were not applied to partial responses to avoid introducing artificial consistency; only complete cases were used in primary analysis. The statistical power is calculated to be as high as 0.98, as shown in Table 2, to showcase robustness. The significance level was put as a default of 5%. This detailed approach to sampling and data quality monitoring provided a comprehensive and reliable dataset that supports the validity of our findings.

**Table 2.** Dependent variable dimension score.

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared | Noncent. Parameter | Observed Power |
|---|---|---|---|---|---|---|---|---|
| Corrected Model | 166,179.8 | 4 | 41,544.96 | 2308.63 | 0.02 | 0.747 | 9234.518 | 0.98 |
| Intercept | 1,165,684 | 1 | 1,165,684 | 64,785.41 | 0.019 | 0.954 | 64,785.41 | 0.981 |
| Category | 166,179.8 | 4 | 41,544.96 | 2308.63 | 0.02 | 0.747 | 9234.518 | 0.98 |
| Error | 56,325.94 | 3130 | 17.996 | | | | | |
| Total | 1,388,352 | 3135 | | | | | | |
| Corrected Total | 222,205.8 | 3134 | | | | | | |

### 2.2. Analysis of Data

The dataset for analysis consists of 10 columns and 628 rows comprising gender, age group, marital status, the highest level of education, and frequency of hotel stays as inputs and the 5 dimensions of SERVQUAL as the dependent variables. The unique value counts of each field are shown in Table 3.

**Table 3.** Value count in each field.

| Category | Values |
|---|---|
| Gender | Male: 399, Female: 228 |
| Age Group | 25–34: 490, 18–24: 129, 35–44: 8 |
| Marital Status | Married: 536, Single: 91 |
| Highest Level of Education | Upto Graduate: 341, Masters and above: 286 |
| Frequency of Hotel Stays (In 1 year) | 1: 123, 3: 106, 0: 102, 2: 82, 6: 75, 4: 72, 5: 67 |

The influence of various parameters, such as gender, age, marital status, highest level of education, and frequency of stays, was analyzed to determine which factor had the most significant impact on each of the 5 SERVQUAL dimensions and the total customer satisfaction score. This investigation was conducted using AutoML and CatBoost, which were implemented within a local Python environment. The selected hyperparameters for CatBoost are presented in Table 4, and the workflow of AutoML in Azure AI studio is shown in Figure 3. The hyperparameters were selected on the basis of (i) highest R square on the validation data and on the k fold (k = 10) cross-validation method, (ii) optimized adjusted R square (the difference between R square and adjusted R square is least), and (iii) least MAPE (mean average percentage error of the predicted vs. actual satisfaction score on the validation data). This ensures the selected model is optimized without being overfit or underfit.

**Table 4.** Hyperparameters considered.

| Hyperparameter | Values |
|---|---|
| Iterations | 500, 600, 700, 800 |
| Depth | 3, 4, 5, 6, 7 |
| Learning Rate | 0.01, 0.03, 0.05 |
| L2 Leaf Regularization | 1, 3, 5, 7 |

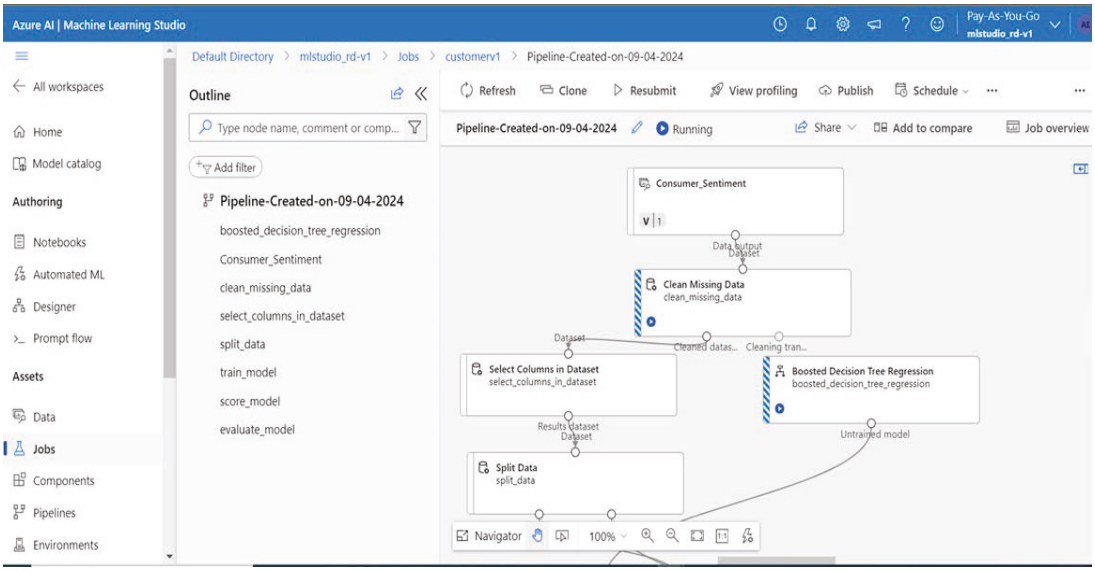



**Figure 3.** Workflow of AutoML in Azure AI Studio.

The configuration of Azure AI studio includes multiple elements that work in parallel with Azure ML and other Azure services. Setting up a managed network is required for setting up Azure AI studio. This helps in creating a boundary of strong network security for the different AI applications. The key components that are used for the setup are as follows: (a) Azure Machine Learning Workspace: this helps in providing the core service where the Azure AI studio is being set up; (b) Azure Key Vaults: this is set up to securely manage encryption keys required to ensure the security of the data and the models built on top of it; (c) Azure Storage: this is set up to ensure that the data marts, intermediate tables, models, and the aggregated output are stored securely in the cloud; (d) Azure Application Insights: this is set up to monitor the performance and usage of the applications to ensure the resources and the nodes are being used in an optimized way.

## 3. Results

### *3.1. CatBoost Results*

The CatBoost model was trained with the hyperparameters shown in Table 3 to extract feature importance for each dependent variable: reliability, assurance, tangibility, empathy, and responsiveness. The feature importance showed that frequency of hotel stays, marital status, age group, and gender strongly influence the dependent variables, with frequency of stays contributing the maximum influence. Similarly, the influence of each of the dimensions, namely, reliability, assurance, tangibility, empathy, and responsiveness, is regressed against the total score. The CatBoost decision tree, as shown in Figure 4, and the splitting rule, shown in Table 5, helps us understand the features in depth. As shown in Figure 5, a correlation heatmap helps us understand the correlation between the SERVQUAL parameters and the total score. Figure 6a–e show the feature importance for each SERVQUAL parameter, and it is concluded that reliability and assurance have the maximum influence on the total score, as evident from the output.

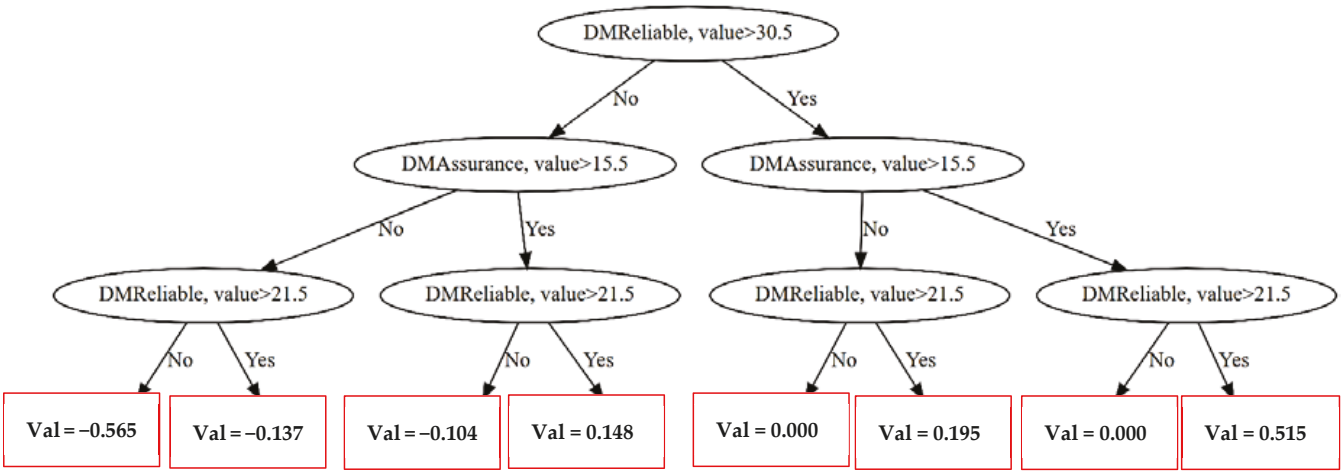



**Figure 4.** Decision tree model for predicting service quality based on reliability and assurance.

**Table 5.** Decision tree splitting rules for service quality based on reliability and assurance.

| Rule | DMReliable Condition | DMAssurance Condition | Value |
|---|---|---|---|
| 1 | DMReliable $\leq$ 30.5 | DMAssurance $\leq$ 15.5 | −0.565 |
| 2 | DMReliable $\leq$ 30.5 | DMAssurance $\leq$ 15.5 | −0.137 |
| 3 | DMReliable $\leq$ 30.5 | DMAssurance > 15.5 | −0.104 |

**Table 5.** *Cont.*

| Rule | DMReliable Condition | DMAssurance Condition | Value |
|---|---|---|---|
| 4 | DMReliable $\leq$ 30.5 | DMAssurance > 15.5 | 0.148 |
| 5 | DMReliable > 30.5 | DMAssurance $\leq$ 15.5 | 0.000 |
| 6 | DMReliable > 30.5 | DMAssurance $\leq$ 15.5 | 0.195 |
| 7 | DMReliable > 30.5 | DMAssurance > 15.5 | 0.000 |
| 8 | DMReliable > 30.5 | DMAssurance > 15.5 | 0.515 |

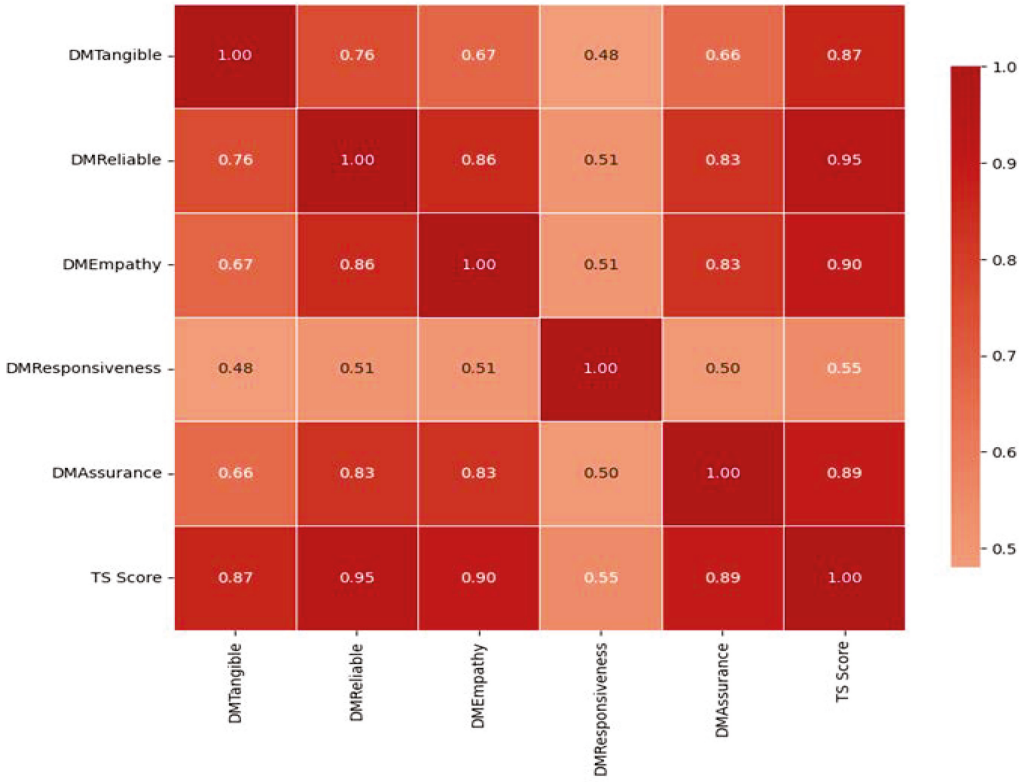



**Figure 5.** Correlation heat map of service quality dimensions and total score.

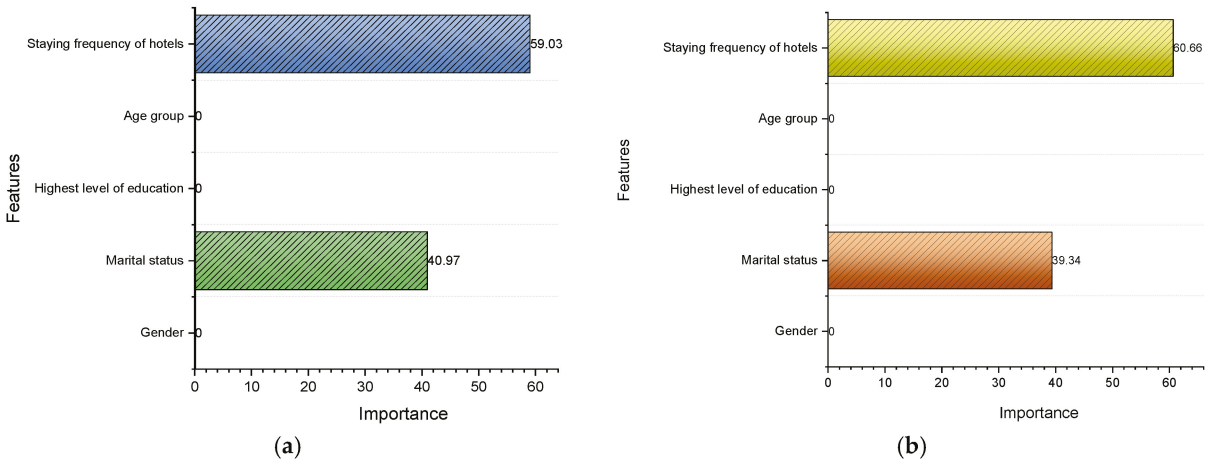



**Figure 6.** *Cont.*

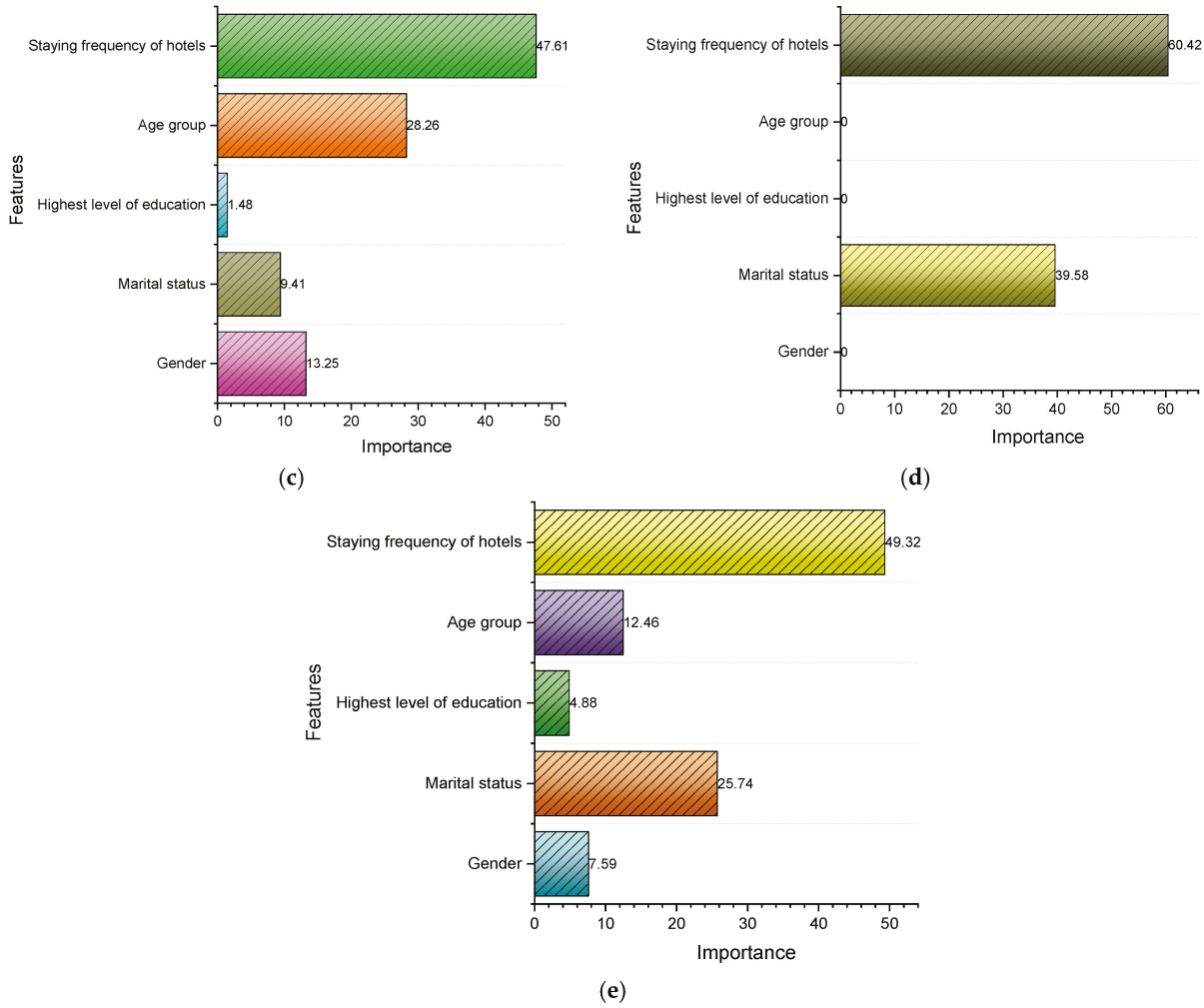



**Figure 6.** Feature importance results from CatBoost for service quality dimensions: (**a**) reliability, (**b**) assurance, (**c**) tangibility, (**d**) empathy, and (**e**) responsiveness.

### *3.2. Azure Auto ML Results*

The feature importance extracted from Azure Auto ML showed that the frequency of stays, marital status, age group, and gender strongly influence the dependent variables, with the frequency of stays contributing the maximum influence, as shown in Figure 7a–e.

A comparison of both models revealed that both CatBoost and Azure AutoML identified the frequency of hotel stays and age group as the dominant features. However, Azure AutoML has assigned slightly higher importance to the marital status, suggesting that it might be more significant in customer preferences than initially expected. Azure AutoML identified additional features that influenced the dependent parameters. This could be an option where in-depth analysis is required, analyzing the weight of each variable against the dependent variable. These insights show how advanced machine learning techniques, like those used here, can also be valuable in other areas [25]. For example, similar models could be applied to materials analytics to study polymer-based materials. By analyzing the influence of different variables—such as stress, strain, or shape recovery—on material behavior, these models could help us better understand how these materials perform under various conditions [26,27]. Table 6 presents the performance metrics of two predictive models, the CatBoost Gradient Boosting Model and the Azure Automated ML Model, evaluated based on their $R^2$, RMSE, and MAPE values. The Azure Automated ML Model demonstrates superior performance, with an $R^2$ value of 98.34%, indicating a stronger cor-

relation between predicted and actual values compared to the CatBoost Gradient Boosting Model, which has an $R^2$ of 91.50%. Additionally, the Azure model achieves a lower RMSE of 1.37 and MAPE of 1.17, highlighting its higher accuracy and precision in predictions. In contrast, the CatBoost model exhibits an RMSE of 2.10 and MAPE of 1.72, showing relatively higher error rates. These results underline the Azure Automated ML Model's effectiveness in delivering more accurate predictions with minimal errors.

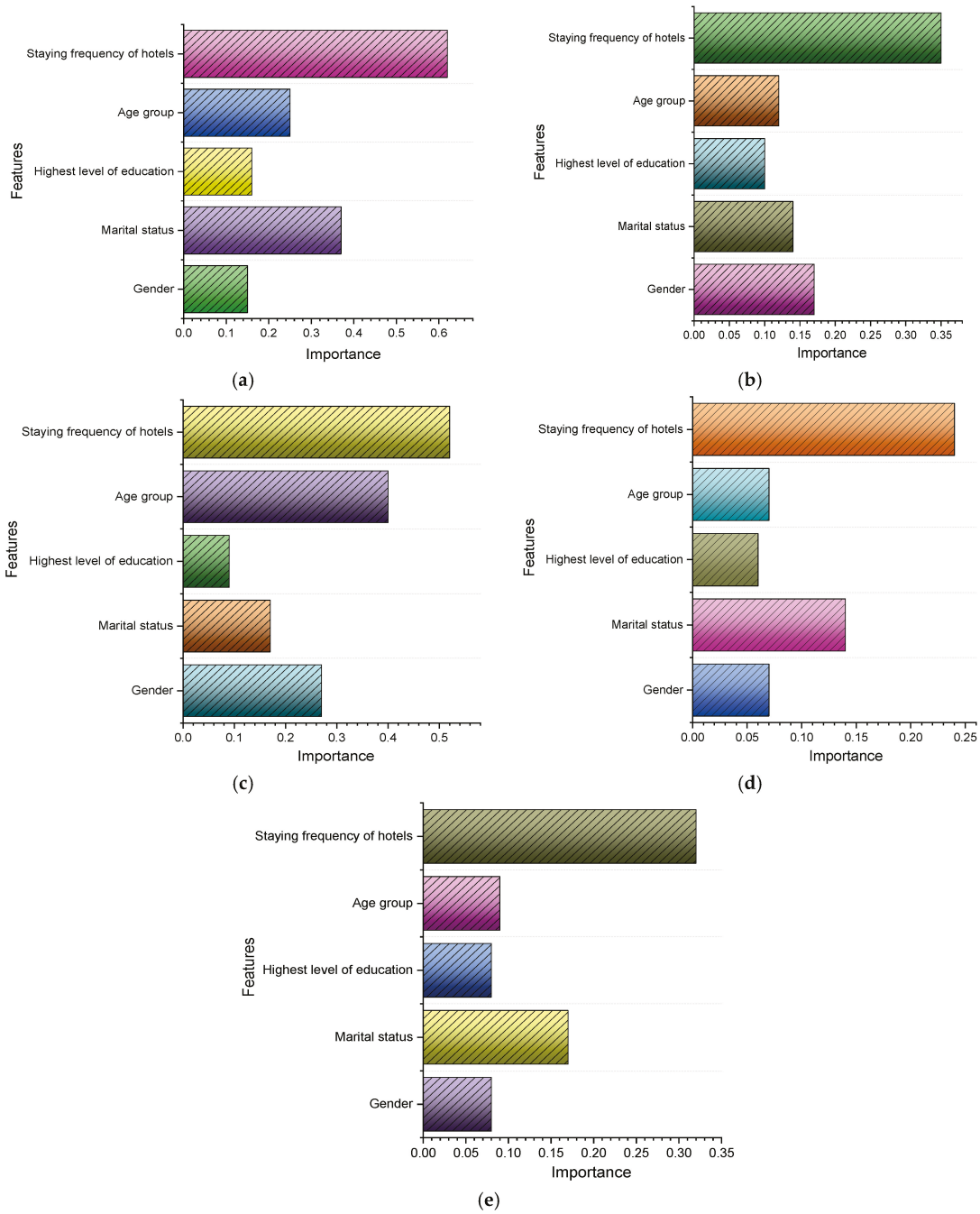



**Figure 7.** AutoML feature importance results for service quality dimensions for (**a**) reliability, (**b**) assurance, (**c**) tangibility, (**d**) empathy, and (**e**) responsiveness.

**Table 6.** Performance metrics of CatBoost Gradient Boosting and Azure Automated ML Model.

| Sr. No. | Models | R Square | RMSE (Root Mean Square Error) | MAPE (Mean Average Percentage Error) |
|---|---|---|---|---|
| 1 | CatBoost Gradient Boosting Model | 91.50% | 2.10 | 1.72 |
| 2 | Azure Automated ML Model | 98.34% | 1.37 | 1.17 |

## 4. Conclusions

Several service marketing researchers have emphasized the need to measure service quality because it greatly impacts customer satisfaction and customer retention in the long run. The present study attempts to evaluate if all the dimensions of SERVQUAL can be considered equal in their impact on the overall quality of the service. The research findings confirm that reliability and assurance are the most important factors affecting the environment of the traveler's decision-making process, especially in service quality. Both reliability and assurance are further influenced by the frequency of stays, marital status, and gender. The investigation results enable decision-makers to determine which dimension depends on specific predictors, allowing them to focus on targeted improvements. This lets the hotel manager focus on cardinal parameters like frequency of stays, gender, and marital status. For example, the hotel can launch a targeted marketing campaign focusing on a particular gender or marital status. This study also focuses on estimating the service quality of a respondent using the different predictor variables, leveraging different nonlinear machine learning models, including ensemble techniques. Its findings enrich the corpus of scholarly work centered on service quality by penetrating antecedents and subsequent repercussions within the Indian hotel industry. The revelations gleaned from this study beckon for further scholarly inquiry, particularly in shifting demographic and geographic parameters, their complex interplay, and their nuanced effect on service quality dynamics.

One of the limitations of this study is that a large-scale quantitative study with larger samples for validating the findings can be expensive and time-consuming, requiring significant resources for planning, data collection, and analysis. As a future study, the generalization of the findings could be tested by replicating the methodology across sectors to compare and validate the insights and findings. Explainable AI (XAI) is normally extensively used to bring AI to nontechnical end users including matrices like LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Additive explanations), TCAV (Testing with Concept Activation Vectors), DALEX (Descriptive Machine Learning Explanations), and DICE (Diverse Counterfactual Explanations). Since our research focuses on ML-based ensemble techniques and less on AI models, explainable AI matrices are not incorporated as part of this submission. However, this has been included as a suggestion for future research.

**Author Contributions:** Conceptualization, A.K. and S.G.K.; methodology, A.K. and S.G.K.; software, S.K.S. and N.D.B.; formal analysis, A.K., S.K.S., and N.D.B.; investigation, A.K., S.K.S., and N.D.B.; resources, S.K.S. and N.D.B.; data curation, S.K.S. and N.D.B.; writing—original draft preparation, A.K. and S.G.K.; writing—review and editing, S.K.S. and N.D.B.; visualization, A.K. and S.K.S.; supervision, S.K.S. and N.D.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The permanent link to the code/repository used for service quality prediction in this study is publicly available at the following link. Repository name: Mendeley Data. https://data.mendeley.com/datasets/h489xpc53b/2 (accessed on 21 January 2025).

**Acknowledgments:** The authors acknowledge Rohan Das, Consultant in E&Y, for his help in the generation of AWS outputs including pipelines.

**Conflicts of Interest:** The author Avisek Kundu is employed by Ernst & Young LLP. The author Nitesh Dhar Badgayan is employed by the company KPMG, India. 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. The other authors declare that they have no conflicts of interest.

## References

1. Rane, N.L.; Mallick, S.K.; Kaya, O.; Rane, J. Tools and frameworks for machine learning and deep learning: A review. In *Applied Machine Learning and Deep Learning: Architectures and Techniques*; Deep Science Publishing: London, UK, 2024; pp. 80–95.
2. Salehin, I.; Islam, M.S.; Saha, P.; Noman, S.M.; Tuni, A.; Hasan, M.M.; Baten, M.A. AutoML: A systematic review on automated machine learning with neural architecture search. *J. Inf. Intell.* **2024**, *2*, 52–81. [CrossRef]
3. Liu, B.; Sun, Y.; Gao, L. Enhancing Groundwater Recharge Prediction: A Feature Selection-Based Deep Forest Model With Bayesian Optimisation. *Hydrol. Process.* **2024**, *38*, e15309. [CrossRef]
4. Singh, P. Systematic review of data-centric approaches in artificial intelligence and machine learning. *Data Sci. Manag.* **2023**, *6*, 144–157. [CrossRef]
5. Choi, W.; Choi, T.; Heo, S. A Comparative Study of Automated Machine Learning Platforms for Exercise Anthropometry-Based Typology Analysis: Performance Evaluation of AWS SageMaker, GCP VertexAI, and MS Azure. *Bioengineering* **2023**, *10*, 891. [CrossRef]
6. Chang, W.; Wang, X.; Yang, J.; Qin, T. An improved CatBoost-based classification model for ecological suitability of blueberries. *Sensors* **2023**, *23*, 1811. [CrossRef]
7. Gazi, M.A.I.; Islam, M.A.; Masud, A.A.; Senathirajah, A.R.B.S.; Biswas, S.; Shuvro, R.A. The moderating impacts of COVID-19 fear on hotel service quality and tourist satisfaction: Evidence from a developing country. *Cogent Soc. Sci.* **2024**, *10*, 2331079. [CrossRef]
8. Sahin, A.; Imamoglu, G.; Murat, M.; Ayyildiz, E. A holistic decision-making approach to assessing service quality in higher education institutions. *Socio-Econ. Plan. Sci.* **2024**, *92*, 101812. [CrossRef]
9. Nur, H.R. The Competence in the Digital Era in Improving Public Service Performance. *Entrep. J. Bisnis Manaj. Kewirausahaan* **2024**, *5*, 61–72. [CrossRef]
10. Khan, A.; Talukder, M.S.; Islam, Q.T.; Islam, A.N. The impact of business analytics capabilities on innovation, information quality, agility and firm performance: The moderating role of industry dynamism. *VINE J. Inf. Knowl. Manag. Syst.* **2024**, *54*, 1124–1152. [CrossRef]
11. Elaigwu, M.; Abdulmalik, S.O.; Talab, H.R. Corporate integrity, external assurance and sustainability reporting quality: Evidence from the Malaysian public listed companies. *Asia-Pac. J. Bus. Adm.* **2024**, *16*, 410–440. [CrossRef]
12. Tuan, A.; Corciolani, M.; Giuliani, E. Being reassuring about the past while promising a better future: How companies frame temporal focus in social responsibility reporting. *Bus. Soc.* **2024**, *63*, 626–667. [CrossRef]
13. Ismail, A.; Bakri, M.H.; Rusli, N.B.; Bakar, M.A.B.A.; Othman, H. Relationship between Service Quality and Customer Satisfaction: A Systematic Literature Review. *Resmilitaris* **2023**, *13*, 262–281.
14. Kim, J.J.; Lee, Y.; Han, H. Exploring competitive hotel selection attributes among guests: An importance-performance analysis. *J. Travel Tour. Mark.* **2019**, *36*, 998–1011. [CrossRef]
15. Parasuraman, A.; Berry, L.L.; Zeithaml, V.A. Perceived service quality as acustomer-based performance measure: An empirical examination of organizational barriers using an extended service quality model. *Hum. Resour. Manag.* **1991**, *30*, 335–364. [CrossRef]
16. Park, J.; Lee, B.K. An opinion-driven decision-support framework for benchmarking hotel service. *Omega* **2021**, *103*, 102415. [CrossRef]
17. German, J.D.; Redi, A.A.N.P.; Prasetyo, Y.T.; Persada, S.F.; Ong, A.K.S.; Young, M.N.; Nadlifatin, R. Choosing a package carrier during COVID-19 pandemic: An integration of pro-environmental planned behavior (PEPB) theory and Service Quality (SERVQUAL). *J. Clean. Prod.* **2022**, *346*, 131123. [CrossRef]

18. Lizarelli, F.L.; Osiro, L.; Ganga, G.M.; Mendes, G.H.; Paz, G.R. Integration of SERVQUAL, Analytical Kano, and QFD using fuzzy approaches to support improvement decisions in an entrepreneurial education service. *Appl. Soft Comput.* **2021**, *112*, 107786. [CrossRef]
19. Adler, A.I.; Painsky, A. Feature importance in gradient boosting trees with cross-validation feature selection. *Entropy* **2022**, *24*, 687. [CrossRef]
20. Rivero, D.M.; Suson, R.; Arnejo, A.; Atibing, N.M.; Aro, J.L.; Wenceslao, C.; Burdeos, A.; Yamagishi, K.; Ocampo, L. Service recovery and innovation on customer satisfaction amidst massive typhoon-induced disruptions: The mediating role of SERVQUAL. *Int. J. Disaster Risk Reduct.* **2023**, *99*, 104130. [CrossRef]
21. Stefano, N.M.; CasarottoFilho, N.; Barichello, R.; Sohn, A.P. A fuzzy SERVQUAL based method for evaluated of service quality in the hotel industry. *Procedia CIRP* **2015**, *30*, 433–438. [CrossRef]
22. Rosário, A.T.; Boechat, A.C. How Automated Machine Learning Can Improve Business. *Appl. Sci.* **2024**, *14*, 8749. [CrossRef]
23. Paladino, L.M.; Hughes, A.; Perera, A.; Topsakal, O.; Akinci, T.C. Evaluating the performance of automated machine learning (AutoML) tools for heart disease diagnosis and prediction. *AI* **2023**, *4*, 1036–1058. [CrossRef]
24. Abdulrab, M.; Hezam, N. Service Quality and Customer Satisfaction in the Hospitality Sector: A paper review and future research directions. *Libr. Prog. Int.* **2024**, *44*, 7486–7503.
25. Kanaparthi, V. Transformational application of Artificial Intelligence and Machine learning in Financial Technologies and Financial services: A bibliometric review. *arXiv* **2024**, arXiv:2401.15710. [CrossRef]
26. Pradhan, S.; Sahu, S.K.; Pramanik, J.; Badgayan, N.D. An insight into mechanical & thermal properties of shape memory polymer reinforced with nanofillers; A critical review. *Mater. Today Proc.* **2022**, *50*, 1107–1112.
27. Sahu, S.K.; Sreekanth, P.R. Artificial neural network for prediction of mechanical properties of HDPE based nanodiamond nanocomposite. *Polymer* **2022**, *46*, 614–620.


**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.

computers

MDPI

Review

# Electromagnetic Field Distribution Mapping: A Taxonomy and Comprehensive Review of Computational and Machine Learning Methods

**Yiannis Kiouvrekis [1,2,3,*] and Theodor Panagiotakopoulos [4]**

[1] Mathematics, Computer Science and Artificial Intelligence Lab, Faculty of Public and One Health, University of Thessaly, 43100 Karditsa, Greece
[2] Department of Information Technologies, University of Limassol, Limassol 3036, Cyprus
[3] Business School, University of Nicosia, Nicosia 2115, Cyprus
[4] Department of Management Science and Technology, University of Patras, 26334 Patras, Greece; tpanagiotakop@upatras.gr
* Correspondence: yiannis.kiouvrekis@gmail.com or kiouvrekis.y@uth.gr; Tel.: +30-6944347515

**Abstract:** Electromagnetic field (EMF) exposure mapping is increasingly important for ensuring compliance with safety regulations, supporting the deployment of next-generation wireless networks, and addressing public health concerns. While numerous surveys have addressed specific aspects of radio propagation or radio environment maps, a comprehensive and unified overview of EMF mapping methodologies has been lacking. This review bridges that gap by systematically analyzing computational, geospatial, and machine learning approaches used for EMF exposure mapping across both wireless communication engineering and public health domains. A novel taxonomy is introduced to clarify overlapping terminology—encompassing radio maps, radio environment maps, and EMF exposure maps—and to classify construction methods, including analytical models, model-based interpolation, and data-driven learning techniques. In addition, the review highlights domain-specific challenges such as indoor versus outdoor mapping, data sparsity, and model generalization, while identifying emerging opportunities in hybrid modeling, big data integration, and explainable AI. By combining perspectives from communication engineering and public health, this work provides a broader and more interdisciplinary synthesis than previous surveys, offering a structured reference and roadmap for advancing robust, scalable, and socially relevant EMF mapping frameworks.

**Keywords:** radio map; path loss function; electromagnetic exposure map

## 1. Introduction

During the past few decades, technological advancements have made electronic and wireless personal communication devices, such as mobile phones, Wi-Fi, and Bluetooth devices, essential to daily life. These devices, which emit non-ionizing electromagnetic radiation, are now almost universally used, not only for personal purposes, but also in public and professional settings [1]. Consequently, there has been a notable surge in the adoption of wireless communication technologies, particularly mobile networks like 5G, which offer extensive coverage throughout the city [2]. This widespread adoption has been driven in part by the development of modern radio communication networks, such as 5G, which offer significantly faster data transfer speeds and greater network capacity compared to previous generations. In addition, the use of devices that generate electromagnetic fields,

such as security scanners, smart meters, and medical equipment, has increased. Together, these factors have contributed to the increase in human-generated electromagnetic radiation in the environment [3]. As wireless networks expand rapidly, accurate estimation of signal propagation has become essential [4]. The effective prediction of signal behavior is critical in multiple scientific domains [5].

In wireless communications and radio frequency (RF) engineering, estimating the path loss function and building radio environmental maps (REM) play a significant role in network optimization and smart city planning [6]. Such tools aid in strategic base station placement, especially for 5G and future cellular network generations [7]. A key challenge in RF engineering is determining the most accurate propagation model or achieving optimal path loss estimation [8]. These estimations are crucial for creating radio maps, cognitive radio networks, and power spectrum maps [9]. The practical application of radio maps in the communications industry includes network optimization, such as guiding the optimal placement of antennas or base stations. For example, a 5G provider could use radio maps to optimize urban coverage by visualizing how buildings and terrain affect signal propagation. Cognitive Radio Networks (CRNs) also benefit from such maps, dynamically adjusting to the environment to prevent interference, particularly in densely populated areas with crowded frequency bands [10].

In environmental science, the focus is on estimating electromagnetic field strength levels across specific geographic regions. Creating precise electromagnetic pollution maps is vital because they reveal how radio-frequency and low-frequency fields spread through urban and natural landscapes, help researchers quantify human and wildlife exposure, guide policymakers in setting evidence-based safety standards, and enable engineers to design greener, interference-free infrastructure [11]. The proliferation of wireless technologies has led to increased RF field density in human environments, raising health concerns about exposure to non-ionizing radiation [12]. This societal issue has prompted scientific efforts to measure and predict electric field strength levels in various settings, both indoor and outdoor. However, most studies have concentrated on direct measurements rather than predictive modeling [13,14]. Current methodologies often emphasize limited geographical areas or rely on conventional geospatial techniques rather than advanced machine learning approaches, highlighting the need for more sophisticated predictive tools in this field.

The International Commission on Non-Ionizing Radiation Protection (ICNIRP) provides guidelines on safe exposure levels to electric, magnetic, and electromagnetic fields. In the EU, member states follow ICNIRP-based regulations, such as Recommendation 1999/519/EC and Directive 2013/35/EU [15]. Despite RF-EMF being non-ionizing, public concern remains high—particularly regarding 5G—due to fears about increased exposure from higher frequencies and dense signal concentrations in urban areas [16]. Given the growing relevance of the issue, awareness raising efforts have intensified: numerous campaigns and initiatives aim to educate the public [17,18], various projects have been funded to expand scientific understanding [13,14,19], and measures to control EMF exposure in homes, public spaces and workplaces continue to be re-assessed in accordance with regulations established by public authorities [20,21].

In recent years, studies related to radio frequency electromagnetic fields (RF-EMFs) have focused mainly on empirical and modeling approaches to ensure safe exposure conditions and to verify that RF-EMF values comply with international and national regulations [13,14]. This surge in new research has come from the need for continuous evidence not only to demonstrate that exposure limits are not exceeded, but also to conduct epidemiological surveillance [22]. Across all sectors, the generation of electric field maps is

considered essential for monitoring radio signal behavior and assessing their application-specific implications.

*Contribution—Objectives*

The objective of this review is not only to provide a comprehensive and structured analysis of current methodologies used in the construction of electric field distribution mapping, but also to address the existing ambiguity and frequent overlap in terminology such as radio maps, path loss functions, electromagnetic exposure maps and electric field strength maps. This paper aims to clarify these concepts by establishing a consistent and unified definitional framework and reviewing all these methodologies.

Furthermore, the review introduces a detailed taxonomy that systematically categorizes the diverse electric field distribution mapping methodologies. This taxonomy forms the foundation for a critical and thorough examination of each methodological category, focusing particularly on computational techniques and the unique characteristics of different application environments, such as indoor versus outdoor scenarios. By structuring the field in this way, the review intends to provide clarity on the current methodological landscape, identify existing knowledge gaps, and offer actionable insights and direction for future research and practical implementations in EMF exposure assessment and modeling.

## 2. Fundamental Concepts and Background

### *2.1. Electric Field Distribution Terminology and Map Types*

The initial step in addressing this issue involves the establishment of a precise and rigorous mathematical framework. This approach not only resolves the inherent ambiguities in problem formulation, but also explicitly delineates the essential variables, the underlying assumptions, and the relevant constraints. A solid mathematical foundation improves precision and serves as a common language in disciplines such as signal processing, machine learning, and electromagnetics. In addition, it strengthens the credibility and transparency of the proposed solutions while supporting both theoretical advances and practical implementation at scale.

The primary challenge researchers face is predicting the level of electromagnetic field strength at specific points or within regions with varying topographical characteristics, without prior measurements at those locations. This task is commonly referred to as propagation modeling [23]. The goal of propagation modeling is to estimate the field or signal strength based on known parameters of the wireless system, such as transmission frequency, characteristics, antecedent heights, and other relevant factors [24]. The problem can be defined as follows:

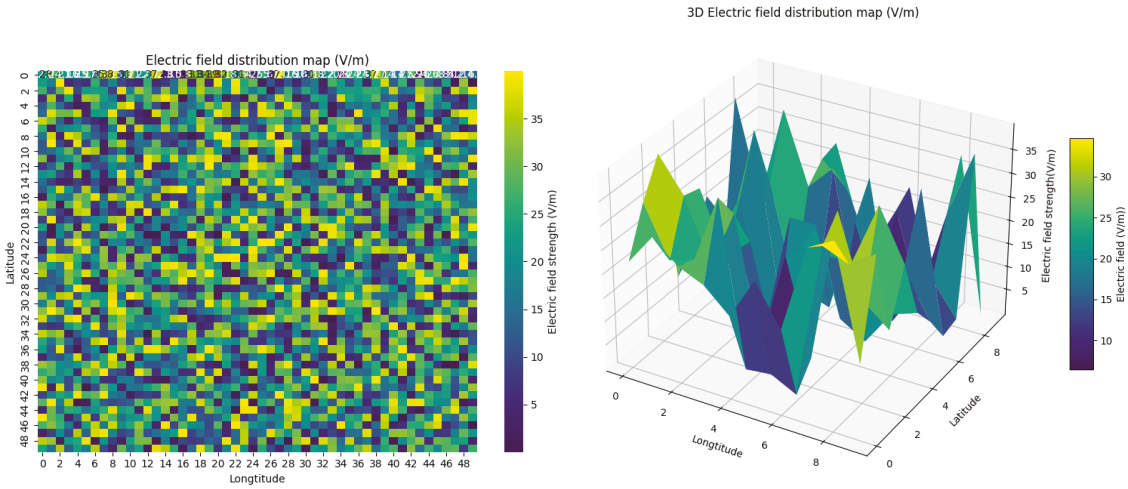
For $n$ discrete points $\vec{x}_i \in \mathbb{D} \subset \mathbb{R}^d$, where $i \in \{1, \ldots, n\} = [n]$, representing the locations in the $d$ dimensional study space $\mathbb{D}$, our aim is to estimate the electromagnetic field (EMF) $E$ at these points, which means calculating $E(\vec{x}_i)$, the corresponding electromagnetic field exposure value that we seek to predict. The objective is to estimate these theoretical values using an appropriate function

$$F : \mathbb{D} \times \mathbb{S} \to \mathbb{R} \tag{1}$$

where $\mathbb{S}$ is a $m$ dimensional space which consists of frequency and other characteristics, such that $F(\vec{x}_i; \vec{s}) = E(x_i)$ $(\vec{s} \in \mathbb{S})$, ensuring that $F$ respects the underlying surface conditions of the study space.

Visualization maps of mathematical interpolations in plane or in space can take various forms depending on the dimensionality and complexity of the data. These representations

may include 2D contour plots or 3D surface plots (Figure 1). Each type serves a specific purpose, allowing researchers to better understand the spatial distribution, gradients, or patterns emerging from the interpolation results. The choice of visualization often depends on the nature of the interpolated data and the insights one aims to extract from it.




(**a**) (**b**)

**Figure 1.** Example of electric field distribution maps: (**a**) Example of two dimensional electric field distribution map. (**b**) Example of three dimensional electric field distribution map.

Within this mathematical framework, various terms are encountered in the literature when referring to this problem. The literature on electric field distribution mapping reveals a striking variety of terms used to describe conceptually similar constructs. Among these are Electric Field Strength Map, Electromagnetic Exposure Map, Electromagnetic Field Exposure Map, Electromagnetic Field Magnitude, Electromagnetic Field Map, and Electromagnetic Pollution Map, all referring to spatial representations of field intensity or exposure. Additionally, terms such as EMF Exposure Map, Path Loss Map, Power Spectrum Map, and Radio Environment Map (REM) are commonly encountered, reflecting variations in focus or application domain. Furthermore, the terminology extends to Radio Map, Radio Map–Path Loss, Radio Network Map, Radio Propagation Map, and Radio Signal Strength Field. All these different terms, which often overlap in meaning and usage, are summarized in Tables 1 and 2. This diversity underscores a lack of standardization in the field, often leading to confusion and inconsistency across studies, and highlights the need for a unified definitional framework.

### 2.1.1. Radio Map and Spectrum Cartography

Radio maps are primarily designed to visualize signal characteristics across space, such as received signal strength (RSS), signal-to-noise ratio (SNR), interference power, power spectral density (PSD), electromagnetic absorption, channel gain, and other related metrics. These maps serve as a fundamental tool within the domain of wireless communication, supporting network planning, optimization, and performance analysis. The typical output format is a heat map or a spatial grid of signal values, with common units including dBm, RSSI, and SNR. Radio maps generally provide medium to high spatial resolution and may be constructed either as static representations or time-averaged snapshots, depending on the specific application requirements. Data sources used to generate radio maps include sensor-based measurements, simulation data, and mobile user reports. These maps are used predominantly by engineers, network planners, and system designers to evaluate coverage, identify dead zones, and improve service quality in

wireless networks [25]. Spectrum cartography encompasses a comprehensive suite of techniques designed to construct and update radio maps, offering detailed insights into the radio frequency environment. These techniques capture and represent key metrics such as the power of the received signal, the intensity of interference, the spectral density of power, the electromagnetic absorption, and the gain of the channel across geographic areas (Table 1). In practice, spectrum cartography often involves constructing multidomain RF maps spanning dimensions such as frequency, space, and time—using only sparse and incomplete measurement data [9,26–28].

**Table 1.** Comparison between the terms "Radio maps" and "Spectrum cartography".

| Feature | Radio Maps | Spectrum Cartography |
|---|---|---|
| **Definition** | Static maps of radio signal features | The technique to estimate & construct maps |
| **Focus** | The map itself | The process of mapping |

2.1.2. Radio Environment Map

The concept of a radio map broadly refers to a spatial representation of the power spectral density (PSD) of wireless signals across geographic, frequency, and temporal dimensions. A more advanced form, the Radio Environment Map, integrates diverse data sources—including geographic information, radio device specifications, propagation models, and real-time spectrum data—to provide a comprehensive cognitive map of the wireless environment. In addition, spatial spectrum databases serve as practical implementations of these concepts, enabling global awareness of spectrum usage. These systems play a pivotal role in applications such as spectrum monitoring, dynamic spectrum access, interference management, and support for adaptive mobile network architectures [29].

A REM aims to capture spectrum usage and contextual environmental information in dynamic and often complex radio environments. It is a core element in cognitive radio systems and dynamic spectrum access frameworks, where understanding the temporal and spatial behavior of spectrum occupancy is essential. REMs are typically structured as databases or spatial maps containing RF metadata, with key metrics such as Signal-to-Interference-plus-Noise Ratio (SINR), occupancy probability, and interference level [30]. These maps maintain a medium to high spatial resolution and are designed to be both dynamic and time-aware, reflecting real-time changes in spectrum activity. Data for REMs are collected through spectrum sensing and network-derived metadata. They are mainly used by cognitive radio agents and spectrum regulators for managing interference, guiding adaptive transmission strategies, and supporting spectrum policy enforcement [25].

2.1.3. Electromagnetic Field Exposure Map

Here, we enter a different field—one at the intersection of public health, medical physics, and radiation protection. In this context, the maps are referred to as electromagnetic field exposure maps. Given the widespread integration of wireless communication systems into modern daily life, the monitoring of associated phenomena, particularly radiofrequency electromagnetic field (RF-EMF) exposure, has become increasingly important. In urban environments, a diverse array of EMF sources, including WiFi networks and cellular technologies such as 5G systems, collectively shape the electromagnetic landscape, necessitating continuous surveillance and assessment.

An EMF Exposure Map is constructed to evaluate and visualize human exposure to electromagnetic field (EMF) radiation. These maps play a vital role in public health assessment and regulatory compliance by providing spatial representations of EMF intensity

levels. The application domain includes studies on environmental health, urban planning, and electromagnetic compatibility. EMF exposure maps present their output as detailed spatial intensity distributions, typically using physical units such as volts per meter (V/m), amperes per meter (A/m), or watts per square meter (W/m$^2$). They are known for offering high spatial resolution, particularly in proximity to transmission sources where exposure gradients can be steep. These maps can be generated in real-time or through periodic sampling, using data derived from EMF meters as well as modeled or measured values. They are widely used by health agencies, municipalities, urban developers and academic researchers to monitor exposure levels, inform the public, and support evidence-based regulatory decisions. Generating such maps typically involves extensive and costly measurement campaigns. To optimize these efforts and enhance the representativeness of the resulting exposure maps, systematic criteria must be established for the selection of measurement locations.

The comparative Table 2 highlights that although Radio Maps, REMs, and EMF Exposure Maps all involve spatial representation of electromagnetic phenomena, they serve different purposes across different domains, using different metrics, data sources, and user audiences. Understanding these differences is essential for selecting the appropriate mapping approach based on the intended application, whether for network optimization, dynamic spectrum management, or public health assessment (See Figure 2).

**Table 2.** Comparison of Radio Map, REM, and EMF Exposure Map: Commonalities and Differences.

| Feature | Radio Map | REM (Radio Environment Map) | EMF Exposure Map |
|---|---|---|---|
| **Primary Purpose** | Visualize signal characteristics (e.g., RSS, SNR) | Capture spectrum usage and environment context | Assess human exposure to EMF radiation |
| **Domain** | Wireless communications | Wireless communications | Public health, EMF compliance |
| **Output Type** | Heatmap or grid of signal values | Database or map with RF metadata | Spatial exposure map with EMF intensity |
| **Units/Metric** | dBm, RSSI, SNR | SINR, occupancy probability, interference level | V/m, A/m, W/m$^2$ |
| **Spatial Resolution** | Medium to high | Medium to high | High (especially near transmitters) |
| **Temporal Aspect** | May be static or time-averaged | Often dynamic and time-aware | Can be real-time or periodic |
| **Sources of Data** | Sensor measurements, simulation, mobile data | Spectrum sensing, network metadata | EMF meters, modeled or measured values |
| **Used By** | Network planners, engineers | Cognitive radio agents, regulators | Health agencies, municipalities, researchers |

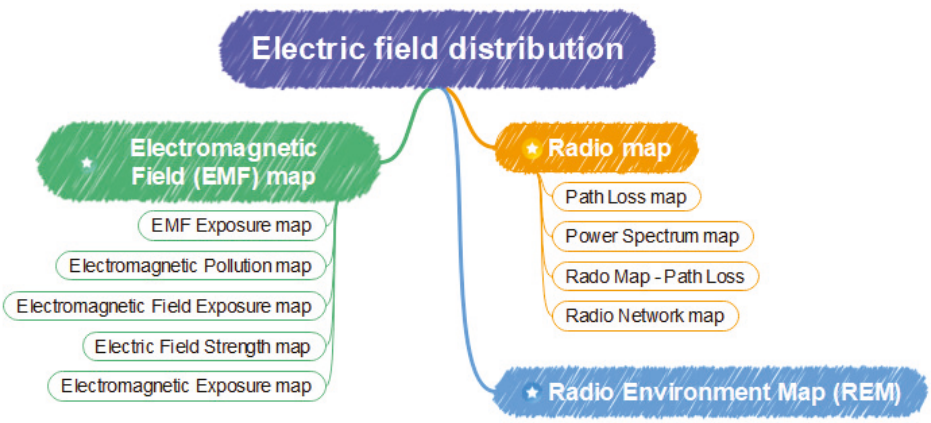



**Figure 2.** Conceptual mind map of the terminology used in the literature for spatial representations of electric-field distribution.

### *2.2. Outdoor and Indoor Maps*

To continue addressing this complexity, we introduce the terms indoor and outdoor as a first key contextual categories within our broader analysis. These terms reflect the environmental settings in which EMF mapping techniques are applied and serve as fundamental dimensions for interpreting methodological differences. Radio mapping and electromagnetic field exposure mapping varies significantly depending on the environmental context in which it is conducted. The propagation characteristics of electromagnetic signals, the complexity of geograchical background, and the availability of measurements all differ between indoor and outdoor environments (Figure 3). A clear distinction between these contexts is essential for understanding the suitability of different computational and machine learning methods (Table 2).

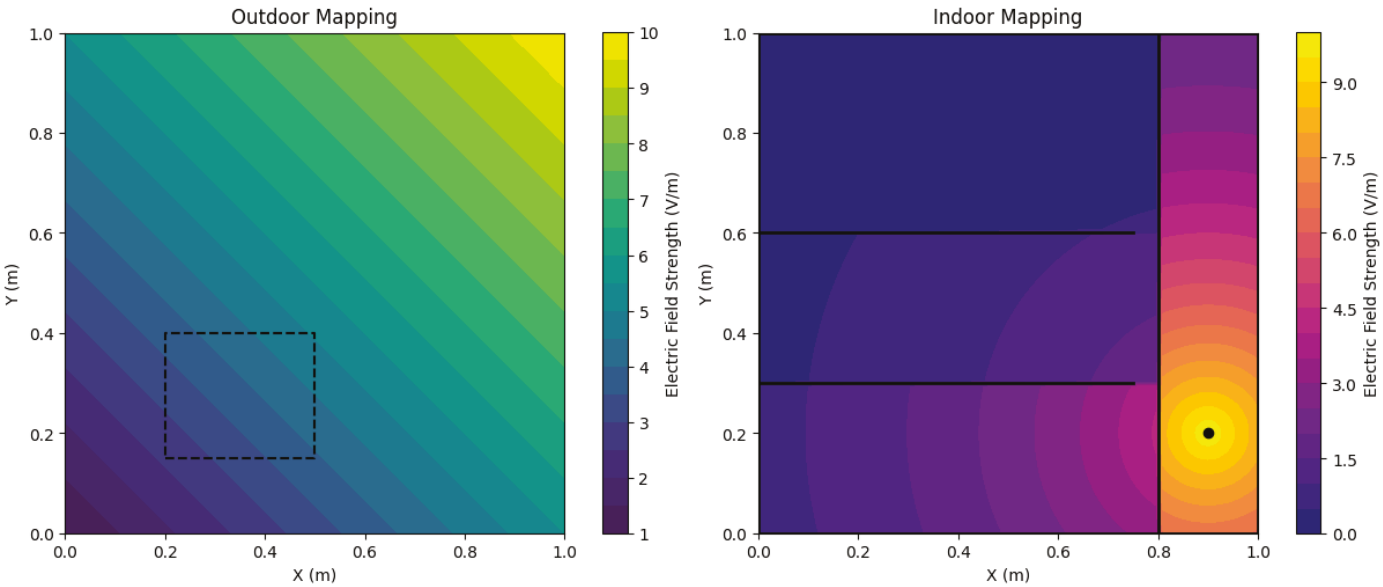



**Figure 3.** Spatial Interpolation of Electromagnetic Field Exposure in Outdoor and Indoor Environments.

#### 2.2.1. Outdoor Environment

Outdoor radio mapping or EMF exposure mapping typically involves wide-area scenarios such as urban, suburban, or rural landscapes. The propagation of electromagnetic waves in outdoor environments is influenced by large-scale structures such as buildings, terrain topology and atmospheric conditions. One of the key challenges in outdoor environments is the sparsity of data due to the vast area and the cost of measurement campaigns. This has led to an increased reliance on spatial interpolation and predictive modeling.

#### 2.2.2. Indoor Environment

In contrast, indoor radio mapping or EMF exposure mapping deals with environments where signal propagation is significantly affected by walls, furniture, floors, and other interior elements. The presence of multipath effects, diffraction, and scattering leads to highly complex and variable EMF distributions. This makes traditional modeling approaches less effective unless detailed structural data and advanced ray-tracing techniques are employed. The indoor setting allows for more controlled data collection and finer granularity in exposure mapping.

This contextual differentiation not only helps structure the discussion but also highlights how environmental characteristics influence the selection of computational methods,

the modeling assumptions adopted, and the practical challenges encountered in deploying electric field distribution mapping solutions.

### 2.3. Data Acquisition Methods

#### 2.3.1. In Situ Measurements

In this method, technicians use handheld electromagnetic field metersesearchers to manually measure the field strength at specific locations. These devices include spectrum analyzers with probes or EMF meters. The process is straightforward and cost-effective, making it suitable for small-scale or preliminary surveys. However, it is time-consuming and provides low spatial resolution unless measurements are taken at a high density across the area [31,32].

#### 2.3.2. Wireless Sensor Networks

This method involves the deployment of distributed sensor arrays installed throughout an area. These stationary sensors continuously monitor the electromagnetic field strength over time. The key advantage of this approach lies in its ability to provide long-term and real-time monitoring, which makes it highly suitable for urban environments. Nevertheless, the initial setup cost can be significant, and data is limited to the locations where the sensors are installed [33,34].

#### 2.3.3. Simulation and Model-Based Data—Synthetic Data

Rather than direct field measurements, this method uses computational electromagnetic tools such as ray tracing to predict the distribution of electromagnetic fields. These models are based on known source parameters and environmental configurations. The primary benefit is the ability to estimate field values in inaccessible areas. However, the precision of the simulation depends heavily on the quality of the input data and the degree to which the model is calibrated with real world measurements [35].

### 2.4. Data Sets

Measurement data is available from the Department of Electronic Communications [36], operating under the Ministry of Research, Innovation, and Digital Policy of Cyprus. The dataset includes measurements from 2017 to the present, with two measurement campaigns conducted annually—one per semester. The methodology involves performing three distinct measurements around each telecommunications site, specifically at locations where electromagnetic radiation exposure is expected to be highest. These points are typically situated at elevated positions near the base stations and aligned with the main signal emission axis. Measurements are taken at different distances to ensure thorough spatial coverage.

Another open data set is provided by the Agence Nationale des Fréquences (ANFR) [37], the French government agency responsible for managing national radio frequencies. ANFR operates accredited laboratories to measure public exposure to radio waves from cell towers and personal devices such as mobile phones. These measurements support public inquiries and contribute to the development of national and international exposure standards. Cartoradio, ANFR's online mapping tool, displays the locations of relay antennas and electromagnetic field measurements. Users can search by address or area and access detailed measurement reports. Since October 2021, Cartoradio also offers statistics on radio sites, including 5G deployment by region, operator, and frequency band.

Bakirtzis et al. [38] released a comprehensive data set of radio maps generated through ray tracing simulations in indoor environments of varying structural complexity, across

multiple frequency bands, and under various antenna radiation pattern assumptions. According to the authors, leveraging this dataset facilitates the development of fully data-driven propagation models capable of generalizing across unseen building layouts, frequency ranges, and antenna configurations—thereby offering a viable path toward replacing traditional radio signal propagation modeling techniques.

Another open-access dataset, known as RadioMapSeer [39,40], is a publicly available collection of radio map datasets generated for dense urban environments. It includes simulated data for path loss, received signal strength (RSS), and time of arrival (ToA) across a wide range of realistic urban scenarios based on actual city maps. The dataset is provided in two configurations: (1) a 2D setting, where transmitters (Tx) are deployed at street level, 1.5 m above ground—matching the height of the 256 × 256 receiver pixel grid; and (2) an extended 3D setting, which incorporates variable building heights and rooftop transmitter deployments. Another dataset creation effort, described in [41], is known as RSRPSet. In this approach, a 320-m by 320-m area is defined based on the location and orientation of the base station's antenna to represent its effective coverage. This area is then divided into a grid, where each cell corresponds to a 5-m by 5-m region. Similar, the data set, called High-Resolution Radio Environment Map Dataset [42] includes two measured REMs of a corridor in an indoor office environment. These high-resolution maps were generated using a semi-automated setup, where an Automated Guided Vehicle (AGV) equipped with a transmitter followed a predefined path.

### *2.5. Related Work*

Although the literature contains several review articles, they tend to address only narrow sub-fields. Abdulkarim et al. [43] review path loss models in the context of machine learning-based algorithms, with a particular focus on models developed between 2000 and 2021. The study examines network parameters, architectural designs, and the applicability of these models, while also identifying key trends and proposing directions for future research in this evolving field. Li et al. [44], for instance, focus on the concept of radio environment map (REM), summarize recent research progress, and highlight the key challenges involved in constructing accurate spectrum patterns. Phillips et al. [35] present a comprehensive and up-to-date survey of path loss prediction methods, covering over six decades of continuous research in the field. They introduce a novel taxonomy to systematically compare and contrast the wide range of existing approaches, offering a concise yet thorough overview of the methods. In addition, the authors provide valuable insights into emerging trends and outline promising directions for future research. Feng et al. [45] offer a systematic taxonomy of radio map estimation (RME) techniques based on their dependence on prior model knowledge, distinguishing three families, model-driven, data-driven, and hybrid approaches. For each category, they survey representative studies, analyzing strengths and limitations in detail. The authors also compile publicly available radio map datasets and emphasize the importance of mixed simulation-and-measurement corpora for advancing RME research. Dare et al. [46] provide a comprehensive review of the Radio Environment Map (REM) concept, detailing the various layers of information it encompasses, the techniques used for its construction, and key contributions from the existing literature. Romero et al. [9] begin by introducing several representative applications of radio maps, then proceed to examine the most prominent radio map estimation (RME) methods. The discussion progresses from basic regression techniques to more advanced and state-of-the-art algorithms, offering a structured view of the methodological landscape. To aid in understanding this diverse set of tools, the authors also include illustrative toy examples that provide intuitive insight into the functioning of each approach. Kurt et al. [47]

present an overview of experimentally validated propagation models for wireless sensor networks (WSNs), offering quantitative comparisons of different models used in WSN research. Their study evaluates these models under various deployment scenarios and frequency bands, providing valuable insights into their relative performance and applicability in real-world WSN environments.

Abbreviations for Table 3: MBML—Model-Based Machine Learning (data-driven predictors that incorporate physical priors); MBI—Model-Based Interpolation (classical spatial-interpolation techniques such as Kriging or IDW guided by propagation models); AM—Analytical Models (closed-form or semi-empirical formulas derived directly from electromagnetic theory).

**Table 3.** Comparison of relevant survey papers.

| References | Taxonomy Domain | Map Category | Map Construction Method | Dataset Categories |
|---|---|---|---|---|
| [43] | Wireless Communication | - | MBML | - |
| [44] | Wireless Communication | Radio Environment Map | - | - |
| [35] | Wireless Communication | Radio Environment Map | MBI + AM | - |
| [45] | Wireless Communication | Radio Map | MBI + MBML + AM | X |
| [46] | Wireless Communication | Radio Environment Map | MBI + MBML | - |
| [9] | Wireless Communication | Radio Map | MBI + MBML + AM | - |
| [47] | Wireless Communication | - | AM | - |
| **Our paper** | Wireless Communication + Public Health | Radio Environment Map Radio Map EMF Exposure Map | MBI + MBML + AM | X |

Compared with previous surveys (Table 3), our review extends the scope of analysis beyond wireless communication to also address the public health domain through the consideration of EMF exposure maps. In addition to covering radio environment maps and radio maps, our paper integrates multiple map construction methods (MBI, MBML, and AM) and discusses dataset categories. This broader taxonomy and interdisciplinary perspective provide a more comprehensive overview of the field and highlight emerging research directions that were not covered in earlier works.

The search and selection process of the reviewed studies is summarized in Figure 4. The systematic search was conducted in the following electronic databases: IEEE Xplore, Scopus, Web of Science, and PubMed. The Boolean query used (adapted to the syntax of each database) was:

```
("radio map" OR "radio environment map" OR "REM" OR "spectrum cartography" OR
 "path loss map" OR "electromagnetic exposure map" OR "electromagnetic field map" OR
 "electric field strength map")
AND
("mapping" OR "estimation" OR "prediction" OR "machine learning" OR "computational model")
```

A total of N = 1216 records were initially identified from major databases (IEEE Xplore, Scopus, Web of Science, and PubMed). The inclusion criteria required that studies be peer-reviewed journal articles or conference papers, focusing explicitly on radio maps, radio environment maps, or electromagnetic field (EMF) exposure maps. Eligible works employed analytical models, model-based interpolation, or data-driven and machine learning approaches for field distribution mapping. Only studies published in English and providing methodological contributions—such as the construction, estimation, or validation of EMF or radio maps—were considered.

Conversely, exclusion criteria applied to papers not directly related to electromagnetic field distribution mapping, non peer-reviewed sources such as reports, theses, or white

papers, and publications in languages other than English. Articles without accessible full text were excluded, as were studies focusing exclusively on hardware design, antenna optimization, or unrelated RF applications without a mapping component.

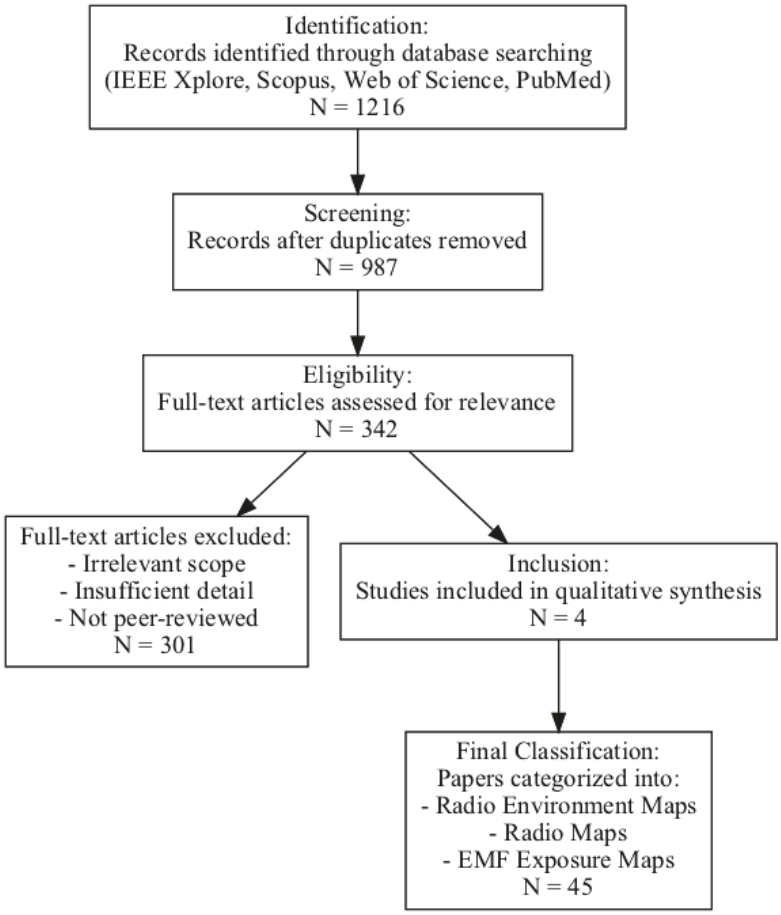



**Figure 4.** PRISMA-style flow diagram illustrating the search and selection process of the reviewed studies.

After removing duplicates, N = 987 records remained. Screening based on title and abstract resulted in N = 342 full-text articles assessed for eligibility. Following the application of inclusion and exclusion criteria, a final set of N = 45 studies was retained and categorized into three groups: radio environment maps, radio maps, and EMF exposure maps.

This section established a clear conceptual and mathematical framework for electromagnetic field distribution mapping, addressing the fragmented and overlapping terminology found in previous literature. By systematically distinguishing between radio maps, radio environment maps, and EMF exposure maps, as well as differentiating indoor and outdoor mapping contexts, our work provides a unified foundation for subsequent methodological analysis. Furthermore, we reviewed the main data acquisition techniques and open-access datasets, which are essential for developing, benchmarking, and validating mapping methods. Unlike earlier studies that focus narrowly on either wireless communication or public health, this integrated framework—encompassing terminology, environmental context, and datasets—bridges both domains, offering a structured basis for comparing approaches and guiding future research in a standardized and reproducible manner.

## 3. Electric Field Distribution Maps Construction Methods

### *3.1. Analytical Expressions for Radio Map Construction*

The concept of analytical expressions for radio map construction is based to the construction of the path loss function, which characterizes the attenuation of an electromagnetic wave as it propagates through space. Specifically, path loss denotes the reduction in wave power density experienced due to factors such as distance, obstacles, and environmental conditions [23]. We refer to it as the path loss function because, in the literature, these methods are specifically designed to address signal attenuation problems in wireless communication and are not typically applied in other domains. The models discussed here are a priori in nature, meaning that they rely solely on predefined parameters and theoretical assumptions, without incorporating direct measurement data in their predictions. The path loss function is a function $PL(d)$ that models the attenuation of a radio signal as a function of distance $d$ between the transmitter and receiver.

We begin with the Free-Space Propagation Model, which represents the simplest scenario of electromagnetic wave propagation under ideal line-of-sight conditions. It is governed by the Friis transmission equation [48], relating the received power $P_r$ to the transmitted power $P_t$, antenna gains $G_t$, $G_r$, wavelength $\lambda$, and the distance $r$ between antennas:

$$P_r = P_t \cdot G_t \cdot G_r \cdot \left(\frac{\lambda}{4\pi r}\right)^2 \tag{2}$$

Although useful for initial estimations and theoretical analysis, the free-space model neglects environmental effects such as reflection and diffraction, which are often significant in real deployments. To account for such effects, the *Two-Ray Reflection Model* extends the free-space formulation by including a reflected path from the ground. This model provides more realistic estimates for scenarios that involve large transmitter-receiver separation and flat reflective surfaces. It is particularly applicable to outdoor line-of-sight conditions, where ground reflection can substantially alter the signal strength. The complete mathematical formulation is provided in the Appendix A. Building on these deterministic approaches, empirical models such as the *Egli Model* offer practical alternatives based on measured data. Designed for point-to-point communication in urban and rural settings, the Egli model operates in the UHF and VHF bands (30 MHz to 3 GHz) and considers parameters such as distance, frequency, and antenna heights. It is especially useful when detailed environmental modeling is impractical. For its mathematical expression, see Appendix B. Finally, the *Hata Model*, an extension of the Okumura model [35], is one of the most widely adopted empirical models for outdoor mobile communication. Applicable to frequencies from 150 MHz to 1500 MHz, it provides correction factors for urban, suburban, and rural environments. Its reliability and simplicity have made it a standard in mobile network planning. Detailed equations and parameterizations are available in Appendix C.

Figure 5 consists of the flowchart illustrating the analytical approach, where electromagnetic field distributions are computed directly from input parameters (frequency, transmitter configuration, environment type) using propagation models, yielding grid-based EMF maps.

The authors in [49] modeled a single narrowband WiFi transmitter using the log-distance path loss model (LDPL). The study in [50] shows that satellite images of buildings and the adjacent spaces can be used effectively to estimate path loss model coefficients for signal propagation along straight, level urban streets. Using easily accessible environmental parameters, the proposed models enable accurate path loss predictions and support the development of efficient node placement strategies to ensure comprehensive urban coverage. The work presented in [51] offers a detailed analysis of large-scale path loss in

wireless networks operating under the IEEE 802.11 g standard [51] at 2.4 GHz, considering both indoor and outdoor environments. The authors propose a new propagation model and conduct a comparative evaluation against existing models. In addition, the study examines the impact of environmental factors, specifically temperature and relative humidity, on signal attenuation. Rizk et al. [52] investigate the application of diffraction theory and geometrical optics to model radio wave propagation in microcellular urban environments. The proposed model adopts a simplified two-dimensional (2D) representation, where building walls are modeled as line segments and buildings are assumed to be of infinite height. Despite this abstraction, the model is capable of handling arbitrary building layouts through an efficient algorithm, which is briefly outlined in the paper.

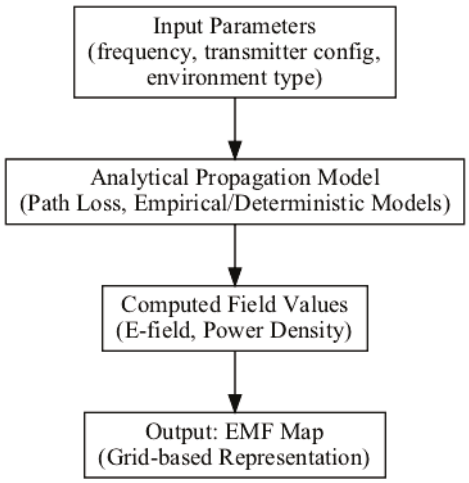



**Figure 5.** Analytical models for EMF mapping.

### *3.2. Model-Based Radio Map Estimation*

According to [25], a radio map is generally characterized by a function that maps a 3-tuple variable $(\mathbf{x}, f, t)$ to the signal power, denoted by $\Phi(\mathbf{x}, f, t)$, where $\mathbf{x} \in \mathbb{R}^3$ represents the location in a three-dimensional Cartesian coordinate system, and $f$ and $t$ denote the frequency point and time instant, respectively. This is a subcase of Equation (1), where $\mathbb{D} = \mathbb{R}^3$ and $\mathbb{S} = F \times T$, with $F$ denoting the frequency domain and $T$ the time domain.

This definition of a radio map builds upon the foundational framework introduced in [53], and it has been extended to encompass a broader range of scenarios. The radio map is modeled as a function of frequency $f$ and geographic location $\boldsymbol{c}$, and is expressed as:

$$\Phi(\boldsymbol{c}, f) = \sum_{i=1}^{n} g_i(\boldsymbol{c}, f)\Psi_i(f) \tag{3}$$

where $n$ is the number of transmitters, $g_i(\boldsymbol{c}, f)$ represents the channel power gain from the $i$-th transmitter, and $\Psi_i(f)$ denotes its power spectral density (PSD) [25]. The channel gain $g_i$ may be explicitly specified based on prior knowledge relevant to specific applications or estimated through sensor data.

In [54], the authors introduce a spline-based methodology for electromagnetic field estimation, utilizing a basis expansion model wherein the field is represented as a combination of known basis functions weighted by unknown coefficients. These coefficients are inferred from noisy field samples. The proposed spline-based framework is particularly motivated by applications in spectrum cartography, where a collection of cognitive radios cooperatively estimate the spatial and spectral distribution of RF power. Sato et al. [55] propose a novel method for radio map construction that performs simultaneous interpolation of received signal

power across both spatial and frequency domains. The approach leverages the observation that shadowing effects exhibit strong correlation over a wide frequency range. The key idea is to utilize shadowing values measured at various frequencies as proxies for those at the target frequency. Using real-world datasets collected at 870, 2115, and 3500 MHz within a cellular network, the authors demonstrate that their method can accurately generate radio maps, even when little or no data is available at the target frequency.

### *3.3. Model-Free Radio Map Reconstruction Interpolation-Geospatial Based Methods*

In contrast to model-based approaches, model-free methods do not rely on predefined propagation models, but instead use spatial neighborhood information. Broadly, model-free techniques can be classified into two categories: interpolation-based methods and learning-based methods, as outlined below.

Interpolation-based methods and machine learning techniques such as $k$-nearest neighbors ($k$-NN) approximate the electromagnetic field distribution at a given location by combining measurements from neighboring points. This can be mathematically expressed as:

$$\Phi(\boldsymbol{c}, f) = \sum_{i=1}^{N_s} w_i(\boldsymbol{c}, f) q_i(f) \tag{4}$$

where $q_i$ denotes the observation from the $i$-th sensor, and $w_i$ represents the associated combination weight [25,53]. A representative example is inverse distance weighted (IDW) interpolation, where the power gain is inversely proportional to the distance between the transmitter and receiver. This structure is mathematically analogous to the classical weighted mean $\bar{f} = \frac{\sum_{b \in \mathcal{B}} f(b) w(b)}{\sum_{b \in \mathcal{B}} w(b)}$, where $f(b)$ is a function defined over a set $\mathcal{B}$ and $w(b)$ is a non-negative weight function. In this context, the spatial weights $w_i$ in the radio map equation play the same role as the weight functions in interpolation theory, such as the Sibson or Laplace weight functions [23]. Beyond linear interpolation, more sophisticated alternatives include radio basis function (RBF) interpolation, where various kernels such as Gaussian Kriging, multiquadrics, or splines are employed. Among these, ordinary Kriging is a classical interpolation technique that optimizes the power spectral density (PSD) by minimizing the variance of estimation errors. Classical interpolation approaches typically assume that each observation influences the surrounding surface only within a certain distance.

Figure 6 consists of the flowchart of the interpolation approach, which uses sparse measurement samples and spatial interpolation methods (e.g., Kriging, IDW, Sibson) to estimate field values at unsampled locations and generate a continuous EMF map.

In [56], a hybrid approach is proposed that combines kriging-based spatial interpolation and recurrent neural networks for temporal prediction. This framework estimates spectrum occupancy probabilities using empirical measurements in cellular cognitive radio (CR) networks. The analysis in [57] evaluates spatial interpolation methods, such as Inverse Distance Weighting (IDW), for indoor radio information field (RIF) estimation. The study demonstrates robust and reliable performance in a prototype REM setting. In a case study conducted in Caracas, Venezuela, ref. [58] investigates interpolation methods for mapping EMF magnitudes. The authors compare SPLINE, IDW, and Kriging, finding IDW to offer the most accurate estimates across the pilot area. Three spatial interpolation techniques for estimating radio environments are compared in [59], with simulations performed for both indoor and outdoor scenarios. The results highlight trade-offs in accuracy and reliability among the methods. Finally, ref. [60] presents a geostatistical modeling approach applied to a 2.5 GHz WiMax network. The study evaluates robust techniques for radio environment and coverage mapping, validated through a case study at the University of Colorado,

Boulder. Interference cartography is introduced in [61] as a means to detect and utilize spectrum opportunities in secondary usage scenarios. The method combines geo-located measurements from various network entities and applies spatial interpolation to estimate interference levels across a target area. Finally, Guillén-Piña et al. [62] employ evolutionary computation to identify an optimal subset of measurement points for constructing high-quality electromagnetic field exposure maps—minimizing error with respect to a reference map—and demonstrate that accurate exposure maps can be achieved with significantly fewer samples. Panagiotakopoulos et al. [63] aim to determine the most effective geospatial interpolation method for constructing electromagnetic field strength maps in large-scale urban environments. The study evaluates five interpolation techniques: four based on Gaussian Process Regression (Kriging) and one using the classical inverse-distance weighted average of nearest neighbors. Using a dataset of 3632 EMF measurements collected across the city of Paris, the authors demonstrate that Kriging—particularly with the exponential variogram model—offers superior performance in capturing spatial autocorrelation compared to the nearest-neighbor approach. Additionally, the removal of outliers was found to significantly enhance the performance of all models, reducing both the mean squared error (MSE) and prediction variability. Also, Kiouvrekis et al. [64] compared five methods for mapping nationwide electric-field strength (30 MHz–6 GHz) using 3621 measurements over 9251 km$^2$. Gaussian-process (Kriging) models generally produce the most accurate spatial predictions, but when outliers are removed, a simple weighted nearest-neighbor approach performs almost as well. This suggests that both advanced Kriging and classical neighbor techniques can be effective, depending on data quality.

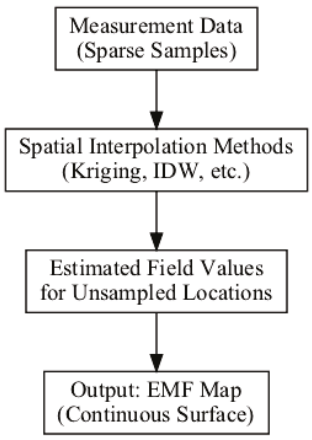



**Figure 6.** Model-based interpolation for EMF mapping.

In [65] the authors proposed a methodology for assessing electromagnetic-field exposure over large areas. It is demonstrated in an urban zone of 2.8 km$^2$ within a 35.11 km$^2$ municipality, with both rural and urban settings being considered. To limit the sampling density, the authors introduce line-of-sight criteria to identify relevant emitters. In the urban case study, they estimate that 8–10 measurement points per square kilometer are sufficient.

### *3.4. Model-Free Radio Map Reconstruction Machine Learning Based Methods*

In this case, the situation is more akin to a black-box modeling approach, where no explicit analytical expression for the electromagnetic field distribution is available. Instead, the mapping is learned directly from data using machine learning techniques. The radio map can therefore be expressed as:

$$\Phi(\boldsymbol{c}, f) = \boldsymbol{F}_{\text{blackbox}}(x_1, \ldots, x_{N_s}, f, t), \tag{5}$$

where $F_{\text{blackbox}}$ denotes a regression function approximated by a data-driven model, such as a neural network, random forest, or Gaussian process regression, trained on a set of sensor observations. Here, $x_i$ represents the $i$-th sensor location or measurement or characteristic vector, $f$ is the frequency, and $t$ is the time instance.

Instead of relying on predefined rules or neighborhood structures like model-based and interpolation approaches, black-box models directly infer complex spatial and spectral relationships from the data, proving useful in dynamic or hard-to-model settings.

Figure 7 consists of the flowchart showing the data-driven approach, where measurements and contextual features are used to train machine learning models (e.g., Random Forests, XGBoost, CNNs, GANs) that learn the mapping between features and field strength, enabling prediction of EMF maps for new locations and conditions.

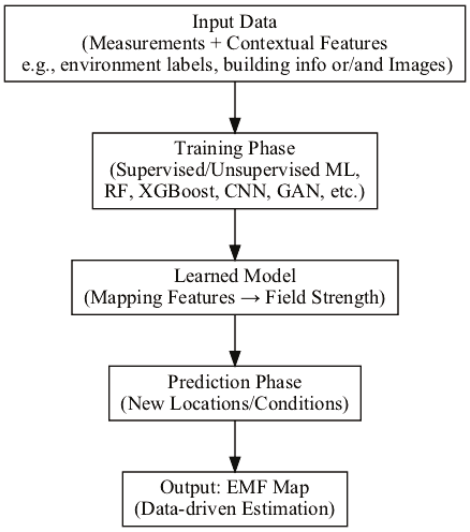



**Figure 7.** Machine learning approaches for EMF mapping.

The study in [66] introduces a disaggregated emitter radio map approach, where deep neural networks (DNNs) model the radio maps of individual emitters. This framework mitigates learning and generalization challenges and incorporates a two-stage spectrum cartography (SC) method that combines non-negative matrix factorization and an optimized iterative algorithm. Theoretical contributions address recoverability, sample complexity, and robustness to noise, properties that are not often formally established in the completion of the deep learning-based radio tensor. In [67], a limited feedback-based SC framework is proposed, which incorporates a deep generative adversarial network (GAN) based on deep feedback to mitigate the effects of heavy shadowing. Unlike conventional real-valued feedback strategies, the authors adopt a random quantization technique with a maximum likelihood estimation (MLE) criterion, theoretically proving radio map recoverability under certain conditions and validating performance through simulations.

Chaves-Villota et al. [68] introduced DeepREM, a framework comprising two deep learning models—U-Net and cGAN—designed to estimate radio environment maps (REMs) from sparse measurement data, without the need for any supplementary information. In [53], a two-phase learning framework for radio map estimation (RME) is developed by combining a conventional propagation model with a conditional GAN (cGAN). The method leverages global propagation patterns and local shadowing features to optimize the generative model, demonstrating improved accuracy in outdoor environments with scarce data.

The work in [69] presents a spectrum cartography algorithm to estimate power spectrum distributions in wide frequency bands. The algorithm is based on a nonparametric

regression framework, leading to a support vector machine (SVM) formulation. Simulations show that the method yields accurate maps using a simplified sensing architecture with reduced feedback. In [70], DNNs are used to predict the strength of radio signals in urban environments. The model incorporates sampled signal strengths and 3D environmental data to account for wave scattering effects, enhancing spatial prediction accuracy. The authors of [6] propose a deep learning model, RadioUNet, to estimate the propagation path loss between arbitrary transmitter-receiver pairs. The model is trained on a simulation dataset and significantly reduces computation time compared to traditional simulation tools, maintaining high accuracy. To address the challenges in cognitive radio networks, Ref. [71] presents MEGANs, a GAN-based algorithm to estimate power spectrum maps (PSMs). Unlike traditional approaches that rely on predefined propagation assumptions, MEGANs learns propagation features directly, yielding more accurate spectrum estimations.

The paper in [72] introduces a deep-complete autoencoder architecture to estimate radio occupancy maps with fewer measurements by learning spatial propagation characteristics. This is the first reported use of DNNs in this context, demonstrating substantial measurement reduction while maintaining accuracy. In [73], a hybrid algorithm is proposed that combines estimators of deep neural networks and traditional models, addressing the trade-off between performance and data requirements. The study emphasizes the potential of combining learning-based and classical approaches to improve efficiency and accuracy. RecNet, a neural network designed to reconstruct fine-grained radio maps, is introduced in [74]. Imagining reconstruction as a super-resolution task, the model generates high-resolution maps from low-resolution data, reducing the error by more than 20% compared to recent algorithms. In [6], the authors propose RadioUNet, a deep neural network trained on simulation data to estimate path loss in urban environments. The model delivers high accuracy while significantly reducing computation time, making it suitable for real-time use.

In [75], the authors conducted field measurements in an urban environment to collect geographic and network-related data from receiving mobile equipment and to assess path loss for radio signals transmitted at 189.25 MHz and 479.25 MHz. Multiple artificial neural network (ANN) models were developed and trained, exploring different input parameter combinations, network architectures, activation functions, and learning algorithms. The aim was to accurately predict path loss values under varying conditions. Model performance was rigorously evaluated using several statistical indicators, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Standard Deviation (SD), and the Regression Coefficient (R), providing a comprehensive assessment of prediction accuracy. In [76], a federated learning model called RadioSRCNet is proposed, incorporating super-resolution techniques for prediction of pathloss. Users locally train the model while a Unmanned aerial vehicle (UAV) aggregates non-sensitive data. The approach balances resolution and computational overhead through an Arbitrary Position Prediction (APP) module.

A transfer learning framework is proposed in [77] to reconstruct radio maps for different antenna tilt configurations. By transferring knowledge between configurations, the model achieves high accuracy with minimal or no new measurements, outperforming traditional machine learning methods. The work in [78] evaluates artificial neural networks (ANNs) for the prediction of macrocell pathloss, comparing them against the ITU-R P.1546 and Okumura–Hata models. The results indicate that ANNs can outperform classical models in certain terrain conditions. A convolutional neural network (CNN)-based propagation prediction model is introduced in [79]. The paper details model architecture and input mapping, demonstrating improved accuracy in modeling complex propagation patterns. Deep learning techniques incorporating satellite imagery are used in [80] to model channel

characteristics. The proposed model is benchmarked against ray-tracing and stochastic approaches, with results showing competitive performance. In [81], the authors introduce FadeNet, a CNN-based model for large-scale fading prediction.

Using topographical features as input, FadeNet achieves root mean square error as low as 5.6 dB across millimeter-wave deployments in multiple cities. Kiouvrekis et al. [82] present an explainable machine learning framework for predicting electric field strength across urban, semi-urban, and rural areas in Cyprus. The model is trained on 6543 EMF measurements collected in 2023 from mobile and digital TV stations. The dataset includes geospatial and environmental features such as antenna distance, population density, urbanization level, and building characteristics. Several models—kNN, neural networks, decision trees, random forests, XGBoost, and LightGBM—were evaluated using a two-semester split. Explainable AI analysis identified antenna distance, building volume, and population density as key predictors of EMF intensity. In [83], the authors propose a novel framework for indoor radio map construction that integrates image sensing, 3D reconstruction, deep learning, and ray tracing. The approach models indoor environments using RGB-D images, representing structural information as a set of cuboids. By converting room materials and obstacles into vector data, the method enables accurate ray-tracing simulations based on detailed indoor geometry.

Zhang et al. [84] propose a deep learning framework for predicting electric field (E-field) levels in complex urban environments. The study begins by describing the measurement campaign and publicly available databases used to construct the training dataset, including a detailed explanation of how these datasets are processed and integrated to optimize their compatibility with Convolutional Neural Network (CNN)-based models. The authors then introduce the proposed model, ExposNet, and provide a comprehensive description of its network architecture and operational workflow. Nyakyi et al. [85] conducted measurements of power density levels at various locations in Dodoma, Dar es Salaam, and Mwanza—areas characterized by high population density—to assess RF-EMF exposure. To estimate the total power density at different sites, the authors developed predictive models based on Artificial Neural Networks (ANN) and Multiple Linear Regression (MLR), using data from multiple RF-EMF sources.

Xiangyu et al. [86] introduce DeepMap, a deep Gaussian process framework designed for indoor radio map construction and location estimation. The proposed method aims to enhance the accuracy of WiFi-based localization, which remains one of the most widely used positioning techniques due to the broad availability of WiFi infrastructure and its cost-effectiveness. In [79], the authors propose a radio propagation prediction model based on deep learning using convolutional neural networks (CNNs). The study provides a detailed explanation of the model and evaluates its performance by analyzing the influence of map-related input parameters on the CNN's predictive behavior. Seretis et al. [87] develop generalizable models for indoor radio wave propagation capable of predicting received signal strength in previously unseen environments. Their approach supports varying transmitter and receiver positions, as well as different operating frequencies. The study demonstrates that a convolutional neural network (CNN) can effectively learn the underlying physics of indoor propagation from ray-tracing simulations of a limited set of training geometries, enabling it to generalize to substantially different layouts.

In [88], the authors constructed a detailed electric field strength map of urban Paris by evaluating 410 machine learning models based on k-nearest neighbors, neural networks, and decision trees. This extensive analysis enabled the optimization of model performance for accurate EMF predictions. SHAP analysis revealed that building volume around antennas is the most influential feature in the kNN model, followed by population density. In [89], the authors propose a conditional generative adversarial network (cGAN)-based

approach to reconstruct electromagnetic field exposure maps using sparse sensor data in outdoor urban environments. The model learns to estimate propagation characteristics based on the surrounding topology. When compared with simple Kriging, the cGAN-based method demonstrated superior accuracy, highlighting its potential as a reliable solution for EMF exposure map reconstruction.

Kiouvrekis et al. [90] conducted a large-scale analysis involving 566 machine learning models across eight French cities, employing six core approaches: k-Nearest Neighbors, XGBoost, Random Forest, Neural Networks, Decision Trees, and Linear Regression. Ensemble methods like Random Forest and XGBoost outperformed other models in predictive accuracy, while simpler models such as Decision Trees and k-NN provided competitive yet slightly less precise results. Neural Networks showed potential for further refinement. All machine learning models significantly outperformed the linear regression baseline. SHAP analysis also revealed notable differences in feature importance rankings between tree-based methods and other model types. M. Mallik et al. [91] propose an algorithm for estimating electromagnetic field exposure maps using a U-Net-based convolutional neural network. The task is formulated as an image reconstruction problem, where each pixel represents received power intensity. The model learns indoor wireless propagation characteristics, accounting for varying positions of Wi-Fi access points. Results demonstrate that realistic indoor environments can be effectively modeled from data, enabling accurate power map estimation.

Also, Kalatzis et al [92] apply explainable-AI (XAI) techniques to the spectral analysis of electromagnetic-field (EMF) measurements collected across Cyprus in the 30 MHz–6 GHz range. The data set combines emissions from mobile telephony, radio broadcasting, and terrestrial digital television, covering FM radio (87.5–108 MHz); VHF and UHF television; and mobile bands at 700, 800, 900, 1800, 2100, 2600, and 3600 MHz—capturing both legacy and emerging wireless technologies. Six machine-learning algorithms (XGBoost, LightGBM, Random Forests, k-nearest neighbours, neural networks, and decision trees) were evaluated for predictive performance in each band. SHAP (Shapley Additive Explanations) was then used to quantify how spatial and demographic factors influence EMF intensity. Ensemble tree models, particularly Random Forests and LightGBM, consistently delivered the highest accuracy and interpretability. SHAP analysis revealed distinct exposure patterns tied to urban form, population density, and built-environment metrics. This XAI-based framework enables data-driven EMF-exposure mapping to support infrastructure planning, compliance monitoring, and public-health risk assessment.

The tables below (Tables 4–6) provide the taxonomy based on computational methods. Neural network architectures, including NN, and CNN, are widely used for spatial prediction and interpolation. Generative models such as GANs and cGANs are used for sparse data reconstruction and adversarial learning. Statistical methods such as kriging, inverse distance weighting (IDW), and spline interpolation remain fundamental in many model-based studies. Several works also employ ensemble learning strategies, combining models such as random forests, XGBoost, and k-nearest neighbors. Hybrid approaches, for instance those blending kriging with Gaussian models or neural networks, exemplify efforts to integrate physical knowledge with data-driven inference. The methodological diversity reflects the interdisciplinary nature of EMF mapping research, which brings together expertise from geostatistics, machine learning, and wireless communications.

**Table 4.** Summary of studies that employ closed-form analytical expressions for radio-map construction.

| Reference | Taxonomy Domain | Map Category | Construction Method–Models–Tools | Environment | Dismnsion of the Map | Units of Metric |
|---|---|---|---|---|---|---|
| [49] | Wireless Communication | Radio Map | Analytical expressions for radio map construction | Indoor | 2D | Mean Distance Error in meter |
| [50] | | | | | | RMSE in dB |
| [51] | | | | | | RMSE in dB |
| [52] | | | | Outdoor | | Mean error in dB |

**Table 5.** Summary of studies that employ model-free radio map based to interpolation geospatial methods for radio-map construction.

| Reference | Taxonomy Domain | Map Category | Construction Method–Model–Tools | Dataset Acquisition | Environment | Dismnsion of the Map | Units of Metric |
|---|---|---|---|---|---|---|---|
| [58] | Public Health | Map | Model-free radio map reconstruction interpolation geospatial based methods. | Measurement points | Outdoor | 2D | MAE and MSE in V/m |
| [62] | | Electromagnetic field exposure maps | | | | | Logarithm of sum of absolute differences in V/m |
| [63] | | | | | | | RMSE in V/m |
| [64] | | | | | | | RMSE in V/m |
| [65] | | | | | | | Absolute error in V/m |
| [57] | Wireless Communication | Radio environment map | Model-free radio map reconstruction interpolation geospatial based methods. | | Indoor | | MAE in dB |
| [59] | | | | | Outdoor and Indoor | | Relative Mean Absolute Error |
| [60] | | | | | Outdoor | | Mean RMSE in dB |
| [61] | | - | - | | - | | Mean error in dBm |

**Table 6.** Summary of studies that employ model-free radio map based machine learning methods for radio-map construction.

| Reference | Taxonomy Domain | Map Category | Construction Method–Models–Tools | Dataset Acquisition | Environment | Dismnsion of the Map | Units of Metric |
|---|---|---|---|---|---|---|---|
| [67] | Wireless Communication | Radio Map | Model-free radio map reconstruction machine learning based methods. | Measurement points | - | 2D | LNMSE |
| | | | | | | | **log-domain normalized mean square error** |
| [69] | | spectrum distribution | | | | | MSE |
| [72] | | Radio Map | | | | | RMSE |
| [74] | Wireless Communication | Radio Map | Model-free radio map reconstruction machine learning based methods. | Measurement points | Indoor | | Mean RSSI Difference |
| [83] | | | | Images | | | RMSE |
| [86] | | | | Measurement points | | | Mean Distance Error, Mean training error |
| [87] | | - | | Map data | Outdoor | | RMSE |
| [68] | Wireless Communication | Radio environment map | Model-free radio map reconstruction machine learning based methods. | **DeepREM dataset** | Outdoor | | RMSE |
| | | | | | | | - |
| | | | | Images | | | And |
| | | | | | | | - |
| | | | | | | | MAE |
| [76] | | Radio Map | | Images | Outdoor | | normalized mean squared error (NMSE) |
| [79] | | - | | Measurement points | | | RMSE |
| [80] | | | | Satellite images | | | RMSE |
| [81] | | | | Images | | | RMSE |

**Table 6.** *Cont.*

| Reference | Taxonomy Domain | Map Category | Construction Method–Models–Tools | Dataset Acquisition | Environment | Dismnsion of the Map | Units of Metric |
|---|---|---|---|---|---|---|---|
| [66] | Wireless Communication | Radio Map | Model-free radio map reconstruction machine learning based methods. | Measurement points | Outdoor | 2D | Squared Reconstruction Error |
| [70] | | - | Model-free radio map reconstruction machine learning based methods. | Measurement points and #D model of environment | | | Mean absolute error |
| [71] | | Power Spectrum Map | | Images + distribution functions | | | average image reconstruction error |
| [73] | | Radio Map | | Measurement points | | | RMSE |
| [75] | | - | | | | | RMSE, MSE, MAE |
| [77] | | Radio Map | | | | | MAPE |
| [78] | | - | | | | | average total hit rate error (AHRE) |
| [82] | Public Helath | Radio Environment Map | Model-free radio map reconstruction machine learning based methods. | Measurement points | Outdoor | | RMSE |
| [84] | | Exposure maps | | Images | | | RMSE and MAPE |
| [85] | | - | | Measurement points | | | R squared and MSE |
| [88] | | - | | Images | Indoor | | - |
| [89] | | Electric Field Strength Map | | Measurement points | Outdoor | | RMSE |
| [90] | | **Electromagnetic Field Exposure Map** | | Measurement points and Images | | | structural similarity index (SSIM) |
| [91] | | Electric Field Strength Map | | Measurement points | | | RMSE |

**Table 6.** *Cont.*

| Reference | Taxonomy Domain | Map Category | Construction Method–Models–Tools | Dataset Acquisition | Environment | Dismnsion of the Map | Units of Metric |
|---|---|---|---|---|---|---|---|
| [92] | Public Helath | EMF Exposure Map | Model-free radio map reconstruction machine learning based methods. | sensor measurement map images | Indoor | 2D | - |
| | | | | | | | - |
| | | | | | | | Average SSIM |
| [93] | | Electric Field Strength Map | | Measurement points | Outdoor | | RMSE |

In this section, we reviewed three primary methodological families for EMF mapping: analytical models, model-based interpolation techniques, and machine learning-driven approaches. While earlier reviews typically concentrate on a single category—often analytical or geospatial methods—our taxonomy captures the full methodological spectrum, including state-of-the-art deep learning models and hybrid techniques that merge physical modeling with data-driven inference. This comprehensive perspective highlights the evolution from traditional propagation models toward adaptive, intelligent systems, underscoring the novelty of our work in unifying these diverse approaches and revealing new opportunities for interdisciplinary applications in both wireless network optimization and public health exposure assessment.

## 4. Discussion

The analysis presented in this review highlights several key trends shaping the current landscape of electromagnetic field (EMF) exposure mapping. First, there is a notable transition from classical analytical propagation models toward more sophisticated, data-driven approaches, particularly those leveraging machine learning and deep learning techniques. These modern methodologies offer significant advantages in terms of adaptability, accuracy, and the ability to handle complex, dynamic environments where traditional approaches often fail.

Another important observation is the substantial variation in terminology used across studies, including terms such as "radio map," "path loss map," "electromagnetic exposure map," and "electric field strength map." This diversity underscores the existing ambiguity and frequent overlap in definitions within the field. Establishing a unified and consistent definitional framework, as proposed in this review, is crucial to advance both academic research and practical applications.

An important distinction must be made between single-source, single-frequency mapping and multi-band exposure surveys. The former is typically motivated by engineering needs, such as optimizing coverage and capacity in wireless networks, and can often be addressed with established propagation models. By contrast, exposure surveys that span multiple sources and frequency bands are vastly more complex, as they must capture the aggregate contribution of heterogeneous transmitters operating simultaneously in dynamic environments. This challenge is further amplified in public health studies, where assessments often require not only cross-band exposure levels at a given point in time but also historical trends to support epidemiological analyses. Developing methods that can integrate such spatial, spectral, and temporal dimensions remains an open and pressing research direction.

In addition, the review emphasizes the distinct methodological considerations necessary for indoor versus outdoor environments. Outdoor EMF mapping typically involves dealing with sparse data, large-scale geographic areas, and environmental variables such as terrain and atmospheric conditions, often necessitating predictive modeling and interpolation. In contrast, indoor mapping presents challenges due to complex signal propagation patterns influenced by building structures and interior elements, highlighting the need for advanced ray tracing and detailed structural modeling techniques.

The comprehensive taxonomy presented in this review categorizes EMF mapping methodologies effectively, providing clarity regarding current practices and identifying existing knowledge gaps. Although computational techniques, including interpolation and kriging methods, remain foundational, rapid advancement in machine learning methodologies, such as neural networks and generative adversarial networks (GANs), indicates a strong direction for future research. Specifically, hybrid models that combine traditional physical models with machine learning show promise, as they can potentially leverage the strengths of both approaches.

Based on the results, the imbalance in domain taxonomy suggests that scientific effort remains chiefly motivated by network design and optimization, with comparatively less emphasis on exposure assessment for epidemiological or regulatory purposes. A stronger integration of exposure mapping with public health research could provide valuable evidence for epidemiological studies and inform risk communication efforts.

Machine-learning techniques are the predominant construction approach, accounting for roughly two thirds of the corpus, while interpolation-based geospatial models contribute a further quarter and purely analytical formulations less than ten per cent. The dominance of data-driven ensembles reflects the availability of powerful regressors and open-source frameworks, but it also raises questions of transparency and generalisability, particularly when comparative baselines are sparse. There is room for hybrid paradigms that combine the physical grounding of analytical models with the flexibility of learning algorithms.

Direct measurement remains the chief data source, representing about sixty-two per cent of studies. Image-based or remotely sensed data now supply close to thirty per cent, highlighting growing interest in fast, low-cost sampling surrogates. Around ten per cent of papers (formerly blank, now labelled input of analytical methods) rely on look-up parameters or synthetic data instead of empirical observations. As image repositories expand and photogrammetric tools mature, the relative weight of the *Images* category can be expected to rise, potentially reducing deployment costs and enabling large-scale exposure mapping.

Two-dimensional outdoor scenarios dominate: two thirds of the investigations concentrate exclusively on outdoor propagation, twenty-one per cent examine indoor exposure, and the remainder either combine the two or leave the field undefined. No study in the present sample produces an explicit three-dimensional map, even though multilayer urban environments are known to exhibit strong vertical gradients. Expanding empirical coverage to complex indoor venues, high-rise districts and mixed-use neighbourhoods would markedly improve the ecological validity of exposure models.

The results underscore a pronounced technical tilt in the field. Most contributions are rooted in wireless-communication engineering, whereas public-health–oriented work occupies a noticeably smaller—but still significant—niche. No additional thematic branches surfaced, suggesting that EMF-mapping research remains anchored in these two primary motivations.

Turning to terminology, the landscape is strikingly fragmented. A sizeable subset of authors leaves their maps without any formal label, hinting either at implicit understanding within sub-communities or at a lack of consensus. Among the studies that do assign a name, "Radio Map" clearly dominates, with "Radio Environment Map" following behind. Labels such as "Electromagnetic-field exposure map" and "Electric-field-strength map" form a middle tier, while an eclectic mix of eight additional terms appears only sporadically, forming a long semantic tail. This dispersion complicates literature searches and reinforces the need for a standardised vocabulary.

With respect to construction methodology, machine-learning pipelines overwhelmingly lead the way. Classic geostatistical interpolation holds a respectable but secondary position, whereas purely analytical formulations now occupy the margins. The methodological hierarchy thus mirrors a broader trend across applied electromagnetics, where data-driven surrogates increasingly supplant first-principles solvers when speed and adaptability matter.

Examining how data are gathered, on-site field measurements remain the dominant evidence base. Remote imagery—be it satellite, aerial, or street-level—constitutes a robust alternative that is gaining traction, especially as open geo-datasets expand. In contrast, studies that forego empirical inputs altogether and rely on synthetic parameters or

look-up tables represent only a small minority, underscoring the community's preference for measurement-anchored validation.

Finally, the environmental taxonomy reveals that outdoor propagation scenarios attract the lion's share of attention. Indoor deployments form a smaller, but vibrant cluster, reflecting the growing importance of connectivity in buildings. Mixed indoor–outdoor studies are still rare, and a nontrivial set of articles leaves the operational setting vague, an omission that limits reproducibility and comparability between studies.

From an evaluative perspective, recent evidence suggests that ensemble learning methods such as Random Forests and XGBoost, together with deep learning hold particular promise for advancing EMF mapping. These approaches are effective in capturing nonlinear spatial patterns, reconstructing sparse datasets, and enabling scalable predictions across complex environments. However, their adoption also faces limitations: ensemble models often sacrifice interpretability, while deep learning frameworks typically require extensive training data and risk reduced generalizability when applied across different urban or regulatory contexts. Beyond the engineering perspective, there remains a notable gap in the public health application of field mapping. Addressing this requires research that systematically integrates multi-band, multi-source assessments with temporal and historical dimensions of exposure, thus strengthening the evidentiary base for epidemiological studies and regulatory decision-making.

A key avenue for future research lies in the systematic comparison of methods for assessing total RF exposure in urban environments. Measurement-based campaigns, simulation-driven approaches, and machine learning models each present distinct trade-offs: direct measurements offer high reliability but are costly and labor-intensive; simulations provide scalability but often rely on simplifying assumptions; and data-driven models can capture complex dynamics at lower cost but may suffer from generalization challenges. A comparative framework that evaluates these approaches in terms of cost, accuracy, scalability, and limitations would be highly valuable for guiding both urban network planning and public health assessments.

## 5. Conclusions

This review has presented a comprehensive and structured synthesis of methodologies for electromagnetic field (EMF) exposure mapping, spanning analytical models, interpolation techniques, and machine learning approaches. Unlike previous surveys that focus primarily on wireless communication engineering, our work integrates perspectives from both communication networks and public health, thereby broadening the scope and societal relevance of EMF mapping research. A unified taxonomy was proposed to clarify the fragmented terminology in the field, distinguishing among radio maps, radio environment maps, and EMF exposure maps, and linking each to its methodological foundations. This framework not only organizes the diverse literature but also highlights knowledge gaps, such as the under representation of health-oriented applications, the limited exploration of three-dimensional and hybrid mapping approaches, and the challenges posed by data sparsity and generalization.

By situating technical innovations within a wider interdisciplinary context, this review underscores the need for future research that combines computational rigor with public health priorities. Integrating big data, edge computing, and explainable AI holds particular promise for developing next-generation EMF mapping frameworks that are robust, scalable, and socially impactful. Ultimately, the contribution of this review lies in providing both a comprehensive state-of-the-art reference and a roadmap for future directions, reinforcing its novelty compared to prior works and supporting the development of reliable, interdisciplinary solutions for EMF exposure assessment.

**Author Contributions:** Conceptualization, Y.K. and T.P.; methodology, Y.K.; validation, Y.K. and T.P.; formal analysis, Y.K. and T.P.; investigation, Y.K.; writing—original draft preparation, Y.K.; writing—review and editing, Y.K. and T.P.; visualization, Y.K.; supervision, Y.K. and T.P.; project administration, Y.K. and T.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

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

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| EMF | Electromagnetic Field |
| RF | Radio Frequency |
| REM | Radio Environment Map |
| UAV | Unmanned aerial vehicle |
| ANN | Artificial Neural Network |
| CNN | Convolutional Neural Network |
| DNN | Deep Neural Network |
| GAN | Generative Adversarial Network |
| DL | Deep Learning |
| ML | Machine Learning |
| IDW | Inverse Distance Weighting |
| SVM | Support Vector Machine |
| KNN | k-Nearest Neighbors |
| XGBoost | Extreme Gradient Boosting |
| SAR | Specific Absorption Rate |
| RSSI | Received Signal Strength Indicator |
| SNR | Signal-to-Noise Ratio |
| NNI | Natural Neighbor Interpolation |
| GIS | Geographic Information System |
| ICNIRP | International Commission on Non-Ionizing Radiation Protection |
| WHO | World Health Organization |
| ITU | International Telecommunication Union |
| IEEE | Institute of Electrical and Electronics Engineers |

## Appendix A. Two-Ray Gound Reflection Model Equations

In this model, the received power $P_r$ is estimated using the following approximation:

$$P_r \approx \frac{P_t \cdot G_t \cdot G_r \cdot h_t^2 \cdot h_r^2}{d^4} \tag{A1}$$

where $P_t$ is transmitted power; $G_t$, $G_r$ are the gains of the transmitting and receiving antennas; $h_t$, $h_r$ are the heights of the transmitter and receiver antennas, respectively and $d$ is the distance between the transmitter and receiver [93–95].

## Appendix B. Egli Model for Radio Propagation

Basic analytics models form a common family of formulas, primarily focused on calculating path loss along a single propagation path. These models typically utilize distance, carrier frequency, and transmitter/receiver heights as inputs and often include

environment-specific correction factors derived from measurements. Our starting point is the Egli model, an empirical radio propagation model designed to predict path loss for point-to-point communication in both urban and rural settings. It is commonly applied to UHF and VHF frequencies (30 MHz to 3 GHz) and considers the influence of transmitter and receiver antenna heights, the operational frequency, and the separation distance

The path loss $L$ [96] in decibels (dB) is given by:

$$L = 117 + 40\log_{10}(d) - 20\log_{10}(h_t h_r) + 20\log_{10}(f) \tag{A2}$$

where $L$ is the attenuation (the path loss) in dB; $d$ is the distance between the transmitter and receiver in miles; $h_t$ is the height of the transmitting antenna in feet; $h_r$ is the height of the receiving antenna in feet and $f$ is the frequency in MHz. The Egli model is particularly useful for providing fast estimates of signal attenuation in wireless communication system design, especially in relatively unobstructed environments [97,98].

## Appendix C. Hata Model Equations

The Hata model, also known as the Hata-Okumura model [35], is an empirical for formulation of the Okumura model. It is widely used for predicting path loss in outdoor environments—urban, suburban, and rural—based on extensive field measurements. This model is particularly applicable to mobile network planning within the frequency range of 150 MHz to 1500 MHz.

The general formula [99] for median path loss in urban areas is given by:

$$L = 69.55 + 26.16\log_{10}(f) - 13.82\log_{10}(h_t) - a(h_r) + \left(44.9 - 6.55\log_{10}(h_t)\right)\log_{10}(d) \tag{A3}$$

where $L$ is the path loss in dB, $f$ is the frequency in MHz ($150 \leq f \leq 1500$), $h_t$ is the transmitter antenna height in meters ($30 \leq h_t \leq 200$), $h_r$ is the receiver antenna height in meters ($1 \leq h_r \leq 10$), $d$ is the distance between antennas in kilometers ($1 \leq d \leq 20$), $a(h_r)$ is the correction factor for receiver antenna height, defined as follows: for small to medium-sized cities:

$$a(h_r) = \left(1.1\log_{10}(f) - 0.7\right)h_r - \left(1.56\log_{10}(f) - 0.8\right) \tag{A4}$$

and for large cities and frequencies $f > 400$ MHz:

$$a(h_r) = 3.2\left(\log_{10}(11.75h_r)\right)^2 - 4.97 \tag{A5}$$

The Hata model also provides modifications for suburban and open rural environments. It is favored for its simplicity, minimal input requirements, and practical accuracy in network planning. However, it is not suitable for frequencies above 1500 MHz or scenarios involving complex terrain or significant obstructions.

## References


1. Wu, Y.; Gao, X.; Zhou, S.; Yang, W.; Polyanskiy, Y.; Caire, G. Massive Access for Future Wireless Communication Systems. *IEEE Wirel. Commun.* **2020**, *27*, 148–156. [CrossRef]
2. Attaran, M. The impact of 5G on the evolution of intelligent automation and industry digitization. *J. Ambient Intell. Humaniz. Comput.* **2023**, *14*, 5977–5993. [CrossRef] [PubMed]
3. Verbeek, J.; Oftedal, G.; Feychting, M.; van Rongen, E.; Rosaria Scarfì, M.; Mann, S.; Wong, R.; van Deventer, E. Prioritizing health outcomes when assessing the effects of exposure to radiofrequency electromagnetic fields: A survey among experts. *Environ. Int.* **2021**, *146*, 106300. [CrossRef]
4. Santra, A.; Wang, P.; Shaker, G.; Mysore, B.S.; Dolmans, G.; Chen, Y.; Shariati, N.; Pandharipande, A. Machine Learning-Powered Radio Frequency Sensing: A Review. *IEEE Sens. J.* **2025**, *25*, 23164–23183. [CrossRef]

5. Sarkar, T.; Ji, Z.; Kim, K.; Medouri, A.; Salazar-Palma, M. A survey of various propagation models for mobile communication. *IEEE Antennas Propag. Mag.* **2003**, *45*, 51–82. [CrossRef]
6. Levie, R.; Yapar, C.; Kutyniok, G.; Caire, G. RadioUNet: Fast Radio Map Estimation With Convolutional Neural Networks. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 4001–4015. [CrossRef]
7. Zhang, S.; Zhang, R. Radio Map Based Path Planning for Cellular-Connected UAV. In Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, USA, 9–13 December 2019; pp. 1–6. [CrossRef]
8. Rafie, I.F.M.; Lim, S.Y.; Chung, M.J.H. Path Loss Prediction in Urban Areas: A Machine Learning Approach. *IEEE Antennas Wirel. Propag. Lett.* **2023**, *22*, 809–813. [CrossRef]
9. Romero, D.; Kim, S.J. Radio Map Estimation: A data-driven approach to spectrum cartography. *IEEE Signal Process. Mag.* **2022**, *39*, 53–72. [CrossRef]
10. Liang, Y.C.; Chen, K.C.; Li, G.Y.; Mahonen, P. Cognitive radio networking and communications: An overview. *IEEE Trans. Veh. Technol.* **2011**, *60*, 3386–3407. [CrossRef]
11. Bennett, N.D.; Croke, B.F.; Guariso, G.; Guillaume, J.H.; Hamilton, S.H.; Jakeman, A.J.; Marsili-Libelli, S.; Newham, L.T.; Norton, J.P.; Perrin, C.; et al. Characterising performance of environmental models. *Environ. Model. Softw.* **2013**, *40*, 1–20. [CrossRef]
12. Frank, J.W. Electromagnetic fields, 5G and health: What about the precautionary principle? *J. Epidemiol. Community Health* **2021**, *75*, 562–566. [CrossRef]
13. Kiouvrekis, Y.; Manios, G.; Tsitsia, V.; Gourzoulidis, G.; Kappas, C. A statistical analysis for RF-EMF exposure levels in sensitive land use: A novel study in Greek primary and secondary education schools. *Environ. Res.* **2020**, *191*, 109940. [CrossRef]
14. Panagiotakopoulos, T.; Kiouvrekis, Y.; Misthos, L.M.; Kappas, C. RF-EMF Exposure Assessments in Greek Schools to Support Ubiquitous IoT-Based Monitoring in Smart Cities. *IEEE Access* **2023**, *11*, 7145–7156. [CrossRef]
15. ICNIRP. Guidelines for limiting exposure to electromagnetic fields (100 kHz to 300 GHz). *Health Phys.* **2020**, *118*, 483–524. [CrossRef]
16. Elzanaty, A.; Chiaraviglio, L.; Alouini, M.S. 5G and EMF Exposure: Misinformation, Open Questions, and Potential Solutions. *Front. Commun. Netw.* **2021**, *2*, 635716. [CrossRef]
17. Korkmaz, E.; Aerts, S.; Coesoij, R.; Bhatt, C.R.; Velghe, M.; Colussi, L.; Land, D.; Petroulakis, N.; Spirito, M.; Bolte, J. A comprehensive review of 5G NR RF-EMF exposure assessment technologies: Fundamentals, advancements, challenges, niches, and implications. *Environ. Res.* **2024**, *260*, 119524. [CrossRef]
18. Gajšek, P.; Ravazzani, P.; Grellier, J.; Samaras, T.; Bakos, J.; Thuróczy, G. Review of Studies Concerning Electromagnetic Field (EMF) Exposure Assessment in Europe: Low Frequency Fields (50 Hz–100 kHz). *Int. J. Environ. Res. Public Health* **2016**, *13*, 875. [CrossRef]
19. Veludo, A.F.; Stroobandt, B.; Van Bladel, H.; Sandoval-Diez, N.; Guxens, M.; Joseph, W.; Röösli, M. Exploring RF-EMF levels in Swiss microenvironments: An evaluation of environmental and auto-induced downlink and uplink exposure in the era of 5G. *Environ. Res.* **2025**, *266*, 120550. [CrossRef] [PubMed]
20. Lee, A.K.; Jeon, S.B.; Choi, H.D. EMF Levels in 5G New Radio Environment in Seoul, Korea. *IEEE Access* **2021**, *9*, 19716–19722. [CrossRef]
21. Ramirez-Vazquez, R.; Escobar, I.; Thielens, A.; Arribas, E. Measurements and Analysis of Personal Exposure to Radiofrequency Electromagnetic Fields at Outdoor and Indoor School Buildings: A Case Study at a Spanish School. *IEEE Access* **2020**, *8*, 195692–195702. [CrossRef]
22. Ramirez-Vazquez, R.; Escobar, I.; Arribas, E.; Vandenbosch, G.A.E. Systematic Review of Exposure Studies to Radiofrequency Electromagnetic Fields: Spot Measurements and Mixed Methodologies. *Appl. Sci.* **2024**, *14*, 11161. [CrossRef]
23. Seybold, J.S. *Introduction to RF Propagation*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2005. [CrossRef]
24. Yun, Z.; Iskander, M.F. Ray Tracing for Radio Propagation Modeling: Principles and Applications. *IEEE Access* **2015**, *3*, 1089–1100. [CrossRef]
25. Bi, S.; Lyu, J.; Ding, Z.; Zhang, R. Engineering Radio Maps for Wireless Resource Management. *IEEE Wirel. Commun.* **2019**, *26*, 133–141. [CrossRef]
26. Romero, D.; Ha, T.N.; Shrestha, R.; Franceschetti, M. Theoretical Analysis of the Radio Map Estimation Problem. *Trans. Wirel. Comm.* **2024**, *23*, 13722–13737. [CrossRef]
27. Bazerque, J.A.; Giannakis, G.B. Distributed spectrum sensing for cognitive radio networks by exploiting sparsity. *Trans. Signal Proc.* **2010**, *58*, 1847–1862. [CrossRef]
28. Yilmaz, H.B.; Tugcu, T.; Alagöz, F.; Bayhan, S. Radio environment map as enabler for practical cognitive radio networks. *IEEE Commun. Mag.* **2013**, *51*, 162–169. [CrossRef]
29. Haykin, S. Cognitive radio: Brain-empowered wireless communications. *IEEE J. Sel. Areas Commun.* **2005**, *23*, 201–220. [CrossRef]

30. Koutlia, K.; Bojović, B.; Lagén, S.; Giupponi, L. Novel radio environment map for the ns-3 NR simulator. In Proceedings of the WNS3 '21, Virtual, 23–24 June 2021; Association for Computing Machinery: New York, NY, USA, 2021; p. 41–48. [CrossRef]
31. Colombi, D.; Ghasemifard, F.; Joshi, P.; Xu, B.; Paola, C.D.; Törnevik, C. Methods and Practices for In Situ Measurements of RF EMF Exposure From 5G Millimeter Wave Base Stations. *IEEE Trans. Electromagn. Compat.* **2022**, *64*, 1986–1993. [CrossRef]
32. Capriglione, D. In-Situ RF Measurements of EMFs for Human Exposure Assessment Due to Modern Cellular Base Stations. *IEEE Instrum. Meas. Mag.* **2021**, *24*, 31–36. [CrossRef]
33. Wang, S.; Wiart, J. Sensor-Aided EMF Exposure Assessments in an Urban Environment Using Artificial Neural Networks. *Int. J. Environ. Res. Public Health* **2020**, *17*, 3052. [CrossRef]
34. Miclăuş, S.; Petriţa, T. Electromagnetic Field Monitoring by Fixed Stations in Romania: Data Processing Methods and Outcomes. *Int. Conf. Knowl.-Based Organ.* **2024**, *30*, 93–98. [CrossRef]
35. Phillips, C.; Sicker, D.; Grunwald, D. A Survey of Wireless Path Loss Prediction and Coverage Mapping Methods. *IEEE Commun. Surv. Tutor.* **2013**, *15*, 255–270. [CrossRef]
36. Department of Electronic Communications, Republic of Cyprus. EMF Measurements—Electromagnetic Fields Website. 2025. Available online: http://www.emf.mcw.gov.cy/emf/?page=emfmeasurements (accessed on 20 May 2025).
37. Agence Nationale des Fréquences (ANFR). ANFR—Agence Nationale des Fréquences. 2025. Available online: https://www.anfr.fr/accueil (accessed on 20 May 2025).
38. Bakirtzis, S.; Çagkan Yapar.; Qui, K.; Wassell, I.; Zhang, J. Indoor Radio Map Dataset. *IEEE DataPort* **2024** . [CrossRef]
39. Yapar, Ç.; Levie, R.; Kutyniok, G.; Caire, G. Dataset of Pathloss and ToA Radio Maps With Localization Application. *arXiv* **2022**, arXiv:2212.11777.
40. Yapar, Ç.; Levie, R.; Kutyniok, G.; Caire, G. Dataset of Pathloss and ToA Radio Maps With Localization Application. *IEEE DataPort* **2022**. [CrossRef]
41. Zheng, Y. RSRPSet_urban: Radio map in dense urban. *IEEE DataPort* **2022**. [CrossRef]
42. Burmeister, F.; Li, Z.; Bizon, I. High-Resolution Radio Environment Map Data Set for Indoor Office Environment. *IEEE DataPort* **2022**. [CrossRef]
43. Abdulkarim, A.; Faruk, N.; Alozie, E.; Sowande, O.A.; Olayinka, I.F.Y.; Usman, A.D.; Adewole, K.S.; Oloyede, A.A.; Chiroma, H.; Garba, S.; et al. Application of Machine Learning Algorithms to Path Loss Modeling: A Review. In Proceedings of the 2022 5th Information Technology for Education and Development (ITED), Abuja, Nigeria, 1–3 November 2022; pp. 1–6. [CrossRef]
44. Li, J.; Ding, G.; Zhang, X.; Wu, Q. Recent Advances in Radio Environment Map: A Survey. In Proceedings of the Machine Learning and Intelligent Communications, Weihai, China, 5–6 August 2017; Gu, X., Liu, G., Li, B., Eds.; Springer: Cham, Switzerland, 2018; pp. 247–257.
45. Feng, B.; Zheng, M.; Liang, W.; Zhang, L. A Recent Survey on Radio Map Estimation Methods for Wireless Networks. *Electronics* **2025**, *14*, 1564. [CrossRef]
46. Dare, O.E.; Okokpujie, K.; Adetiba, E. Radio Environment Map Construction: A Mini-Review. In Proceedings of the 2023 2nd International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS), Abuja, Nigeria, 1–3 November 2023; Volume 1, pp. 1–7. [CrossRef]
47. Kurt, S.; Tavli, B. Path-Loss Modeling for Wireless Sensor Networks: A review of models and comparative evaluations. *IEEE Antennas Propag. Mag.* **2017**, *59*, 18–37. [CrossRef]
48. Friis, H. A Note on a Simple Transmission Formula. *Proc. IRE* **1946**, *34*, 254–256. [CrossRef]
49. Lee, M.; Han, D. Voronoi Tessellation Based Interpolation Method for Wi-Fi Radio Map Construction. *IEEE Commun. Lett.* **2012**, *16*, 404–407. [CrossRef]
50. Herring, K.T.; Holloway, J.W.; Staelin, D.H.; Bliss, D.W. Path-Loss Characteristics of Urban Wireless Channels. *IEEE Trans. Antennas Propag.* **2010**, *58*, 171–177. [CrossRef]
51. de Souza, R.S.; Lins, R.D. A new propagation model for 2.4 GHz wireless LAN. In Proceedings of the 2008 14th Asia-Pacific Conference on Communications, Akihabara, Japan, 14–16 October 2008; pp. 1–5.
52. Rizk, K.; Wagen, J.F.; Gardiol, F. Two-dimensional ray-tracing modeling for propagation prediction in microcellular environments. *IEEE Trans. Veh. Technol.* **1997**, *46*, 508–518. [CrossRef]
53. Zhang, S.; Wijesinghe, A.; Ding, Z. RME-GAN: A Learning Framework for Radio Map Estimation Based on Conditional Generative Adversarial Network. *IEEE Internet Things J.* **2023**, *10*, 18016–18027. [CrossRef]
54. Bazerque, J.A.; Mateos, G.; Giannakis, G.B. Group-Lasso on Splines for Spectrum Cartography. *IEEE Trans. Signal Process.* **2011**, *59*, 4648–4663. [CrossRef]
55. Sato, K.; Suto, K.; Inage, K.; Adachi, K.; Fujii, T. Space-Frequency-Interpolated Radio Map. *IEEE Trans. Veh. Technol.* **2021**, *70*, 714–725. [CrossRef]

56. Agarwal, A.; Gangopadhyay, R. Predictive spectrum occupancy probability-based spatio-temporal dynamic channel allocation map for future cognitive wireless networks. *Trans. Emerg. Telecommun. Technol.* **2018**, *29*, e3442. [CrossRef]
57. Denkovski, D.; Atanasovski, V.; Gavrilovska, L.; Riihijärvi, J.; Mähönen, P. Reliability of a radio environment Map: Case of spatial interpolation techniques. In Proceedings of the 2012 7th International ICST Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM), Stockholm, Sweden, 18–20 June 2012; pp. 248–253. [CrossRef]
58. Azpurua, M.A.; Dos, Ramos, K. A comparison of spatial interpolation methods for estimation of average electromagnetic field magnitude. *Prog. Electromagn. Res. M* **2010**, *14*, 135–145. [CrossRef]
59. Angjelicinoski, M.; Atanasovski, V.; Gavrilovska, L. Comparative analysis of spatial interpolation methods for creating radio environment maps. In Proceedings of the 2011 19thTelecommunications Forum (TELFOR) Proceedings of Papers, Belgrade, Serbia, 22–24 November 2011; pp. 334–337. [CrossRef]
60. Phillips, C.; Ton, M.; Sicker, D.; Grunwald, D. Practical radio environment mapping with geostatistics. In Proceedings of the 2012 IEEE International Symposium on Dynamic Spectrum Access Networks, Bellevue, WA, USA, 16–19 October 2012; pp. 422–433. [CrossRef]
61. Alaya-Feki, A.B.H.; Jemaa, S.B.; Sayrac, B.; Houze, P.; Moulines, E. Informed spectrum usage in cognitive radio networks: Interference cartography. In Proceedings of the 2008 IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications, Cannes, France, 15–18 September 2008; pp. 1–5. [CrossRef]
62. Guillén-Pina, J.; Pérez-Aracil, J.; del Cerro, R.C.; Sánchez-Montero, R.; López-Espí, P.L.; Salcedo-Sanz, S. Efficient design of electromagnetic field exposure maps with multi-method evolutionary ensembles. *Environ. Res.* **2025**, *278*, 121636. [CrossRef]
63. Panagiotakopoulos, T.; Kiouvrekis, Y.; Ploussi, A.; Spyratou, E.; Efstathopoulos, E.P. Comparative Analysis of Interpolation Models for RF-EMF Exposure Mapping in Urban Areas. *IEEE Access* **2025**, *13*, 67006–67016. [CrossRef]
64. Kiouvrekis, Y.; Zikas, S.; Katis, I.; Tsilikas, I.; Filippopoulos, I. Development of electromagnetic pollution maps utilizing Gaussian process spatial models. *Sci. Total Environ.* **2024**, *955*, 176907. [CrossRef] [PubMed]
65. López-Espí, P.; Sánchez-Montero, R.; Guillén-Pina, J.; del Cerro, R.C.; Rojas, J. Optimal design of electromagnetic field exposure maps in large areas. *Environ. Impact Assess. Rev.* **2024**, *106*, 107525. [CrossRef]
66. Shrestha, S.; Fu, X.; Hong, M. Deep Spectrum Cartography: Completing Radio Map Tensors Using Learned Neural Models. *IEEE Trans. Signal Process.* **2022**, *70*, 1170–1184. [CrossRef]
67. Timilsina, S.; Shrestha, S.; Fu, X. Deep Spectrum Cartography Using Quantized Measurements. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5. [CrossRef]
68. Chaves-Villota, A.; Viteri-Mera, C.A. DeepREM: Deep-Learning-Based Radio Environment Map Estimation From Sparse Measurements. *IEEE Access* **2023**, *11*, 48697–48714. [CrossRef]
69. Romero, D.; Kim, S.J.; López-Valcarce, R.; Giannakis, G.B. Spectrum cartography using quantized observations. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 3252–3256. [CrossRef]
70. Krijestorac, E.; Hanna, S.; Cabric, D. Spatial Signal Strength Prediction using 3D Maps and Deep Learning. In Proceedings of the ICC 2021—IEEE International Conference on Communications, Montreal, QC, Canada, 14–23 June 2021; pp. 1–6. [CrossRef]
71. Han, X.; Xue, L.; Shao, F.; Xu, Y. A Power Spectrum Maps Estimation Algorithm Based on Generative Adversarial Networks for Underlay Cognitive Radio Networks. *Sensors* **2020**, *20*, 311. [CrossRef]
72. Teganya, Y.; Romero, D. Deep Completion Autoencoders for Radio Map Estimation. *IEEE Trans. Wirel. Commun.* **2022**, *21*, 1710–1724. [CrossRef]
73. Shrestha, R.; Ha, T.N.; Viet, P.Q.; Romero, D. Radio Map Estimation in the Real-World: Empirical Validation and Analysis. In Proceedings of the 2023 IEEE Conference on Antenna Measurements and Applications (CAMA), Genoa, Italy, 15–17 November 2023; pp. 169–174. [CrossRef]
74. Niu, Q.; Nie, Y.; He, S.; Liu, N.; Luo, X. RecNet: A Convolutional Network for Efficient Radiomap Reconstruction. In Proceedings of the 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, USA, 20–24 May 2018; pp. 1–7. [CrossRef]
75. Popoola, S.I.; Jefia, A.; Atayero, A.A.; Kingsley, O.; Faruk, N.; Oseni, O.F.; Abolade, R.O. Determination of Neural Network Parameters for Path Loss Prediction in Very High Frequency Wireless Channel. *IEEE Access* **2019**, *7*, 150462–150483. [CrossRef]
76. Tan, Z.; Xiao, L.; Tang, X.; Zhao, M.; Li, Y. A FL-Based Radio Map Reconstruction Approach for UAV-Aided Wireless Networks. *Electronics* **2023**, *12*, 2817. [CrossRef]
77. Parera, C.; Liao, Q.; Malanchini, I.; Tatino, C.; Redondi, A.E.C.; Cesana, M. Transfer Learning for Tilt-Dependent Radio Map Prediction. *IEEE Trans. Cogn. Commun. Netw.* **2020**, *6*, 829–843. [CrossRef]
78. Ostlin, E.; Zepernick, H.J.; Suzuki, H. Macrocell Path-Loss Prediction Using Artificial Neural Networks. *IEEE Trans. Veh. Technol.* **2010**, *59*, 2735–2747. [CrossRef]

79. Imai, T.; Kitao, K.; Inomata, M. Radio Propagation Prediction Model Using Convolutional Neural Networks by Deep Learning. In Proceedings of the 2019 13th European Conference on Antennas and Propagation (EuCAP), Krakow, Poland, 31 March–5 April 2019; pp. 1–5.
80. Thrane, J.; Zibar, D.; Christiansen, H.L. Model-Aided Deep Learning Method for Path Loss Prediction in Mobile Communication Systems at 2.6 GHz. *IEEE Access* **2020**, *8*, 7925–7936. [CrossRef]
81. Ratnam, V.V.; Chen, H.; Pawar, S.; Zhang, B.; Zhang, C.J.; Kim, Y.J.; Lee, S.; Cho, M.; Yoon, S.R. FadeNet: Deep Learning-Based mm-Wave Large-Scale Channel Fading Prediction and its Applications. *IEEE Access* **2021**, *9*, 3278–3290. [CrossRef]
82. Kiouvrekis, Y.; Panagiotakopoulos, T.; Nousi, E.; Filippopoulos, I.; Ploussi, A.; Spyratou, E.; Efstathopoulos, E.P. Explainable Machine Learning for Radio Environment Mapping: An Intelligent System for Electric Field Strength Monitoring. *IEEE Access* **2025**, *13*, 75100–75118. [CrossRef]
83. Suga, N.; Maeda, Y.; Sato, K. Indoor Radio Map Construction via Ray Tracing With RGB-D Sensor-Based 3D Reconstruction: Concept and Experiments in WLAN Systems. *IEEE Access* **2023**, *11*, 24863–24874. [CrossRef]
84. Zhang, Y.; Wang, S.; Wiart, J. ExposNet: A Deep Learning Framework for EMF Exposure Prediction in Complex Urban Environments. *arXiv* **2025**, arXiv:2503.02966. [CrossRef]
85. Nyakyi, C.P.; Mpeshe, S.C.; Dida, M.A. Optimization of RF-EMF exposure to public in Tanzania using Artificial Neural Network and multi linear regression models. *Environ. Chall.* **2025**, *18*, 101066. [CrossRef]
86. Wang, X.; Wang, X.; Mao, S.; Zhang, J.; Periaswamy, S.C.G.; Patton, J. Indoor Radio Map Construction and Localization With Deep Gaussian Processes. *IEEE Internet Things J.* **2020**, *7*, 11238–11249. [CrossRef]
87. Seretis, A.; Sarris, C.D. Toward Physics-Based Generalizable Convolutional Neural Network Models for Indoor Propagation. *IEEE Trans. Antennas Propag.* **2022**, *70*, 4112–4126. [CrossRef]
88. Kiouvrekis, Y.; Psomadakis, I.; Vavouranakis, K.; Zikas, S.; Katis, I.; Tsilikas, I.; Panagiotakopoulos, T.; Filippopoulos, I. Explainable Machine Learning-Based Electric Field Strength Mapping for Urban Environmental Monitoring: A Case Study in Paris Integrating Geographical Features and Explainable AI. *Electronics* **2025**, *14*, 254. [CrossRef]
89. Mallik, M.; Tesfay, A.A.; Allaert, B.; Kassi, R.; Egea-Lopez, E.; Molina-Garcia-Pardo, J.M.; Wiart, J.; Gaillot, D.P.; Clavier, L. Towards Outdoor Electromagnetic Field Exposure Mapping Generation Using Conditional GANs. *Sensors* **2022**, *22*, 9643. [CrossRef]
90. Kiouvrekis, Y.; Givisis, I.; Panagiotakopoulos, T.; Tsilikas, I.; Ploussi, A.; Spyratou, E.; Efstathopoulos, E.P. A Comparative Analysis of Explainable Artificial Intelligence Models for Electric Field Strength Prediction over Eight European Cities. *Sensors* **2025**, *25*, 53. [CrossRef]
91. Mallik, M.; Kharbech, S.; Mazloum, T.; Wang, S.; Wiart, J.; Gaillot, D.P.; Clavier, L. EME-Net: A U-net-based Indoor EMF Exposure Map Reconstruction Method. In Proceedings of the 2022 16th European Conference on Antennas and Propagation (EuCAP), Madrid, Spain, 27 March–1 April 2022; pp. 1–5. [CrossRef]
92. Kalatzis, D.; Ploussi, A.; Spyratou, E.; Panagiotakopoulos, T.; Efstathopoulos, E.P.; Kiouvrekis, Y. Explainable AI for Spectral Analysis of Electromagnetic Fields. *IEEE Access* **2025**, *13*, 113407–113427. [CrossRef]
93. Sommer, C.; Joerer, S.; Dressler, F. On the applicability of Two-Ray path loss models for vehicular network simulation. In Proceedings of the 2012 IEEE Vehicular Networking Conference (VNC), Seoul, Republic of Korea, 14–16 November 2012; pp. 64–69. [CrossRef]
94. Sommer, C.; Dressler, F. Using the Right Two-Ray Model? A Measurement-based Evaluation of PHY Models in VANETs. In Proceedings of the MobiCom 2011, Las Vegas, NV, USA, 19–23 September 2011.
95. Giordano, E.; Frank, R.; Ghosh, A.; Pau, G.; Gerla, M. Two Ray or not Two Ray this is the price to pay. In Proceedings of the 2009 IEEE 6th International Conference on Mobile Adhoc and Sensor Systems, Macau, China, 12–15 October 2009; pp. 603–608.
96. Egli, J.J. Radio Propagation above 40 MC over Irregular Terrain. *Proc. IRE* **1957**, *45*, 1383–1391. [CrossRef]
97. Jawhly, T.; Tiwari, R.C. The special case of Egli and Hata model optimization using least-square approximation method. *SN Appl. Sci.* **2020**, *2*, 1296. [CrossRef]
98. Mlinar, T.; Podgrajšek, U.; Batagelj, B. Comparison of Coverage-Prediction Models for Modern Mobile Radio Networks. *Electronics* **2023**, *12*, 4554. [CrossRef]
99. Deme, A.; Dajab, D.D.; Bajoga, B.G.; Mu'azu, M.A.; Choji, D.N. Hata-Okumura Model Computer Analysis for Path Loss Determination at 900MHz for Maiduguri, Nigeria. *Math. Theory Model.* **2013**, *3*, 1–9.


**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.

computers

MDPI

Review

# Leveraging Artificial Intelligence in Project Management: A Systematic Review of Applications, Challenges, and Future Directions

**Dorothea S. Adamantiadou * and Loukas Tsironis ***

Department of Business Administration, University of Macedonia, 54636 Thessaloniki, Greece
* Correspondence: d.adamantiadou@uom.edu.gr (D.S.A.); loukas@uom.edu.gr (L.T.)

**Abstract:** This article presents a systematic literature review exploring the integration of Artificial Intelligence (AI) methodologies in project management (PM). Key applications include cost estimation, duration forecasting, and risk assessment, which are critical factors for project success. This review synthesizes findings from 97 peer-reviewed studies published between 2011 and 2024, using the PRISMA methodology to ensure rigor and transparency. AI techniques such as machine learning, deep learning, and hybrid models have exhibited their potential to enhance PM techniques across projects' phases, including planning, execution, and monitoring. Decision trees are created to represent the application of AI methodologies in various PM stages and tasks to facilitate understanding and real-world implementation. Among these are hybrid AI models that enhance risk assessment, duration forecasting, and cost estimation, as well as categorization based on project phases to optimize AI integration. Despite these advancements, there are still gaps in addressing dynamic project environments, validating AI models with real-world data, and expanding research into underexplored phases like project closure.

**Keywords:** project management (PM); Artificial Intelligence (AI); cost estimation; duration forecasting; risk assessment; machine learning (ML); hybrid models; decision trees; PRISMA methodology; systematic literature review

## 1. Introduction

Project management (PM), according to PMBOK GUIDE 7th edition [1], is defined as the application of knowledge, skills, tools, and techniques to manage and lead activities. PM methodology describes project success as the attainment of project objectives and the completion and acceptance of all project deliverables by the client [2]. In addition, Turner and Xue [3] argue that a project can be successful when it offers meaningful benefits, aligns with the agreed deliverables, and follows the specified schedule and budget. The knowledge of suitable information and technology is essential for achieving success in projects [4]. Due to technological improvements, the field of PM has seen rapid evolution in recent years. Artificial Intelligence (AI) is one of the most revolutionary technologies making a significant impact in this industry. The study of intelligent computational agents, or "computational intelligence", is a component of AI [5,6]. AI has demonstrated the ability to direct projects, automate PM duties, and support decision-making processes [7,8]. The applications of AI in projects are diverse, serving to reduce risks, monitor their progress, and identify abnormalities and exceptions in them [7,9]. Costing, schedule control, activity

management, and resource planning can be supported by leveraging AI in PM [7,10]. Martínez and Fernadez-Rodriguez [11], through their research, indicated that AI tools exhibit greater accuracy compared to traditional methods, yet they remain supplementary to traditional tools. Project managers can greatly benefit from the integration of AI tools in overseeing and managing their projects.

This research offers a systematic literature review (SLR) to thoroughly examine the role of AI in PM. By analyzing the utilization of AI methodologies in critical project success factors (CSFs), such as cost estimation, duration forecasting, and risk assessment, this study focuses on enlightening researchers, stakeholders, and practitioners about the potentials of AI in advancing PM techniques. Cost estimation, duration forecasting, and risk assessment are often considered CSFs in PM. These elements are vital for the successful delivery of a project across its lifecycle. According to the research by Pinto and Slevin [12], these factors are consistently highlighted as crucial during the planning and execution phases, as they directly influence a project's ability to meet its objectives within the defined constraints of time and cost [13].

The present article delves deeply into the literature on the application of AI in PM, encompassing various approaches, tools, and challenges. It aims to analyze the applications of AI in PM, identify gaps and weaknesses in the existing literature, and consolidate knowledge through the methodical collection and evaluation of research and studies. Additionally, it seeks to highlight the advantages of AI—such as increased productivity, decreased risk, and optimized processes—while examining the limitations and challenges organizations must overcome. The findings will assist professionals and researchers in making informed decisions, developing a theoretical framework for integrating AI, and encouraging innovation and sustainable growth in PM.

In the rapidly growing field of Artificial Intelligence (AI) in project management, various studies have explored distinct aspects of this intersection. Notable contributions include the SLRs conducted by Nenni et al. [14] and Prasetyo et al. [15]. These reviews, along with studies by additional researchers such as Costa et al. [7] and Taboada et al. [16], provide a foundation for examining the adoption, integration, and impact of AI technologies in project management processes. Nenni et al. applied an SLR with a bibliometric approach, analyzing 215 studies using PRISMA and VOSviewer to identify trends in AI-driven project management. This study is more focused on mapping AI applications across different Project Risk Management processes and evaluating AI categories and tools commonly used in project risk mitigation. Prasetyo et al. focused on AI adoption in open innovation project management by employing an SLR that emphasized AI's role in enhancing collaboration, decision-making, and knowledge sharing. Taboada et al. examined AI's impact on Industry 5.0 project management, concentrating on AI's application in planning, measurement, and uncertainty performance domains. Their analysis suggested that AI, especially machine learning, has been notably effective in construction and IT projects for forecasting and decision-making. The research of Costa et al. (2022) followed a technology-driven approach, categorizing AI techniques (e.g., machine learning, deep learning, NLP, fuzzy logic) and their specific project management applications. This study focuses more on technical AI methodologies rather than their practical implementation within PM frameworks.

This study is an SLR using the PRISMA framework, ensuring transparency and replicability. It integrates decision tree methodologies to classify AI applications across different project management knowledge areas, including cost estimation, duration forecasting, and risk assessment. By framing the analysis within the PMBOK framework, this research ensures a transparent and replicable process for identifying, screening, and synthesizing the relevant literature and provides a novel roadmap for future AI development, addressing

both technical and qualitative challenges, thereby extending the focus beyond quantitative efficiencies emphasized in earlier studies. Moreover, by integrating decision tree methodologies to classify AI applications across PM knowledge areas, this article aims to provide a visual and analytical framework to understand the distribution of AI applications, highlight gaps, and identify opportunities for future research. These methodological advancements distinguish this study by offering a comprehensive and integrative perspective that bridges theoretical and practical insights.

The structure of this article is as follows: It begins by outlining the methodology used in this research, detailing the steps for collecting and evaluating articles, and explaining how the results are presented through diagrams. Next, the paper provides a comprehensive literature review, summarizing existing studies and their findings on the integration of AI in PM, categorized by the areas of knowledge in PM. In order to effectively present the results of the literature review, decision trees are developed to classify and visually organize the above-mentioned theories, models, and findings. Following this, the article identifies research gaps in the current literature, highlighting areas that require further study. Finally, it concludes with a review of the findings and offers recommendations for future research directions.

## 2. Materials and Methods

This study conducts a comprehensive review of the global literature on the utilization of AI in PM. Before the start of the search, a review protocol was entered into INPLASY (INPLASY202510041). An SLR, as defined by Tranfield et al. [17] and Denicol et al. [18], is a transparent, rigorous, and detailed methodology that facilitates decision-making. This approach is pivotal for theory-building through the synthesis of knowledge derived from the analysis of studies and methodologies, thereby enhancing the coherence of results and conclusions. It employs explicit and systematic procedures to minimize bias in searching, identifying, evaluating, synthesizing, analyzing, and summarizing the literature. When executed meticulously, this study has the potential to yield reliable findings and robust conclusions, guiding decision-makers and scientific practitioners in making informed decisions and taking appropriate actions [19,20].

The present SLR adopts the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology [7,21]. The PRISMA methodology encompasses a series of steps aimed at achieving desired outcomes, with an information flow similar to that depicted in Figure 1.

The reason behind adopting the PRISMA system lies in its recognized comprehensiveness, its global, widespread use across various fields beyond medical domains, and its ability to enhance consistency among assessments [22], aiding in the conduct of a rigorous systematic review [7].

For data collection, it was decided to utilize two platforms, Web of Science and Scopus, in order to minimize duplicate results. The search terms used on each platform were combinations of "Artificial Intelligence" and "Project Management"; "Machine Learning" and "Project Management"; "Deep Learning" and "Project Management"; "Neural Networks" and "Project Management"; "Natural Language Processing" and "Project Management"; "Fuzzy Logic" and "Project Management"; "Heuristics" and "Project Management"; "Forecasting" and "Project Management"; and finally "Data Analytics" and "Project Management". Quotation marks were used in all searches to ensure exact matching with these terms. From these searches, a database containing 10,043 records was initially obtained.

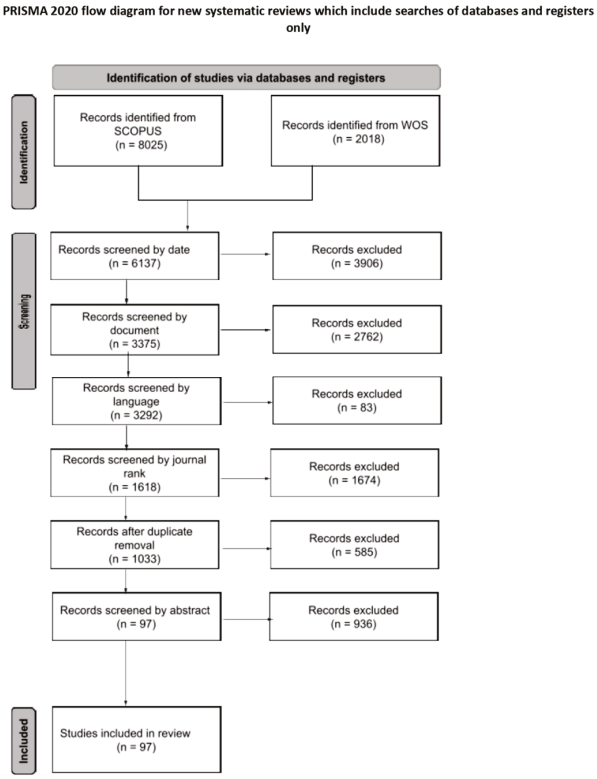



**Figure 1.** AI in PM—modified PRISMA Flow Diagram [22]. The figure illustrates the systematic review process following the PRISMA 2020 guidelines. Studies were identified from multiple databases and underwent a series of screening steps, including filtering by publication date, document type, language, journal quality, duplicate removal, and abstract relevance. Each step narrowed down the pool of studies to ensure that only the most relevant and high-quality research was included in the final review.

To optimize the dataset to include articles relevant to the research subject of this study, the PRISMA methodology was continued. The first criterion was the publication year of the findings. It was decided to include only findings from 2011 to February 2024 due to the transformation trend in PM resulting from rapid technological advancements in the Fourth Industrial Revolution [23]. This restriction reduced the database by 3906 records, leaving 6137 data points. Next, priority was given to the type of data, retaining only those that were published articles in scientific journals. As a result, 3375 data points were retained (2762 were excluded). Subsequently, for further analysis of abstracts, articles not written in English were excluded (83), resulting in the database containing 3292 data points. Significance was attributed to journal credibility. The criteria chosen were Scimago rankings meeting the conditions for the scientific prestige of journals from Q1 to Q4. To achieve this selection, all journals from the 3292 articles were researched on Scimago and classified, eliminating unrecognized or "non-assigned" journals. Consequently, the remaining articles were published in Q1-category journals, totaling 1618 (1674 were excluded). Duplicate articles that remained were then removed, resulting in a final dataset of 1033 data points (585 were excluded). The remaining articles were read and evaluated based on each article's title, abstract, and full content, and the process was completed with a final database of 97 entries.

Figure 2, along with Tables 1 and 2, collects the bibliometric results of the selected papers. Figure 2 can be analyzed by examining the points along each axis, which stand for data values for particular years. The scale of the values is shown by the distance from the chart's center; greater values are represented by larger distances. By comparing changes

over the years shown in the chart, the connecting lines aid in the visualization of trends over time. For example, since 2020, more articles have focused on the involvement of AI in PM. As depicted in Table 1, the natures of the journals of the identified works are diverse: they cover construction and engineering (*Journal of Construction Engineering and Management* is the most cited, followed by *Automation in Construction*), management (*International Journal of PM* is the second most cited in that discipline), and computer science (*Neural Computing and Applications*, *Expert Systems with Applications*, and *Scientific reports*). Table 2 shows that engineering is the main field of the selected papers, with construction industry, computer science, and business and economics following.

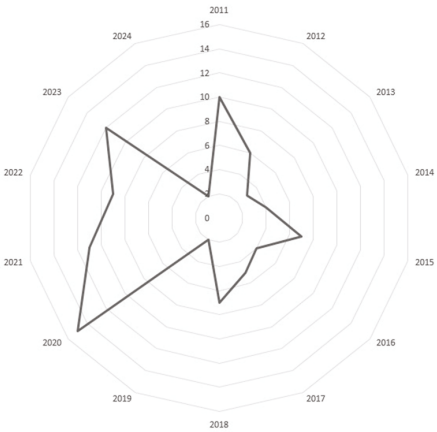



**Figure 2.** Bibliometric results. Publications per year.

**Table 1.** Bibliometric results. Articles per journal.

| JOURNAL | VALUES |
|---|---|
| *JOURNAL OF CONSTRUCTION ENGINEERING AND MANAGEMENT* | 13 |
| *INTERNATIONAL JOURNAL OF PROJECT MANAGEMENT* | 10 |
| *SUSTAINABILITY (SWITZERLAND)* | 6 |
| *AUTOMATION IN CONSTRUCTION* | 5 |
| *JOURNAL OF MANAGEMENT IN ENGINEERING* | 4 |
| *IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT* | 4 |
| *DECISION SUPPORT SYSTEMS* | 3 |
| *SCIENTIFIC REPORTS* | 3 |
| *IEEE ACCESS* | 3 |
| *AUTOMATION IN CONSTRUCTION* | 3 |

**Table 2.** Bibliometric results. Articles per industry.

| INDUSTRY | VALUES |
|---|---|
| ENGINEERING | 53 |
| CONSTRUCTION AND BUILDING TECHNOLOGY | 27 |
| COMPUTER SCIENCE | 25 |
| BUSINESS AND ECONOMICS | 24 |
| OPERATIONS RESEARCH AND MANAGEMENT SCIENCE | 14 |
| SCIENCE AND TECHNOLOGY | 10 |
| ENVIRONMENTAL SCIENCES AND ECOLOGY | 6 |
| TELECOMMUNICATIONS | 4 |
| MANAGEMENT | 2 |
| ARCHITECTURE CONSTRUCTION AND BUILDING TECHNOLOGY | 1 |
| AUTOMATION AND CONTROL SYSTEMS | 1 |
| PUBLIC ADMINISTRATION | 1 |

## 3. Literature Review

This literature review will identify and present the current knowledge, theories, models, and discoveries in the field of research on the integration of AI in PM, through the examination of selected articles, which are categorized into project management areas of knowledge. The ten knowledge areas described in the PMBOK guide are Project Integration Management, Project Scope Management, Project Schedule Management, Project Cost Management, Project Quality Management, Project Resource Management, Project Communications Management, Project Risk Management, Project Procurement Management, and Project Stakeholder Management [24].

This classification aims to highlight the evolving role of intelligent systems and methodologies in enhancing decision-making, predictive accuracy, and overall project outcomes. In addition to identifying gaps, contradictions, or difficulties in current research, this chronological analysis will aid in understanding the field's evolution and contributing to identifying areas where further research is needed.

### *3.1. Project Integration Management*

This literature review describes 13 articles that could be categorized under the area of Project Integration Management. More specifically, focusing on monitoring the progress of construction projects, the authors review pattern recognition techniques, which include neural network pattern recognition (NN-PR) techniques and statistical pattern recognition (S-PR). Integrating AI into pattern recognition has enabled project managers to access advanced features for decision-making within the construction industry [25]. In the same industry, a hybrid model was developed, the evolutionary fuzzy hybrid neural network (EFHNN), that merges four AI approaches: NNs, high-order NNs, FL, and GAs. This model is used in the execution phase of projects in order to assist project managers in controlling project performance and in implementing corrective measures necessary for project success [26]. From planning to execution and monitoring, a new proposed multivariate model suggested identifying performance issues and helping project managers to implement corrective measures [27]. Fuzzy time series forecasting and data envelopment analysis were used to create a theoretical framework to improve the prediction of project performances. Futures research has suggested the incorporation of ML algorithms in order to improve predictive capabilities [28]. For software projects, an entailment-based intelligent system was proposed (EIS), which contains NLP, entailment recognition, contradiction recognition, a recognition result merger, and a performance decision-maker and aims to monitor and manage projects in the monitoring and control phase of projects [29]. Focusing on the creation of a model capable of predicting the performance of construction projects under unit price contracts, another study suggested the use of the System Dynamics (SD) approach. Futures research has incorporated the integration of real-time data and sensor technologies to enhance the accuracy and timeliness of performance forecasting in construction projects [30]. Cheng et al. [31] created the Hybrid Gaussian Process Inference Model, which consists of data input, the Gaussian process and Bayesian inference, and hyper-parameter optimization. This model offers accurate forecasts regarding the results of construction projects to help in the decision-making process. FL and hybrid fuzzy approaches have been applied to address uncertainty in construction variables. Fuzzy set theory has been used to represent transitions between concepts, improving the realism of predictions. By combining fuzzy arithmetic with ML, this approach seeks to boost decision-making and project outcomes in construction engineering and management. Future studies could explore methods like conditional fuzzy clustering, collaborative fuzzy clustering, DL, and hierarchical learning to manage the complexity of construction problems caused by

high dimensionality, to enhance the accuracy of fuzzy hybrid models [32]. A new framework combined Building Information Modeling (BIM), ML algorithms, Virtual Reality (VR) technology, and the Unity game engine to create a game-like hybrid application for the on-demand monitoring of construction projects. By focusing on enhancing integration, real-time data processing, automation, scalability, user experience, interoperability, and validation, future research could improve PM and decision-making in the construction industry [33]. Aiming to automate the alignment of long-term and short-term plans in construction projects, a hybrid model has been proposed that consists of NLP techniques. By using this model, managers can improve the productivity, effectiveness, and precision of construction projects [34]. Moreover, a new model was created, which integrates GAs and probabilistic models, to forecast the outcomes of agile projects within global software development enterprises. According to the creators, this model could be used by project managers in the planning stage of a project to forecast its success [35]. The same year, Uddin et al. [36] proposed five ML models—logistic regression, the SVM, K-Nearest Neighbors, the Random Forest, and Extreme Gradient Boosting—to describe the connections between project attributes, networks, and the "Iron Triangle", which contains project cost, time, and quality. Different models outperformed others, based on the inputs. These models could be used by project managers in different stages of projects. More specifically, by using the models in the planning stage, managers could predict duration and project cost; in the execution phase, they could control project progress and risks; and in the control phase, they could minimize risks and modify project plans. Kim et al. [37] developed an alternative method that can be used to measure and monitor the performance of construction projects in situations where credible data are collected for performance indicators such as cost duration using GAs.

### *3.2. Project Scope Management*

Project Scope Management includes two articles. The first one focuses on suggesting a hybrid model called BABE, which combines the Artificial Bee Colony (ABC) and Analogy-Based Estimation (ABE) approaches. Focusing on software development projects, this model could be used in the early stages of a project in order to foresee the effort needed for development [38]. Another study explores the use of the Extreme Learning Machine (ELM) for estimating software development effort and concludes that it outperforms other methods in accuracy, based on metrics like the mean absolute error (MAE) and the magnitude of the relative error (MRE). Future research could focus on improving the model through parameter tuning, adding more machine learning techniques, and testing on larger, more varied datasets [39].

### *3.3. Project Schedule Management*

In 2011, research was conducted in order to propose an ant colony optimization algorithm that could improve existing software quality assessment models by adapting them to new software systems. Ant colony optimization (ACO) is a swarm intelligence algorithm that takes inspiration from the foraging behavior observed in real ant colonies [40]. This model excels in comparison to other approaches, such as the C4.5 ML algorithm and random guessing, while maintaining the transparency of the models it generates [41]. The same year, in order to forecast the duration for the reconstruction project following the Wenchuan earthquake, two ML models were developed, using Bromilow's time–cost (BTC) model and the Elman network (EN). Even though the EN model was more accurate in reconstruction time than the BTC model, the authors concluded that the BTC model is easier to use in real-life projects [42]. Another study used an ANN to predict project schedules

in the Indian construction industry. The developed model exhibited strong predictive skills [43]. Irfan et al. [44] decided to develop mathematical models in the planning stage in order to estimate the duration of highway projects by creating mathematical relationships between the duration and the factors that affect it, such as project cost, project type, and contract type. The same year, a neurogenetic approach was proposed, which combines the strengths of GAs and NNs, focusing on project scheduling. This hybrid approach improves the quality of solutions and contributes to the advancement of the field of project scheduling by enhancing the efficiency of scheduling processes [45]. Chou et al. [46] suggested a hybrid intelligence model, called ESIM, Evolutionary SVM Inference, which consists of an SVM and a fmGA. The model was designed for application in the initial stage of software development projects and surpassed traditional models in terms of estimating the time required to complete the development of a software project. Sadeghi et al. [47] developed a fuzzy discrete event simulation (FDES) framework for the construction industry. This framework combines FL with discrete event simulation to refine the precision of project time estimation. The accuracy of the proposed FDES model was attributed to the approach for calculating event times. For future research, the authors suggested the addition of algorithms and approaches in order to improve project scheduling and management. Wauters and Vanhoucke [48] recommended a hybrid model, which contains the Nearest Neighbor (NN) method, decision trees (DTs), Bagging, the Random Forest (RF), Boosting, and the SVM. During the planning phase of a project, this model aims to predict the duration of the project. Focusing on dam projects, Golizadeh et al. [49], in order to predict the duration of a construction project, created a multilayer perceptron network based on the backpropagation learning method using an ANN. This model was used in the pre-contracting phase of a project. For software projects, Choetkiertikul et al. [50] used a variety of ML approaches, such as Random Forests, NNs, decision trees, the Naïve Bayes method, DNNs with dropouts, and Gradient Boosting Machines, in order to build models that could forecast whether an issue was at risk of being delayed. In the year 2018, Pospieszny et al. [51] suggested the combined use of three ML algorithms, the SVM, the multilayer perceptron (MLP), and generalized linear models (GLMs). This research aimed to estimate software project duration and effort in the initial stages of a project. In addition, a hybrid model, called the firefly-tuned least squares SVM (FLSVM), which combines the least squares SVM (LS-SVM) and the firefly algorithm (FA), was suggested in order to forecast the duration of the construction of a diaphragm wall [52]. In the construction field, in order to predict delays in construction projects, the impact of factors causing delays can be measured and the time performance of a project can be forecasted according to risk levels. The model named the Random Forest classifier with genetic algorithm optimization (RF-GA) could be used in the planning and scheduling stages of a project [53]. Gondia et al. [54] developed and compared the performance of two models: one based on the decision tree algorithm and the other based on the Naïve Bayesian algorithm. The evaluation and comparison of these two models allowed the researchers to conclude that the Naïve Bayesian model provided superior predictive performance for analyzing project delays in the construction industry compared to the decision tree model. Research could focus on creating dynamic risk analysis models that could adjust to evolving project conditions and offer real-time risk assessments throughout a project lifecycle. In 2021, a new model was created in order to forecast delays in construction projects. This model used Random Forest classification and real-time data to assist in updating managers on a project's schedule in real time [55]. In the context of construction schedule management, a subset simulation-based Probabilistic Critical Path Method (PCPM) was introduced, in order to provide to project managers the ability to monitor timetables and reduce the risk of project delays [56].

The same year, a DL-based approach that combined long short-term memory (LSTM) and a gated recurrent unit (GRU) proposed to predict difficulties in the scheduling of engineering projects in order to help managers take corrective measures to guarantee the on-time completion of them [57]. The scheduling risk assessment framework (SRAF) addresses uncertainties in project duration and resource availability by capturing, analyzing, and integrating probabilistic distributions related to activity durations and resource breakdown scenarios. The focus of this framework is on enhancing project schedules by incorporating buffer time, adjusting the objective function, and applying the Enhanced Move-Based Local Search Heuristic (EMBLSH) algorithm to manage uncertainties and enhance project results. Future studies could aim to find a way for how this model could help practitioners predict future budgets more accurately, reduce financial and time losses, and improve project-planning strategies [58]. By using eight ML algorithms to train regression models, researchers aimed to predict the effort and duration of tasks in software projects. After creating 48 different regression models, they concluded that the Random Forest, the Extra Trees Regressor, and XGBoost consistently outperformed the other algorithms [59]. In order to monitor and forecast a Graphical Evaluation and Review Technique (GERT) project, a hybrid model was proposed, which combined stochastic EDM (Earned Duration Management) and ML algorithms, including Ada Boost, the Bagging Regressor, and Gradient Boosting. The model used the Monte Carlo simulation to collect information for the project completion time [60]. Short-term project progress could be predicted by a comprehensive framework that combined data analytics techniques with agility practices, as explored by Jafaar et al. [61]. Moreover, another model developed in a study used fuzzy PERT (FPERT) and the critical path method (CPM) to provide accurate timeline and budget estimates for digitized startups, specifically focusing on the case study of the startup e-Karsaz in Pakistan. Future research in this area could focus on validating and refining the proposed model by conducting comparative analyses with other startups and integrating additional PM techniques to enhance accuracy and applicability across various startup contexts [62]. Yang and Chen [63] focused on elevating Earned Schedule Methods (ESMs) in order to forecast project duration for delayed projects by proposing the ESmin+ method. Future studies should focus on validating this method, using actual projects to assess its effectiveness in more complex and real-world scenarios. These articles focused on enhancing Project Schedule Management.

### *3.4. Project Cost Management*

Many authors have tried to explore the area of Project Cost Management. Cheng and Roy [64] created a hybrid Support Vector Machine (SVM) model, EFSIM T (Evolutionary Fuzzy SVM Inference Model for Time Series Data), to predict cash flow and control project performance. This model was used in the planning and implementation phases of a project with better outcomes than other models and consisted of three AI methodologies, FL, weighted SVMs, and fast messy GAs (fmGA). Hwang [65] developed the time series models ARMA (5,5) and VAR(12) in order to forecast construction costs. Utilizing time series index data for cost estimation enhances forecasting accuracy, supports informed decision-making, and enables proactive cost management strategies based on historical cost trends and market dynamics. Moreover, another study suggested the use of a novel time-dependent evolutionary fuzzy SVM inference model that combined FL, a weighted SVM (wSVM), and a fmGA. By combining these methodologies, the authors aimed to enhance the performance of Estimation at Completion (EAC) in construction projects. The use of additional ML methodologies could enhance the predictive capacity of the model and the use of real-time data could provide more precise data to improve cost estimation [66].

In the computer science community, a study was published, which presented a hybrid model, a long-linear regression and multilayer perceptron (MLP) NN model. In the early stages of software projects, this model aims to offer early cost estimation [67]. In the year 2014, a new research work forecasted the accuracy of projects' Earned Value metrics by developing a mathematical modeling procedure and combining time series and linear regression analysis. The authors suggested the combination of AI, ML, or big data analytics in order to enhance the accuracy of the results [68]. A year after, the Second Moment Bayesian (SMB) model was created, which could be applied in the execution phase of a project in order to predict the final project cost, using data collected during the project's execution [69]. The authors tried to create a model for Taiwanese construction projects. More specifically, a hybrid computational model for forecasting the Taiwanese Construction Cost Index (CCI) was proposed, which consisted of multivariate adaptive regression splines (MARS), a radial basis function neural network (RBFNN), an Artificial Bee Colony (ABC) algorithm, a generalized likelihood ratio (GLR) test, and an Extreme Learning Support Vector Machine (ELSVM). The hybrid model was created in order to assist cost engineers in dealing with the variability of the CCI and improving the accuracy of the CCI forecast model [70]. Another study compared four different NN models for software development effort estimation: the multilayer perceptron (MLP), general regression neural network (GRNN), radial basis function neural network (RBFNN), and cascade correlation neural network (CCNN). The findings indicated that the cascade correlation neural network (CCNN) and radial basis function neural network (RBFNN) tended to enhance the accuracy of software development effort estimation in comparison to the multilayer perceptron (MLP) and general regression neural network (GRNN) models. Further investigation involving alternative validation methods such as leave-one-out (LOO) validation could offer more information on the performance of these models [71]. In order to enhance cost forecast accuracy, another study suggested the combined use of Earned Value Management (EVM) models, Reference Class Forecasting (RCF), and the Exponential Smoothing Method (ESM). The study introduced the Exponential Smoothing with Reference Class Forecasting (XSM) method [72]. For the construction field, and specifically for expressway projects, a convolutional neural network (CNN) model was created, as a tool for the estimation of projects' cost. Before the determination of funding, a manager could apply this model in order to minimize capital risks and make considered decisions [73]. In another study focused on predicting project cost overrun levels, an ensemble-learning model employed the Ripple-Down Rule Learner (RIDOR) and K-star algorithms as foundational learners. The RIDOR is capable of automatically generating classification rules from input data, whereas the K-star model is a lazy classifier that retains training instances for classification purposes. These algorithms were chosen based on their proven reliability and accuracy in previous construction data studies, thus enhancing the effectiveness of the ensemble model in predicting cost overrun levels. The incorporation of real-time data, such as project progress updates, market trends, and economic indicators, into the predictive models could enhance their accuracy and responsiveness to changing conditions [74]. In the field of construction, a hybrid model, consisting of multiple ANN models that were trained by the Lower Upper-Bound Estimation (LUBE) method, was created to forecast material prices [75]. Xue et al. [76] discussed integrating the Sparrow Search Algorithm (SSA) with a backpropagation neural network (BPNN) to enhance cost prediction accuracy for substation projects. By optimizing the BPNN's weights and thresholds, the SSA-BP model outperformed traditional approaches, achieving lower error metrics. Future research should focus on incorporating qualitative factors, improving data collection, and broadening the model's application to better analyze and predict costs in the energy sector.

Kim and Cha [77] introduced a hybrid expert system known as LD-MCS, which integrates a rule-based system, logic-based design (LD) with a case-based reasoning system, and a memory-based case system (MCS) for the automated prediction of costs in apartment renovation projects. This system is engineered to enhance the precision and efficiency of cost estimation through the fusion of expert insights and historical data. Focusing on cost estimation, Castro Miranda et al. [78] published a systematic literature review in order to analyze predictive analytics methods that can improve the effectiveness of estimating construction costs during the initial phases of projects. Lee et al. [79] proposed an advanced ANN-based cost estimation model that could integrate ensemble modeling and factor analysis to enhance the precision and robustness of preliminary cost estimates for large-scale construction projects with limited historical data availability. Futures research has suggested the use of additional ML techniques and algorithms for cost prediction models with limited data samples. The evaluation of ML-based techniques in calculating the effort and duration of individual tasks or issues in software projects was studied in another study. Regression analysis was conducted and a parametric model was used in order to estimate the cost of new product development projects [80]. A combination of the fuzzy analytic hierarchy process (FAHP) and traditional ML techniques, such as K-Nearest Neighbors (KNNs), Random Forests (RFs), and XGBoost, was proposed to create a Construction Cost Index (CCI) specific to Jordan's construction industry and predict upcoming CCI values. Future research could focus on new ML algorithms or hybrid models for improved accuracy and in predicting the integration of the construction costs of advanced technologies such as AI, the Internet of Things (IoT), and big data analytics [81].

### *3.5. Project Quality Management*

Project Quality Management was analyzed by Chen et al. [82], who created a decision tree framework that was designed to enhance the evaluation and quality assurance of software processes. In 2019, a hybrid model was developed that consisted of FL and NNs, the weighted neuro-fuzzy framework. This model is applied in the early and later stages of projects in order to forecast errors in the software and to enhance the quality of projects by using different data [83]. A Bayesian Belief Network (BBN) model was used to forecast project performance, based on historical data, as it was more accurate in identifying complexity, uncertainty, and performance factors [84]. With the use of data mining, researchers aimed to detect elements that participate in project success. Using these data, an ANN was proposed that could predict the success of projects based on these factors [85].

### *3.6. Project Resource Management*

Project Resource Management was explored in 2011, when ANNs, Gas, and FL were combined to create the neuro-fuzzy genetic system and improve the accuracy of selecting competent project managers for construction projects [86]. A multilayer feedforward NN trained with a backpropagation algorithm was used in the implementation, planning, and design stages of a project to measure and forecast workforce efficiency in construction projects [87]. Patel and Jha [88] proposed an ANN model that aimed to predict the safe work behavior of employees in construction projects. In another study, a hybrid model based on the ANN and activation and transfer functions was created in order to forecast, improve, and manage labor productivity in construction projects. The model was trained to use everyday data from the employees' work hours [89]. In addition, an NLP was applied in order to create a text classification model that could automatically allocate building service requests to maintenance personnel. The NLP was combined with three

ML methodologies to find the most accurate model. Between logistic regression, the Naïve Bayes method, and the SVM, the last model surpassed the other two [90]. In the year 2022, a hybrid neuro-fuzzy system that was optimized using an innovative EEBAT algorithm was applied in projects that used scrum agile PM techniques to help managers in selecting suitable resources. This was achieved by minimizing the difference between actual and estimated effort while ensuring timely delivery within defined timelines [91]. FL and metaheuristic approaches have been combined to investigate the strength, energy efficiency, and cost-effectiveness of building materials in construction management. ML algorithms with existing hybrid metaheuristic approaches could develop the accuracy and efficiency of material optimization in construction projects in future studies, while developing a real-time monitoring and control system would help the algorithm to monitor building material usage based on changing environmental conditions and energy demands [92]. A model was presented that offered a mathematical framework, which addressed the intricate interplay among construction tasks, resource allocation, financial flows, and the uncertainties arising from various influencing factors. The model was based on a partially observable Markov decision process. Furthermore, the model harnessed deep reinforcement learning (DRL) to achieve the ongoing adaptive optimization of labor and material management, leading to the enhancement of work and cash flows. Additionally, a discrete event simulation-based simulator was created to replicate dynamic project characteristics and external conditions, contributing to the training of deep reinforcement learning [93]. Another study used a total of seven ML models to predict organizational agility. These models included the SVM, decision trees (DTs), K-Nearest Neighbors (KNNs), the Gradient Boosting Machine (GBM), the Random Forest (RF), the Naïve Bayes (NB) method, and logistic regression (LR). The results of the study indicated high test accuracy percentages for most of the models, with the GBM, RF, SVM, and KNNs ranking higher than the other methods [94].

### *3.7. Project Risk Management*

The ANN was used to examine the decision process regarding risk in public–private partnership (PPP) projects. Even though the model produced satisfactory results, the authors suggested, as future research, the use of other approaches to create more accurate models [95]. In the year 2012, another hybrid model was created, which consisted of a casual map and a Bayesian Network (BN), in order to forecast how combinations of alterations in variables of interest (leadership style, agility, project performance) influence overall project performance [96]. In the same year, a study was published that focused on early planning. More specifically, a hybrid model was suggested that combined NNs and SVMs with traditional logistic regression models. Project teams could use these models in the early planning stage of a project to better predict final results [97]. Fang and Marle [98] focused on developing a simulation-based risk network model within a decision support system. They created a framework in order to offer to project managers the ability to identify, analyze, model, and respond to risks effectively. The authors suggested the use of real-time data to develop a real-time risk monitoring system for future research. In the year 2013, a framework for software project risk analysis and planning was proposed that combined risk analysis and planning in the early stages of a project. The framework used ML algorithms (Random Forest, NN, and logistic regression) to achieve risk analysis and estimations [99]. Shi et al. [100] suggested FL theory, as a method, in order to analyze and manage delivery risks in construction programs. Futures research combines the integration of advanced technologies to improve decision-making in program management. Focusing on FL and risk management, like the previous study, another study aimed to enhance project performance by optimizing risk allocations [101]. Gunduz

et al. [102] proposed a model, which contained FL, as a decision support tool for contractors to identify delays in Turkish construction projects before the bidding stage. In 2016, a new method was proposed, which combined the Analytic Hierarchy Process (AHP) and fuzzy inference system (FIS) in order to offer a more adaptable method for risk evaluation in IT projects [103]. A hybrid model called the BN-HFACS combines the Human Factors Analysis and Classification System (HFACS) with Bayesian Network (BN) modeling to forecast safety performance in construction projects [104]. Targeting risk management, FL combined with mathematical models aimed to assess the probability of risk events and their consequences on the financial stability of a company, particularly in the context of promoting software products [105]. Focusing on Build–Operate–Transfer (BOT) projects, authors proposed a fuzzy analytic hierarchy process (FAHP) model that joined fuzzy theory with the analytic hierarchy process (AHP) to determine the importance of stakeholders' risks [106]. A combined approach using Social Network Analysis (SNA) and the Matrice d'Impacts Croisés Multiplication Appliquée à un Classement (MICMAC) method was employed to investigate and comprehend factors influencing project delays. Futures research has involved AI methods like DL and reinforcement learning to improve the predictive abilities of models used for analyzing project delays [107]. Assaad et al. [108] developed a framework that focused on predicting project performance based on project risks. The model contained mathematical and statistical analysis techniques. Further research is needed in order to determine the importance of project risks in cost and schedule overruns. Practical application and validation with real-world data are required in order to validate the model and compare it with existing predictive models. In the same scientific area of construction, focusing on forecasting financial losses due to natural processes, a deep neural network (DNN) was developed, using the convolutional neural network approach [109]. ML techniques were used in order to create models that could predict conflicts in construction projects. Among all the models, the SVM was the most accurate. Project managers could benefit from this model by identifying risks and factors that cause dispute appearances and proceeding to measures that minimize the problems [110]. A hybrid approach combining machine learning techniques, specifically a genetic algo-rithm (GA) integrated with K-means clustering, focused on identifying high-risk factors in large-scale construction projects. By analyzing expert-sourced data and employing the synthetic minority over-sampling technique (SMOTE) for better data representation, the study effectively categorized risks based on their severity and impact on project performance, enhancing stakeholder awareness and risk allocation [111]. For the purpose of forecasting cost overruns in thermal power plant projects, focusing on the most critical risk factors, a hybrid model was created that combined genetic programming (GP) and Monte Carlo simulation (MCS). By using real data from projects, this model could be applied in the early stages of a project to help managers confront risk efficiently [112]. Moreover, another group of researchers proposed a hybrid model, a combination of wavelet analysis and an artificial neural network (W-ANN), to forecast the incidence of accidents in the construction industry [113]. Researchers utilized the DNN algorithm to develop a model, which could predict financial damages from accidents in apartment construction sites. The model was trained by historical data and the predictions differed based on the inputs that the user included [114]. In addition, Santos et al. [115], focusing on assisting project managers in understanding the relationships between project activities and outcomes, in order to distribute resources and manage risks, created a new method that involved ML techniques, specifically Monte Carlo sampling and the triad method, alongside the Shapley value approach from cooperative game theory. Finally, another study expanded the development of a construction site dynamic risk ensemble model, based on ML algorithms, that

could offer forecasts safety risk indicators in various workplace settings. The ensemble model, which consisted of the Naïve Bayes model, decision trees, Random Forests, SVMs, and ANNs, combined the prediction capabilities of the individual algorithms to make predictions, which could achieve the same level of performance as the top-performing base algorithms or potentially outperform them [116]. De Andrade et al. [117] provided project managers with a framework for assessing and enhancing project forecasting accuracy by leveraging the concept of project consistency. Advanced data analytics techniques, such as ML algorithms, could be added in the future to improve the prediction accuracy of project duration forecasting based on project regularity indicators. All the articles above could be categorized under the area of Project Risk Management.

### 3.8. Project Stakeholder Management

Project Stakeholder Management was analyzed by focusing on improving decision-making and conflict management in construction projects. The study proposed a framework that integrates SVMs, FL, and the fast and messy genetic algorithm (fmGA). This hybrid AI system outperforms the use of Support Vector Machines (SVMs) [118]. In the same industry, a study suggested a multilayer perceptron (MLP) NN model, which could improve the decision-making process for arbitration, litigation, and contract administration by concerning differences in contracts [119]. Emphasizing the use of data and ML, a data-driven framework was suggested to assist project managers in decision-making and problem solving. This framework could adapt to different projects, and its accuracy and effectiveness depend on its application in specific problems and datasets [4].

### 3.9. Project Cost and Schedule Management

Project Cost and Schedule Management, combined in the construction industry, were studied by utilizing a framework of Earned Value Management. Future studies could focus on adding ML algorithms and AI methodologies to enhance the accuracy of project duration and cost predictions [120]. In addition, an SVM regression model was created to control project time and cost outcome forecasting. This framework helped project managers make decisions about planning, execution, and monitoring [121]. Habibi et al. [122] proposed a new hybrid model that they aimed to apply in the planning and control stages of projects in order to improve time and cost estimation and project scheduling. The model combined classical PERT and CPM methods with FL and appeared to offer more accurate estimations than traditional methods. A model described in another study suggested the use of actual project schedule information to assess how well Earned Value Management (EVM) and Earned Duration Management (EDM) could predict project outcomes. The study analyzed the effectiveness of project duration forecasting techniques [123].

### 3.10. Evolution Path

By examining the literature, Figure 3 explores how AI's contributions to PM have developed over time, highlighting key milestones and emerging trends in research. More specifically, between 2011 and 2012, research focused on introducing basic AI techniques like ACO and hybrid neural networks for scheduling and cost estimation. From the year 2013 to 2016, studies expanded into hybrid systems by analyzing the growth of advanced algorithms combining fuzzy logic, neural networks, and genetic algorithms for risk and resource management. Later, in 2017–2019, researchers marked a shift towards integrating AI with real-time data and multivariate analysis for execution monitoring and performance improvement. The integration of NLP, the IoT, and DL for enhanced communication, prediction, and quality management was expanded between the years 2020 and 2021, while

current trends (2022–present) focus on holistic frameworks that integrate multiple PM knowledge areas while addressing challenges in stakeholder and integration management.

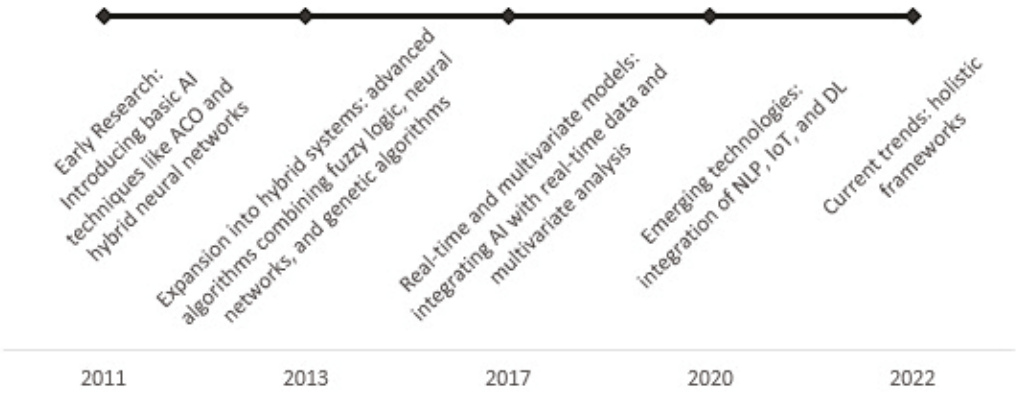



**Figure 3.** Evolution path of AI techniques in PM.

### *3.11. Decision Tree Analyses*

Focusing on the literature above, this study introduces a set of decision trees that were developed to offer a guide for comprehending the ways in which different AI methodologies are used during the planning, execution, and monitoring stages of a project. They provide a summary of the relationships among several AI models, including machine learning, deep learning, and hybrid models, as well as the applications of them in projects.

The first decision tree, shown in Figure 4, aims to illustrate how several hybrid AI models combine strategies to enhance PM outcomes, such as cost, duration, and risk. A "hybrid model" in machine learning is a method that combines multiple algorithms or techniques to improve performance by leveraging the strengths of each component. These models are valuable in fields requiring high accuracy, like finance and engineering, where they enhance predictive power and data handling [124]. The tree in Figure 4 highlights the application of AI and hybrid methods in PM, targeting enhanced efficiency in project duration, cost, and risk management. For duration optimization, models combine neural networks, genetic algorithms, fuzzy logic, and machine learning, aiming to predict or optimize scheduling and forecasting. For cost optimization, combinations include Bayesian techniques, fuzzy logic, and regression models to enhance cost prediction accuracy. Risk analysis methods incorporate simulation (Monte Carlo) and AI for robust risk assessment, with predictive models such as decision trees and logistic regression applied to identify and classify risk factors. Figure 4 expands in detail in Figures 5–7. In Figure 8, AI models are categorized according to the project phase they are most effectively used in (planning, execution, monitoring). Through this structure, readers can understand the context and timing of applying different models to achieve the best results, along with an organized plan for effectively applying AI across the project lifetime.

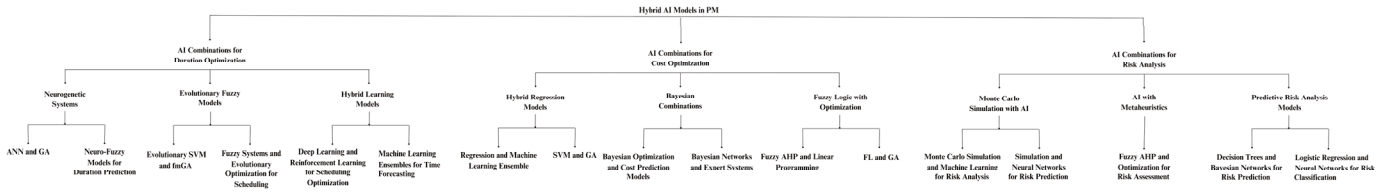



**Figure 4.** Different hybrid AI models combine techniques for PM outcomes.

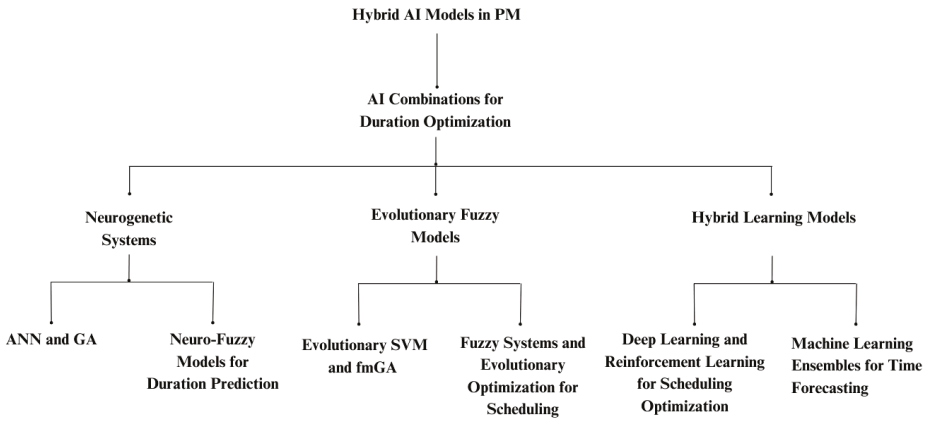



**Figure 5.** Part of Figure 3—hybrid AI combinations for duration optimization.

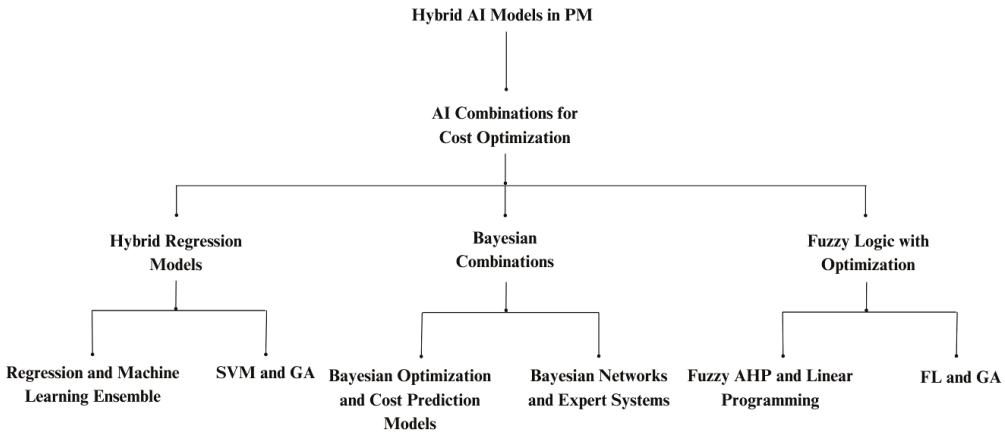



**Figure 6.** Part of Figure 3—hybrid AI combinations for cost optimization.

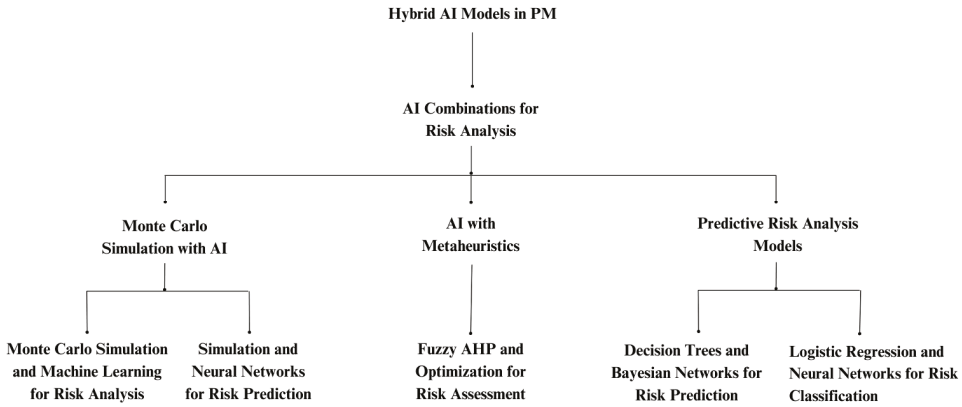



**Figure 7.** Part of Figure 3—hybrid AI combinations for risk analysis.

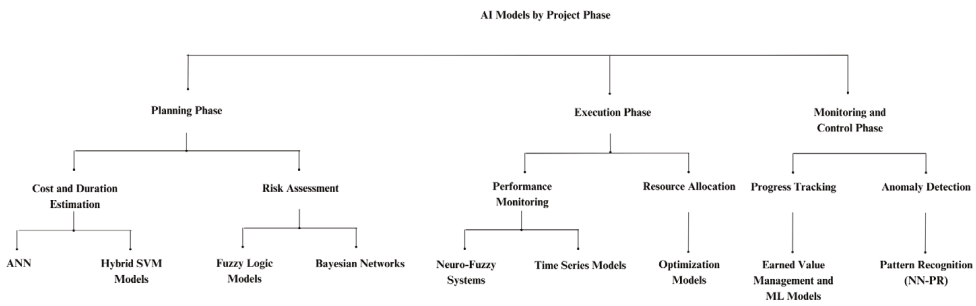



**Figure 8.** AI models according to the project phase they are most effectively used in.

Figures 9–11 focus on cost, duration, and risk forecasting, respectively. Figure 9 illustrates the various AI methodologies used for cost estimation in PM. This tree provides project managers with a comprehensive understanding of which methodologies are most appropriate for their particular requirements and project contexts by classifying models into categories such as regression, neural networks, Bayesian methods, and expert/statistical models. Figure 10's decision tree classifies the AI models used to predict project durations. It makes a distinction between hybrid, deep learning, and more conventional machine learning techniques, such regression and time series models. Each branch focuses on particular methods and how they might be applied. The AI models used in Project Risk Management and forecasting are highlighted in Figure 11. It illustrates how several AI methodologies are applied to identify, assess, and minimize risks. From FL systems for risk prediction to complex neural networks for performance forecasting, the decision tree shows how techniques for risk management evolve as a result of the application of AI methods.

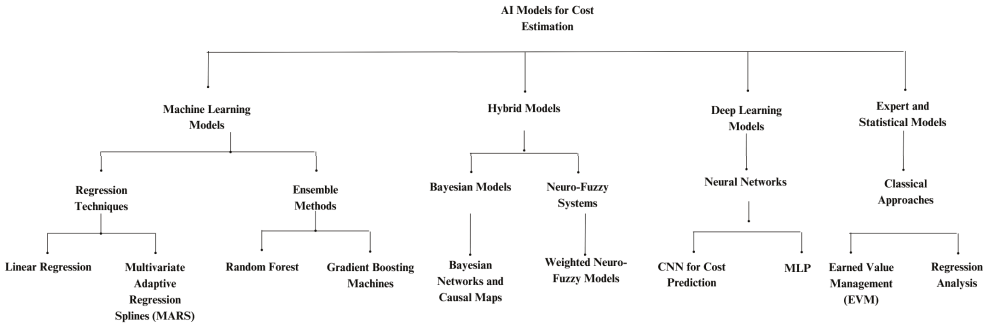



**Figure 9.** AI methodologies used for cost estimation in PM.

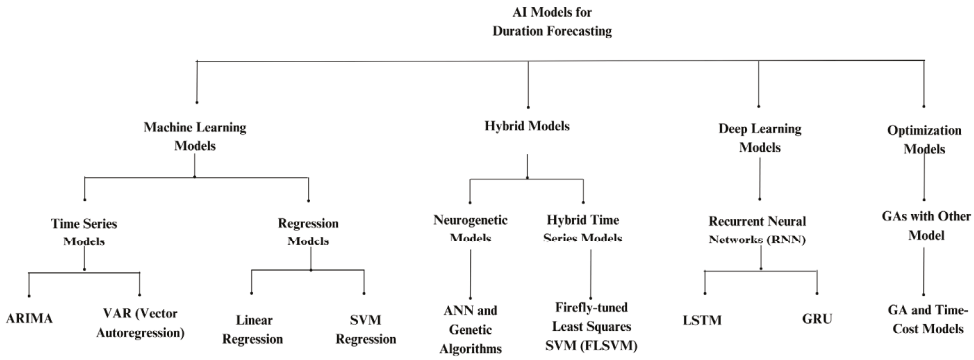



**Figure 10.** AI models used for forecasting project durations.

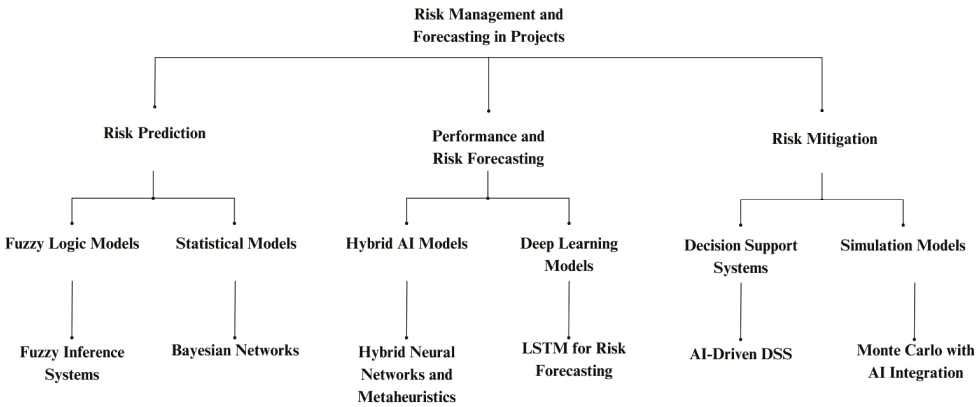



**Figure 11.** AI models used for risk management and forecasting within projects.

### *3.12. Data Quality and Sources in AI-Driven Project Management Studies*

The success of AI methodologies is heavily contingent upon the quality of the data utilized, highlighting challenges such as data's completeness, accuracy, representativeness, and temporal relevance [125]. The studies in this systematic review have used a range of data sources, such as industry reports, structured datasets, and case studies of real-world projects. Many studies rely on publicly accessible databases in order to ensure transparency and consistency in their findings. However, when it comes to actual project restrictions, these statistics frequently lack detail. Although they provide high-quality, accurate information, industry datasets are sometimes inaccessible because of confidentiality agreements. Certain studies have chosen to use predictive models to simulate data, which enables controlled testing but might not accurately represent the unpredictability of the real world. Additionally, expert interviews and survey-based data have been used to obtain qualitative insights into project management, although subjectivity and bias may have been included. Several studies have highlighted issues with completeness, where missing data points affect the reliability of AI-driven predictions, particularly in cost estimation and risk assessment models. Accuracy remains a significant concern, as project parameters such as budget, timeline, and risk factors often fail to align with actual project outcomes. While some studies incorporate real-time IoT-driven data collection, many still rely on static historical datasets. This can create discrepancies in how AI-driven predictions align with evolving project environments. Studies leveraging a hybrid approach, combining structured data with real-time and expert-driven insights, tend to provide the most comprehensive datasets. Focusing on the articles contained in this literature review, it is highlighted that the cross-industry validation and benchmarking of AI models remain largely underutilized. This undermines the real-world applicability of the suggested AI-driven project management tools.

## 4. Conclusions, Discussion, and Future Work

This literature review presents the development of PM methodologies by examining the use of hybrid models, such as ML and FL, to improve project outcomes and predictive accuracy. Research has indicated that models such as the weighted neuro-fuzzy framework and the BABE model can be helpful for project effort and software error predictions. Furthermore, in the software development and construction industries, advanced methodologies like deep learning and Bayesian Networks have proved crucial for risk assessment and performance forecasts. Models such as the Random Forest and XGBoost consistently outperform other models, highlighting the trend toward real-time data integration for accurate predictions.

AI's contributions to project management have been substantial, particularly in Project Schedule Management, Project Cost Management, and Project Risk Management. Predictive models such as machine learning algorithms and hybrid systems have participated in improving forecasting accuracy, optimizing schedules, and identifying risks. By classifying the literature into the 10 areas of knowledge in PM, it is concluded that limited research has been conducted on Project Integration Management, Project Scope Management, Project Quality Management, Project Communications Management, Project Procurement Management, and Project Stakeholder Management.

Despite notable improvements in hybrid models and ML applications for PM, several gaps and weaknesses become visible in the literature. A large number of studies are not fully validated using real-world data, which reduces their practical relevance. Furthermore, while AI is increasingly used in planning, scheduling, and execution, there is little focus on post-project performance. Additionally, many current models neglect the interdependencies between different risk indicators in favor of a more fragmented approach to risk assessment.

Knowledge areas like Project Communications Management and Project Stakeholder Management require skills, emotional intelligence, and contextual awareness, qualities that are challenging for AI systems to replicate. On the other hand, research has focused on quantitative aspects of project management, such as cost estimation and schedule management, as these align more effectively with the capabilities of existing AI models. Implementing AI in underexplored areas often demands significant investments in infrastructure, skilled personnel, and organizational change, which can create barriers for both researchers and practitioners. Additionally, resistance to adopting AI in areas involving human interaction and decision-making has further slowed progress. Concerns over job displacement and the interpretability of AI-driven decisions exacerbate this resistance, contributing to the persistent gaps in AI's application across all project management knowledge areas

Several limitations remain in the application of AI to project management. One challenge is the overfitting of machine learning models, particularly when models are trained on historical project data that may not reflect the complexities and uncertainties of real-world projects. Another significant challenge is the generalization of AI-driven project management methodologies across industries. While AI applications in construction and IT project management have been well-explored, their applicability in other industries—such as healthcare, finance, or public administration—remains underdeveloped. Differences in requirements, project structures, and stakeholder interactions create barriers to widespread AI adoption. Moreover, the integration of AI into project management requires substantial investment in data infrastructure and expertise, which may be a barrier for smaller organizations.

Future research should expand on unexplored knowledge areas in PM by developing AI tools focused on Project Communications Management to analyze communication patterns, on Project Procurement Management to analyze contracts, and on Project Stakeholder Management to analyze preferences and concerns, while industry-specific AI adaptations and cross-domain validations should be formed to enhance AI's reliability in diverse project environments. AI-based structured frameworks could be developed for analyzing project success factors and feedback loops to enable continuous improvement. No previous research has investigated the application of a framework capable of integrating real-world data across multiple PM knowledge areas, such as risk, cost, and schedule management, to provide a new approach to project decision-making. Questions are raised for further research. The first question focuses on enhancing project forecasting and decision-making in PM by analyzing and incorporating real-world data into current AI models. The second question addresses the difficulties in maintaining the accuracy and consistency of AI models when constantly updating them with real-world data. The third question concerns how to improve decision-making and project performance in PM using predictive models based on AI. These models could simultaneously focus on developing the prediction and optimization of cost, duration, risk, and task prioritization.

**Author Contributions:** Conceptualization, D.S.A. and L.T.; methodology, D.S.A.; validation, D.S.A. and L.T.; formal analysis, D.S.A.; investigation, D.S.A.; resources, D.S.A.; data curation, D.S.A.; and L.T.; writing—original draft preparation, D.S.A.; writing—review and editing, D.S.A. and L.T.; visualization, D.S.A. and L.T.; supervision, L.T.; project administration, L.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** All data are included in the article.

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

## References


1. Project Management Institute. A Guide to the Project Management Body of Knowledge (PMBOK Guide)—Seventh Edition and the Standard for Project Management. 2021. Available online: https://ibimone.com/PMBOK%207th%20Edition%20(iBIMOne.com).pdf (accessed on 10 November 2024).
2. EU. Project Management Methodology Guide Open Edition. 2018. Available online: http://eeas.europa.eu/delegations/index_en.htm (accessed on 10 November 2024).
3. Turner, J.R.; Xue, Y. On the success of megaprojects. *Int. J. Manag. Proj. Bus.* **2018**, *11*, 783–805. [CrossRef]
4. Uddin, S.; Ong, S.; Lu, H.H. Machine learning in project analytics: A data-driven framework and case study. *Sci. Rep.* **2022**, *12*, 15252. [CrossRef] [PubMed]
5. Poole, D.L.; Mackworth, A.K. Artificial Intelligence: Foundations of Computational Agents. 2010. Available online: https://books.google.gr/books?hl=el&lr=&id=B7khAwAAQBAJ&oi=fnd&pg=PR7&dq=5.+Poole,+D.+L.,+&+Mackworth,+A.+K.+(2010).+Artificial+Intelligence:+Foundations+of+Computational+Agents&ots=_5d-NKEPiv&sig=2T1FNOa8LUJqqVraLwBeRMNNmvw&redir_esc=y#v=onepage&q&f=false (accessed on 11 November 2024).
6. Shang, G.; Low, S.P.; Lim, X.Y.V. Prospects, drivers of and barriers to artificial intelligence adoption in project management. *Built Environ. Proj. Asset Manag.* **2023**, *13*, 629–645. [CrossRef]
7. Costa, R.L.; da Dias, Á.L.; Gonçalves, R.; Pereira, L.; Abreu, S. Artificial Intelligence in Project Management: Systematic Literature Review. *Int. J. Technol. Intell. Plan.* **2022**, *13*, 143–163. [CrossRef]
8. Munir, M. How Artificial Intelligence can help Project Managers. *Glob. J. Manag. Bus. Res.* **2019**, *19*, 29–35.
9. Ong, S.; Uddin, S. Data science and artificial intelligence in project management: The past, present and future. *J. Mod. Proj. Manag.* **2020**, *7*, 26–33. [CrossRef]
10. Somasundaram, M.; Mohamed Junaid, K.A.; Mangadu, S. Artificial intelligence (AI) enabled intelligent quality management system (IQMS) for personalized learning path. *Procedia Comput. Sci.* **2020**, *172*, 438–442. [CrossRef]
11. Magaña Martínez, D.; Fernandez-Rodriguez, J.C. Artificial Intelligence Applied to Project Success: A Literature Review. *Int. J. Interact. Multimed. Artif. Intell.* **2015**, *3*, 77. [CrossRef]
12. Pinto, J.K. Critical Success Factors Across the Project Life Cycle. 1988. Available online: https://www.researchgate.net/publication/236175751 (accessed on 16 December 2024).
13. Abylova, L. Critical Success Factors in Project Management: A Comprehensive Review. In Proceedings of the 8th Scientific Conference on Project Management in the Baltic States, Riga, Latvia, 25–26 April 2019; Available online: www.pmworldlibrary.net (accessed on 6 November 2024).
14. Nenni, M.E.; de Felice, F.; de Luca, C.; Forcina, A. How artificial intelligence will transform project management in the age of digitization: A systematic literature review. *Manag. Rev. Q.* **2024**, 1–48. [CrossRef]
15. Taboada, I.; Daneshpajouh, A.; Toledo, N.; de Vass, T. Artificial Intelligence Enabled Project Management: A Systematic Literature Review. *Appl. Sci.* **2023**, *13*, 5014. [CrossRef]
16. Tranfield, D.; Denyer, D.; Smart, P. Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review* Introduction: The need for an evidence- informed approach. *Br. J. Manag.* **2003**, *14*, 207–222. [CrossRef]
17. Denicol, J.; Davies, A.; Krystallis, I. What Are the Causes and Cures of Poor Megaproject Performance? A Systematic Literature Review and Research Agenda. *Proj. Manag. J.* **2020**, *51*, 328–345. [CrossRef]
18. Mengist, W.; Soromessa, T.; Legese, G. Method for conducting systematic literature review and meta-analysis for environmental science research. *MethodsX* **2020**, *7*, 100777. [CrossRef]
19. 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. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
20. Prasetyo, M.L.; Peranginangin, R.A.; Martinovic, N.; Ichsan, M.; Wicaksono, H. Artificial intelligence in open innovation project management: A systematic literature review on technologies, applications, and integration requirements. *J. Open Innov. Technol. Mark. Complex.* **2025**, *11*, 100445. [CrossRef]
21. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; Antes, G.; Atkins, D.; Barbour, V.; Barrowman, N.; Berlin, J.A.; Clark, J.; et al. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Ann. Intern. Med.* **2009**, *151*, 264–269. [CrossRef]
22. Liberati, A.; Altman, D.G.; Tetzlaff, J.; Mulrow, C.; Gøtzsche, P.C.; Ioannidis, J.P.; Clarke, M.; Devereaux, P.J.; Kleijnen, J.; Moher, D. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Med.* **2009**, *6*, e1000100. [CrossRef]
23. Gemuenden, H.G.; Schoper, Y. Future Trends in Project Management. *J. View Proj. Relatsh. Mark. Teams View Proj.* **2015**.

24. A guide to the Project Management Body of Knowledge (PMBOK Guide). Project Management Institute. 2017. Available online: https://prothoughts.co.in/wp-content/uploads/2022/06/a-guide-to-the-project-management-body-of-knowledge-6e.pdf (accessed on 20 December 2024).
25. Elazouni, A.; Salem, O.A. Progress monitoring of construction projects using pattern recognition techniques. *Constr. Manag. Econ.* **2011**, *29*, 355–370. [CrossRef]
26. Cheng, M.-Y.; Tsai, H.-C.; Sudjono, E. Evolutionary fuzzy hybrid neural network for dynamic project success assessment in construction industry. *Autom. Constr.* **2012**, *21*, 46–51. [CrossRef]
27. Colin, J.; Martens, A.; Vanhoucke, M.; Wauters, M. A multivariate approach for top-down project control using earned value management. *Decis. Support Syst.* **2015**, *79*, 65–76. [CrossRef]
28. Salari, M.; Khamooshi, H. A better project performance prediction model using fuzzy time series and data envelopment analysis. *J. Oper. Res. Soc.* **2016**, *67*, 1274–1287. [CrossRef]
29. Chang, Y.C.; Shih, C.W.; Hsu, W.L. Entailment-Based Intelligent System for Software Project Monitoring and Control. *IEEE Syst. J.* **2018**, *12*, 216–227. [CrossRef]
30. Leon, H.; Osman, H.; Georgy, M.; Elsaid, M. System Dynamics Approach for Forecasting Performance of Construction Projects. *J. Manag. Eng.* **2018**, *34*, 04017049. [CrossRef]
31. Cheng, M.-Y.; Wu, Y.-W.; Huang, C.-C. Hybrid Gaussian Process Inference Model for Construction Management Decision Making. *Int. J. Inf. Technol. Decis. Mak.* **2020**, *19*, 1015–1036. [CrossRef]
32. Fayek, A.R. Fuzzy Logic and Fuzzy Hybrid Techniques for Construction Engineering and Management. *J. Constr. Eng. Manag.* **2020**, *146*, 04020064. [CrossRef]
33. Pour Rahimian, F.; Seyedzadeh, S.; Oliver, S.; Rodriguez, S.; Dawood, N. On-demand monitoring of construction projects through a game-like hybrid application of BIM and machine learning. *Autom. Constr.* **2020**, *110*, 103012. [CrossRef]
34. Amer, F.; Jung, Y.; Golparvar-Fard, M. Transformer machine learning language model for auto-alignment of long-term and short-term plans in construction. *Autom. Constr.* **2021**, *132*, 103929. [CrossRef]
35. Shameem, M.; Nadeem, M.; Zamani, A.T. Genetic algorithm based probabilistic model for agile project success in global software development. *Appl. Soft Comput.* **2023**, *135*, 109998. [CrossRef]
36. Uddin, S.; Ong, S.; Lu, H.; Matous, P. Integrating machine learning and network analytics to model project cost, time and quality performance. *Prod. Plan. Control.* **2020**, *12*, 1475–1489. [CrossRef]
37. Kim, C.W.; Yoo, W.S.; Lim, H.; Yu, I.; Cho, H.; Kang, K.I. Early-warning performance monitoring system (EPMS) using the business information of a project. *Int. J. Proj. Manag.* **2018**, *36*, 730–743. [CrossRef]
38. Shah, M.A.; Jawawi DN, A.; Isa, M.A.; Younas, M.; Abdelmaboud, A.; Sholichin, F. Ensembling Artificial Bee Colony With Analogy-Based Estimation to Improve Software Development Effort Prediction. *IEEE Access* **2020**, *8*, 58402–58415. [CrossRef]
39. de Carvalho, H.D.P.; Fagundes, R.; Santos, W. Extreme Learning Machine Applied to Software Development Effort Estimation. *IEEE Access* **2021**, *9*, 92676–92687. [CrossRef]
40. Mavrovouniotis, M.; Anastasiadou, M.N.; Hadjimitsis, D. Measuring the Performance of Ant Colony Optimization Algorithms for the Dynamic Traveling Salesman Problem. *Algorithms* **2023**, *16*, 545. [CrossRef]
41. Azar, D.; Vybihal, J. An ant colony optimization algorithm to improve software quality prediction models: Case of class stability. *Inf. Softw. Technol.* **2011**, *53*, 388–393. [CrossRef]
42. Sun, C.; Xu, J. Estimation of Time for Wenchuan Earthquake Reconstruction in China. *J. Constr. Eng. Manag.* **2011**, *137*, 179–187. [CrossRef]
43. Jha, K.N.; Chockalingam, C. Prediction of schedule performance of Indian construction projects using an artificial neural network. *Constr. Manag. Econ.* **2011**, *29*, 901–911. [CrossRef]
44. Irfan, M.; Khurshid, M.B.; Anastasopoulos, P.; Labi, S.; Moavenzadeh, F. Planning-stage estimation of highway project duration on the basis of anticipated project cost, project type, and contract type. *Int. J. Proj. Manag.* **2011**, *29*, 78–92. [CrossRef]
45. Agarwal, A.; Colak, S.; Erenguc, S. A Neurogenetic approach for the resource-constrained project scheduling problem. *Comput. Oper. Res.* **2011**, *38*, 44–50. [CrossRef]
46. Chou, J.S.; Cheng, M.Y.; Wu, Y.W.; Wu, C.C. Forecasting enterprise resource planning software effort using evolutionary support vector machine inference model. *Int. J. Proj. Manag.* **2012**, *30*, 967–977. [CrossRef]
47. Sadeghi, N.; Fayek, A.R.; Seresht, N.G. A fuzzy discrete event simulation framework for construction applications: Improving the simulation time advancement. *J. Constr. Eng. Manag.* **2016**, *142*, 04016071. [CrossRef]
48. Wauters, M.; Vanhoucke, M. A Nearest Neighbour extension to project duration forecasting with Artificial Intelligence. *Eur. J. Oper. Res.* **2017**, *259*, 1097–1111. [CrossRef]
49. Golizadeh, H.; Banihashemi, S.; Sadeghifam, A.N.; Preece, C. Automated estimation of completion time for dam projects. *Int. J. Constr. Manag.* **2017**, *17*, 197–209. [CrossRef]

50. Choetkiertikul, M.; Dam, H.K.; Tran, T.; Ghose, A. Predicting the delay of issues with due dates in software projects. *Empir. Softw. Eng.* **2017**, *22*, 1223–1263. [CrossRef]
51. Pospieszny, P.; Czarnacka-Chrobot, B.; Kobylinski, A. An effective approach for software project effort and duration estimation with machine learning algorithms. *J. Syst. Softw.* **2018**, *137*, 184–196. [CrossRef]
52. Cheng, M.-Y.; Hoang, N.-D. Estimating construction duration of diaphragm wall using firefly-tuned least squares support vector machine. *Neural Comput. Appl.* **2018**, *30*, 2489–2497. [CrossRef]
53. Yaseen, Z.M.; Ali, Z.H.; Salih, S.Q.; Al-Ansari, N. Prediction of Risk Delay in Construction Projects Using a Hybrid Artificial Intelligence Model. *Sustainability* **2020**, *12*, 1514. [CrossRef]
54. Gondia, A.; Siam, A.; El-Dakhakhni, W.; Nassar, A.H. Machine Learning Algorithms for Construction Projects Delay Risk Prediction. *J. Constr. Eng. Manag.* **2020**, *146*, 04019085. [CrossRef]
55. Awada, M.; Srour, F.J.; Srour, I.M. Data-Driven Machine Learning Approach to Integrate Field Submittals in Project Scheduling. *J. Manag. Eng.* **2021**, *37*, 04020104. [CrossRef]
56. Zhang, S.; Wang, X. Dynamic Probability Analysis for Construction Schedule Using Subset Simulation. *Adv. Civ. Eng.* **2021**, *2021*, 1567261. [CrossRef]
57. Sai, L.; Hao, W.Q. Forecasting the scheduling issues in engineering project management: Applications of deep learning models. *Future Gener. Comput. Syst.-Int. J. Esci.* **2021**, *123*, 85–93. [CrossRef]
58. Chakrabortty, R.K.; Abbasi, A.; Ryan, M.J. A Risk Assessment Framework for Scheduling Projects With Resource and Duration Uncertainties. *IEEE Trans. Eng. Manag.* **2022**, *69*, 1917–1931. [CrossRef]
59. Sousa, A.O.; Veloso, D.T.; Gonçalves, H.M.; Faria, J.P.; Mendes-Moreira, J.; Graça, R.; Gomes, D.; Castro, R.N.; Henriques, P.C. Applying Machine Learning to Estimate the Effort and Duration of Individual Tasks in Software Projects. *IEEE Access* **2023**, *11*, 89933–89946. [CrossRef]
60. Liang, A.; Tao, L.; Lei, H. Combined machine-learning and EDM to monitor and predict a complex project with a GERT-type network: A multi-point perspective. *Comput. Ind. Eng.* **2023**, *180*, 109256. [CrossRef]
61. Jaafar, K.; Watfa, M.; Aloran, A. Framework for a Predictive Progress Model—Case of infrastructure projects. *Int. J. Manag. Sci. Eng. Manag.* **2023**, *18*, 157–169. [CrossRef]
62. Khan, U.U.; Ali, Y.; Garai-Fodor, M.; Csiszárik-Kocsir, A. Application of Project Management Techniques for Timeline and Budgeting Estimates of Startups. *Sustainability* **2023**, *15*, 15526. [CrossRef]
63. Yang, J.-B.; Chen, Y.-H. Enhanced ESM approach for duration forecasting in delayed projects. *Ain Shams Eng. J.* **2024**, *15*, 102472. [CrossRef]
64. Cheng, M.Y.; Roy, A.F.V. Evolutionary fuzzy decision model for cash flow prediction using time-dependent support vector machines. *Int. J. Proj. Manag.* **2011**, *29*, 56–65. [CrossRef]
65. Hwang, S. Time series models for forecasting construction costs using time series indexes. *J. Constr. Eng. Manag.* **2011**, *137*, 656–662. [CrossRef]
66. Cheng, M.-Y.; Hoang, N.-D.; Roy AF, v.; Wu, Y.-W. A novel time-depended evolutionary fuzzy SVM inference model for estimating construction project at completion. *Eng. Appl. Artif. Intell.* **2012**, *25*, 744–752. [CrossRef]
67. Nassif, A.B.; Ho, D.; Capretz, L.F. Towards an early software estimation using log-linear regression and a multilayer perceptron model. *J. Syst. Softw.* **2013**, *86*, 144–160. [CrossRef]
68. Chen, H.L. Improving Forecasting Accuracy of Project Earned Value Metrics: Linear Modeling Approach. *J. Manag. Eng.* **2014**, *30*, 135–145. [CrossRef]
69. Kim, B.C. Integrating Risk Assessment and Actual Performance for Probabilistic Project Cost Forecasting: A Second Moment Bayesian Model. *IEEE Trans. Eng. Manag.* **2015**, *62*, 158–170. [CrossRef]
70. Cao, M.-T.; Cheng, M.-Y.; Wu, Y.-W. Hybrid computational model for forecasting Taiwan construction cost index. *J. Constr. Eng. Manag.* **2015**, *141*, 04014089. [CrossRef]
71. Nassif, A.B.; Azzeh, M.; Capretz, L.F.; Ho, D. Neural network models for software development effort estimation: A comparative study. *Neural Comput. Appl.* **2016**, *27*, 2369–2381. [CrossRef]
72. Batselier, J.; Vanhoucke, M. Improving project forecast accuracy by integrating earned value management with exponential smoothing and reference class forecasting. *Int. J. Proj. Manag.* **2017**, *35*, 28–43. [CrossRef]
73. Xue, X.; Jia, Y.; Tang, Y. Expressway Project Cost Estimation with a Convolutional Neural Network Model. *IEEE Access* **2020**, *8*, 217848–217866. [CrossRef]
74. Moon, H.; Williams, T.P.; Lee, H.-S.; Park, M. Predicting project cost overrun levels in bidding stage using ensemble learning. *J. Asian Archit. Build. Eng.* **2020**, *19*, 586–599. [CrossRef]
75. Mir, M.; Kabir, H.M.D.; Nasirzadeh, F.; Khosravi, A. Neural network-based interval forecasting of construction material prices. *J. Build. Eng.* **2021**, *39*, 102288. [CrossRef]

76. Xu, X.; Peng, L.; Ji, Z.; Zheng, S.; Tian, Z.; Geng, S. Research on substation project cost prediction based on sparrow search algorithm optimized BP neural network. *Sustainability* **2021**, *13*, 13746. [CrossRef]
77. Kim, J.; Cha, H.S. Expediting the Cost Estimation Process for Aged-Housing Renovation Projects Using a Probabilistic Deep Learning Approach. *Sustainability* **2022**, *14*, 564. [CrossRef]
78. Castro Miranda, S.L.; del Rey Castillo, E.; Gonzalez, V.; Adafin, J. Predictive Analytics for Early-Stage Construction Costs Estimation. *Buildings* **2022**, *12*, 1043. [CrossRef]
79. Lee, J.G.; Lee, H.S.; Park, M.; Seo, J. Early-stage cost estimation model for power generation project with limited historical data. *Eng. Constr. Archit. Manag.* **2022**, *29*, 2599–2614. [CrossRef]
80. Disch, M. Estimating the Costs of Product Development Projects Based on External Data: The New Product Development Cost Benchmarking Method. *IEEE Trans. Eng. Manag.* **2023**, *71*, 5117–5128. [CrossRef]
81. Al Kailani, H.; Sweis, G.J.; Sammour, F.; Maaitah, W.O.; Sweis, R.J.; Alkailani, M. Predicting construction cost index using fuzzy logic and machine learning in Jordan. *Constr. Innov.* **2024**. [CrossRef]
82. Chen, N.; Hoi, S.C.H.; Xiao, X. Software process evaluation: A machine learning framework with application to defect management process. *Empir. Softw. Eng.* **2013**, *19*, 1531–1564. [CrossRef]
83. Juneja, K. A fuzzy-filtered neuro-fuzzy framework for software fault prediction for inter-version and inter-project evaluation. *Appl. Soft Comput. J.* **2019**, *77*, 696–713. [CrossRef]
84. Dikmen, I.; Qazi, A.; Erol, H.; Birgonul, M.T. Meta-Modeling of Complexity-Uncertainty-Performance Triad in Construction Projects. *EMJ—Eng. Manag. J.* **2020**, *33*, 30–44. [CrossRef]
85. Dumitrascu-Băldău, I.; Dumitrascu, D.D.; Dobrota, G. Predictive model for the factors influencing international project success: A data mining approach. *Sustainability* **2021**, *13*, 3819. [CrossRef]
86. Rashidi, A.; Jazebi, F.; Brilakis, I. Neurofuzzy genetic system for selection of construction project managers. *J. Constr. Eng. Manag.* **2011**, *137*, 17–29. [CrossRef]
87. Heravi, G.; Eslamdoost, E. Applying Artificial Neural Networks for Measuring and Predicting Construction-Labor Productivity. *J. Constr. Eng. Manag.* **2015**, *141*, 04015032. [CrossRef]
88. Patel, D.A.; Jha, K.N. Neural network model for the prediction of safe work behavior in construction projects. *J. Constr. Eng. Manag.* **2015**, *141*, 04014066. [CrossRef]
89. El-Gohary, K.M.; Aziz, R.F.; Abdel-Khalek, H.A. Engineering Approach Using ANN to Improve and Predict Construction Labor Productivity under Different Influences. *J. Constr. Eng. Manag.* **2017**, *143*, 04017045. [CrossRef]
90. Mo, Y.; Zhao, D.; Du, J.; Syal, M.; Aziz, A.; Li, H. Automated staff assignment for building maintenance using natural language processing. *Autom. Constr.* **2020**, *113*, 103150. [CrossRef]
91. Arora, M.; Verma, S.; Kavita Wozniak, M.; Shafi, J.; Ijaz, M.F. An efficient ANFIS-EEBAT approach to estimate effort of Scrum projects. *Sci. Rep.* **2022**, *12*, 7974. [CrossRef]
92. Ronghui, S.; Liangrong, N. An intelligent fuzzy-based hybrid metaheuristic algorithm for analysis the strength, energy and cost optimization of building material in construction management. *Eng. Comput.* **2022**, *38*, 2663–2680. [CrossRef]
93. Jiang, C.; Li, X.; Lin, J.-R.; Liu, M.; Ma, Z. Adaptive control of resource flow to optimize construction work and cash flow via online deep reinforcement learning. *Autom. Constr.* **2023**, *150*, 104817. [CrossRef]
94. Shafiabady, N.; Hadjinicolaou, N.; Din, F.U.; Bhandari, B.; Wu, R.M.X.; Vakilian, J. Using Artificial Intelligence (AI) to predict organizational agility. *PLoS ONE* **2023**, *18*, e0283066. [CrossRef]
95. Jin, X.H.; Zhang, G.M. Modelling optimal risk allocation in PPP projects using artificial neural networks. *Int. J. Proj. Manag.* **2011**, *29*, 591–603. [CrossRef]
96. de Oliveira, M.A.; Valentina, L.; Possamai, O. Forecasting project performance considering the influence of leadership style on organizational agility. *Int. J. Product. Perform. Manag.* **2012**, *61*, 653–671. [CrossRef]
97. Wang, Y.R.; Yu, C.Y.; Chan, H.H. Predicting construction cost and schedule success using artificial neural networks ensemble and support vector machines classification models. *Int. J. Proj. Manag.* **2012**, *30*, 470–478. [CrossRef]
98. Fang, C.; Marle, F. A simulation-based risk network model for decision support in project risk management. *Decis. Support Syst.* **2012**, *52*, 635–644. [CrossRef]
99. Hu, Y.; Du, J.; Zhang, X.; Hao, X.; Ngai, E.W.T.; Fan, M.; Liu, M. An integrative framework for intelligent software project risk planning. *Decis. Support Syst.* **2013**, *55*, 927–937. [CrossRef]
100. Shi, Q.; Zhou, Y.K.; Xiao, C.; Chen, R.Y.; Zuo, J. Delivery risk analysis within the context of program management using fuzzy logic and DEA: A China case study. *Int. J. Proj. Manag.* **2014**, *32*, 341–349. [CrossRef]
101. Nasirzadeh, F.; Khanzadi, M.; Rezaie, M. Dynamic modeling of the quantitative risk allocation in construction projects. *Int. J. Proj. Manag.* **2014**, *32*, 442–451. [CrossRef]

102. Gunduz, M.; Nielsen, Y.; Ozdemir, M. Fuzzy Assessment Model to Estimate the Probability of Delay in Turkish Construction Projects. *J. Manag. Eng.* **2015**, *31*, 04014055. [CrossRef]
103. Rodríguez, A.; Ortega, F.; Concepción, R. A method for the evaluation of risk in IT projects. *Expert Syst. Appl.* **2016**, *45*, 273–285. [CrossRef]
104. Xia, N.; Zou, P.X.W.; Liu, X.; Wang, X.; Zhu, R. A hybrid BN-HFACS model for predicting safety performance in construction projects. *Saf. Sci.* **2018**, *101*, 332–343. [CrossRef]
105. Kataev, M.; Bulysheva, L.; Xu, L.D.; Ekhlakov, Y.; Permyakova, N.; Jovanovic, V. Fuzzy model estimation of the risk factors impact on the target of promotion of the software product. *Enterp. Inf. Syst.* **2020**, *14*, 797–811. [CrossRef]
106. Aladağ, H.; Işik, Z. The Effect of Stakeholder-Associated Risks in Mega-Engineering Projects: A Case Study of a PPP Airport Project. *IEEE Trans. Eng. Manag.* **2020**, *67*, 174–186. [CrossRef]
107. Bashir, H.; Ojiako, U.; Mota, C. Modeling and Analyzing Factors Affecting Project Delays Using an Integrated Social Network-Fuzzy MICMAC Approach. *EMJ—Eng. Manag. J.* **2020**, *32*, 26–36. [CrossRef]
108. Assaad, R.; El-Adaway, I.H.; Abotaleb, I.S. Predicting Project Performance in the Construction Industry. *J. Constr. Eng. Manag.* **2020**, *146*, 04020030. [CrossRef]
109. Kim, J.M.; Bae, J.; Son, S.; Son, K.; Yum, S.G. Development of Model to Predict Natural Disaster-Induced Financial Losses for Construction Projects Using Deep Learning Techniques. *Sustainability* **2021**, *13*, 5304. [CrossRef]
110. Ayhan, M.; Dikmen, I.; Talat Birgonul, M. Predicting the Occurrence of Construction Disputes Using Machine Learning Techniques. *J. Constr. Eng. Manag.* **2021**, *147*, 04021022. [CrossRef]
111. Chattapadhyay, D.B.; Putta, J.; Rama Mohan Rao, P. Risk identification, assessments, and prediction for mega construction projects: A risk prediction paradigm based on cross analytical-machine learning model. *Buildings* **2021**, *11*, 172. [CrossRef]
112. Islam, M.S.; Mohandes, S.R.; Mahdiyar, A.; Fallahpour, A.; Olanipekun, A.O. A Coupled Genetic Programming Monte Carlo Simulation-Based Model for Cost Overrun Prediction of Thermal Power Plant Projects. *J. Constr. Eng. Manag.* **2022**, *148*, 04022073. [CrossRef]
113. Koc, K.; Ekmekcioğlu, Ö.; Gurgun, A.P. Accident prediction in construction using hybrid wavelet-machine learning. *Autom. Constr.* **2022**, *133*, 103987. [CrossRef]
114. Kim, J.M.; Bae, J.; Park, H.; Yum, S.G. Predicting financial losses due to apartment construction accidents utilizing deep learning techniques. *Sci. Rep.* **2022**, *12*, 5365. [CrossRef]
115. Santos, J.I.; Pereda, M.; Ahedo, V.; Galán, J.M. Explainable machine learning for project management control. *Comput. Ind. Eng.* **2023**, *180*, 109261. [CrossRef]
116. Gondia, A.; Moussa, A.; Ezzeldin, M.; El-Dakhakhni, W. Machine learning-based construction site dynamic risk models. *Technol. Forecast. Soc. Chang.* **2023**, *189*, 122347. [CrossRef]
117. de Andrade, P.A.; Vanhoucke, M.; Martens, A. An empirical project forecasting accuracy framework using project regularity. *Ann. Oper. Res.* **2023**, *2*, 501–521. [CrossRef]
118. Chou, J.-S.; Cheng, M.-Y.; Wu, Y.-W. Improving classification accuracy of project dispute resolution using hybrid artificial intelligence and support vector machine models. *Expert Syst. Appl.* **2013**, *40*, 2263–2274. [CrossRef]
119. Chaphalkar, N.B.; Iyer, K.C.; Patil, S.K. Prediction of outcome of construction dispute claims using multilayer perceptron neural network model. *Int. J. Proj. Manag.* **2015**, *33*, 1827–1835. [CrossRef]
120. Soltan, S.; Ashrafi, M. Predicting project duration and cost, and selecting the best action plan using statistical methods for earned value management. *J. Proj. Manag.* **2020**, *5*, 157–166. [CrossRef]
121. Wauters, M.; Vanhoucke, M. Support Vector Machine Regression for project control forecasting. *Autom. Constr.* **2014**, *47*, 92–106. [CrossRef]
122. Habibi, F.; Birgani, O.T.; Koppelaar, H.; Radenovic, S.N. Using fuzzy logic to improve the project time and cost estimation based on Project Evaluation and Review Technique (PERT). *J. Proj. Manag.* **2018**, *3*, 183–196. [CrossRef]
123. de Andrade, P.A.; Martens, A.; Vanhoucke, M. Using real project schedule data to compare earned schedule and earned duration management project time forecasting capabilities. *Autom. Constr.* **2019**, *99*, 68–78. [CrossRef]
124. Giabbanelli, P.J.; Schneider, B.I.; Wainer, G.A. Hybrid Models That Combine Machine Learning and Simulations. *Comput. Sci. Eng.* **2022**, *24*, 72–76. [CrossRef]
125. Whang, S.E.; Roh, Y.; Song, H.; Lee, J.-G. Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective. *VLDB J.* **2021**, *32*, 791–813. [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.

computers

MDPI

Systematic Review

# Artificial Intelligence and the Future of Mental Health in a Digitally Transformed World

**Aggeliki Kelly Fanarioti and Kostas Karpouzis ***

Department of Communication, Media and Culture, Panteion University of Social and Political Sciences, GR-176 71 Athens, Greece; kellyyfanarioti@gmail.com
* Correspondence: kkarpou@panteion.gr

**Abstract:** Artificial Intelligence (AI) is reshaping mental healthcare by enabling new forms of diagnosis, therapy, and patient monitoring. Yet this digital transformation raises complex policy and ethical questions that remain insufficiently addressed. In this paper, we critically examine how AI-driven innovations are being integrated into mental health systems across different global contexts, with particular attention to governance, regulation, and social justice. The study follows the PRISMA-ScR methodology to ensure transparency and methodological rigor, while also acknowledging its inherent limitations, such as the emphasis on breadth over depth and the exclusion of non-English sources. Drawing on international guidelines, academic literature, and emerging national strategies, it identifies both opportunities, such as improved access and personalized care, and threats, including algorithmic bias, data privacy risks, and diminished human oversight. Special attention is given to underrepresented populations and the risks of digital exclusion. The paper argues for a value-driven approach that centers equity, transparency, and informed consent in the deployment of AI tools. It concludes with actionable policy recommendations to support the ethical implementation of AI in mental health, emphasizing the need for cross-sectoral collaboration and global accountability mechanisms.

**Keywords:** artificial intelligence; digital health systems; ethical AI; health informatics; AI governance; human-centered computing; policy-aware computing

## 1. Introduction

Artificial Intelligence is rapidly changing the field of mental health, offering methods and tools that support diagnosis, personalize treatment, and extend access to care through digital platforms. With applications ranging from conversational agents that deliver cognitive-behavioral interventions to predictive models using biometric data, AI is enabling new forms of mental health support [1]; these advances are situated within the broader digital transformation of healthcare, marked by data-driven technologies, remote monitoring, and decision-support tools in clinical and community settings [2].

This transformation comes at a time of acute systemic strain: mental health conditions are rising globally, exacerbated by economic stress, environmental instability, and sociotechnical disruption. The COVID-19 pandemic highlighted both the urgency of improving mental health services and the potential of digital tools to address access gaps [3]. However, despite these advances, major questions remain about how AI can be ethically and equitably deployed in mental health systems.

Concerns include the protection of sensitive personal data, transparency of algorithmic decision making, the potential for bias and harm, and the social impact of replacing or augmenting human care with automated systems [4,5]. These challenges are especially acute in mental health, where context, trust, and human judgment are often integral to effective care. Furthermore, deployment environments are deeply unequal: while some countries are experimenting with AI-assisted diagnostics, others struggle with basic digital infrastructure or mental health workforce shortages.

Policy responses are beginning to address these risks. The World Health Organization's Global Strategy on Digital Health 2020–2025 promotes the ethical use of digital tools in health, emphasizing human rights, inclusive governance, and equitable access [6], while the European Union's Artificial Intelligence Act classifies AI systems used in healthcare as "high-risk," requiring extensive transparency, oversight, and compliance with fundamental rights [7]. Yet, many frameworks remain aspirational or underspecified in how ethical principles translate into practice—particularly in low-resource or decentralized health systems.

In this paper, we offer a strategic review of AI in mental health, viewed through a policy and ethics lens. Unlike studies that focus on technical capabilities or clinical validation, this analysis discusses how AI systems are governed, the values they encode, and the means that international institutions and national governments can deploy to respond to their challenges, aiming to identify both the opportunities and structural limitations that shape AI integration in mental health services. The structure of the paper is as follows. Section 2 surveys related work and conceptual frameworks from both the computer science and health ethics literature, Section 3 describes current AI applications in mental health, including diagnostic support, virtual care platforms, and affective computing tools, and Section 4 examines international and national policy strategies, with a case study on Greece. Then, Sections 5 and 6 focus on ethical and governance challenges, including data protection, algorithmic bias, and digital exclusion; the paper concludes with recommendations for ethically grounded and policy-aware deployment of AI in mental healthcare.

*Methodology*

This paper is based on a strategic review originally developed as part of a graduate dissertation examining the intersection of AI, digital transformation, and mental health policy. In this paper, we have extended the original treatise with the additional steps included in the PRISMA [8] extension for scoping reviews (PRISMA-ScR) (PRISMA for Scoping Reviews (PRISMA-ScR), https://www.prisma-statement.org/scoping, last accessed: 31 May 2025). This section presents information related to the methods section of the checklist, results are outlined in Section 4, and a discussion of the results takes place in Sections 5 and 6.

The review process combined structured literature searches with thematic analysis of international and national policy documents. Academic sources were identified through targeted searches on PubMed, Scopus, and Google Scholar using the following keywords: "artificial intelligence," "mental health," "digital transformation," "algorithmic bias," and "health ethics". We narrowed the search to peer-reviewed articles published after 2018 to increase relevance to our research scope ($n$ = 7577), and removed entries repeating across the three databases ($n$ = 5402).

During screening, we removed papers which referred to only a subset of query keywords ($n$ = 1910) and assessed the remaining 260 papers for relevance. During this process, we found that a large number of them ($n$ = 210) mentioned Digital Transformation briefly,

while others utilized conventional rule-based methods, e.g., prescripted interaction ($n$ = 30). Finally, we separated the remaining 20 entries to 16 new studies and 4 reviews or meta-reports. This process is summarized in Figure 1.

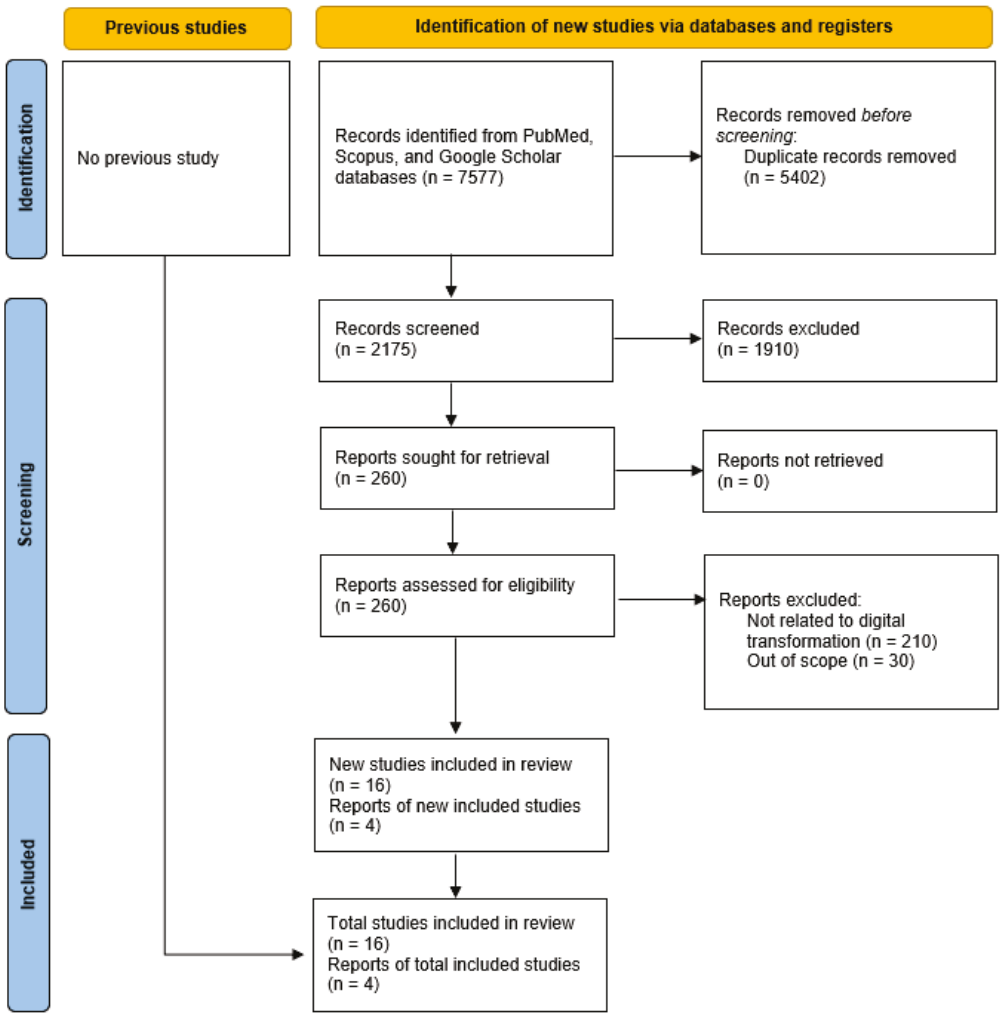



**Figure 1.** Summary of PRISMA-related information.

In parallel, current policy materials were drawn from the World Health Organization, European Commission, UNICEF, and national health strategies, with a focus on documents that explicitly reference digital mental health or AI regulation. Grey literature [9] from civil society and intergovernmental bodies was also reviewed to capture emerging concerns around equity, surveillance, and participatory design.

Then, sources were analyzed thematically with particular attention to values of fairness, transparency, inclusivity, and accountability. These themes informed the structure of the paper, which moves from a review of technological applications to an analysis of ethical and governance challenges, concluding with strategic recommendations for responsible AI integration in mental health care.

## 2. Related Work and Conceptual Background

The possible impact of artificial intelligence on mental health treatment has generated growing academic interest, particularly in relation to diagnostics, treatment delivery, and self-monitoring applications ([10,11]). A substantial body of work explores how ma-

chine learning (ML) and natural language processing (NLP) are applied in psychiatry and digital phenotyping, offering tools that aim to improve prediction, classification, and personalization of mental healthcare [2,12]. These developments have been supported by the proliferation of mobile health (mHealth) applications and wearable technologies that capture behavioral and physiological data relevant to mental health assessment.

Despite these advances, the integration of AI into mental health systems has raised a number of ethical, legal, and policy-related concerns. Researchers have questioned the adequacy of existing data protection regimes, the risks of algorithmic discrimination or bias, the transparency of decision-making processes, and the broader social implications of substituting human interaction with automated tools [4,5]. In particular, mental health presents distinctive ethical challenges due to the sensitivity of patient data, the potential for stigma, and the vulnerability of user populations. The literature has called for frameworks that center human dignity, autonomy, and the principle of "do no harm" in AI design and deployment [7].

From a policy perspective, several institutions have developed guidance documents and strategic roadmaps aimed at governing digital mental health technologies. The World Health Organization's Global Strategy on Digital Health 2020–2025 underscores the importance of global coordination and equity in digital health innovation, while emphasizing the need to strengthen digital governance systems [6]. At the European level, the Artificial Intelligence Act proposes a risk-based regulatory framework that treats AI tools in healthcare as high-risk applications, requiring thorough documentation, algorithmic transparency, and human oversight.

Beyond institutional policy, critical scholars have highlighted the limitations of technocentric approaches and called for a more nuanced understanding of the political and social contexts in which AI systems operate. Sharon [13], for instance, argued that digital health infrastructures risk amplifying existing inequities unless questions of trust, participation, and power asymmetries are explicitly addressed. Likewise, the governance of digital mental health must contend with issues such as cross-border data flows, private sector involvement, and the digital divide—especially when implemented in low-resource settings.

In the following sections, we build upon these contributions by offering a strategic review that synthesizes ethical discourse, regulatory initiatives, and implementation challenges in the context of AI-driven mental health. Our work positions itself at the intersection of computer science, global health policy, and applied ethics, proposing a framework for assessing readiness and responsibility in the deployment of mental health AI technologies.

## 3. Digital Transformation and AI in Mental Health

Digital transformation in mental health refers to the integration of information and communication technologies, including AI-based systems, into mental health services. This transformation is not limited to the digitization of existing practices, but increasingly involves the development of novel tools that expand, automate, or decentralize care delivery, and sometimes even the redesign of established practices and workflows from scratch. AI systems now support a range of functions including diagnosis, therapy, monitoring, triage, and patient engagement [2,14]. Among the most widely adopted technologies are telehealth platforms that facilitate synchronous (e.g., video consultations) and asynchronous (e.g., app-based messaging) interactions between mental health professionals and patients. These systems have been particularly effective in increasing access for underserved populations in rural or remote areas, especially during the COVID-19 pandemic [1].

AI-assisted diagnostics represent a second major area of growth. Machine learning models are used to analyze speech, facial expressions, social media activity, or wearable data to detect early signs of depression, anxiety, and other mood disorders [2,12]. Natural language processing (NLP) and affective computing techniques have been applied to identify biomarkers of distress or suicidal ideation in textual and audio inputs.

Digital therapeutics and chatbot-based interventions form a third important category. Tools such as Woebot and Wysa offer conversational cognitive behavioral therapy (CBT), mindfulness prompts, and psychoeducation without human mediation [15]. These systems are increasingly used in low-threshold, self-help contexts and for population-level mental health promotion, though their clinical validation remains limited.

Beyond emotional detection, some AI systems now simulate companionship and intimacy, creating the illusion of affective reciprocity. Applications like Replika and Character.AI illustrate how generative models can engage users in sustained, emotionally charged interactions that feel socially meaningful, despite lacking genuine understanding or care. These systems raise new ethical questions about attachment, dependency, and emotional vulnerability. If such a system is reprogrammed, monetized, or withdrawn, as occurred with Replika's removal of romantic features (The confusing reality of AI friends, https://www.theverge.com/c/24300623/ai-companions-replika-openai-chatgpt-assistant-romance, last accessed: 31 May 2025), users may experience real psychological distress. The boundaries between therapeutic support, entertainment, and manipulation become increasingly blurred.

**Monitoring and prediction systems** use mobile sensors, app data, and biometric inputs to track mood patterns, sleep quality, and behavioral anomalies. These can be combined with predictive analytics to anticipate depressive episodes or treatment dropouts. Predictive systems are being explored as components of stepped-care models and relapse prevention strategies [16]. Notable examples of this category include Stepped Care 2.0 (Stepped Care 2.0, https://mentalhealthcommission.ca/what-we-do/access/stepped-care-2-0/, last accessed: 31 May 2025), an evidence-informed model that integrates in-person and virtual care, SilverCloud (SilverCloud, https://www.silvercloudhealth.com/uk/programmes/tools/stepped-care-model, last accessed: 31 May 2025), a clinically tested digital platform which provides patient support from self-management to guided, evidence-based programs and secondary care, and TourHeart+, a web-based stepped-care platform which supports mental well-being by offering internet-based interventions [17].

Virtual reality (VR) and augmented reality (AR) environments are also being explored for use in exposure therapy, cognitive training, and social skill development. These immersive technologies have shown early promise in the treatment of anxiety disorders and PTSD [18], but the cost and the sociotechnical complexity of scaling to multiple users, remote locations, and varied curriculum may be a limiting factor [19].

Generative AI systems capable of producing realistic text, images, audio, and video introduce new mental health risks and regulatory challenges. While these tools do offer creative and therapeutic opportunities, they also enable the generation of potentially disturbing, violent, or delusional content, with individuals misusing them to reinforce negative self-images, simulate traumatic events, or isolate themselves within artificial realities. These dynamics can be especially harmful to users with existing vulnerabilities, including depression or psychosis, where distinctions between generated and lived experience may become psychologically destabilizing. Moreover, there is growing concern that generative AI contributes to disembodied identity formation, reality distortion, and compulsive content creation, all of which may require new forms of digital mental health literacy and governance [20].

Finally, integrated hybrid care models combine AI-supported interventions with traditional services. These models aim to maintain human oversight while leveraging automation to reduce costs, personalize care, and extend reach. However, hybridization raises its own challenges, including interoperability, continuity of care, and clinician workload management.

Despite their potential, most of these tools face critical implementation barriers, including lack of robust clinical validation, limited regulation of safety and efficacy, data governance challenges, and inconsistent integration into national health systems. Moreover, many applications are developed by private sector actors whose interests may not align with public health priorities [13].

*Case Studies*

In this subsection, we briefly review prominent digital mental health tools that have received public attention and funding, as a basis for discussion on the strengths and limitations of current AI-based interventions.

**Woebot**, a well-known chatbot grounded in principles of cognitive behavioral therapy, has demonstrated short-term effectiveness in alleviating anxiety and depressive symptoms, especially among younger users. Its user-friendly interface and availability outside traditional care settings have been lauded as important steps toward accessible psychological support. However, the system's reliance on scripted interactions and its inability to recognize or respond to crisis language (e.g., expressions of suicidality) have prompted concerns about safety and accountability in clinical contexts. Despite its therapeutic framing, Woebot does not operate under formal regulatory oversight as a medical device in most jurisdictions, and there is limited transparency about the boundaries of its functionality.

A similar pattern emerges with **Wysa**, an AI-enabled self-help platform increasingly integrated into workplace wellness programs and primary care settings. While it includes optional human coaching, the default experience is automated, and questions remain about the long-term efficacy of such hybrid models, particularly when deployed in underresourced systems. Ethical critiques have also highlighted insufficient clarity on data governance, especially in corporate use cases where sensitive mental health interactions may occur without fully informed consent or user awareness of third-party data access.

**Mindstrong** was a venture-backed U.S. startup that aimed to use smartphone metadata, such as typing speed and sleep patterns, to detect early signs of mental health deterioration. Their platform was eventually shut down in 2022 after internal concerns about scientific validity and a lack of demonstrable outcomes; this case underscores the difficulty of operationalizing predictive analytics in real-world mental health contexts, where noisy data, ethical ambiguity, and the need for clinical nuance often outpace the promises of AI-driven optimization.

These examples illustrate a broader tension in the digital mental health space: AI systems are increasingly positioned as scalable solutions to long-standing mental health access gaps, yet many of them are introduced without independent evaluation, longitudinal evidence, or mechanisms for participatory user feedback. The asymmetry between technical development and ethical infrastructure risks eroding trust among users and practitioners alike. If these systems are to play a meaningful role in future mental health ecosystems, they must be embedded within robust frameworks of clinical accountability, regulatory clarity, and co-designed development processes that center the experiences and rights of users—not just the ambitions of innovation.

When taken together, these three cases reveal a structural pattern in the deployment of AI systems for mental health care: while technical innovation is often rapid, visible, and

well-funded, the ethical, clinical, and regulatory infrastructures necessary to support these systems tend to lag behind. Woebot demonstrates how even well-intentioned tools can inadvertently give a false sense of safety and therapeutic adequacy when deployed without formal oversight or safeguards for crisis scenarios, while Wysa highlights concerns about data transparency and user agency, particularly in commercial or institutional contexts where the platform may serve multiple, and potentially conflicting, stakeholders. Mindstrong, by contrast, shows the consequences of overpromising the predictive capabilities of AI in high-stakes domains without robust scientific validation or sustainable implementation models. Despite their different approaches and outcomes, all three underscore the same core lesson: without synchronized development between technological capability and ethical governance, such tools risk undermining user trust, misrepresenting clinical value, and reinforcing systemic gaps in care. This persistent tension between the speed of innovation and the slowness of institutional adaptation is not an incidental problem, but a defining characteristic of contemporary digital mental health, requiring urgent attention from regulators, developers, and mental health professionals alike.

## 4. Policy Frameworks and Strategic Initiatives

The rapid adoption of AI in mental healthcare has prompted international organizations and national governments to articulate policies aimed at guiding development, deployment, and oversight. These policy frameworks vary significantly in scope, specificity, and enforceability, but collectively reflect growing awareness of the need for governance mechanisms that ensure safety, equity, and ethical compliance.

At the international level, the World Health Organization (WHO) has been instrumental in promoting strategic coordination through its Global Strategy on Digital Health 2020–2025 [6]. The document outlines goals such as strengthening national digital health ecosystems, ensuring interoperability, and establishing governance structures that promote trust. It explicitly calls for capacity-building, data protection, and inclusive stakeholder participation—especially in low- and middle-income countries. Mental health is identified as a priority area where digital interventions, including AI, could enhance reach and efficiency, provided ethical safeguards are in place.

The European Union (EU) has taken a more regulatory approach. The proposed Artificial Intelligence Act (2021) classifies AI systems used in healthcare as "high-risk" and requires developers to meet stringent transparency, accountability, and data management criteria [21]. The Act mandates impact assessments, documentation of training datasets, and human oversight mechanisms. It also includes prohibitions on certain uses of AI deemed incompatible with fundamental rights. In parallel, the EU's Comprehensive Approach to Mental Health (2023) links digital innovation to the broader goal of reducing the annual cost of inaction on mental illness, estimated at over EUR 600 billion [22].

Other bodies such as UNICEF and the Council of Europe have also weighed in. UNICEF's Mental Health Innovation Portfolio emphasizes the potential of chatbot-driven platforms, VR interventions, and big data analytics, especially for children and adolescents. It also supports partnerships with tech companies and governments to scale promising digital tools, but highlights the need for co-creation and evidence-based implementation [23]. The Council of Europe, while lacking legislative authority in health policy, has issued human rights-based recommendations stressing data privacy, consent, and algorithmic fairness—particularly in light of the General Data Protection Regulation (GDPR) and the European Convention on Human Rights.

At the national level, policies differ according to digital maturity, health infrastructure, and political will. For example, countries such as Finland and Germany have integrated

digital mental health services into public health insurance schemes, while others rely heavily on NGO or private-sector initiatives. In many contexts, national strategies lack specific provisions for mental-health-related AI tools, relying instead on general digital health laws or AI ethics guidelines. Greece is a prominent example here: despite recent digital modernization efforts, mental healthcare remains underresourced, with long-standing gaps in service coverage, workforce distribution, and public funding. While Greece has adopted the European Digital Compass and participates in EU-wide digital health initiatives, there is no dedicated national framework for AI in mental health. Pilot programs exist—mainly driven by NGOs or university consortia—but they remain fragmented and dependent on external funding. In contrast, countries such as Finland and the United Kingdom have adopted more coordinated and future-oriented approaches to integrating AI in health systems: Finland's national AI strategy emphasizes trust, transparency, and citizen engagement, including sector-specific ethical guidelines and strong public sector capacity [24], while the UK's NHS AI Lab [25] has created targeted funding instruments and evaluation protocols to support safe experimentation and public-value alignment in AI-driven health services. While these models are not without challenges, they illustrate how proactive governance and institutional coherence can facilitate ethical and scalable AI adoption.

These differences highlight a central challenge: while international frameworks increasingly recognize the importance of ethical and inclusive AI in mental health, national implementation remains uneven. There is a risk that global best practices will remain aspirational unless translated into enforceable local policy, supported by institutional capacity and stakeholder engagement: the former includes not only legal mandates and technical infrastructure, but also adequately resourced regulatory agencies, coordinated digital health strategies, and trained personnel within public health systems, while stakeholder engagement refers to the meaningful inclusion of mental health professionals, service users, civil society organizations, and local communities in the design, evaluation, and governance of AI applications. The next section addresses the key ethical and governance issues that must be resolved to ensure that such translation is not only technically feasible but normatively robust.

## 5. Ethical Challenges and Governance Gaps

As AI technologies become increasingly embedded in mental health services, ethical questions regarding their development, deployment, and oversight grow more urgent. These concerns are not only technical but structural, intersecting with broader questions of fairness, inclusion, and human rights. This section identifies three core ethical domains—privacy and consent, algorithmic accountability, and access and equity—where significant governance gaps persist.

### *5.1. Data Privacy and Informed Consent*

AI-driven mental health tools rely on the continuous collection and processing of sensitive personal data, including behavioral, emotional, and biometric indicators. In mobile health (mHealth) and digital phenotyping contexts, users may be unaware of the extent to which their data is being tracked, repurposed, or shared with third parties [5], while informed consent processes are often reduced to passive, opt-in digital agreements that do not reflect a nuanced understanding of the potential risks—particularly in populations experiencing psychological distress. Even when consent is, indeed, obtained, data protection mechanisms are frequently inadequate. Mental health data is especially vulnerable to misuse, whether for discriminatory purposes (e.g., in employment or insurance) or through reidentification from anonymized datasets. While the General Data Protection Regulation

(GDPR) offers a strong regulatory foundation in the European context, enforcement remains uneven, and protections are less robust in many low- and middle-income countries.

### *5.2. Algorithmic Bias and Clinical Accountability*

AI models used in mental health are typically trained on limited datasets that may not capture the cultural, linguistic, or socio-demographic diversity of real-world populations. As a result, these systems risk reproducing existing health disparities or introducing new forms of bias—particularly in diagnosis and triage [4]. For instance, emotion recognition systems may perform poorly across different ethnic groups or misinterpret affective signals due to cultural variation [26]. In addition to this, the opacity of many machine learning systems complicates clinical accountability. When AI systems produce recommendations or risk scores, it is often unclear how these outputs are derived. This "black box" problem limits clinicians' ability to interrogate the rationale behind algorithmic decisions, raising concerns about liability and due process in clinical care [27]. Calls for explainable AI (XAI) in healthcare have emerged as a response, but there is limited consensus on what constitutes meaningful explanation in complex psychiatric contexts, in addition to doubts related to algorithmic innovation and protection of intellectual property.

In addition, as AI systems are increasingly integrated into clinical and policy workflows, there is a risk that decisions with profound emotional and social consequences will become dehumanized. AI triage systems, automated referrals, and mental health chatbots may expedite service delivery, but they also risk creating a "bureaucratic distancing effect" where human judgment and empathy are outsourced to algorithmic systems. If not carefully designed, such systems may obscure responsibility, limit recourse for appeals, and weaken the relational aspect of care. These trends reflect a broader concern with "AI-mediated bureaucracy", in which institutions adopt algorithmic logic at the expense of human discretion and emotional attunement [28].

### *5.3. Digital Exclusion and Structural Inequity*

Digital mental health tools, including AI-based interventions, often assume reliable internet access, digital literacy, and the presence of enabling technologies such as smartphones or wearables. These assumptions may not hold across diverse populations, leading to the exclusion of already marginalized groups—including older adults, refugees, low-income individuals, and those in remote areas [29]. In these cases, AI tools may inadvertently reinforce systemic inequities by allocating resources based on data availability rather than clinical need. Furthermore, many tools are developed and deployed by private companies whose commercial priorities may conflict with public health goals. In the absence of strong regulatory frameworks or public accountability mechanisms, there is a risk of data commodification, lock-in to proprietary ecosystems, and the erosion of clinician–patient relationships. Addressing these structural barriers also requires sustainable funding and strategic partnerships: scaling equitable AI deployment in mental health will not be feasible without long-term investment in infrastructure, open-source tool development, and local capacity-building. Public–private partnerships (PPPs) often do combine resources with technical expertise, but they must be carefully governed to align with public interest goals and prevent commercial capture. For example, the UK's NHSX initiative (NHS Transformation Directorate, https://transform.england.nhs.uk/digitise-connect-transform/our-strategy-to-digitise-connect-and-transform/, last accessed: 31 May 2025) has piloted AI-driven mental health tools while embedding governance structures that promote transparency and ethical oversight, while UNICEF's AI for Children (UNICEF, AI for Children, https://www.unicef.org/innocenti/projects/ai-for-children, last accessed:

31 May 2025) initiative illustrates how cross-sector collaboration can address equity and data protection risks in technology deployments affecting vulnerable populations. To ensure that such partnerships contribute meaningfully to mental health equity, transparent procurement frameworks, regulatory scrutiny, and stakeholder participation are essential.

## 6. Strategic Preconditions for Ethical AI Integration

The effective and ethical integration of AI into mental health systems requires more than technical robustness or regulatory compliance: it calls for a multi-level approach that addresses governance, design, evaluation, and capacity-building. In this section, we identify three strategic preconditions essential to the responsible deployment of AI in mental healthcare: values-based design, oversight and evaluation, and digital readiness through education and training.

### *6.1. Values-Based Design and Participatory Approaches*

The design of AI systems for mental health should be informed by core ethical values such as autonomy, fairness, dignity, and transparency. While these principles are often cited in guidelines, their operationalization remains inconsistent. Translating high-level ethics into design features requires interdisciplinary collaboration across software engineering, clinical psychology, user experience design, and social sciences [7]. Participatory approaches can help bridge this gap. Co-design methods that involve clinicians, patients, caregivers, and marginalized users in the development process enhance contextual sensitivity and trustworthiness [13]. Such methods are particularly important in mental health, where lived experience plays a key role in evaluating the relevance, tone, and impact of digital interventions.

Translation of abstract ethical values into concrete design choices remains a critical challenge in AI development. For example, the principle of autonomy can be embodied in mental health apps by allowing users to set personalized interaction thresholds, choose between different types of support (e.g., structured CBT prompts vs. open-ended journaling), and control when and how data is collected or shared. A commitment to fairness, meanwhile, might involve building diverse and representative training datasets, as well as testing system outputs for performance disparities across demographic groups. Tools like differential privacy and model interpretability can also help ensure that fairness and transparency are not afterthoughts, but embedded throughout the system lifecycle. These kinds of design choices require collaboration not only between engineers and clinicians, but also with end-users, especially those from vulnerable or marginalized populations, whose perspectives are crucial to identifying ethical blind spots and defining what "value" truly means in context.

### *6.2. Evaluation, Oversight, and Accountability Structures*

The deployment of AI systems in clinical or community settings must be accompanied by continuous evaluation mechanisms. These should assess not only accuracy and efficacy but also social impact, accessibility, and potential harms. Randomized controlled trials (RCTs) remain a gold standard, but real-world monitoring and post-deployment auditing are also essential. Related oversight structures should be both internal and external. Internally, developers and healthcare institutions must implement ethics review boards, bias testing protocols, and documentation standards. Externally, public regulators should enforce independent evaluations, safety certifications, and grievance redress mechanisms. The concept of "algorithmic impact assessment," already emerging in digital governance, could serve as a model for high-risk mental health applications [4].

### *6.3. Education, Literacy, and Workforce Development*

Digital and AI-based mental health tools can only be scaled equitably if clinicians, support staff, and users are equipped with the necessary competencies. Mental health professionals must understand the capabilities and limitations of AI systems to use them critically and responsibly. Similarly, users need adequate digital literacy to make informed choices and protect their rights [29]. National digital health strategies should, therefore, include targeted training programs, ethical AI curricula in medical and computer science education, and public awareness campaigns. Special attention should be paid to non-specialist health workers and underresourced regions, where AI tools may be used to extend care in the absence of formal psychiatric services.

Together, these three domains—design, governance, and literacy—form a strategic foundation for ethically aligned AI in mental health. While policy and regulation provide essential guardrails, ethical integration ultimately depends on how these principles are embedded into everyday practices of development, deployment, and care.

## 7. Discussion

In this paper, we have examined the integration of artificial intelligence into mental health systems in the context of ethics and policy. While much of the existing literature emphasizes technical feasibility, diagnostic accuracy, or clinical potential, this analysis has foregrounded the broader social, regulatory, and institutional challenges that shape AI implementation. The discussion now turns to two broader implications: the reframing of AI innovation through a public interest lens, and the need for future research to address unresolved governance questions.

AI is often framed as a fix for systemic gaps in mental health care; however, this narrative risks obscuring the political and economic roots of these challenges. For AI to improve mental health at scale, its development and deployment must be aligned not only with individual autonomy and privacy, but with collective values such as social justice, democratic oversight, and solidarity [13,27]. This requires shifting from a market-driven model of digital innovation toward one rooted in the public interest.

Future research must bridge the gap between abstract ethical principles and practical implementation. What institutional models can turn values into enforceable standards? How can algorithmic impact assessments be adapted for mental health contexts? What forms of participatory governance are feasible in underresourced settings? These questions require interdisciplinary approaches that combine technical insight with policy analysis, qualitative research, and legal scholarship.

When it comes to institutional models able to translate values into enforceable standards, future research may use other high-risk AI domains such as biometric surveillance, automated credit scoring, and employment screening as an example. Emerging governance frameworks in these areas include algorithmic auditing procedures, red-teaming protocols [30], and regulatory sandboxes [31], i.e., controlled environments in which AI systems are evaluated in practice before deployment. For example, the European Commission's proposed AI Liability Directive (European Commission, Liability Rules for Artificial Intelligence, https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en, last accessed: 11 June 2025) and the OECD's AI Risk Classification Framework both outline mechanisms for accountability, including documentation obligations, third-party oversight, and domain-specific risk thresholds. In the mental health context, such mechanisms could support the operationalization of algorithmic impact assessments by linking them to institutional review boards, health data custodians, or public registries of AI tools in clinical use. Exploring

these cross-sectoral analogues offers a promising foundation for designing domain-sensitive governance structures that are both flexible and enforceable.

Moreover, national strategies must be context-specific. As the case of Greece illustrates, digital mental health policy cannot be separated from broader questions of health infrastructure, administrative capacity, and public trust. International frameworks may offer useful templates, but their local uptake depends on political will, stakeholder engagement, and institutional readiness.

*Limitations*

This study presents a strategic review informed by the PRISMA-ScR methodology, combining a scoping review of academic literature with thematic analysis of policy and grey sources. While this approach offers a broad and interdisciplinary perspective, it also presents specific limitations, starting with the selection process for academic literature: guided by inclusion criteria and keyword-based searches, our search was constrained by database coverage (limited to PubMed, Scopus, and Google Scholar), language (English-only sources), and publication date (post-2018). As a result, relevant contributions in non-indexed journals or non-English contexts may have been excluded, potentially biasing the sample toward Western or Global North perspectives.

In addition, while the PRISMA-ScR methodology facilitates transparency in reporting and emphasizes breadth over depth, it does not require critical appraisal of study quality. Therefore, while we identified key themes and trends, we did not conduct formal assessments of evidence robustness, clinical efficacy, or methodological rigor across included studies, effectively limiting the extent to which our findings can support causal claims or fine-grained comparative evaluations of AI applications in mental health.

Another limitation has to do with the thematic synthesis of policy documents and grey literature, which is essential for capturing current governance landscapes, but introduces subjectivity in interpretation. These documents vary in scope, specificity, and political context, and our coding framework, while value-driven (e.g., fairness, accountability), reflects certain normative assumptions that may not be universally shared. Moreover, given the dynamic nature of both AI technologies and mental health governance, our review offers a snapshot in time that may soon be outdated.

Finally, the inclusion of illustrative case studies (e.g., Woebot, Wysa, Mindstrong) highlights important ethical tensions, but these examples were chosen for their prominence and accessibility rather than systematic representativeness. Future research should extend this work with empirical studies, stakeholder interviews, and comparative policy analysis to test the generalizability of the conclusions presented here.

## 8. Conclusions

Artificial intelligence has shown the potential to support more personalized, efficient, and accessible mental health services. However, realizing this potential requires ethical integration, strategic foresight, and regulatory clarity. In this paper, we argued that the responsible deployment of AI in mental health depends on three core preconditions: values-based and participatory design, robust oversight and accountability mechanisms, and investment in digital literacy and workforce training. Current international strategies, including those from the WHO and EU, reflect growing recognition of these needs, but practical implementation remains uneven across national contexts. Without concerted efforts to bridge this gap, AI risks reinforcing existing inequalities, eroding trust in mental health services, and compromising fundamental rights.

To address these risks and promote inclusive innovation, the paper offers the following policy recommendations:

- National governments should incorporate AI governance into mental health policy, with attention to data protection, bias mitigation, and inclusive design;
- Health ministries and public regulators should mandate impact assessments and post-deployment audits for AI tools used in clinical or community mental health;
- Training programs in ethical AI should be embedded in medical, psychological, and computing education [32], alongside public digital literacy initiatives;
- Funding schemes should support participatory design research that centers marginalized voices and lived experience in mental health AI development.

As AI systems become more prevalent in mental healthcare, their legitimacy will depend not only on technical performance, but on whether they uphold the rights, dignity, and diverse needs of those they are meant to serve.

**Author Contributions:** Conceptualization, A.K.F. and K.K.; methodology, K.K.; investigation, A.K.F.; writing—original draft preparation, A.K.F.; writing—review and editing, A.K.F. and K.K.; supervision, K.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Dataset available on request from the authors.

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

## References

1. Shimada, K. The role of artificial intelligence in mental health: A review. *Sci. Insights* **2023**, *43*, 1119–1127. [CrossRef]
2. Bzdok, D.; Meyer-Lindenberg, A. Machine learning for precision psychiatry: Opportunities and challenges. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **2018**, *3*, 223–230. [CrossRef] [PubMed]
3. Holmes, E.A.; O'Connor, R.C.; Perry, V.H.; Tracey, I.; Wessely, S.; Arseneault, L.; Ballard, C.; Christensen, H.; Silver, R.C.; Everall, I.; et al. Multidisciplinary research priorities for the COVID-19 pandemic: A call for action for mental health science. *Lancet Psychiatry* **2020**, *7*, 547–560. [CrossRef] [PubMed]
4. Morley, J.; Machado, C.C.; Burr, C.; Cowls, J.; Joshi, I.; Taddeo, M.; Floridi, L. The ethics of AI in health care: A mapping review. *Soc. Sci. Med.* **2020**, *260*, 113172. [CrossRef]
5. Vayena, E.; Blasimme, A.; Cohen, I.G. Machine learning in medicine: Addressing ethical challenges. *PLoS Med.* **2018**, *15*, e1002689. [CrossRef]
6. World Health Organization. *Global Strategy on Digital Health 2020–2025*; World Health Organization: Geneva, Switzerland, 2021.
7. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef]
8. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.; Horsley, T.; Weeks, L.; et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473. [CrossRef]
9. Benzies, K.M.; Premji, S.; Hayden, K.A.; Serrett, K. State-of-the-evidence reviews: Advantages and challenges of including grey literature. *Worldviews-Evid.-Based Nurs.* **2006**, *3*, 55–61. [CrossRef]
10. Watkins, E.R.; Warren, F.C.; Newbold, A.; Hulme, C.; Cranston, T.; Aas, B.; Bear, H.; Botella, C.; Burkhardt, F.; Ehring, T.; et al. Emotional competence self-help app versus cognitive behavioural self-help app versus self-monitoring app to prevent depression in young adults with elevated risk (ECoWeB PREVENT): An international, multicentre, parallel, open-label, randomised controlled trial. *Lancet Digit. Health* **2024**, *6*, e894–e903.
11. Bralee, E.; Mostazir, M.; Warren, F.C.; Newbold, A.; Hulme, C.; Cranston, T.; Aas, B.; Bear, H.; Botella, C.; Burkhardt, F.; et al. Brief use of behavioral activation features predicts benefits of self-help app on depression symptoms: Secondary analysis of a selective prevention trial in young people. *J. Consult. Clin. Psychol.* **2025**, *93*, 293. [CrossRef]
12. Milne-Ives, M.; Selby, E.; Inkster, B.; Lam, C.; Meinert, E. Artificial intelligence and machine learning in mobile apps for mental health: A scoping review. *PLoS Digit. Health* **2022**, *1*, e0000079. [CrossRef] [PubMed]

13. Sharon, T. When digital health meets digital capitalism, how many common goods are at stake? *Big Data Soc.* **2018**, *5*, 2053951718819032. [CrossRef]
14. Keles, S. Navigating in the moral landscape: Analysing bias and discrimination in AI through philosophical inquiry. *AI Ethics* **2023**, *5*, 555–565. [CrossRef]
15. Fitzpatrick, K.K.; Darcy, A.; Vierhile, M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Ment. Health* **2017**, *4*, e7785. [CrossRef] [PubMed]
16. Birk, R.H.; Samuel, G. Digital phenotyping for mental health: Reviewing the challenges of using data to monitor and predict mental health problems. *Curr. Psychiatry Rep.* **2022**, *24*, 523–528. [CrossRef]
17. Mak, W.W.; Ng, S.M.; Leung, F.H. A web-based stratified stepped care platform for mental well-being (TourHeart+): User-centered research and design. *JMIR Form. Res.* **2023**, *7*, e38504. [CrossRef]
18. Valmaggia, L.R.; Latif, L.; Kempton, M.J.; Rus-Calafell, M. Virtual reality in the psychological treatment for mental health problems: An systematic review of recent evidence. *Psychiatry Res.* **2016**, *236*, 189–195. [CrossRef]
19. Allers, S.; Carboni, C.; Eijkenaar, F.; Wehrens, R. A Cross-Disciplinary Analysis of the Complexities of Scaling Up eHealth Innovation. *J. Med. Internet Res.* **2024**, *26*, e58007. [CrossRef]
20. Elyoseph, Z.; Gur, T.; Haber, Y.; Simon, T.; Angert, T.; Navon, Y.; Tal, A.; Asman, O. An ethical perspective on the democratization of mental health with generative AI. *JMIR Ment. Health* **2024**, *11*, e58011. [CrossRef]
21. European Commission. *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*; COM/2021/206 Final; European Commission: Brussels, Belgium, 2021.
22. European Commission. *A Comprehensive Approach to Mental Health*; COM(2023) 298 Final; European Commission: Brussels, Belgium, 2023.
23. UNICEF Office of Innovation. *Mental Health Innovation Portfolio*; UNICEF: New York, NY, USA, 2022.
24. Ministry of Economic Affairs and Employment. *Leading the Way into the Age of Artificial Intelligence: Final Report of Finland's Artificial Intelligence Programme 2019*; Ministry of Economic Affairs and Employment: Helsinki, Finland, 2019.
25. Kerstein, R. NHS X-plained. *Bull. R. Coll. Surg. Engl.* **2020**, *102*, 48–49. [CrossRef]
26. Pantic, M.; Caridakis, G.; André, E.; Kim, J.; Karpouzis, K.; Kollias, S. Multimodal emotion recognition from low-level cues. In *Emotion-Oriented Systems: The Humaine Handbook*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 115–132.
27. London, A.J. Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Cent. Rep.* **2019**, *49*, 15–21. [CrossRef] [PubMed]
28. Eubanks, V. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*; St. Martin's Press: New York, NY, USA, 2018.
29. Karizat, N.; Vinson, A.H.; Parthasarathy, S.; Andalibi, N. Patent applications as glimpses into the sociotechnical imaginary: Ethical speculation on the imagined futures of emotion AI for mental health monitoring and detection. *Proc. ACM Hum.-Comput. Interact.* **2024**, *8*, 1–43. [CrossRef]
30. Feffer, M.; Sinha, A.; Deng, W.H.; Lipton, Z.C.; Heidari, H. Red-Teaming for generative AI: Silver bullet or security theater? In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, San Jose, CA, USA, 21–23 October 2024; Volume 7, pp. 421–437.
31. Truby, J.; Brown, R.D.; Ibrahim, I.A.; Parellada, O.C. A sandbox approach to regulating high-risk artificial intelligence applications. *Eur. J. Risk Regul.* **2022**, *13*, 270–294. [CrossRef]
32. Panagopoulou, F.; Parpoula, C.; Karpouzis, K. Legal and ethical considerations regarding the use of ChatGPT in education. *arXiv* **2023**, arXiv:2306.10037.


**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

*Computers* Editorial Office
E-mail: computers@mdpi.com
www.mdpi.com/journal/computers

Disclaimer/Publisher's Note: The title and front matter of this reprint are at the discretion of the Guest Editors. 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 Editors 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 978-3-7258-8742-2