diagnostics

Special Issue Reprint

Machine-Learning-Based Disease Diagnosis and Prediction

Edited by
Fakhar Abbas

mdpi.com/journal/diagnostics

MDPI

Machine-Learning-Based Disease Diagnosis and Prediction

Machine-Learning-Based Disease Diagnosis and Prediction

Guest Editor

Fakhar Abbas

MDPI

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

Guest Editor
Fakhar Abbas
Centre for Trusted Internet and
Community
National University of Singapore
Singapore
Singapore

Editorial Office
MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Special Issue, published open access by the journal *Diagnostics* (ISSN 2075-4418), freely accessible at: https://www.mdpi.com/journal/diagnostics/special_issues/68ATM6BG1S.

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 978-3-7258-8637-1 (Hbk)
ISBN 978-3-7258-8638-8 (PDF)
https://doi.org/10.3390/books978-3-7258-8638-8

© 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 Editor . . . . . **ix**

Preface . . . . . **xi**

Fakhar Abbas
Machine-Learning-Based Disease Diagnosis and Prediction: Progress, Perspectives, and the Path Forward
Reprinted from: *Diagnostics* **2026**, *16*, 1844, https://doi.org/10.3390/diagnostics16121844 . . . . **1**

A. A. Abd El-Aziz, Mahmood A. Mahmood and Sameh Abd El-Ghany
Advanced Deep Learning Fusion Model for Early Multi-Classification of Lung and Colon Cancer Using Histopathological Images
Reprinted from: *Diagnostics* **2024**, *14*, 2274, https://doi.org/10.3390/diagnostics14202274 . . . . **9**

Sameh Abd El-Ghany, Mahmood A. Mahmood and A. A. Abd El-Aziz
FracFusionNet: A Multi-Level Feature Fusion Convolutional Network for Bone Fracture Detection in Radiographic Images
Reprinted from: *Diagnostics* **2025**, *15*, 2212, https://doi.org/10.3390/diagnostics15172212 . . . . **37**

Erdem Yanar, Fırat Hardalaç and Kubilay Ayturan
CELM: An Ensemble Deep Learning Model for Early Cardiomegaly Diagnosis in Chest Radiography
Reprinted from: *Diagnostics* **2025**, *15*, 1602, https://doi.org/10.3390/diagnostics15131602 . . . . **91**

Md. Tanvir Hayat, Yazan M. Allawi, Wasan Alamro, Salman Md Sultan, Ahmad Abadleh, Hunseok Kang and and Aymen I. Zreikat
A Hybrid Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM)–Attention Model Architecture for Precise Medical Image Analysis and Disease Diagnosis
Reprinted from: *Diagnostics* **2025**, *15*, 2673, https://doi.org/10.3390/diagnostics15212673 . . . . **124**

Farzeen Khalid Khan, Waleed Bin Tahir, Mu Sook Lee, Jin Young Kim, Shi Sub Byon, Sun-Woo Pi and Byoung-Dai Lee
Leveraging Large-Scale Public Data for Artificial Intelligence-Driven Chest X-Ray Analysis and Diagnosis
Reprinted from: *Diagnostics* **2026**, *16*, 146, https://doi.org/10.3390/diagnostics1610146 . . . . . **156**

Karl Ludger Radke, Anja Müller-Lutz, Daniel B. Abrar, Marius Vach, Christian Rubbert, David Latz, et al.
Precision Through Detail: Radiomics and Windowing Techniques as Key for Detecting Dens Axis Fractures in CT Scans
Reprinted from: *Diagnostics* **2025**, *15*, 2599, https://doi.org/10.3390/diagnostics15202599 . . . . **178**

Seokhwan Bang, Heehwan Wang, Hoyoung Bae, Sung-Hoo Hong, Jiook Cha and Moon Hyung Choi
Computed Tomography-Based Radiomics Diagnostic Model for Fat-Poor Small Renal Tumor Subtypes
Reprinted from: *Diagnostics* **2025**, *15*, 1365, https://doi.org/10.3390/diagnostics15111365 . . . . **194**

Cansel Ficici, Ziya Telatar, Osman Erogul and Onur Kocak
Temporal Lobe Epilepsy Focus Detection Based on the Correlation Between Brain MR Images and EEG Recordings with a Decision Tree
Reprinted from: *Diagnostics* **2024**, *14*, 2509, https://doi.org/10.3390/diagnostics14222509 . . . . **215**

Valentin A. Nechaev, Nataliya Y. Kashtanova, Evgenii V. Kopeykin, Umamat M. Magomedova, Maria S. Gribkova, Anton V. Hardin, et al.
Diagnostic Accuracy of a Multi-Target Artificial Intelligence Service for the Simultaneous Assessment of 16 Pathological Features on Chest and Abdominal CT
Reprinted from: *Diagnostics* **2025**, *15*, 2778, https://doi.org/10.3390/diagnostics15212778 . . . . **232**

Eliodoro Faiella, Stefania Lamja, Rebecca Casati, Michele Tondo, Raffaele Ragone, Adriano Redi, et al.
Feasibility of Artificial Intelligence Models for Longitudinal CT Analysis of Epicardial Adipose Tissue After Immunotherapy
Reprinted from: *Diagnostics* **2026**, *16*, 852, https://doi.org/10.3390/diagnostics16060852 . . . . . **252**

Pin-Chieh Wu, Yun-Ju Wu, Chiao-Lin Hsu, Hsien-Chung Yu, Chi-Shen Chen and Fu-Zong Wu
Nomogram for Osteoporosis Risk Using LDCT Trabecular Parameters
Reprinted from: *Diagnostics* **2026**, *16*, 1429, https://doi.org/10.3390/diagnostics16101429 . . . . **262**

Ivan Petrović, Serena Broggi, Monika Killer-Oberpfalzer, Johannes A. R. Pfaff, Christoph J. Griessenauer, Isidora Milosavljević, et al.
Predictors of In-Hospital Mortality after Thrombectomy in Anterior Circulation Large Vessel Occlusion: A Retrospective, Machine Learning Study
Reprinted from: *Diagnostics* **2024**, *14*, 1531, https://doi.org/10.3390/diagnostics14141531 . . . . **274**

Lorenz A. Kapsner, Manuel Feißt, Ariawan Purbojo, Hans-Ulrich Prokosch, Thomas Ganslandt, Sven Dittrich, et al.
Using Machine Learning and Feature Importance to Identify Risk Factors for Mortality in Pediatric Heart Surgery
Reprinted from: *Diagnostics* **2024**, *14*, 2587, https://doi.org/10.3390/diagnostics14222587 . . . . **291**

Eduardo Guzmán-Muñoz, Manuel Vásquez-Muñoz, Yeny Concha-Cisternas, Rodrigo Olivares-Ordenes, Vicente Clemente-Suárez, Antonio Castillo-Paredes and Rodrigo Yáñez-Sepúlveda
Supervised Machine Learning-Based Prediction of In-Hospital Mortality Following Hip Fracture in Older Adults
Reprinted from: *Diagnostics* **2026**, *16*, 612, https://doi.org/10.3390/diagnostics1640612 . . . . . **314**

Ying-Qiang Liu, Tzu-Wei Chang, Lung-Chun Lee, Chia-Yu Chen, Pi-Shan Hsu, Yu-Tse Tsan, et al.
Use of Machine Learning to Predict the Incidence of Type 2 Diabetes Among Relatively Healthy Adults: A 10-Year Longitudinal Study in Taiwan
Reprinted from: *Diagnostics* **2025**, *15*, 72, https://doi.org/10.3390/diagnostics1510072 . . . . . **333**

Elisabetta D'Aversa, Bianca Antonica, Miriana Grisafi, Rosanna Asselta, Elvezia Maria Paraboschi, Angelina Passaro, et al.
Integrating Host Genetics and Clinical Setting in Machine Learning Models: Predicting COVID-19 Prognosis for Healthcare Decision-Making (The FeMiNa Study)
Reprinted from: *Diagnostics* **2026**, *16*, 583, https://doi.org/10.3390/diagnostics1640583 . . . . . **344**

Hasnain Iftikhar, Atef F. Hashem, Moiz Qureshi and Paulo Canas Rodrigues
Clinical Application of Machine Learning Models for Early-Stage Chronic Kidney Disease Detection
Reprinted from: *Diagnostics* **2025**, *15*, 2610, https://doi.org/10.3390/diagnostics15202610 . . . . **367**

Hee Soo Yoon, Min Jin Kim, Kang Hyeon Lim, Min Suk Kim, Byung Jae Kang, Yoon Chan Rah and June Choi
Evaluating Prediction Models with Hearing Handicap Inventory for the Elderly in Chronic Otitis Media Patients
Reprinted from: *Diagnostics* **2024**, *14*, 2000, https://doi.org/10.3390/diagnostics14182000 . . . . **385**

Suliman Aladhadh
An Explainable Ensemble and Deep Learning Framework for Accurate and Interpretable Parkinson's Disease Detection from Voice Biomarkers
Reprinted from: *Diagnostics* **2025**, *15*, 2892, https://doi.org/10.3390/diagnostics15222892 . . . . **396**

Bolaji A. Omodunbi, David B. Olawade, Omosigho F. Awe, Afeez A. Soladoye, Nicholas Aderinto, Saak V. Ovsepian and Stergios Boussios
Stacked Ensemble Learning for Classification of Parkinson's Disease Using Telemonitoring Vocal Features
Reprinted from: *Diagnostics* **2025**, *15*, 1467, https://doi.org/10.3390/diagnostics15121467 . . . . **418**

Chih-Yueh Chang, Dee Pei, Yen-Liang Kuo, Li-Na Lee, Chung-Ze Wu, Ta-Wei Chu, et al.
Explainable Machine Learning Models for Predicting FEV_1 in Non-Smoking Taiwanese Men Aged 45–55 Years
Reprinted from: *Diagnostics* **2025**, *15*, 3152, https://doi.org/10.3390/diagnostics15243152 . . . . **436**

Blanca Vazquez, Mariano Rojas-García, Jocelyn Isabel Rodríguez-Esquivel, Janeth Marquez-Acosta, Carlos E. Aranda-Flores, Lucely del Carmen Cetina-Pérez, et al.
Machine and Deep Learning for the Diagnosis, Prognosis, and Treatment of Cervical Cancer: A Scoping Review
Reprinted from: *Diagnostics* **2025**, *15*, 1543, https://doi.org/10.3390/diagnostics15121543 . . . . **453**

Siyami Aydın, Mehmet Ağar, Muharrem Çakmak and Mesut Toğaçar
Diagnosis of Mesothelioma Using Image Segmentation and Class-Based Deep Feature Transformations
Reprinted from: *Diagnostics* **2025**, *15*, 2381, https://doi.org/10.3390/diagnostics15182381 . . . . **495**

Zne-Jung Lee, Jing-Xun Cai, Liang-Hung Wang and Ming-Ren Yang
Ensemble Algorithm Based on Gene Selection, Data Augmentation, and Boosting Approaches for Ovarian Cancer Classification
Reprinted from: *Diagnostics* **2024**, *14*, 2772, https://doi.org/10.3390/diagnostics14242772 . . . . **522**

Maimoona Khalid, Muhammad Zaheer Sajid, Ayman Youssef, Nauman Ali Khan, Muhammad Fareed Hamid and Fakhar Abbas
CAD-EYE: An Automated System for Multi-Eye Disease Classification Using Feature Fusion with Deep Learning Models and Fluorescence Imaging for Enhanced Interpretability
Reprinted from: *Diagnostics* **2024**, *14*, 2679, https://doi.org/10.3390/diagnostics14232679 . . . . **536**

Wanzi Su, Damon Hoad, Leandro Pecchia and Davide Piaggio
Validation of an Eye-Tracking Algorithm Based on Smartphone Videos: A Pilot Study
Reprinted from: *Diagnostics* **2025**, *15*, 1446, https://doi.org/10.3390/diagnostics15121446 . . . . **561**

Catalina Bennasar, Antonio Nadal-Martínez, Sebastiana Arroyo, Yolanda Gonzalez-Cid, Ángel Arturo López-González and Pedro Juan Tárraga
Integrating Machine Learning and Deep Learning for Predicting Non-Surgical Root Canal Treatment Outcomes Using Two-Dimensional Periapical Radiographs
Reprinted from: *Diagnostics* **2025**, *15*, 1009, https://doi.org/10.3390/diagnostics15081009 . . . . **581**

Ali J. B. Al-Sharqi, Mohammed Taha Ahmed Baban, Nada K. Imran, Sarhang S. Gul and Ali A. Abdulkareem
Comparison of Supervised Machine Learning Models to Logistic Regression Model Using Tooth-Related Factors to Predict the Outcome of Nonsurgical Periodontal Treatment
Reprinted from: *Diagnostics* **2025**, *15*, 2333, https://doi.org/10.3390/diagnostics15182333 . . . . **595**

Ziad Khamaysi, Mahdi Awwad, Badea Jiryis, Naji Bathish and Jonathan Shapiro
The Role of ChatGPT in Dermatology Diagnostics
Reprinted from: *Diagnostics* **2025**, *15*, 1529, https://doi.org/10.3390/diagnostics15121529 . . . . **610**

Sita Rani, Raman Kumar, B. S. Panda, Rajender Kumar, Nafaa Farhan Muften, Mayada Ahmed Abass and Jasmina Lozanović
Machine Learning-Powered Smart Healthcare Systems in the Era of Big Data: Applications, Diagnostic Insights, Challenges, and Ethical Implications
Reprinted from: *Diagnostics* **2025**, *15*, 1914, https://doi.org/10.3390/diagnostics15151914 . . . . **631**

About the Editor

Fakhar Abbas

Fakhar Abbas is a Senior Research Fellow at the Centre for Trusted Internet and Community, National University of Singapore, and a Senior Member of IEEE. His research focuses on trustworthy artificial intelligence, medical imaging, deepfake and multimodal misinformation detection, and secure machine learning for edge and vehicular systems, alongside a sustained interest in the application of machine learning to disease diagnosis and clinical prediction. He has authored numerous peer-reviewed publications, including contributions to *IEEE Transactions* journals, and has served as a Reviewer and Guest Editor for several international journals in artificial intelligence and biomedical informatics. In addition to his research activities, he has substantial teaching experience and has received recognition for excellence in instruction, including a Best Teacher Award. As Guest Editor of this Special Issue, he worked closely with authors and reviewers across Asia, Europe, the Middle East, and Latin America to assemble a collection that reflects both the methodological diversity and the clinical relevance of machine-learning-based diagnostic research.

Preface

This Reprint compiles the Editorial together with thirty original research articles and reviews from the Special Issue "Machine Learning Based Disease Diagnosis and Prediction." It provides a consolidated, accessible resource for clinicians, biomedical engineers, and data scientists working at the interface of artificial intelligence and clinical medicine. The contributions cover medical imaging, structured clinical prediction, explainable AI, oncology, ophthalmology, dentistry, and system level healthcare applications, reflecting both the breadth of current research and the methodological rigor required for trustworthy clinical translation. This collection targets researchers developing diagnostic models, clinicians assessing evidence for AI assisted tools, and graduate students seeking a structured introduction to the field. We hope that assembling these studies will foster interdisciplinary collaboration and accelerate the responsible adoption of machine learning for disease diagnosis and prediction.

Fakhar Abbas
Guest Editor

diagnostics

MDPI

Editorial

Machine-Learning-Based Disease Diagnosis and Prediction: Progress, Perspectives, and the Path Forward

Fakhar Abbas

Centre for Trusted Internet & Community, National University of Singapore, Singapore 119077, Singapore; fakhar.5@nus.edu.sg or fakhar.14@hotmail.com

1. Introduction

The convergence of machine learning (ML) and clinical medicine has significantly reshaped modern healthcare. Over the past decade, the rapid expansion of health-related data—including electronic health records, medical imaging, genomic data, wearable sensor outputs, and multimodal biosignals—has created both substantial analytical challenges and unprecedented opportunities. While traditional statistical and diagnostic approaches remain foundational, they are often limited in their ability to extract actionable insights from high-dimensional and heterogeneous data at scale. In this context, ML and deep learning (DL) methods have emerged as powerful tools capable of identifying complex, nonlinear patterns and, in certain applications, achieving diagnostic performance comparable to that of experienced clinicians [1–4].

Despite this progress, several critical challenges remained unresolved at the inception of this Special Issue. These include the generalizability of ML models across diverse populations, the interpretability of black-box systems, the integration of multimodal data, and the limited evaluation of ML approaches across a broad spectrum of diseases.

This Special Issue of *Diagnostics*, entitled "Machine-Learning-Based Disease Diagnosis and Prediction", was conceived to address these challenges by bringing together original research, review articles, and methodological studies spanning diverse ML-assisted diagnostic applications. Following a rigorous peer-review process, 30 papers were accepted, representing contributions from research groups across Asia, Europe, the Middle East, and Latin America. The breadth of topics—including oncology, neurology, cardiology, nephrology, respiratory medicine, ophthalmology, dentistry, and infectious diseases—highlights both the versatility of ML methodologies and their growing translational relevance. Collectively, these contributions provide a timely snapshot of a rapidly advancing field that continues to face important methodological and clinical barriers.

2. Overview of the Published Papers

The thirty papers featured in this Special Issue span multiple clinical domains and methodological approaches. The contributions are summarized below according to thematic clusters.

2.1. Medical Imaging and Radiological AI

Medical imaging represents the most prominent domain in this Special Issue, with 11 contributions covering radiography, histopathology, computed tomography (CT), and ophthalmological imaging.

Abd El-Aziz et al. (contribution 1) propose a deep learning fusion model for multiclass classification of lung and colon cancer using histopathological images. By integrating

ResNet-101V2, NASNetMobile, and EfficientNet-B0, the model achieved 99.94% accuracy on the LC25000 dataset, demonstrating the effectiveness of feature fusion across complementary architectures.

Abd El-Ghany et al. (contribution 2) introduce FracFusionNet, a multi-level feature fusion convolutional network for bone fracture detection in radiographic images. By combining hierarchical feature representations, the model improved sensitivity across fracture types and offered a computationally efficient approach to automated radiographic screening in emergency and orthopaedic settings.

Yanar et al. (contribution 3) present CELM, an ensemble deep learning model for early cardiomegaly diagnosis in chest radiography. Importantly, the study included external validation across multiple centres, underscoring that strong internal performance alone is insufficient to establish clinical readiness for imaging AI systems.

Hayat et al. (contribution 4) have developed a hybrid CNN–LSTM–Attention architecture for medical image analysis and disease diagnosis. By combining spatial feature extraction, temporal context modelling, and attention mechanisms, the model improved performance across multi-disease image classification tasks.

Likewise, Khan et al. (contribution 5) examine the use of large-scale public datasets for AI-driven chest X-ray analysis and diagnosis. Their findings suggest that training on heterogeneous public data can substantially improve robustness and generalizability, offering a practical strategy for addressing the limitations of single-centre datasets.

Radke et al. (contribution 6) investigate radiomics combined with optimised windowing techniques for detecting dens axis fractures in CT scans. The study showed that preprocessing choices, particularly windowing parameters, can meaningfully influence model performance in challenging skeletal trauma detection tasks.

Bang et al. (contribution 7) have developed a CT-based radiomics model for classifying fat-poor small renal tumour subtypes. Using XGBoost as the optimal classifier, the model achieved an average AU-PRC of 0.757 across five tumour subtypes from single-phase CT, supporting the value of radiomics-based ML in renal oncology.

Similarly, Ficici et al. (contribution 8) introduce a decision-tree-based system for temporal lobe epilepsy focus detection using correlated brain MRI and EEG data. The proposed approach achieved complete concordance with physicians on EEG findings and outperformed physician-only assessment in the combined MRI-EEG task, highlighting the clinical value of multimodal integration.

Nechaev et al. (contribution 9) evaluate the diagnostic accuracy of a multi-target AI service for simultaneous assessment of 16 pathological features on chest and abdominal CT. Their results suggest that integrated AI systems may help streamline radiological workflows and reduce reporting burden.

Faiella et al. (contribution 10) examined AI models for longitudinal CT analysis of epicardial adipose tissue changes following immunotherapy. The study supports the feasibility of AI-assisted serial imaging analysis as a non-invasive tool for monitoring cardiac toxicity during cancer treatment.

Finally, Wu et al. (contribution 11) develop a nomogram for osteoporosis risk stratification using trabecular parameters derived from low-dose CT. Their work supports opportunistic screening from routine CT data as a practical approach to improving osteoporosis detection.

2.2. Prediction Using Structured Clinical and Laboratory Data

A substantial cluster of contributions demonstrated the value of ML for risk stratification and outcome prediction using structured clinical, laboratory, and epidemiological data.

Petrović et al. (contribution 12) apply interpretable ML to predict in-hospital mortality after mechanical thrombectomy for anterior circulation large vessel occlusion. Their Pre-MT and Post-MT models achieved AUCs of 0.79 and 0.84, respectively, with baseline NIHSS and age emerging as the most important predictors.

Kapsner et al. (contribution 13) use ML and feature importance analysis to identify mortality risk factors after surgery for congenital heart defects in paediatric patients. Their random survival forest and XGBoost models achieved C-indices of 0.85 and 0.79, with postoperative serum creatinine consistently identified as a key predictor.

Guzmán-Muñoz et al. (contribution 14) a supervised ML model for predicting in-hospital mortality after hip fracture in older adults. Using SHAP-based explainability, the study identified key clinical predictors and demonstrated the value of ML for targeted risk stratification in a high-risk orthopaedic population.

Liu et al. (contribution 15) apply ML to predict the incidence of type 2 diabetes in a 10-year longitudinal cohort from Taiwan. Random forest, logistic regression, and XGBoost all achieved 98–99% accuracy, and free thyroxine emerged as a previously underappreciated predictor.

D'Aversa et al. (contribution 16) integrate host genetics and clinical variables in ML models for predicting COVID-19 prognosis in the FeMiNa Study. Their findings show that combining genetic and clinical data improves prognostic performance over clinical data alone, offering a useful framework for personalised infectious disease prediction.

Iftikhar et al. (contribution 17) apply ensemble ML models to early-stage chronic kidney disease detection using routine laboratory data. The study demonstrated high sensitivity at early disease stages, supporting ML-based screening for earlier referral and intervention.

Yoon et al. (contribution 18) evaluate prediction models incorporating the Hearing Handicap Inventory for the Elderly in patients with chronic otitis media. Logistic regression outperformed random forest, with an AUC of 0.73 and accuracy of 73.56%, highlighting both the utility and current limitations of predictive modelling in otology.

2.3. Explainability, Interpretability, and Validation Methodology

Interpretability emerged as a cross-cutting and recurring theme, with several contributions specifically addressing explainable AI frameworks and the methodological rigor required for trustworthy clinical ML.

Aladhadh (contribution 19) developed an explainable ensemble and deep learning framework for Parkinson's disease detection from voice biomarkers. By combining SHAP analysis with ensemble learning, the study showed that high accuracy and interpretability can be achieved simultaneously.

Omodunbi et al. (contribution 20) apply stacked ensemble learning to Parkinson's disease classification using telemonitoring vocal features, with particular emphasis on subject-wise validation. The marked drop in performance from conventional cross-validation to subject-wise evaluation clearly illustrates the risk of data leakage in recording-based approaches.

Chang et al. (contribution 21) apply explainable ML models to predict FEV1 in non-smoking Taiwanese men aged 45–55 years. SHAP-based analysis identified key pulmonary predictors within this population, illustrating how XAI can support population-level screening and early intervention.

2.4. Oncology Applications

Several contributions extended ML across the cancer care continuum, including diagnosis, prognosis, and molecular classification.

Vazquez et al. (contribution 22) conduct a scoping review of ML and DL applications in cervical cancer diagnosis, prognosis, and treatment. The review found the strongest evidence for image-based diagnostic tasks, while prognostic and treatment applications remain emerging areas of growth.

Likewise, Aydın et al. (contribution 23) address mesothelioma diagnosis using image segmentation and class-based deep feature transformations. Their approach improved classification accuracy for this rare and diagnostically challenging thoracic malignancy.

Finally, Lee et al. (contribution 24) present an ensemble algorithm based on gene selection, data augmentation, and boosting for ovarian cancer classification. Their 10-gene ensemble achieved 98.21% accuracy overall and 100% accuracy for subtype discrimination.

2.5. Ophthalmology and Specialised Diagnostics

Two contributions illustrate the application of ML in ophthalmological screening and the emergence of low-cost device-based diagnostic tools.

Khalid et al. (contribution 25) introduce CAD-EYE, an automated multi-eye disease classification system using feature fusion and fluorescence imaging for improved interpretability. Trained on 65,871 fundus images, the model achieved 98% accuracy and outperformed several widely used architectures.

Su et al. (contribution 26) validate a smartphone-based eye-tracking algorithm using the CHT_TM approach. The method reduced execution time substantially while maintaining high accuracy, supporting the feasibility of low-cost eye-movement tracking for scalable screening applications.

2.6. Dental and Oral Health Applications

Two contributions extend the reach of ML into structured dental clinical data, addressing endodontic and periodontal treatment outcome prediction.

Bennasar et al. (contribution 27) integrate ML and DL to predict non-surgical root canal treatment outcomes using two-dimensional periapical radiographs. Their combined DL-LR model outperformed clinician prognosis in sensitivity, negative predictive value, and accuracy.

Al-Sharqi et al. (contribution 28) compare unsupervised ML models with logistic regression for predicting nonsurgical periodontal treatment outcomes using tooth-related clinical factors. Their findings support the use of ML for structured dental data and evidence-based treatment planning.

2.7. Emerging and System-Level Applications

The final cluster of contributions addresses emerging AI paradigms and broader healthcare system applications, signalling directions that are likely to define the field's next phase of development.

Khamaysi et al. (contribution 29) review the role of ChatGPT in dermatology diagnostics, exploring its application in natural language processing for clinical data interpretation, differential diagnosis assistance, and patient communication. While the review highlights ChatGPT's considerable promise for NLP-based diagnostic support, it also identifies the critical limitation of the inability to perform direct image analysis, suggesting that the integration of large language models with vision-based convolutional neural network tools represents a compelling and imminent frontier for comprehensive dermatological AI.

Rani et al. (contribution 30) provide a comprehensive review of ML-powered smart healthcare systems in the era of big data, mapping applications across diagnostic ecosystems while systematically examining challenges and ethical implications. Their analysis identifies federated learning and robust ethical AI governance frameworks as foundational

priorities, framing privacy-preserving distributed learning not merely as a technical option but as a structural necessity for the responsible scaling of healthcare AI.

3. Future Directions

Several research priorities emerge from the body of work assembled in this Special Issue. First, external validation and prospective clinical evaluation remain the most urgent unmet needs. Many published models were trained and tested on single-centre datasets or public benchmarks, with limited evidence of sustained performance under real-world conditions characterized by population shifts, device variability, and incomplete data. Rigorous multicentre and multinational validation is therefore essential before such systems can be considered ready for routine clinical use [5].

Second, interpretability should be treated as a core design principle rather than a post hoc addition. The studies presented here show that explainable AI methods, particularly SHAP-based analyses, can strengthen clinical insight, identify actionable predictors, and build user confidence when incorporated thoughtfully into model development pipelines. Third, privacy-preserving and federated learning approaches deserve substantially greater emphasis as data-governance requirements tighten globally and cross-institutional collaboration becomes increasingly necessary to train robust, generalizable models.

Fourth, multimodal data integration—combining imaging, genomics, laboratory findings, and patient-reported information—remains one of the most promising and underexplored frontiers for improving diagnostic precision. However, robust methods for handling missing or incomplete modalities in real clinical settings are still needed. Fifth, stronger standardization in benchmarking, reporting, and study design is necessary to reduce the persistent problem of inflated or non-transferable performance claims. Frameworks such as TRIPOD + AI provide important guidance in this regard [5]. Finally, future research should expand further into underrepresented domains, including mental health, autoimmune disease, and other clinically challenging areas where diagnostic uncertainty remains high, and ML may offer genuine benefit.

4. Conclusions

The contributions compiled in this Special Issue demonstrate that ML-based disease diagnosis and prediction have evolved from an emerging research area into a credible and increasingly influential paradigm across multiple clinical domains. Spanning medical imaging, clinical risk prediction, oncology, ophthalmology, dentistry, and healthcare systems, the 30 papers collectively reflect both the remarkable breadth of current ML applications in medicine and the persistent methodological challenges that must be addressed before these systems can be fully integrated into clinical practice. This collection underscores that continued progress in the field must be guided by methodological rigor, interpretability, external validation, and sustained attention to fairness and equitable implementation.

The author thanks all contributing authors for their valuable work, the reviewers for their careful and constructive evaluations, and the editorial team at *Diagnostics* for their support throughout the development of this Special Issue. It is hoped that the studies gathered here will stimulate further interdisciplinary research and advance intelligent diagnostic systems that are not only accurate but also trustworthy, equitable, and clinically meaningful.

Funding: This research received no external funding.

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

List of Contributions:

1. Abd El-Aziz, A.A.; Mahmood, M.A.; Abd El-Ghany, S. Advanced Deep Learning Fusion Model for Early Multi-Classification of Lung and Colon Cancer Using Histopathological Images. *Diagnostics* **2024**, *14*, 2274. https://doi.org/10.3390/diagnostics14202274.
2. Abd El-Ghany, S.; Mahmood, M.A.; Abd El-Aziz, A.A. FracFusionNet: A Multi-Level Feature Fusion Convolutional Network for Bone Fracture Detection in Radiographic Images. *Diagnostics* **2025**, *15*, 2212. https://doi.org/10.3390/diagnostics15172212.
3. Yanar, E.; Hardalaç, F.; Ayturan, K. CELM: An Ensemble Deep Learning Model for Early Cardiomegaly Diagnosis in Chest Radiography. *Diagnostics* **2025**, *15*, 1602. https://doi.org/10.3390/diagnostics15131602.
4. Hayat, M.T.; Allawi, Y.M.; Alamro, W.; Sultan, S.M.; Abadleh, A.; Kang, H.; Zreikat, A.I. A Hybrid Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM)–Attention Model Architecture for Precise Medical Image Analysis and Disease Diagnosis. *Diagnostics* **2025**, *15*, 2673. https://doi.org/10.3390/diagnostics15212673.
5. Khan, F.K.; Tahir, W.B.; Lee, M.S.; Kim, J.Y.; Byon, S.S.; Pi, S.-W.; Lee, B.-D. Leveraging Large-Scale Public Data for Artificial Intelligence-Driven Chest X-Ray Analysis and Diagnosis. *Diagnostics* **2026**, *16*, 146. https://doi.org/10.3390/diagnostics16010146.
6. Radke, K.L.; Müller-Lutz, A.; Abrar, D.B.; Vach, M.; Rubbert, C.; Latz, D.; Antoch, G.; Wittsack, H.-J.; Nebelung, S.; Wilms, L.M. Precision Through Detail: Radiomics and Windowing Techniques as Key for Detecting Dens Axis Fractures in CT Scans. *Diagnostics* **2025**, *15*, 2599. https://doi.org/10.3390/diagnostics15202599.
7. Bang, S.; Wang, H.; Bae, H.; Hong, S.-H.; Cha, J.; Choi, M.H. Computed Tomography-Based Radiomics Diagnostic Model for Fat-Poor Small Renal Tumor Subtypes. *Diagnostics* **2025**, *15*, 1365. https://doi.org/10.3390/diagnostics15111365.
8. Ficici, C.; Telatar, Z.; Erogul, O.; Kocak, O. Temporal Lobe Epilepsy Focus Detection Based on the Correlation Between Brain MR Images and EEG Recordings with a Decision Tree. *Diagnostics* **2024**, *14*, 2509. https://doi.org/10.3390/diagnostics14222509.
9. Nechaev, V.A.; Kashtanova, N.Y.; Kopeykin, E.V.; Magomedova, U.M.; Gribkova, M.S.; Hardin, A.V.; Sekacheva, M.I.; Sanikovich, V.D.; Chernina, V.Y.; Gombolevskiy, V.A. Diagnostic Accuracy of a Multi-Target Artificial Intelligence Service for the Simultaneous Assessment of 16 Pathological Features on Chest and Abdominal CT. *Diagnostics* **2025**, *15*, 2778. https://doi.org/10.3390/diagnostics15212778.
10. Faiella, E.; Lamja, S.; Casati, R.; Tondo, M.; Ragone, R.; Redi, A.; Vergantino, E.; Zobel, B.B.; Grasso, F.; Santucci, D. Feasibility of Artificial Intelligence Models for Longitudinal CT Analysis of Epicardial Adipose Tissue After Immunotherapy. *Diagnostics* **2026**, *16*, 852. https://doi.org/10.3390/diagnostics16060852.
11. Wu, P.-C.; Wu, Y.-J.; Hsu, C.-L.; Yu, H.-C.; Chen, C.-S.; Wu, F.-Z. Nomogram for Osteoporosis Risk Using LDCT Trabecular Parameters. *Diagnostics* **2026**, *16*, 1429. https://doi.org/10.3390/diagnostics16101429.
12. Petrović, I.; Broggi, S.; Killer-Oberpfalzer, M.; Pfaff, J.A.R.; Griessenauer, C.J.; Milosavljević, I.; Balenović, A.; Mutzenbach, J.S.; Pikija, S. Predictors of In-Hospital Mortality after Thrombectomy in Anterior Circulation Large Vessel Occlusion: A Retrospective, Machine Learning Study. *Diagnostics* **2024**, *14*, 1531. https://doi.org/10.3390/diagnostics14141531.
13. Kapsner, L.A.; Feißt, M.; Purbojo, A.; Prokosch, H.-U.; Ganslandt, T.; Dittrich, S.; Mang, J.M.; Wällisch, W. Using Machine Learning and Feature Importance to Identify Risk Factors for Mortality in Pediatric Heart Surgery. *Diagnostics* **2024**, *14*, 2587. https://doi.org/10.3390/diagnostics14222587.
14. Guzmán-Muñoz, E.; Vásquez-Muñoz, M.; Concha-Cisternas, Y.; Olivares-Ordenes, R.; Clemente-Suárez, V.; Castillo-Paredes, A.; Yáñez-Sepúlveda, R. Supervised Machine Learning-Based Prediction of In-Hospital Mortality Following Hip Fracture in Older Adults. *Diagnostics* **2026**, *16*, 612. https://doi.org/10.3390/diagnostics16040612.

15. Liu, Y.-Q.; Chang, T.-W.; Lee, L.-C.; Chen, C.-Y.; Hsu, P.-S.; Tsan, Y.-T.; Yang, C.-T.; Chu, W.-M. Use of Machine Learning to Predict the Incidence of Type 2 Diabetes Among Relatively Healthy Adults: A 10-Year Longitudinal Study in Taiwan. *Diagnostics* **2025**, *15*, 72. https://doi.org/10.3390/diagnostics15010072.
16. D'Aversa, E.; Antonica, B.; Grisafi, M.; Asselta, R.; Paraboschi, E.M.; Passaro, A.; Volpato, S.; Remelli, F.; Castellazzi, M.; Marra, A.M.; et al. Integrating Host Genetics and Clinical Setting in Machine Learning Models: Predicting COVID-19 Prognosis for Healthcare Decision-Making (The FeMiNa Study). *Diagnostics* **2026**, *16*, 583. https://doi.org/10.3390/diagnostics16040583.
17. Iftikhar, H.; Hashem, A.F.; Qureshi, M.; Rodrigues, P.C. Clinical Application of Machine Learning Models for Early-Stage Chronic Kidney Disease Detection. *Diagnostics* **2025**, *15*, 2610. https://doi.org/10.3390/diagnostics15202610.
18. Yoon, H.S.; Kim, M.J.; Lim, K.H.; Kim, M.S.; Kang, B.J.; Rah, Y.C.; Choi, J. Evaluating Prediction Models with Hearing Handicap Inventory for the Elderly in Chronic Otitis Media Patients. *Diagnostics* **2024**, *14*, 2000. https://doi.org/10.3390/diagnostics14182000.
19. Aladhadh, S. An Explainable Ensemble and Deep Learning Framework for Accurate and Interpretable Parkinson's Disease Detection from Voice Biomarkers. *Diagnostics* **2025**, *15*, 2892. https://doi.org/10.3390/diagnostics15222892.
20. Omodunbi, B.A.; Olawade, D.B.; Awe, O.F.; Soladoye, A.A.; Aderinto, N.; Ovsepian, S.V.; Boussios, S. Stacked Ensemble Learning for Classification of Parkinson's Disease Using Telemonitoring Vocal Features. *Diagnostics* **2025**, *15*, 1467. https://doi.org/10.3390/diagnostics15121467.
21. Chang, C.-Y.; Pei, D.; Kuo, Y.-L.; Lee, L.-N.; Wu, C.-Z.; Chu, T.-W.; Shen, H.-S.; Huang, C.-Y.; Liang, Y.-J. Explainable Machine Learning Models for Predicting FEV1 in Non-Smoking Taiwanese Men Aged 45–55 Years. *Diagnostics* **2025**, *15*, 3152. https://doi.org/10.3390/diagnostics15243152.
22. Vazquez, B.; Rojas-García, M.; Rodríguez-Esquivel, J.I.; Marquez-Acosta, J.; Aranda-Flores, C.E.; Cetina-Pérez, L.D.; Soto-López, S.; Estévez-García, J.A.; Bahena-Román, M.; Madrid-Marina, V.; et al. Machine and Deep Learning for the Diagnosis, Prognosis, and Treatment of Cervical Cancer: A Scoping Review. *Diagnostics* **2025**, *15*, 1543. https://doi.org/10.3390/diagnostics15121543.
23. Aydın, S.; Ağar, M.; Çakmak, M.; Toğaçar, M. Diagnosis of Mesothelioma Using Image Segmentation and Class-Based Deep Feature Transformations. *Diagnostics* **2025**, *15*, 2381. https://doi.org/10.3390/diagnostics15182381.
24. Lee, Z.-J.; Cai, J.-X.; Wang, L.-H.; Yang, M.-R. Ensemble Algorithm Based on Gene Selection, Data Augmentation, and Boosting Approaches for Ovarian Cancer Classification. *Diagnostics* **2024**, *14*, 2772. https://doi.org/10.3390/diagnostics14242772.
25. Khalid, M.; Sajid, M.Z.; Youssef, A.; Khan, N.A.; Hamid, M.F.; Abbas, F. CAD-EYE: An Automated System for Multi-Eye Disease Classification Using Feature Fusion with Deep Learning Models and Fluorescence Imaging for Enhanced Interpretability. *Diagnostics* **2024**, *14*, 2679. https://doi.org/10.3390/diagnostics14232679.
26. Su, W.; Hoad, D.; Pecchia, L.; Piaggio, D. Validation of an Eye-Tracking Algorithm Based on Smartphone Videos: A Pilot Study. *Diagnostics* **2025**, *15*, 1446. https://doi.org/10.3390/diagnostics15121446.
27. Bennasar, C.; Nadal-Martínez, A.; Arroyo, S.; Gonzalez-Cid, Y.; López-González, Á.A.; Tárraga, P.J. Integrating Machine Learning and Deep Learning for Predicting Non-Surgical Root Canal Treatment Outcomes Using Two-Dimensional Periapical Radiographs. *Diagnostics* **2025**, *15*, 1009. https://doi.org/10.3390/diagnostics15081009.
28. Al-Sharqi, A.J.B.; Baban, M.T.; Imran, N.K.; Gul, S.S.; Abdulkareem, A.A. Comparison of Supervised Machine Learning Models to Logistic Regression Model Using Tooth-Related Factors to Predict the Outcome of Nonsurgical Periodontal Treatment. *Diagnostics* **2025**, *15*, 2333. https://doi.org/10.3390/diagnostics15182333.
29. Khamaysi, Z.; Awwad, M.; Jiryis, B.; Bathish, N.; Shapiro, J. The Role of ChatGPT in Dermatology Diagnostics. *Diagnostics* **2025**, *15*, 1529. https://doi.org/10.3390/diagnostics15121529.

30. Rani, S.; Kumar, R.; Panda, B.S.; Kumar, R.; Muften, N.F.; Abass, M.A.; Lozanović, J. Machine Learning-Powered Smart Healthcare Systems in the Era of Big Data: Applications, Diagnostic Insights, Challenges, and Ethical Implications. *Diagnostics* **2025**, *15*, 1914. https://doi.org/10.3390/diagnostics15151914.

References

1. Topol, E.J. High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nat. Med.* **2019**, *25*, 44–56. [CrossRef] [PubMed]
2. Rajpurkar, P.; Irvin, J.; Ball, R.L.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.P. Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists. *PLoS Med.* **2018**, *15*, e1002686. [CrossRef] [PubMed]
3. Lackowska, A.; Kazimierczak, N.; Chwarścianek, N.; Sultani, N.; Serafin, Z.; Kazimierczak, W. Impact of Imaging Modality on AI-Based Detection of Incidental Maxillary Sinus Pathology: Comparison of Panoramic Radiography and CBCT. *Diagnostics* **2026**, *16*, 1667. [CrossRef] [PubMed]
4. Warner, E.; Lee, J.; Hsu, W.; Syeda-Mahmood, T.; Kahn, C.E., Jr.; Gevaert, O.; Rao, A. Multimodal Machine Learning in Image-Based and Clinical Biomedicine: Survey and Prospects. *Int. J. Comput. Vis.* **2024**, *132*, 3753–3769. [CrossRef] [PubMed]
5. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M. TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods. *BMJ* **2024**, *385*, e078378. [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.

diagnostics

MDPI

Article

Advanced Deep Learning Fusion Model for Early Multi-Classification of Lung and Colon Cancer Using Histopathological Images

A. A. Abd El-Aziz *, Mahmood A. Mahmood and Sameh Abd El-Ghany

Department of Information Systems, College of Computer and Information Sciences, Jouf University, Sakaka 72388, Saudi Arabia; mamahmood@ju.edu.sa (M.A.M.); saabdelwahab@ju.edu.sa (S.A.E.-G.)
* Correspondence: aaeldamarany@ju.edu.sa or zizoah2003@gmail.com; Tel.: +966-545276425

Abstract: Background: In recent years, the healthcare field has experienced significant advancements. New diagnostic techniques, treatments, and insights into the causes of various diseases have emerged. Despite these progressions, cancer remains a major concern. It is a widespread illness affecting individuals of all ages and leads to one out of every six deaths. Lung and colon cancer alone account for nearly two million fatalities. Though it is rare for lung and colon cancers to co-occur, the spread of cancer cells between these two areas—known as metastasis—is notably high. Early detection of cancer greatly increases survival rates. Currently, histopathological image (HI) diagnosis and appropriate treatment are key methods for reducing cancer mortality and enhancing survival rates. Digital image processing (DIP) and deep learning (DL) algorithms can be employed to analyze the HIs of five different types of lung and colon tissues. Methods: Therefore, this paper proposes a refined DL model that integrates feature fusion for the multi-classification of lung and colon cancers. The proposed model incorporates three DL architectures: ResNet-101V2, NASNetMobile, and EfficientNet-B0. Each model has limitations concerning variations in the shape and texture of input images. To address this, the proposed model utilizes a concatenate layer to merge the pre-trained individual feature vectors from ResNet-101V2, NASNetMobile, and EfficientNet-B0 into a single feature vector, which is then fine-tuned. As a result, the proposed DL model achieves high success in multi-classification by leveraging the strengths of all three models to enhance overall accuracy. This model aims to assist pathologists in the early detection of lung and colon cancer with reduced effort, time, and cost. The proposed DL model was evaluated using the LC25000 dataset, which contains colon and lung HIs. The dataset was pre-processed using resizing and normalization techniques. Results: The model was tested and compared with recent DL models, achieving impressive results: 99.8% for precision, 99.8% for recall, 99.8% for F1-score, 99.96% for specificity, and 99.94% for accuracy. Conclusions: Thus, the proposed DL model demonstrates exceptional performance across all classification categories.

Keywords: colon; ResNet-101V2; NASNetMobile; EfficientNet-B0; deep learning; HIs; features fusion; fine-tuning; tumor; cancer

1. Introduction

According to the World Health Organization (WHO), cancer is one of the most lethal and widespread diseases globally, second only to heart disease in terms of its impact on public health. Cancer refers to conditions in which the body produces abnormal cells due to random mutations. Cancer cells grow independently, become genetically unstable, are characterized by the uncontrolled growth and division of abnormal cells, and spread throughout the organs. The American Cancer Society projects that over 606,000 individuals will pass on from the cancer disease. In addition, more than 1.8 million new cases were accounted for in 2020. The brain, breasts, lungs, colon, liver, rectum, skin, prostate, and stomach are the most frequently cancer-affected organs [1].

Cancer can be caused by various factors, including genetic, biological, and physical carcinogens like ultraviolet rays and radiation exposure, smoking, and alcohol consumption, as well as behavioral characteristics like a high body mass index (BMI). However, the reason may differ from patient to patient. Fatigue, pain, difficulty breathing, persistent cough, nausea, weight loss, bleeding, bruising, and muscle pain are all common signs of cancer. Firstly, patients frequently have few or no symptoms, and most of the time, symptoms do not show up until it is too late. None of the symptoms are exclusive to cancer patients [1,2].

Lung cancer causes 18.4% of cancer-related deaths and 11.6% of all cancer cases, whereas colon cancer accounts for 9.2% of all deaths caused by cancer worldwide. The simultaneous occurrence of both lung and colon cancers, known as synchronous lung and colon cancers, is a relatively uncommon occurrence, at about 17% [3]. Although this is rare, metastasis, the spread of cancer cells between the lung and colon, is extremely common without an early diagnosis [4]. Also, colon cancer accounts for 9.2% of all cancer-related deaths worldwide [5,6]. Malignant tumor rates have been increasing worldwide in ongoing patterns, which could be credited to an expansion in the population. All people can get cancer; however, people aged 50 to 60 are more likely to get it. If current trends continue, passing rates could increase by 60% in 2035, as certain evaluations indicate [7,8].

Lung cancer, like many other types of cancer, often begins when normal lung cells undergo genetic mutations that cause them to grow and divide uncontrollably. These mutated cells can then form tumor clusters or masses within the lungs [9]. The most significant factors contributing to the global rise in cancer incidence are the increased number of older people and increased lung exposure to hazardous substances. It is difficult to treat these conditions because their symptoms typically do not show up until they have spread to other organs [10].

Non-smokers can also be affected by lung cancer, but the risk is higher for smokers. There are two main types of lung cancer: Non-small-cell lung cancer (NSCLC), which includes adenocarcinoma and squamous cell carcinoma, and small-cell lung cancer (SCLC) [11].

NSCLC can occur in people who currently smoke, have a smoking history, or have recently quit smoking. While smoking is a major risk factor for lung cancer, including adenocarcinoma, this type of lung cancer is often associated with a broader range of risk factors. It is less related to smoking. Adenocarcinoma of the lung typically originates in the cells that line the smaller airways of the lungs and can form tumors in the lung tissue. It is primarily affecting young women.

SCLC is commonly associated with a history of smoking. It often develops in individuals who are current smokers or have a history of tobacco use, and it is strongly linked to cigarette smoking as a major risk factor. It can develop in any part of the lungs. It grows and spreads so quickly that it is difficult to treat [12,13].

Tumors in the rectum or colon often originate when healthy cells in the lining of the gastrointestinal tract's inner wall (the mucosa) begin to undergo uncontrolled growth and division. This uncontrolled growth can lead to the formation of a mass or lump, which is commonly referred to as a tumor. Most of the time, this kind is malignant [14]. Colon adenocarcinomas are the most common type of colorectal cancer, and they usually originate in the epithelial cells that line the inner surface of the large intestine (colon and rectum). Mucinous adenocarcinoma and signet ring cell adenocarcinoma are two less common subtypes of adenocarcinoma. However, their extreme aggressiveness makes them hard to treat [15]. Your age, ethnicity, financial situation, smoking habits, gender, and other factors can all affect your body's age. However, mutations can occur as quickly as several months if an individual has a rare genetic condition [16].

One way to reduce cancer mortality is by early diagnosis and providing the appropriate patient treatment. When cancer is diagnosed early, the patient's chances of recovery and survival rates are increased. Diagnostic techniques such as MRI, PET scan, CT scan, ultrasound, and biopsy are vital in detecting cancer cells, assessing the extent of cancer,

and aiding in treatment choices. To diagnose and identify the various types and subtypes of cancer, experienced pathologists must examine the microscopic histopathology slides during the biopsy [3]. HIs are crucial in determining a patient's likelihood of survival and are widely used by medical professionals for diagnosis. Traditionally, health professionals undergo a lengthy procedure to diagnose cancer by examining HIs. However, the technological tools that are currently available make it possible to carry out this procedure in less effort and time [3].

The traditional diagnostic methods for lung and colon cancer detection, despite their widespread use, come with several limitations in terms of accuracy, invasiveness, cost, and early detection. These limitations include [17]:

1. **Chest X-rays and CT scans for Lung Cancer:**
   - **Low Sensitivity for Early Detection**: Chest X-rays and CT (computed tomography) scans, commonly used for lung cancer screening, often fail to detect small or early-stage tumors, limiting their ability to identify cancer at an early, more treatable stage.
   - **Radiation Exposure**: Both X-rays and CT scans expose patients to radiation, which can accumulate with repeated screenings and potentially increase cancer risk, especially in patients requiring long-term monitoring.
   - **False Positives**: These imaging techniques can sometimes show benign lesions or abnormalities (like scar tissue or infections) that mimic cancer, leading to unnecessary biopsies, anxiety, or treatments.
   - **Lack of Specificity:** CT scans can detect abnormal lung masses but cannot definitively determine whether they are cancerous. Further invasive procedures, such as biopsies, are required for confirmation.
2. **Sputum Cytology for Lung Cancer:**
   - **Limited Effectiveness for Small or Peripheral Tumors**: Sputum cytology (the examination of mucus for cancer cells) is often ineffective at detecting tumors that are located in the outer parts of the lungs.
   - **Low Sensitivity**: The test has a high chance of missing cancer cells, particularly in the early stages. It is also less effective in detecting non-small-cell lung cancers.
3. **Biopsy for Lung Cancer:**
   - **Invasiveness**: A biopsy, where a tissue sample is removed from the lung for testing, is an invasive procedure that can result in complications like bleeding or infection.
   - **Delay in Diagnosis**: Biopsy results can take time to process, leading to delays in diagnosis and treatment initiation.
4. **Colonoscopy for Colon Cancer:**
   - **Invasiveness and Discomfort**: Colonoscopy is the most common method for diagnosing colon cancer, but it is an invasive procedure that requires bowel preparation and sedation, and involves risks such as bowel perforation, bleeding, and infection.
   - **Patient Reluctance**: Due to the invasive nature and preparation requirements, many patients avoid or delay undergoing a colonoscopy, which can prevent early detection.
   - **False Negatives**: In some cases, small polyps or flat lesions may be missed during a colonoscopy, especially if the bowel preparation is inadequate.
5. **Sigmoidoscopy Colon Cancer:**
   - **Limited Coverage**: Sigmoidoscopy only examines the lower part of the colon, potentially missing cancers in the upper regions of the colon (proximal cancers).
   - **Invasiveness**: Although less invasive than a full colonoscopy, it still involves discomfort and carries some risk of complications like bleeding or perforation.

The shortcomings of conventional diagnostic methods underscore the necessity for more sophisticated, non-invasive, and sensitive diagnostic technologies. Emerging methods, such as molecular imaging, liquid biopsies, and artificial intelligence (AI)-powered diagnostic tools, are being developed to address these limitations by enabling earlier and more accurate detection of lung and colon cancers.

AI technologies have demonstrated tremendous potential in recent years and provided us with a viable alternative to conventional methods. They have gained a reputation for the speed at which they analyze data and make decisions. Machine learning (ML) is used in many ways in pathology, like finding diseases and creating intelligent systems that prescribe conventional medicines based on a patient's symptoms [18]. Several types of biomedical data have been classified and predicted using ML methods. Machines can analyze high-dimensional data like images using DL techniques. DL, a subset of ML, deals with algorithms based on the brain's anatomy and function. Recently, colon and lung cancer diagnosis utilizing DL techniques has become an increasingly popular area of research [19].

This paper proposes a robust fine-tuned DL model of ResNet-101V2, NASNetMobile, and EfficientNet-B0 for multi-classification of colon and lung cancer. The ImageNet dataset was used to pre-train ResNet-101V2, NASNetMobile, and EfficientNet-B0. We used the GlobalAveragePooling2D simultaneously to smooth all layers into a feature vector by obtaining the average for each input channel. Moreover, we used a concatenate layer to fuse all of the pre-trained individual feature vectors of ResNet-101V2, NASNetMobile, and EfficientNet-B0 into a single feature vector. We fine-tuned the fusion single feature vector on the training set of the LC25000 dataset.

The proposed fine-tuned DL model aims to assist pathologists in efficiently de-tecting colon and lung cancer. This model has significant potential to be a valuable tool in pathology, thereby providing appropriate treatment for patients. The proposed DL model has the potential to reduce the effort, time, and cost for pathologists while enhancing early detection and treatment.

The proposed DL model was evaluated on the LC25000 dataset, which contains HIs of the lungs and colon. Normalization and resizing techniques were used for pre-processing the LC25000 dataset. The proposed model achieved 99.8%, 99.8%, 99.8%, 99.9%, and 99.99% for precision, recall, F1-score, specificity, and accuracy, respectively. Therefore, the proposed DL model delivered cutting-edge performance across all classes. A summary of the research's contributions is provided below:

We propose a fine-tuned DL model based on ResNet-101V2, NASNetMobile, and EfficientNet-B0 to detect lung and colon cancer accurately.

We used the feature fusion process to help with learning the features of HIs for the depiction of their rich inner data.

The proposed fine-tuned DL model was not overfitted.

We compared the proposed fine-tuned model to more recent convolutional neural network (CNN) models for a highly accurate multi-classification.

The proposed fine-tuned model could detect colon and lung cancers with high accuracy and less effort and time. It will help pathologists to detect colon and lung cancers early, hence providing appropriate treatment for patients.

The proposed fine-tuned model achieved 99.8%, 99.8%, 99.8%, 99.9%, and 99.99%, for precision, recall, F1-score, specificity, and accuracy.

The remaining sections of this research are structured as follows: Section 2 provides an overview of the existing research on diagnosing lung and colon diseases. Section 3 offers a comprehensive explanation of the model's materials and architecture. The implementation and evaluation of the proposed DL model are discussed in Section 4. Section 5 provides the conclusion of the research.

2. Literature Review

A. H. Chehade1 et al. [1] used five ML methods: random forest (RF), support vector machine (SVM), eXtreme gradient boosting (XGBoost), linear discriminant analysis (LDA), multilayer perceptron (MLP), feature engineering, and image processing techniques to classify the HIs of lung and colon cancer of the LC25000 dataset. They used an image enhancement technique, unsharp masking, to pre-process the dataset. The obtained experimental results demonstrated that ML models accurately identify classes of lung and

colon cancer subtypes and produce satisfactory results. With an F1-score and an accuracy of 98.8% and 99%, respectively, the XGBoost model performed the best.

M. Ali and R. Ali [20] proposed an improved computerized-aided diagnosis (CAD) system for predicting lung and colon adenocarcinomas and squamous cell carcinomas. They developed the CAD system using a capsule network with multiple inputs and digital HIs. They used standard convolutional layers (CLs) in the convolutional layer block (CLB), whereas they used separable CLs in the separable CLB. The CAD system was measured on the LC25000 dataset. The authors pre-processed the dataset using gamma, multi-scale fusion, color balancing, correction, and image sharpening. The accuracy and F1-score of the proposed CAD were 99.58% and 99.04%, respectively; hence, it delivered cutting-edge performance across all classes for colon and lung irregularities.

Sakr, A.S. et al. [21] proposed a new lightweight DL method based on a CNN. The authors pre-processed the LC25000 dataset by a normalization technique. The analysis of the results showed that the proposed DL for detecting colon cancer had an F1-score, precision, recall, and accuracy of 99.49%, 99%, 100%, and 99.50%, respectively.

M. Masud et al. [22] analyzed the HIs of five distinct types of colon and lung tissues, two benign and three malignant, and proposed a CNN classification framework. They used the LC25000 dataset for training and validation of the proposed method. They used two domain transformations to extract four sets of features for image muti-classification. The results demonstrated that the accuracy and F1-score of the proposed CNN classification framework were 96.33% and 96.38%, respectively, for identifying cancerous tissues.

Naresh Kumar et al. [23] compared two methods for feature extraction of the multi-classification tasks of the LC25000 dataset. The first method executed six handcrafted feature extraction methods based on color, texture, shape, and structure. The authors trained and tested four ML techniques: SVM-RBF, RF, MLP, and gradient boosting (GB) for the multi-classification of lung and colon cancers. In the second method, the authors used transfer learning for training seven DL methods for deep feature extraction from HIs with lung and colon cancers. The deeply extracted features were input attributes in SVM-RBF, MLP, RF, and GB techniques to classify lung and colon cancers. According to the findings, the proposed method achieved excellent precision, accuracy, ROC-AUC, F1-score, and recall. The RF model can identify the colon and lung cancer tissue using deep features extracted from DenseNet-121 with a recall and accuracy of 98.60%, F1-score of 98.5%, ROC-AUC of 01, and precision of 98.63%.

M. Shahid et al. [24] used AlexNet and tuned it over the LC25000 colon and lung cancer HI dataset. The proposed model was efficient and accurate for rapid colon and lung cancer diagnosis. Before training ALexNet on the LC25000 dataset, its four layers were altered. The HIs of the LC2500 dataset were pre-processed using resizing, normalizing, and data augmentation methods. With an overall accuracy of 98.4%, the implemented methodology has proven computationally efficient.

Md. A. Talukder et al. [25] proposed a hybrid ensemble model based on a feature extraction approach to detect colon and lung cancers. The proposed model used ensemble techniques and feature extraction. VGG16, VGG1, MobileNet, DenseNet169, and DenseNet201 were used for feature extraction. The authors used the LC25000 dataset to evaluate the model. The accuracy of the hybrid model was 99.30%, 100%, and 99.05% for detecting lung and colon, colon, and lung cancer, respectively.

O. Attallah et al. [26] proposed a framework for detecting lung and colon cancers. The proposed framework used MobileNet, SqueezeNet, and ShuffleNet models to extract features of the images from the LC25000 dataset. The extracted features from the DL models were reduced by principal component analysis (PCA), a feature selection technique, and the fast Walsh–Hadamard transform. Hence, the discrete wavelet transform combined reduced features from the fast Walsh–Hadamard. Moreover, the PCA features are combined. Finally, the reduced features from PCA, the fast Walsh–Hadamard, and the discrete wavelet were fed to several ML techniques. The accuracy of the proposed framework was 99.6%.

J. D. Akinyemi et al. [27] proposed a computer-aided diagnostic (CAD) system based on the EfficientNet model to detect lung and colon cancers. The LC25000 dataset was utilized to evaluate the EfficientNet model. The accuracy of EfficientNet was 99.72%.

In our proposed model, we combined the integrative features from ResNet-101V2, NASNetMobile, and EfficientNet-B0 to enhance the detection of colon and lung tumors. Each model has its strengths in extracting different types of features due to their unique designs. This fusion reduces the weaknesses of individual models while maximizing their strengths, resulting in improved generalization. ResNet-101V2: This model is known for its deep structure and use of residual connections, making it effective at capturing high-level semantic features. NASNetMobile: This model is designed for mobile devices, offering a balance between accuracy and computational efficiency. EfficientNet-B0: This model employs a compound scaling method, allowing it to achieve top-tier performance with fewer parameters.

By merging ResNet-101V2, NASNetMobile, and EfficientNet-B0:

- We created a more comprehensive set of features that can capture both global and local image details across multiple scales and resolutions. This improves the model's ability to detect both small, subtle tumors (early-stage cancer) and large, well-defined ones (advanced-stage cancer).
- The combination of these networks reduces the likelihood of missing important features or making incorrect classifications, thus boosting both sensitivity and specificity in detecting lung and colon tumors.
- By combining multi-scale features, the integrated model can detect tumors of various sizes and appearances, which is crucial in medical imaging, where tumors can be highly variable.
- This hybrid approach ensures that the final model can be deployed in real-world clinical settings, including resource-constrained environments, without sacrificing performance.
- Combining features from these pre-trained models allows for faster convergence and better performance when adapted to lung and colon cancer detection.
- The integrative model can learn from existing knowledge while adapting to the unique characteristics of medical images.
- The integrated model is better suited to handle class imbalance, leading to improved performance in detecting rare cancerous lesions in large datasets dominated by healthy samples.

3. Materials and Methods

3.1. Materials

This research used the LC25000 HI dataset to train and evaluate the proposed DL model. The LC25000 HIs dataset divided colon and lung cancer tissues into five categories: benign lung tissues (lung_b_t), adenocarcinoma (NSCLC) (lung_aca), squamous cell carcinoma (SCLC) (lung_scc), benign colonic tissues (colon_b_t), and colon adenocarcinoma (colon_aca). The HIs of the LC25000 dataset were taken by Andrew A. Borkowski et al. [28]. It has 25,000 HIs, of which 15,000 are for the three lung classes and 10,000 are for the two colon classes. Each class of the five classes has 5000 images [23]. Figure 1 depicts samples from the LC25000 HI dataset.

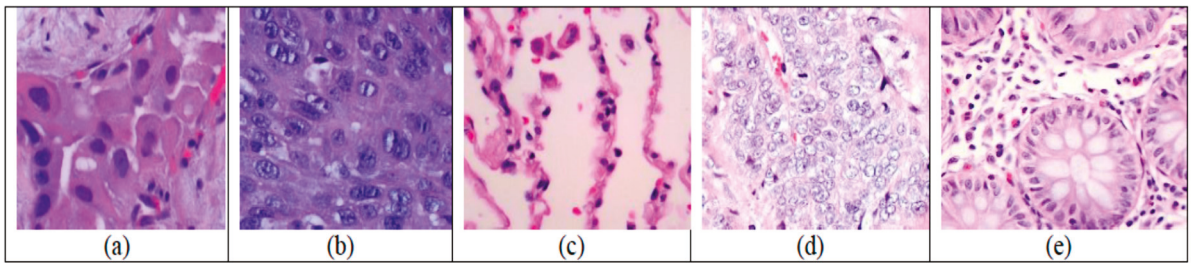


Figure 1. 40× Tissue samples of the LC25000 dataset: (**a**) NSCLC, (**b**) SCLC, (**c**) benign lung tissue, (**d**) colon cancer tissue, and (**e**) benign colon tissue.

Pathology slides contained 1250 images of cancer tissues, with 250 images for each class. Andrew A. Borkowski et al. [28] implemented data augmentation by left and right rotations (with a probability of 1.0 up to 25 degrees) and vertical and horizontal flips (with a probability of 0.5) to bring the dataset up to 25,000 HIs. The original 1024 × 768 pixels were resized to 768 × 768 pixels to reshape the images to a square shape [23].

Colon adenocarcinoma represents over 95% of colon cancers. It is caused by a polyp (tissue growth) known as an adenoma, which later develops into cancer and is mostly found in the large intestine. The most prevalent lung cancer type is lung adenocarcinoma (NSCLC), which accounts for over 40% of all lung cancer cases. It first develops primarily in epithelial cells before spreading to the lungs' alveoli. The second type of lung cancer is squamous cell carcinoma (SCLC), which accounts for nearly 30% of all cases. The bronchi, or airways of the lungs, are where this cancer grows. Although benign colon and lung tumors are typically not life-threatening, surgical removal must be recommended [23].

Table 1 presents the features of the five classes: benign lung tissues, NSCLC, SCLC, benign colonic tissues, and colon cancer tissue [29]:

Table 1. The features of the five classes of the LC25000 dataset.

Class	Features
NSCLC	• Large and irregular nuclei. • Atypical cells with varying shapes and sizes. • Hyperchromatic and prominent nucleoli. • Disorganized tissue structure and evident nuclear pleomorphism.
SCLC	• Small, round to oval cells with scant cytoplasm. • Finely granular chromatin, often called "salt-and-pepper" chromatin. • High nuclear-to-cytoplasmic ratio. • Dense clusters of cancer cells with frequent mitosis and necrosis.
Benign lung tissue	• Well-organized structure of lung tissue with clearly defined alveoli. • Normal-sized nuclei without irregularities. • Lack of abnormal cell growth, typical cell architecture, and absence of mitosis. • Regular shapes and sizes of the cells.
Colon cancer tissue	• Disorganized and distorted glandular structures. • Abnormal nuclear shapes and sizes with prominent nucleoli. • High cellular density and possible necrotic areas. • Loss of normal tissue architecture, with cells invading the surrounding stroma.
Benign colon tissue	• Well-organized glandular structures and epithelial cells. • Uniform nuclei and cells with no signs of atypia. • No invasion into surrounding tissues, and normal tissue architecture is intact. • Lack of inflammatory or necrotic changes.

Cancerous tissues, such as NSCLC, SCLC, and colon cancer, exhibit several distinct features compared to their benign counterparts. Nuclear atypia: Cancerous tissues display a high degree of abnormality in the appearance and structure of the cell nuclei, which are the control centers of the cells. Mitotic activity: Cancerous tissues show an increased and disorganized pattern of cell division, with an abnormal number of cell divisions occurring. Structural disorganization: Cancerous tissues have a disrupted and disorganized internal structure, in contrast to the normal, uniform cell structure observed in benign lung and colon tissues. These differences in nuclear appearance, cell division, and overall tissue organization are key characteristics that distinguish cancerous tissues from their benign counterparts.

3.2. Methodology

We proposed a DL model using three CNN models: ResNet-101V2, NASNetMobile, and EfficientNet-B0. We chose the three DL models for their high performance in analyzing high-dimensional data like images. Each CNN model of ResNet-101V2, NASNetMobile, and EfficientNet-B0 needed to be revised regarding the varieties of texture and shape of the input HI images. However, the proposed model of the three DL models resulted in a notable accomplishment in multi-classification, as it effectively utilized the strengths of each model to enhance the overall accuracy. The proposed model's objective is to analyze the HIs of the LC25000 dataset to detect colon and lung cancers. The phases of the proposed model of the three DL models are depicted in Figure 2. The proposed DL model consists of the following phases:

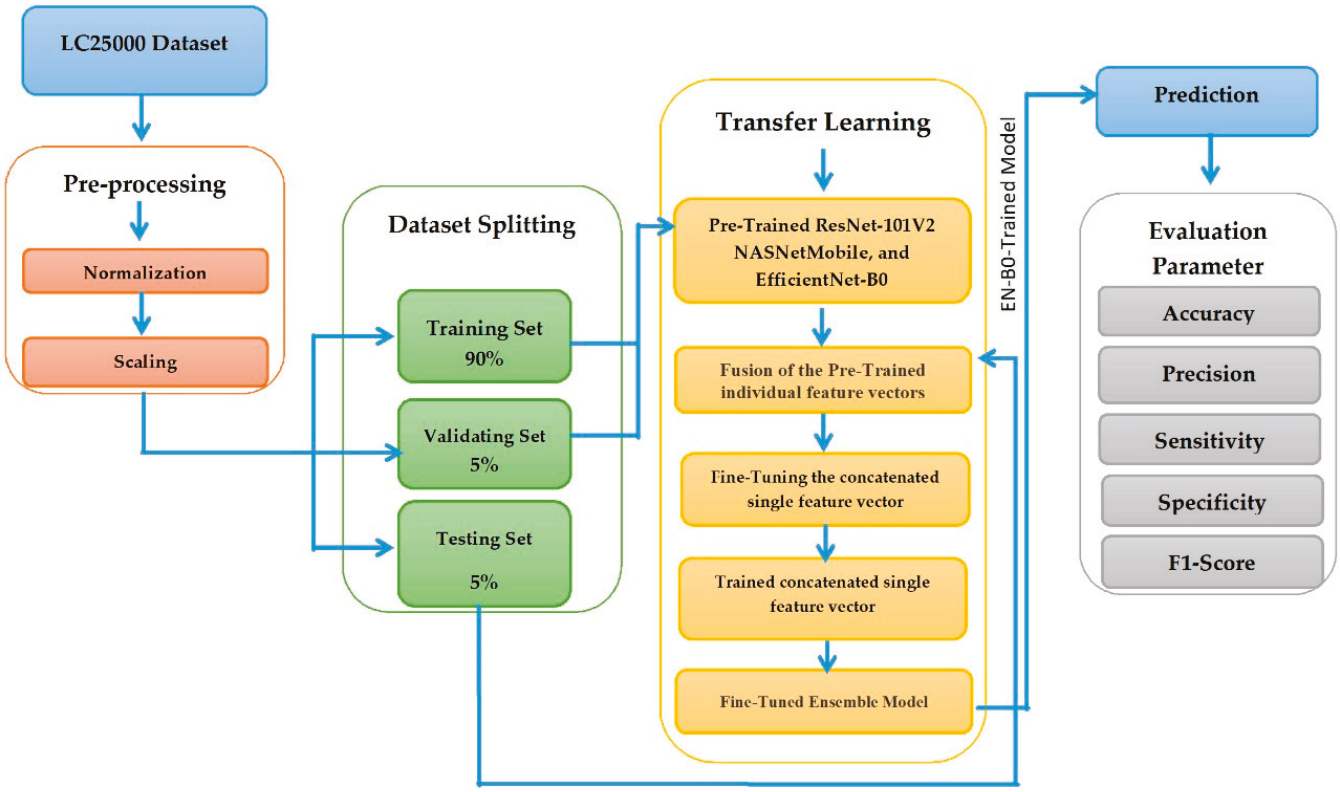


Figure 2. The steps of the proposed DL model.

Phase 1: The LC25000 dataset pre-processing: The LC25000 dataset [28] was downloaded from the Kaggle environment and pre-processed in phase 1. Pre-processing is essential in dealing with the results' accuracy because it effectively yields accurate results. HIs were pre-processed by normalization and resizing methods.

Phase 2: The LC25000 dataset division: In phase 2, the LC25000 dataset was split into 90% (11250 HIs) as a training set, 5% (625 HIs) as a test set, and 5% (625 HIs) as a validation set.

Phase 3: Pre-training for the ResNet-101V2, NASNetMobile, and EfficientNet-B0 models: The three CNN models were pre-trained on the ImageNet dataset. In phase 3, we executed the supervised pre-training phase of the transfer learning process. We trained ResNet-101V2, NASNetMobile, and EfficientNet-B0 on the ImageNet dataset. The pre-trained models used GlobalAveragePooling2D simultaneously to smooth all layers into a feature vector by obtaining the average for each input channel.

Phase 4: Fusion of the pre-trained individual feature vectors: In phase 4, a concatenate layer was used to concatenate all of the pre-trained individual feature vectors of ResNet-101V2, NASNetMobile, and EfficientNet-B0 into a single feature vector.

Phase 5: Fine-tuning the concatenated single feature vector: In phase 5, we fine-tuned the concatenated single feature vector and trained it on the training set of the LC25000 dataset. With this concatenation, more features can be caught; hence, the accuracy of the fine-tuned technique was increased.

Phase 6: The proposed model evaluation: Finally, the proposed technique was measured on the test set of the LC25000 dataset using the measured metrics, like accuracy, F1-score, sensitivity, and precision.

Figure 3 depicts the overall model architecture, Algorithm 1 depicts the algorithm of the proposed model, and Algorithm 2 depicts the training algorithm of a DL mode.

Algorithm 1: The proposed fine-tuned DL model

```
Input → LC25000 dataset DL
Output ← Fine-tuned DL model for lung and colon prediction
BEGIN
    STEP 1: Pre-Processing of HIs
        FOR EACH image IN the DL DO
         Resize HI to 224 × 224
         Normalize HI pixel values from [0, 255] to [0, 1]
        END FOR
  STEP 2: DL Splitting
      SPLIT DL INTO
         Training set → 90%
         Testing set → 5%
         Validation set → 5%
    STEP 3: Model Pre-Training
        FOR EACH ML IN [ResNet-101V2, NASNetMobile, EfficientNet-B0] DO
         Load and pre-train ML on the ImageNet dataset
         Remove the layer of classification from ML for feature extraction
        END FOR
        FOR EACH pre-trained ML DO
         Feed ML with 224 × 224 × 3 HIs
         Feed GlobalAveragePooling2D with the features of ML to smooth them into vectors
        END FOR
        Connect the smoothed feature vectors from the three models into one vector
        Add a SoftMax layer for classification to the connected vector
    STEP 4: Proposed Model Training and Validating
        FOR I = 1 toNwhereN = 5
         FOR EACH I DO
          Fine-tune the connected feature vector on the training set
          Use a validation set for early halting, according to the best model performance
          Assess the performance on the test set
          Record the test accuracies
         END FOR
         Get the average of the recorded test accuracies
     END FOR
    STEP 5: Proposed Model Evaluation
        Evaluate the accuracy of the proposed DL model using the average test accuracies
END
```

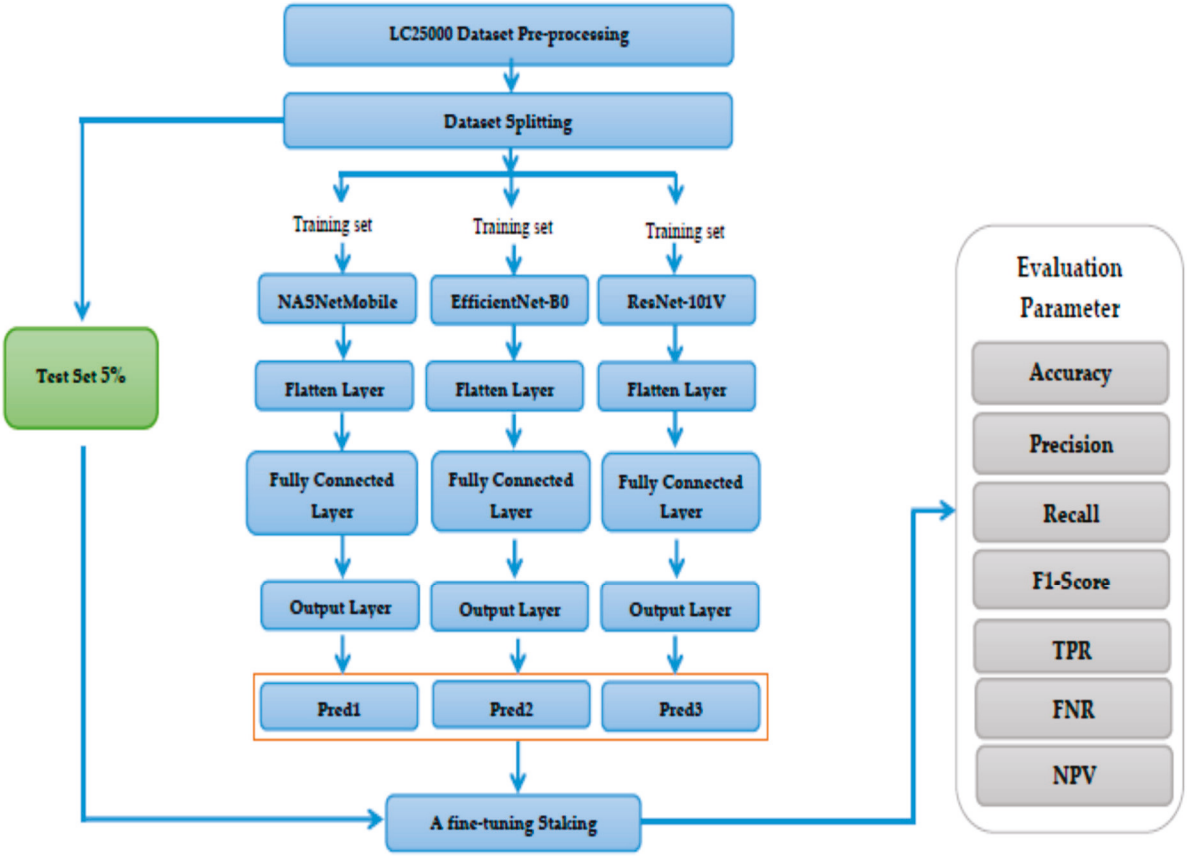


Figure 3. The overall model architecture.

Algorithm 2: The training algorithm of a DL model

Input → LC25000 dataset *DL*
Output ← Trained model
BEGIN
  Inputs→*224* × *224* × *3*
  Batch_S → 25
  Lr → 0.0001
  Dropout → 0.4
  Batch_N → {momentum = 0.99, epsilon = 0.001}
  L →{0,1,2,3,4}.
 G_A_P → Global average pooling
 Dense → = 4
 FOR EACH image IN DL
  Data_images. append(image)
 END FOR
 Splitting DL
 Model.fc → {G_A_P, DROP, Dense}
 Model → Model (inputs = X.inputs, outputs = Model.fc)
 OPT → Adamx (0.0001)
 FOR EACH lay in Model.layers [-20:]
  IF not instance (lay, lay. Batch_N)
    lay.trainable = True
  END IF
 END FOR
 Model.compile(OPT, loss = "sparse_categorical_crossentropy").
END

As depicted in Figure 3, the steps are:

Dataset Pre-Processing.

The first step was dataset pre-processing. This stage involved preparing the LC25000 dataset, which contains images of lung and colon cancer tissues. Pre-processing included resizing and normalization to ensure that the models receive standardized input, which helps improve their learning capabilities.

Dataset Splitting.

The second step was dataset splitting. The dataset was divided into a training set and a test set. The training set was used to train the models, while the test set was held out for final performance evaluation after training.

Deep Feature Fusion.

The third step was combining the DL models. The framework combined three pre-trained DL models: NASNetMobile, EfficientNet-B0, and ResNet-101V2. NASNetMobile was chosen to handle fine texture variations well in HI images due to its unique architecture found through neural architecture search (NAS). EfficientNet-B0's compound scaling method makes it robust in managing both the spatial resolution and complex details of HI images, contributing to the model's effectiveness in recognizing intricate variations in shape and texture. ResNet-101V2 is particularly adept at handling deeper architectures while learning complex features in HI images, such as the structural distinctions between cancerous and benign tissues. Each model processed the images individually, extracting different feature representations of the data.

Model Training Process.

The fourth step involved the model processing layers. In this step, the flattening layer converted the multi-dimensional outputs from the pre-trained models into a 1D vector, making it suitable for input into a fully connected layer. Each flattened output was fed into a fully connected layer, which learned to combine the extracted features from the pre-trained models. This layer acted as a decision-making unit for classification. The output of each model was a prediction that corresponded to the classification of the input image into one of the multiple classes.

Fine-Tuning.

The fifth step was the fine-tuning stage. In this step, the individual model predictions underwent fine-tuning through various techniques to improve performance. Fine-tuning allows the system to adjust the weights of each model's contribution and combine the models' strengths in different aspects of image classification, thereby boosting the final classification accuracy.

Evaluation Metrics.

The final step involved the evaluation metrics. After training, the combined model's performance was evaluated using a series of evaluation parameters.

The combination of NASNetMobile, EfficientNet-B0, and ResNet-101V2 allowed the model to effectively learn different aspects of the HI images. NASNetMobile handled fine texture patterns, EfficientNet focused on overall structural variations, and ResNet-101V2 learned deep features.

Our architecture used three advanced models to improve the classification performance on the LC25000 dataset. The combination of models, fine-tuning, and evaluation metrics ensures the system can accurately detect different classes, including cancerous and benign tissues. This integrated approach will help to overcome the limitations of using a single model by combining the strengths of each model, leading to better accuracy and robustness in cancer detection.

3.2.1. Material Pre-Processing

The pre-processing process was necessary for this research because HIs differed at the pixel level. The processed HIs had varying degrees of pixel density. Loss values at higher image values may differ from those at lower-range values. As a result, the LC25000 dataset

needed to be normalized. Hence, we normalized HI pixel values from [0, 255] to [0, 1] to align the HIs and resized the HIs to 224 × 224.

3.2.2. Transfer Learning

Transfer learning is an ML method where a model is trained on a source task and it is adapted for the target task. It leverages the features and knowledge gained during the training of the source model to improve the model's performance on the target task. Transfer learning has become a common and effective approach in various fields of ML and DL due to its ability to save time, reduce data requirements, and improve the performance of models. Transfer learning allows models to perform well with fewer training data because they start with knowledge acquired from a different but related task. Training CNNs from scratch can be computationally expensive. Transfer learning saves computational resources by using pre-trained models as a starting point. Pre-trained models often capture general features, patterns, and representations useful for various tasks. These learned features can be valuable for a new task. Fine-tuning and feature extraction are the most frequently encountered forms of transfer learning. When performing feature extraction, the initial layers of a pre-existing model are employed to extract features. The extracted features are then input for a new model trained for the target task. Fine-tuning involves taking a pre-trained model and further training some or all of its layers on the target task with a new dataset. Fine-tuning allows the model to adapt to task-specific features while retaining some of the knowledge from the pre-trained model [30].

3.2.3. ResNet-101V2

A residual network (ResNet-101V2) is a CNN with 101 layers. ResNet-101V2 is designed to solve the vanishing gradient problem. The skip connections and residual block ideas are presented in ResNet-101V2. The skip connection associates layers with additional layers by skipping a few in the middle. The skip connection is used to skip a layer if it degrades the performance. ResNets are designed by combining the residual blocks. This model is well known for its deep architecture and residual connections, which help with learning complex features over many layers. ResNet-101V2 is particularly effective at identifying fine details and preserving features across multiple levels, which is useful for detecting texture variations in HI images. ResNet-101V2 was trained on more than 1,000,000 images of the ImageNet dataset. The input images to ResNet-101V2 are 224 × 224 in size [30]. ResNet-101V2's architecture is depicted in Figure 4.

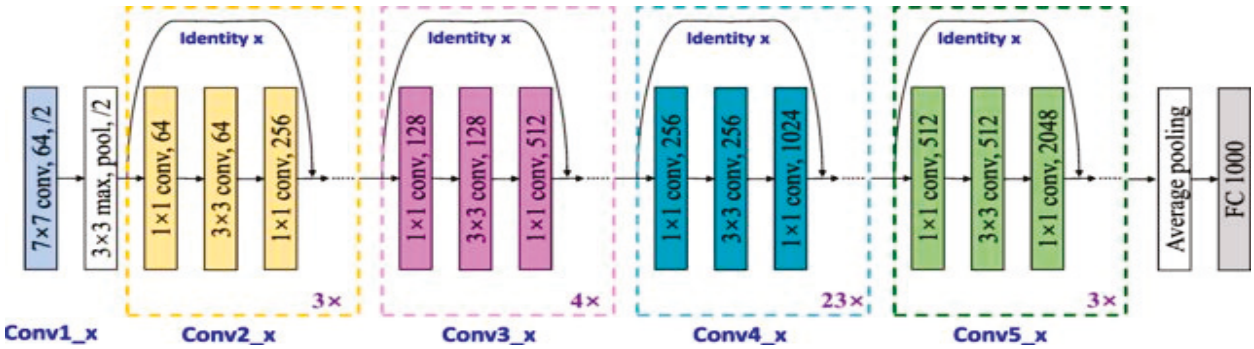


Figure 4. The ResNet-101V2's architecture.

The mathematical model of ResNet-101V2 is represented as follows: Let us denote the input to the block as X, and $F(X,\ W_i)$ is the mapping function of the block parameterized by W_i. The output Y of the residual block is given by:

$$Y = F(X,\ W_i) + X \tag{1}$$

where W_i represents the set of parameters (weights) to be learned during training. The addition operation + is element-wise, and it is performed on the output of the mapping function and the original input. The mapping function $F(X,\ W_i)$ typically consists of the following operations:

1. **Batch Normalization**: Normalize the input:

$$X' = BatchNorm\ (X) \tag{2}$$

2. **ReLU Activation**: Apply the rectified linear unit activation function:

$$X'' = ReLU\left(X'\right) \tag{3}$$

3. **Convolution**: Apply a convolution operation with learnable filters:

$$F(X,\ W_i) = Conv(X'', W_i) \tag{4}$$

4. **Another Batch Normalization**: Normalize the output of the convolution:

$$F'(X,\ W_i) = BatchNorm(F(X,\ W_i)) \tag{5}$$

5. **ReLU Activation**: Apply the rectified linear unit activation function to the output of the second batch normalization:

$$F''\,(X, W_i) = ReLU\left(F'(X, W_i)\right) \tag{6}$$

6. **Convolution**: Apply another convolution operation with learnable filters:

$$F'''\,(X, W_i) = Conv\left(F''\,(X, W_i), W_i'\right) \tag{7}$$

where W_i' represents the set of parameters for the second convolution.

Finally, the output Y is obtained by adding the output of the mapping function to the original input:

$$\mathrm{Y} = F'''\,(X, W_i) + X \tag{8}$$

3.2.4. NASNetMobile

NASNetMobile [31] is a NAS network developed by researchers at Google. It utilizes a unique architecture composed of normal and reduction cells as its core functionalities. The normal cells are designed to maintain the spatial resolution of the input feature maps. This is crucial for preserving detailed information about the input data, such as subtle patterns and textures in the HIs. They focus on learning intricate patterns and feature representations at each layer while keeping the input dimensions unchanged. This allows the network to maintain fine-grained details, which is especially important for differentiating between normal and cancerous cells in HI images, where subtle differences in texture and structure are critical [32].

The reduction cells reduce the spatial dimensions of the input feature maps by downsampling the input. This reduces the computational cost and allows the network to focus on more global features by abstracting higher-level patterns from the data. They capture higher-level, more abstract features from the input data, such as overall shape and structural variations. This is important for distinguishing large-scale features in complex images, such as differentiating between different types of tissues in colon or lung HI images. By pooling features and reducing the spatial resolution, the network can learn hierarchical feature representations, helping to manage complexity without overwhelming the model with high-dimensional data [32].

By alternating between normal and reduction cells, the NASNet architecture strikes a balance between preserving detailed, local features (via normal cells) and capturing global, abstract features (via reduction cells). This balance is crucial when dealing with high-

resolution HI images that require both local texture analysis, such as detecting abnormal cell growth, and broader context recognition, such as differentiating between various tissue types.

The combination of normal and reduction cells allows NASNet to efficiently manage the complexity of the input data. For HI images, which contain both fine texture details and large structural variations, this design helps extract a wide range of relevant features. The reduction cells help prevent overfitting by reducing spatial resolution at deeper layers, while the normal cells ensure that the fine details needed for accurate classification are not lost.

The NASNetMobile model was introduced as part of the NAS project, aiming to automatically discover the optimal network architecture for a given task. Its efficient design and ability to balance accuracy and computational requirements make it well suited for mobile applications and devices with limited resources. The NASNetMobile network is built upon reinforcement learning-optimized fundamental building cells. These cells incorporate multiple convolution, separable-convolution, and pooling operations, enhancing the model's reliability and performance. To further improve NASNetMobile's efficiency, a modified droppath technique called scheduled droppath is employed for effective regularization. The automated architecture search capabilities of NASNetMobile allow it to identify the best-performing network structure for specific tasks, such as image classification. This adaptability and optimization for mobile platforms has contributed to NASNetMobile's widespread adoption and success in various applications. For texture and shape variations, NASNetMobile is particularly effective due to its flexible architecture that can capture diverse features from different image scales. Figure 5 depicts the original NASNetMobile architecture, in which reduction and normal cells are specifically utilized and the number of cells is not predetermined. The control framework within NASNetMobile can anticipate and forecast the complete network layout. To train the NASNetMobile model, around 5.3 million parameters were utilized, specifically tailored for input images that are 224 × 224 pixels in size.

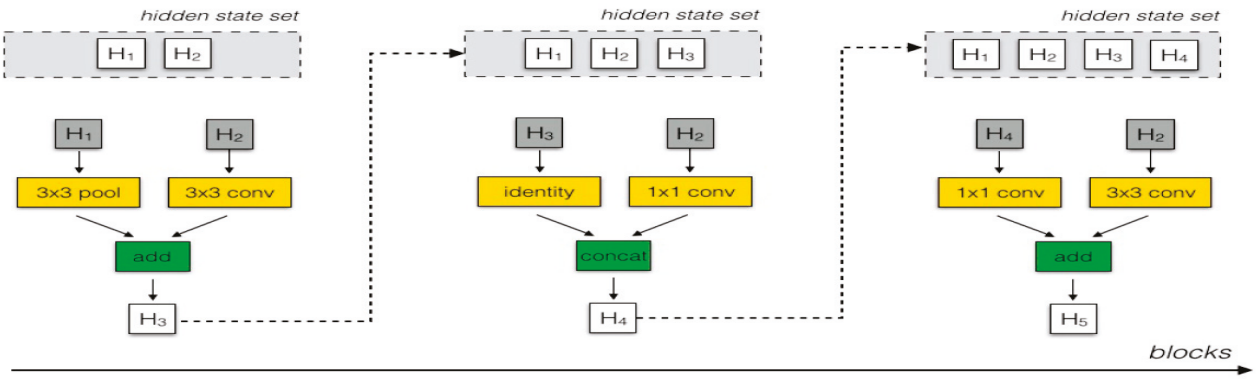


Figure 5. The architecture of the reduction cell and NASNet normal.

The mathematical model of NASNetMobile is represented as follows: Let us denote the input to the cell as X, and $F(X,\ W_i)$ is the mapping function of the cell parameterized by W_i. The output Y of the cell is given by:

$$Y = F(X,\ W_i) + X \tag{9}$$

where W_i represents the set of parameters (weights) to be learned during training. The addition operation + is element-wise, and it is performed on the output of the mapping function and the original input. The mapping function $F(X,\ W_i)$ typically consists of the following operations:

1. **Separable Convolution**: Apply a separable convolution operation:

$$F(X,\ W_i) = SepConv(X,\ W_i) \tag{10}$$

2. **Batch Normalization**: Normalize the output of the separable convolution:

$$X' = BatchNorm\ (X) \tag{11}$$

3. **Convolution**: Apply a convolution operation with learnable filters:

$$F'(X,\ W_i) = BatchNorm(F(X,\ W_i)) \tag{12}$$

4. **ReLU Activation**: Apply the rectified linear unit activation function to the output of the batch normalization:

$$F''\,(X,\ W_i) = ReLU\big(F'(X, W_i)\big) \tag{13}$$

5. **Pooling:** Apply the pooling operation:

$$F'''\,(X,\ W_i) = Pooling(F''\,(X, W_i)) \tag{14}$$

Finally, the output Y is obtained by adding the output of the mapping function to the original input:

$$\mathrm{Y} = F'''\,(X, W_i) + X \tag{15}$$

3.2.5. EfficientNet-B0

EfficientNet is a compound scaling method that scales network width, length, and resolution using a group of fixed coefficients. Scaling width increases feature maps in each layer. Scaling depth increases the layers of the network. Scaling resolution increases the resolution of input images [33]. Tan and Le [33] proposed seven models, ranging from EfficientNet-B0 to EfficientNet-B7. When applied to the ImageNet dataset, EfficientNet CNN models demonstrated superior performance in terms of parameters and the number of Top-1 accuracy [34].

The most crucial part of the EfficientNet model family is the mobile inverted bottleneck convolution (MBConv). MBConv was developed based on the concepts of the MobileNet models [35]. One of the main ideas is the use of depth-wise separable convolutions, which necessitates a depth-wise convolution and a layering pointwise. The two additional ideas were born out of MobileNet-V2: linear bottlenecks and inverted residual connections.

EfficientNet-B0 aims to achieve improved classification accuracy while using fewer computational resources. It accomplishes this by scaling the depth, width, and resolution of the network in a balanced manner. The compound scaling approach of EfficientNet-B0 helps capture both large-scale and fine-grained features related to shape and texture, making it well suited for the task of HI image classification.

EfficientNet-B1 is the first scaled version, with roughly twice as much capacity. There are 237 layers in EfficientNet-B0 and 813 layers in EfficientNet-B7.

Table 2 displays the specifics of the scale for the models EfficientNet-B0, and Figure 6 shows the architecture of EfficientNet-B0 [36].

Table 2. EfficientNet-B0.

Model Name	W Coeff.	D. Coeff.	Resolution	Dropout Rate
EfficientNet-B0	1.0	1.0	224	0.2

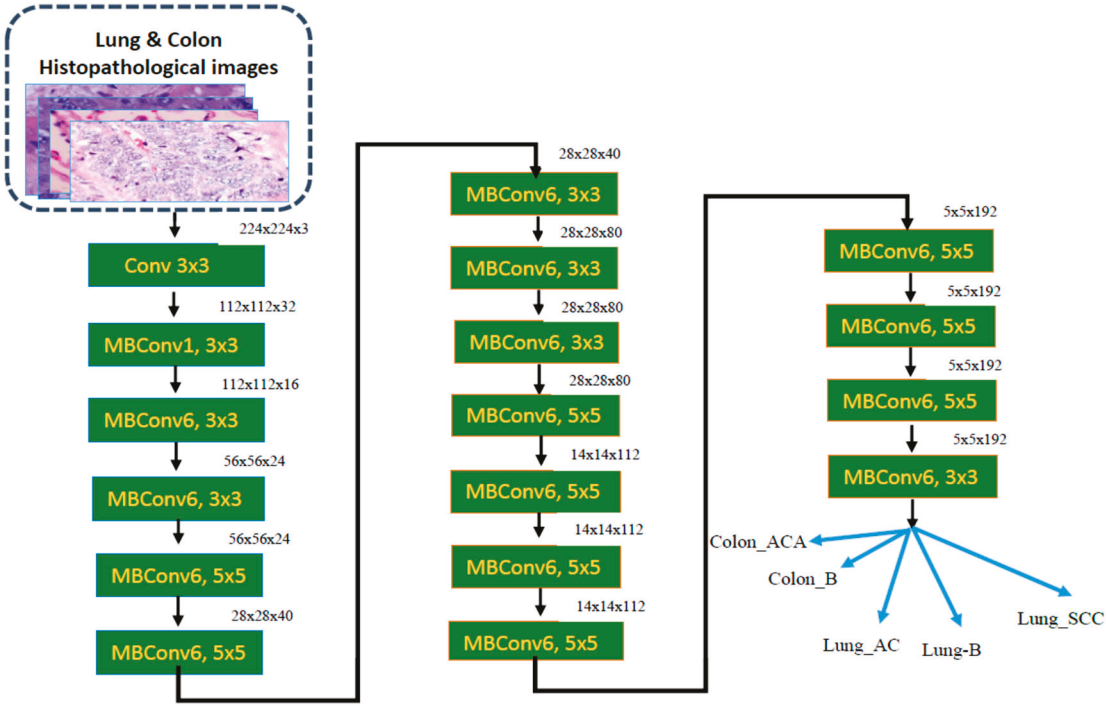


Figure 6. The architecture of EfficientNet-B0.

The mathematical model of EfficientNet-B0 is represented as follows: Let us denote the input to the block as X, and $F(X,\ W_i)$ is the mapping function of the block parameterized by W_i. The output Y of the block is given by:

$$Y = F(X,\ W_i) + X \tag{16}$$

where W_i represents the set of parameters (weights) to be learned during training. The addition operation + is element-wise, and it is performed on the output of the mapping function and the original input. The mapping function $F(X,\ W_i)$ typically consists of the following operations:

1. **Depth-wise Separable Convolution**: Apply a depth-wise separable convolution operation:

$$F(X,\ W_i) = DepthwiseSepConv\ (X,\ W_i) \tag{17}$$

2. **Batch Normalization**: Normalize the output of the depth-wise separable convolution:

$$F'(X,\ W_i) = BatchNorm(F(X,\ W_i)) \tag{18}$$

3. **Swish Activation**: Apply the swish activation function to the output of batch normalization:

$$F''(X, W_i) = Swich(F'(X, W_i)) \tag{19}$$

4. **Squeeze-and-Excitation**: Apply the squeeze-and-excitation operation to capture channel-wise dependencies:

$$F'''(X,\ W_i) = SE(F''(X, W_i)) \tag{20}$$

5. **Linear Projection**: Project the features using a linear projection:

$$F'''(X, W_i) = LinerProjection(F'''(X, W_i)) \tag{21}$$

Finally, the output Y is obtained by adding the output of the mapping function to the original input:

$$\mathrm{Y} = F'''(X, W_i) + X \tag{22}$$

3.2.6. Integration of CNN Features

Integrating features from CNNs means combining the strengths of various CNN designs to enhance the performance of a DL model. This process, called integrative feature fusion, can result in better accuracy, generalization, and robustness. Each CNN extracts features from the input data, capturing different levels of abstraction. These features range from low-level details like edges and textures to high-level semantic information. By merging the features extracted from different CNNs, we create a richer and more comprehensive representation of the data. A new model architecture is then developed to leverage these fused features, often leading to improved performance [37].

In our model, we used three DL models: ResNet-101V2, NASNetMobile, and EfficientNet-B0. To handle the texture complexities in HI images, the combination of ResNet-101V2 and EfficientNet-B0 played a crucial role. ResNet-101V2's depth and ability to extract detailed patterns across many layers helped with recognizing fine-grained textures, while EfficientNet-B0's efficient scaling enabled the model to generalize well across varying texture resolutions. NASNetMobile, with its ability to dynamically adjust its architecture during training, allowed the model to better capture variations in the shape of HI structures. This, combined with the hierarchical feature extraction of ResNet-101V2, made the ensemble capable of distinguishing between different shape variations more effective.

The outputs from the intermediate layers of each model were extracted and fused to create a more robust representation of the image. This ensured that the model benefitted from the strengths of each architecture: ResNet-101V2's detailed texture extraction, NASNetMobile's adaptable architecture, and EfficientNet-B0's balanced feature scaling. Attention layers were added after the feature fusion to ensure the model focused on the most important regions in the images, thus improving performance when texture and shape variations were subtle. After the integrative feature extraction, a custom dense classifier was used to better distinguish between the classes. The classifier was fine-tuned using techniques like dropout and batch normalization to ensure it generalized well across different classes. The integrative approach led to a noticeable improvement in multiclassification accuracy. By fusing features from diverse architectures, the model captured a richer set of representations, addressing the issues of texture and shape variations more effectively than individual models. Additionally, the use of attention mechanisms and custom classifiers further boosted performance, helping the model better discriminate between different classes. This approach not only enhanced the model's ability to handle the complex variations in HI images but also raised the overall classification accuracy, making it a more robust solution for detecting lung and colon cancer.

In our implementation, we initially used the ImageNet dataset, which has more than millions of images, to pre-train the three CNN models. Each model had an input of images of $244 \times 244 \times 3$ in size. ResNet-101V2 extracted 2048 features, NASNetMobile extracted 1956 features, and EfficientNet-B0 extracted 1280 features. After the pre-training phase, the pre-trained models used the GlobalAveragePooling2D simultaneously to smooth all layers into a feature vector by obtaining the average for each input channel. Moreover, a concatenate layer was used to fuse the pre-trained individual feature vectors of ResNet-101V2, NASNetMobile, and EfficientNet-B0 into a single feature vector. In addition, the fusion of a single feature vector was fine-tuned on the training set of the LC25000 dataset. In the fine-tuning phase, we fine-tuned the concatenated single feature vector on the test set of the LC25000 dataset by using six layers (dropout, batch_normalization, dense, dropout_1, batch_normalization_1, and dense_1). Since the LC25000 dataset has five classes, the last dense layer had five neurons. Table 3 shows the structure of the proposed fine-tuned DL model.

Table 3. Structure of the proposed fine-tuned DL model.

Layer	Input	Output
InputLayer	(None,224,224,3)	(None,224,224,3)
EfficientNet-B0	(None,224,224,3)	(None,7,7,1280)
NASNetMobile	(None,224,224,3)	(None,7,7,1056)
ResNet-101V2	(None,224,224,3)	(None,7,7,2048)
GlobalAveragePooling2D	(None,7,7,1280)	(None,1280)
GlobalAveragePooling2D	(None,7,7,1056)	(None,1056)
GlobalAveragePooling2D	(None,7,7,2048)	(None,2048)
Concatenate	(None,1280), (None,1056), (None,2048)	(None,4384)
Dropout	(None,4384)	(None,4384)
BatchNormailzation	(None,4384)	(None,4384)
Dense	(None,4384)	(None,128)
Dropout	(None,128)	(None,128)
BatchNormailzation	(None,128)	(None,128)
Dense	(None,128)	(None,5)

4. Model Implementation and Evaluation

4.1. Measured Metrics

The LC25000 dataset was utilized to measure our proposed model's specificity, accuracy, FNR, NPV, precision, recall, and F1-score, as mentioned in Equations (23)–(29):

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{23}$$

$$\text{Precision} = \frac{TP}{(TP + FP)} \tag{24}$$

$$\text{Sensivity} = \frac{TP}{(TP + FN)} \tag{25}$$

$$\text{Specifity} = \frac{TN}{(TN + FP)} \tag{26}$$

$$\text{F1} - \text{score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{27}$$

$$\text{False negative rate (FNR)} = \frac{FN}{TP + FN} \tag{28}$$

$$\text{Negative predictive value (NPV)} = \frac{TN}{TN + FN} \tag{29}$$

NPV means the ratio of TN predictions that are normal. FNR means the ratio of FN. TP + FN is the sum of the patients who had tumors. True positive (TP) indicates a positive real value and a positive classification. False negative (FN) indicates that the actual value is positive even though it is categorized as negative. False positive (FP) means that the true value is negative when it is classified positively. True negative (TN) indicates a negative real value and a negative classification.

The accuracy is the percentage of the number of correct predictions out of the total number of predictions. Precision is the ratio of true positive predictions to the total number of positive predictions (both true positives and false positives). The proportion of actual cases with positive outcomes that were predicted to be positive is called sensitivity. The proportion of predicted negatives that occurred is known as specificity. The harmonic mean of precision and sensitivity is the F1-score.

4.2. The Proposed Fine-Tuned DL Model Implementation

In this research, we implemented the proposed fine-tuned ensemble model using the Kaggle platform. Due to time constraints during the study, we did not employ formal hyperparameter optimization methods, such as grid search or Bayesian optimization. Instead, we opted for standard default values based on their established reliability in the literature—specifically, learning rate, dropout rate, and batch size. The values of 0.001 for the learning rate, 0.4 for the dropout rate, and 25 for the batch size were selected, as they are commonly used in similar models and tasks. These defaults were chosen to ensure stable model performance without requiring extensive tuning. The hyperparameter values used in the implementation are shown in Table 4.

Table 4. The values of the hyperparameters.

Parameter	Value
img_size	224×224
Number of epochs	15
Batch size	25
Activation	Softmax
Optimizer	Adam
Initial learning rate	1×10^{-3}
Decay	0.00001
Factor to reduce LR	0.5
Dropout	0.4

4.3. The Proposed DL Model Evaluation

ResNet-101V2, NASNetMobile, EfficientNet-B0, and the proposed fine-tuned DL model were evaluated over the LC25000 dataset. It was separated into a 90% training set with 11,250 HIs, a 5% test set with 625 HIs, and a 5% validation set with 625 HIs.

We applied the transfer learning process that pre-trained the three CNN models on the ImageNet dataset. After the supervised pre-training phase, they used GlobalAveragePooling2D simultaneously to smooth all layers into a feature vector by obtaining the average for each input channel. A concatenate layer was used to fuse the pre-trained individual feature vectors of ResNet-101V2, NASNetMobile, and EfficientNet-B0 into a single feature vector. In addition, the fusion single feature vector was fine-tuned on the training set of the LC25000 dataset. Finally, the evaluated metrics in Equations (23)–(29) were applied to ResNet-101V2, NASNetMobile, EfficientNet-B0, and the proposed fine-tuned DL model.

We implemented five experiments for the multi-classification of the LC25000 dataset using ResNet101, NASNetMobile, EfficientNet-B0, the proposed fusion DL model, and the stacked ensemble model.

In the first three experiments, ResNet101, NASNetMobile, and EfficientNet-B0 were pre-trained over the ImageNet dataset and fine-tuned on the training set of the LC25000 dataset to fine-tune their parameters. They used GlobalAveragePooling2D simultaneously to smooth all layers into a feature vector by obtaining the average value for each input channel. Finally, each CNN model was used for the multi-classification and measured by the measured metrics.

In the fourth experiment, we used the stacking algorithm to stack ResNet-101V2, NASNetMobile, and EfficientNet-B0 based on the majority voting technique. The stacked ensemble model was used for the multi-classification and evaluated by the evaluation metrics.

In the five experiments, we combined each CNN model's last fully connected layers, fine-tuned the fusion single feature vector on the training set of the LC25000 dataset, and evaluated it by the measured metrics.

Tables 5–9 demonstrate the multi-classification outcomes of the five experiments on the test set of the LC25000 using the proposed fusion DL model, the stacked ensemble model, EfficientNet-B0, NASNetMobile, and ResNet-101V2, respectively. The five experiments clas-

sified the five tissue classes of the LC25000 dataset: benign lung tissues, adenocarcinoma, squamous cell carcinoma, benign colonic tissues, and colon adenocarcinoma.

Table 5. The proposed fusion results from the multi-classification.

Class	Accuracy (%)	Specificity (%)	FNR (%)	NPV (%)	Precision (%)	Recall (%)	F1-Score (%)
Colon_ACA	100	100	0	100	100	100	100
Colon_B_T	100	100	0	100	100	100	100
Lung_ACA	99.84	99.80	0	100	99.25	100	99.62
Lung_B_T	100	100	0	100	100	100	100
Lung_SCC	99.84	100	0.80	99.80	100	99.20	99.60
Average	**99.94**	**99.96**	**0.16**	**99.96**	**99.84**	**99.84**	**99.84**

Table 6. The stacked ensemble model results from the multi-classification process.

Class	Accuracy (%)	Specificity (%)	FNR (%)	NPV (%)	Precision (%)	Recall (%)	F1-Score (%)
Colon_ACA	100	100	0	100	100	100	100
Colon_B	100	100	0	100	100	100	100
Lung_ACA	98.88	99.39	3.03	99.19	97.71	96.97	97.34
Lung-B	99.84	100	0.85	99.80	100	99.15	99.57
Lung_SCC	99.04	99.20	1.60	99.60	96.85	98.40	97.62
Average	**99.55**	**99.72**	**1.10**	**99.72**	**98.89**	**98.88**	**98.88**

Table 7. The results of EfficientNet-B0 for the multi-classification process.

Class	Accuracy (%)	Specificity (%)	FNR (%)	NPV (%)	Precision (%)	Recall (%)	F1-Score (%)
Colon_ACA	100	100	0	100	100	100	100
Colon_B	100	100	0	100	100	100	100
Lung_ACA	92.96	99.80	32.58	91.96	98.89	67.42	80.18
Lung-B	99.84	100.00	0.85	99.80	100.00	99.15	99.57
Lung_SCC	93.12	91.40	0.00	100.00	74.40	100.00	85.32
Average	**97.18**	**98.24**	**6.68**	**98.35**	**94.65**	**92.96**	**92.80**

Table 8. The results of NASNetMobile for the multi-classification process.

Class	Accuracy (%)	Specificity (%)	FNR (%)	NPV (%)	Precision (%)	Recall (%)	F1-Score (%)
Colon_ACA	89.44	100	54.55	88.42	100	45.45	62.50
Colon_B	100	100	0	100	100	100	100
Lung_ACA	85.92	82.96	3.03	99.03	60.38	96.97	74.42
Lung-B	96.96	100.00	16.10	96.39	100.00	83.90	91.24
Lung_SCC	96.00	97.00	8	97.98	88.46	92.00	90.20
Average	**93.66**	**95.99**	**16.34**	**96.36**	**89.32**	**84.16**	**83.72**

Table 9. The results of ResNet-101V2 for the multi-classification process.

Class	Accuracy (%)	Specificity (%)	FNR (%)	NPV (%)	Precision (%)	Recall (%)	F1-Score (%)
Colon_ACA	100	100	0	100	100	100	100
Colon_B	100	100	0	100	100	100	100
Lung_ACA	99.36	99.59	1.52	99.59	98.48	98.48	98.48
Lung-B	100	100	0.00	100	100	100	100
Lung_SCC	99.36	99.60	1.60	99.60	98.40	98.40	98.40
Average	**99.74**	**99.84**	**0.62**	**99.84**	**99.36**	**99.36**	**99.36**

Tables 5–9 demonstrate the average measured metrics for the proposed fusion DL model, the stacked ensemble model, EfficientNet-B0, NASNetMobile, and ResNet-101V2, respectively. The average accuracy was 99.94%, 99.55%, 97.18%, 93.66%, and 99.74%, respectively, for the proposed fusion DL model, the stacked ensemble model, EfficientNet-B0, NASNetMobile, and ResNet-101V2. **Hence, the fusion DL model had the best results, with an accuracy of 99.94%.**

The proposed fusion model recorded averages of 99.96%, 0.16%, 99.96%, 99.84%, 99.84%, and 99.84% for specificity, FNR, NPV, precision, recall, and F1-score, respectively. The average specificity, FNR, NPV, precision, recall, and F1-score for the stacked ensemble model were 99.20%, 1.60%, 99.60%, 96.85%, 98.40%, and 97.62%, respectively. EfficientNet-B0 yielded 98.24%, 6.68%, 98.35%, 94.65%, 92.96%, and 92.80% for the specificity, FNR, NPV, precision, recall, and F1-score averages, respectively. NASNetMobile achieved 95.99%, 16.34%, 96.36%, 89.32%, 84.16%, and 83.72% for the specificity, FNR, NPV, precision, recall, and F1-score averages, respectively. ResNet-101V2 recorded 99.84%, 0.62%, 99.84%, 99.36%, 99.36%, and 99.36% for the specificity, FNR, NPV, precision, recall, and F1-score averages, respectively.

Therefore, the proposed fusion DL model recorded the highest averages of specificity, NPV, precision, recall, and F1-score, at 99.96%, 99.96%, 99.84%, 99.84%, and 99.84%, respectively, and it achieved the lowest FNR, at 0.16%. NASNetMobile achieved the highest FNR, at 16.34%. Hence, **the proposed fusion DL model performed well, with relatively high values for most of the evaluation criteria, making it a strong performer for the multi-classification task of lung and colon tumors**.

Tables 5–9 provide an in-depth look at the multi-classification process's outcomes. The LC25000 was divided into five categories through this procedure: Colon_Adenocarcinoma (Colon_ACA), Colon_Benign_Tissue (Colon_B_T), Lung-Benign_Tissue (Lung_B_T), Lung_Adenocarcinoma (Lung_ACA), and Lung_Squamous_Cell_Carcinoma (Lung_SCC). Over the LC25000 dataset, the proposed fusion DL model, the stacked ensemble model, EfficientNet-B0, NASNetMobile, and ResNet-101V2 were evaluated through the calculations of the accuracy, precision, specificity, recall, FNR, NPV, and F1-score for each class.

For the **Colon_Adenocarcinoma** class, the accuracy, precision, specificity, recall, NPV, and F1-score were 100% for the proposed fusion model, the stacked ensemble model, EfficientNet-B0, and ResNet-101V2, and they achieved 0% for FNR.

For the **Colon_Benign_Tissue** class, the accuracy, precision, specificity, recall, NPV, and F1-score were 100% for the proposed fusion model, the stacked ensemble model, EfficientNet-B0, NASNetMobile, and ResNet-101V2, and they achieved 0% for FNR.

For the **Lung_Adenocarcinoma** class, the proposed fusion model achieved the highest accuracy, specificity, NPV, precision, recall, and F1-score at 99.84%, 99.80%, 100%, 99.25%, 100%, and 99.62%, respectively, and it recorded the lowest FNR at 0%. Moreover, EfficientNet-B0 had the highest specificity at 99.80%.

For the **Lung-Benign_Tissue** class, the proposed fusion model and ResNet-101V2 achieved the highest accuracy, specificity, NPV, precision, recall, and F1-score at 100%, and it recorded the lowest FNR at 0%. In addition, the proposed stacked ensemble model, EfficientNet-B0 and NASNetMobile, had the highest specificity and precision at 100%.

For the **Lung_Squamous_Cell_Carcinoma** class, the proposed fusion recorded the highest precision, specificity, accuracy, and F1-score at 100%, 100%, 99.84%, and 99.60%, respectively. EfficientNet-B0 had the highest NPV and recall at 100% and the lowest FNR at 0%.

The training and validation loss of the proposed fusion DL model, EfficientNet-B0, NASNetMobile, and ResNet-101V2, is depicted in Figure 7. For the proposed fusion DL model, the training loss began at a very low value and remained consistently low throughout all training epochs. This suggests that the model quickly minimized the loss on the training data and maintained a very low error, indicating a good fit to the training data. The validation loss started at a significantly high value (approximately 7) and decreased sharply during the initial few epochs. After about five epochs, the validation loss stabilized

and remained very low. The rapid drop in validation loss followed by stabilization suggests that the model quickly learned to generalize well on unseen validation data. There was no evidence of overfitting, as both training and validation losses were low and aligned after the initial few epochs. Moreover, the model efficiently reduced both training and validation losses early in the training process, indicating that it generalized well to new data after only a few epochs.

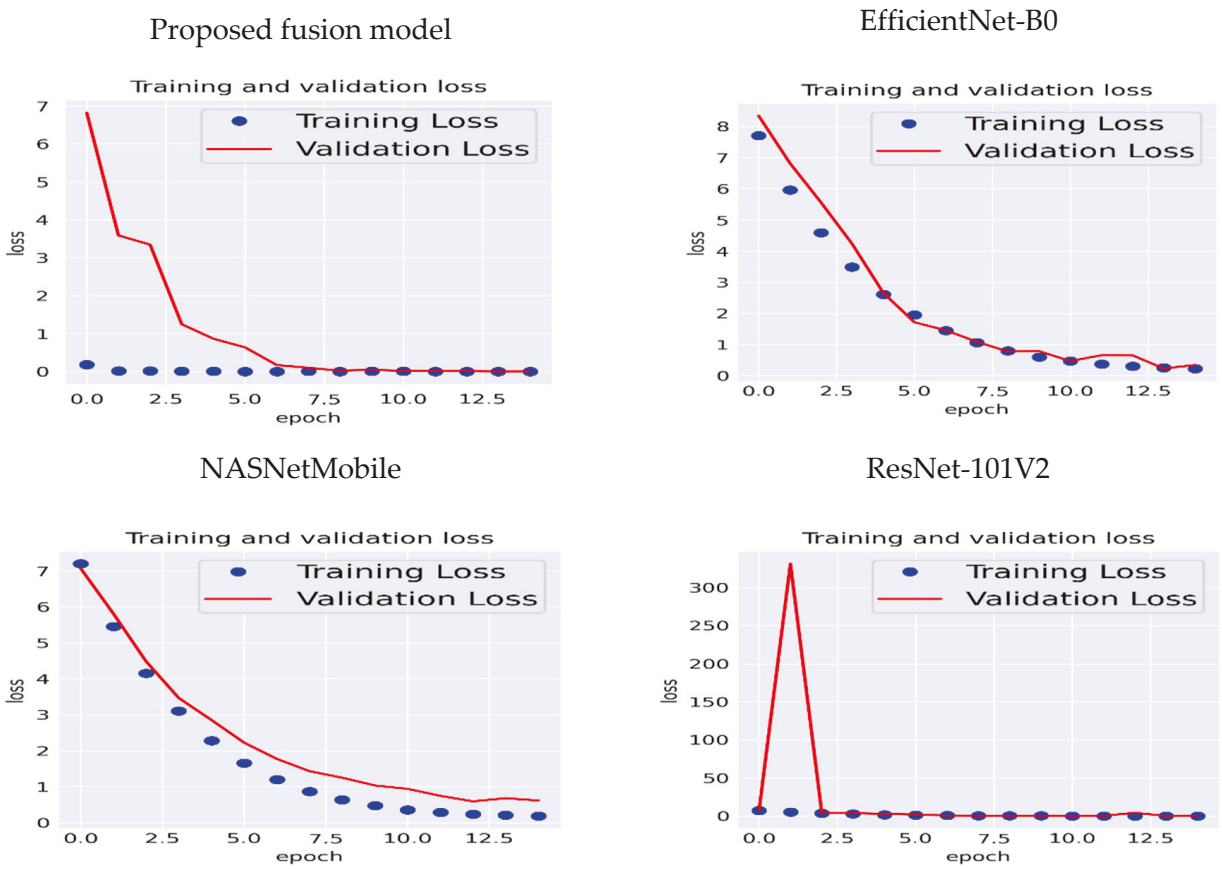


Figure 7. Training and validation loss of the three CNN models and the proposed fusion model.

For EfficientNet-B0 and NASNetMobile, the training and validation loss were very close and decreased as the number of epochs increased. Hence, they were not biased or variances. For the ResNet-101V2, at the beginning of the experiment, the validation loss was very high, but after the second epoch, it decreased and equaled 0%, similar to the training loss. Hence, it was not biased and displayed no variance.

Figure 8 depicts the validation and training accuracy of the proposed model, EfficientNet-B0, NASNetMobile, and ResNet-101V2. The training accuracy for the proposed fusion model was 100% after the 2nd epoch, and the validation accuracy was 100% after the 10th epoch. For the EfficientNet-B0 model, the training accuracy was 100% after the fifth epoch, and the validation accuracy was between 90% and 100% after the fifth epoch. For the NASNetMobile model, the training accuracy was 100% after the fifth epoch, and the maximum validation accuracy was 85%. ResNet-101V2's training and validation accuracy was 100% after the fifth epoch.

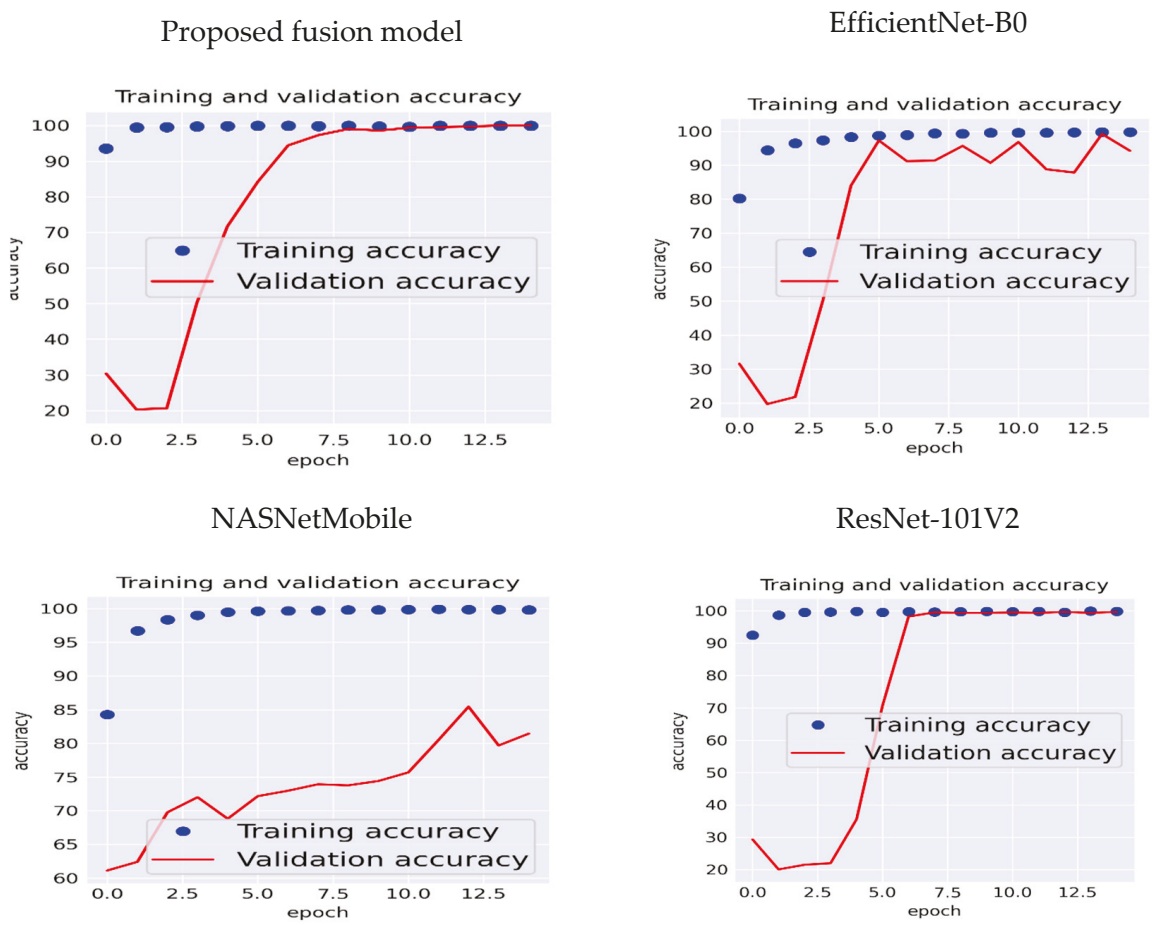


Figure 8. Training and validation accuracy of the three CNN models and the proposed fusion model.

The observed rapid convergence, where the proposed model achieved 100% training accuracy by the second epoch, can be attributed to the ensemble of three highly effective architectures: NASNetMobile, EfficientNet-B0, and ResNet-101V2. Each of these models brings distinct strengths in feature extraction. NASNetMobile is known for its ability to balance accuracy and computational efficiency. EfficientNet-B0 scales efficiently with minimal computational resources while still maintaining accuracy. ResNet-101V2 is particularly effective at preserving feature information over deeper layers, preventing vanishing gradients.

This fusion model accelerated the learning process by efficiently capturing diverse features and reducing the time required for the classifier to optimize. The complementary nature of these models enabled quick convergence while maintaining robust performance across training and validation sets. Furthermore, to ensure robust generalization, we extended the analysis to include precision, recall, F1-score, and confusion matrix evaluations. These metrics, alongside validation accuracy, demonstrated that the model maintained high performance on unseen data, further confirming that the rapid training convergence did not lead to overfitting.

For the EfficientNet-B0 model, the training accuracy was 100% after the fifth epoch, and the validation accuracy was between 90% and 100% after the fifth epoch. For the NASNetMobile model, the training accuracy was 100% after the fifth epoch, and the maximum validation accuracy was 85%. ResNet-101V2's training and validation accuracy was 100% after the fifth epoch.

Figure 9 depicts the confusion matrix for the proposed fusion model, EfficientNet-B0, NASNetMobile, and ResNet-101V2. There were five classes of the test set of the LC25000 dataset: Colon_Adenocarcinoma (121 HIs), Colon_Benign_Tissue (129 HIs), Lung-Benign_

Tissue (118 HIs), Lung_Adenocarcinoma (132 HIs), and Lung_Squamous_Cell_Carcinoma (125 HIs).

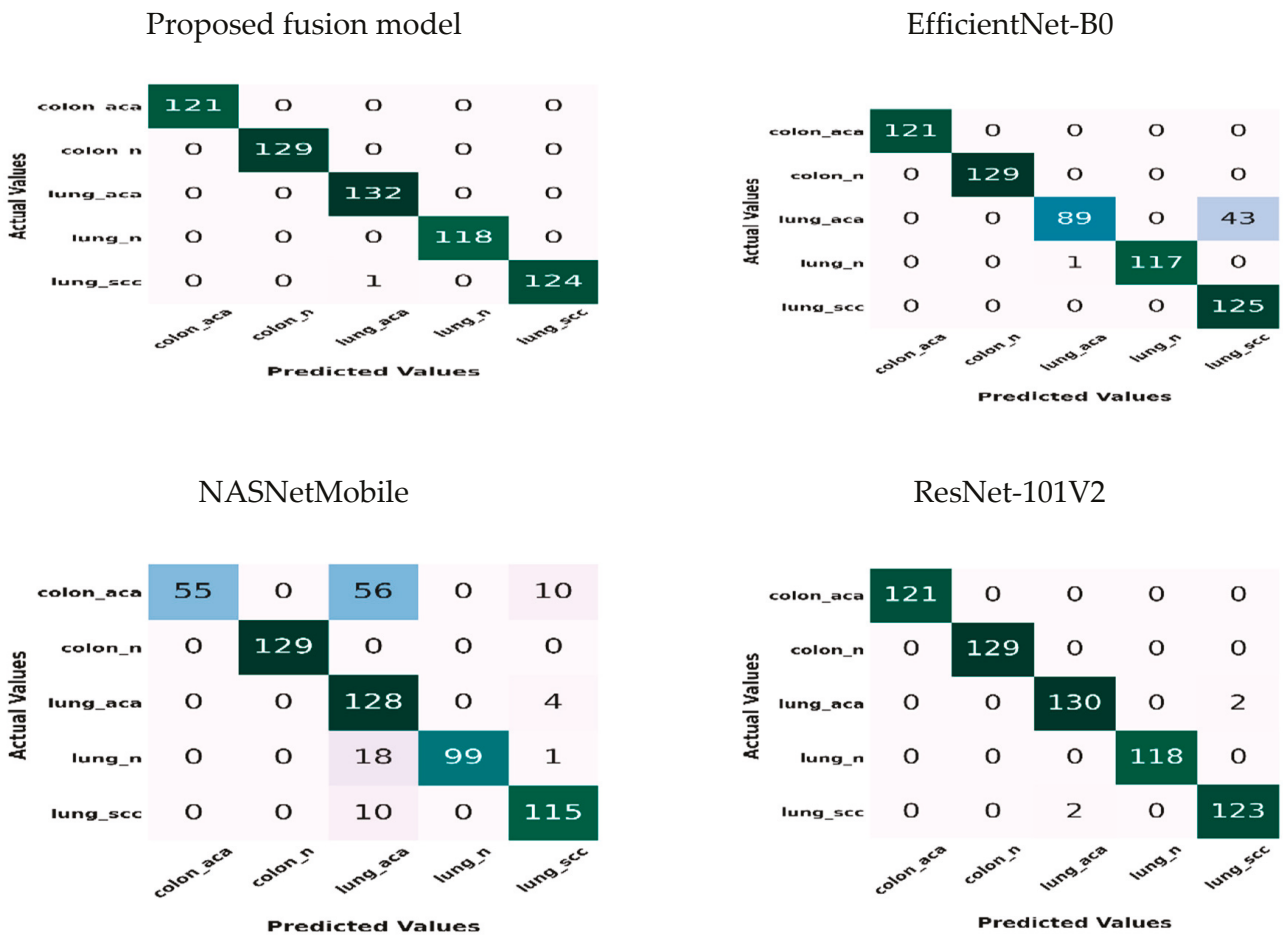


Figure 9. The confusion matrix for the three CNN models and the proposed fusion model on the test set.

For the Colon_Adenocarcinoma class, the proposed fusion model, stacked ensemble model, EfficientNet-B0, and ResNe-t101 classified 121 of 121 HIs and achieved an accuracy of 100%. NASNetMobile classified 55 of 121 HIs and achieved an accuracy of 45.4%. Hence, the proposed fusion model, the stacked ensemble model, EfficientNet-B0, and ResNet-101V2 had the best performance and followed NASNetMobile.

For the Colon_Benign_Tissue class, the accuracy of the proposed fusion model, the stacked ensemble model, EfficientNet-B0, NASNetMobile, and ResNet-101V2 was 100%, as they classified 129 of 129 HIs.

For the Lung_Adenocarcinoma class, the proposed fusion model classified 132 of 132 HIs and recorded an accuracy of 100%. The stacked ensemble and NASNetMobile models classified 128 of 132 HIs and achieved an accuracy of 96.9%. EfficientNet-B0 predicted 89 of 132 HIs and recorded an accuracy of 67.4%. ResNet-101V2 predicted 130 of 132 HIs and achieved an accuracy of 98.4%. Hence, the proposed fusion model performed best, followed by ResNet-101V2, the stacked ensemble, NASNetMobile, and EfficientNet-B0.

For the Lung-Benign_Tissue class, the proposed fusion model and ResNet-101V2 classified 118 of 118 HIs and recorded an accuracy of 100%. The stacked ensemble and EfficientNet-B0 models classified 117 of 118 HIs and achieved an accuracy of 99.1%. NASNetMobile classified 99 of 118 HIs and recorded an accuracy of 83.8%. Hence, the proposed

fusion model and ResNet-101V2 recorded the best performance, followed by the stacked ensemble, the EfficientNet-B0 model, and NASNetMobile.

For the Lung_Squamous_Cell_Carcinoma class, the proposed fusion model classified 124 of 125 HIs and recorded an accuracy of 99.2%. The stacked ensemble and ResNet-101V2 classified 123 of 125 HIs and achieved an accuracy of 98.4%. EfficientNet-B0 classified 125 of 125 HIs and recorded an accuracy of 100%. NASNetMobile classified 115 of 125 HIs and recorded an accuracy of 92%. Hence, EfficientNet-B0 recorded the best performance, followed by the proposed fusion model, the stacked ensemble, ResNet-101V2, and NASNetMobile.

4.4. Model Result Comparison with the Literature

Our proposed model based on feature fusion outperformed the most recent research, displayed in Table 10. The accuracy of the proposed fine-tuned DL model recorded was 99.94%. From the patient's point of view, the proposed fine-tuned model will reduce the cost and time for colon and lung diagnosis, hence providing appropriate treatment. Compared to a stacked ensemble of the same base learners and other recent studies, the assessment of the proposed fusion model uncovered that, with tuning, it is an unmatched discerning outcome.

Table 10. Comparing the proposed fine-tuned technique and recent research.

Reference	Methodology	Accuracy	Dataset
A. H. Chehade1 et al. [1]	XGBoost, SVM, RF, LDA, and MLP	99%	LC25000
M. Ali and R. Ali [20]	CNNs	99.58%	LC25000
Sakr, A.S. et al. [21]	CNN	99.50%	LC25000
M. Masud et al. [22]	CNN	96.33%	LC25000
Naresh Kumar et al. [23]	RF and deep features from DenseNet-121	98.60%	LC25000
M. Shahid et al. [24]	AlexNet	98.4%	LC25000
Md. A. Talukder et al. [25]	VGG16, VGG1, MobileNet, DenseNet169, DenseNet201, Ensemble, and feature extraction	99.05%	LC25000
O. Attallah et al. [26]	MobileNet, SqueezeNet, ShuffleNet, PCA, fast Walsh–Hadamard, and discrete wavelet	99.6%	LC25000
J. D. Akinyemi et al. [27]	EfficientNet-B7	99.72%.	LC25000
The stacked ensemble model	**EfficientNet-B0, NASNetMobile, and ResNet-101V2.**	**99.55%**	**LC25000**
The proposed model	**Fine-tuned model based on feature fusion**	**99.94%**	**LC25000**

5. Conclusions

We proposed a fine-tuned fusion DL model of EfficientNet-B0, NASNetMobile, and ResNet-101V2 for the multi-classification task of colon and lung cancers. The proposed fine-tuned model was based on feature fusion. The proposed fine-tuned model will assist pathologists in detecting lung and colon cancer early with less effort, time, and expense. Each of EfficientNet-B0, NASNetMobile, and ResNet-101V2 needs to be improved regarding varieties in the texture and shape of the input images. However, the proposed fusion model recorded a high accuracy in the multi-classification task of lung and colon cancer since it joined the benefits of the three CNN models to increase the accuracy of the results. We pre-trained the three CNN models on the ImageNet dataset. The three pre-trained models used a GlobalAveragePooling2D method to smooth all layers into a vector. The GlobalAveragePooling2D method computes the average of each feature map across its

whole spatial extent, providing a single value per feature map. We combined the pre-trained individual vectors of the pre-trained EfficientNet-B0, NASNetMobile, and ResNet-101V2 models into a single vector by a concatenate layer. The concatenated single feature vector was fine-tuned on the training set of the LC25000 dataset. We evaluated the proposed model over the LC25000 dataset, which includes colon and lung HIs. The LC25000 dataset was pre-processed by normalization and resizing techniques. The LC25000 dataset was split into a training set (90%), a test set (5%), and a validation set (5%). The proposed fine-tuned model was measured and compared with the stacked ensemble and recent DL classifies. The proposed fusion model recorded 99.84%, 99.84%, 99.84%, 99.94%, 99.96%, 99.84%, 0.16%, and 99.96%, for precision, recall, F1-score, accuracy, specificity, FNR, and NPV, respectively. Therefore, the proposed fine-tuned fusion model delivered cutting-edge performance across all classes. The current limitation of our model is the processing speed, which poses a challenge for our developing system. In future research, we plan to explore the proposed model on a variety of human cancer types and employ hyperoptimization algorithms to automatically enhance the hyperparameterization process.

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

Funding: This work was funded by the Deanship of Graduate Studies and Scientific Research at Jouf University under grant No. (DGSSR-2023-02-02404).

Data Availability Statement: The dataset referenced in this article is the LC25000 dataset, obtained from https://opendatalab.com/OpenDataLab/LC25000 accessed on 21 August 2024. This dataset was issued on 2019 and it is available at no cost to all researchers and scientists for the purpose of conducting experiments.

Acknowledgments: The authors extend their appreciation to the Deanship of Scientific Research at Jouf University for funding this work through research grant No. (DGSSR-2023-02-02404).

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

References

1. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics. *CA Cancer J. Clin.* **2020**, *70*, 7–30. [CrossRef] [PubMed]
2. WHO. Cancer. Available online: https://www.who.int/news-room/fact-sheets/detail/cancer (accessed on 10 August 2020).
3. Chehade, A.H.; Abdallah, N.; Marion, J.-M.; Oueidat, M.; Chauvet, P. Lung and colon cancer classification using medical imaging: A feature engineering approach. *Phys. Eng. Sci. Med.* **2022**, *45*, 729–746. [CrossRef]
4. Togaçar, M. Disease type detection in lung and colon cancer images using the complement approach of inefficient sets. *Comput. Biol. Med.* **2021**, *137*, 104827. [CrossRef]
5. Bermúdez, A.; Arranz-Salas, I.; Mercado, S.; López-Villodres, J.A.; González, V.; Ríus, F.; Ortega, M.V.; Alba, C.; Hierro, I.; Bermúdez, D. Her2-positive and microsatellite instability status in gastric cancer—Clinicopathological implications. *Diagnostics* **2021**, *11*, 944. [CrossRef] [PubMed]
6. Zhu, D.; Ding, R.; Ma, Y.; Chen, Z.; Shi, X.; He, P. Comorbidity in lung cancer patients and its association with hospital readmission and fatality in China. *BMC Cancer* **2021**, *21*, 557. [CrossRef]
7. Schüz, J.; Espina, C. The eleventh hour to enforce rigorous primary cancer prevention. *Mol. Oncol.* **2021**, *15*, 741. [CrossRef] [PubMed]
8. Raman, S. Can curcumin along with chemotherapeutic drugs and lipid provide an effective treatment of metastatic colon cancer and alter multidrug resistance? *Med. Hypotheses* **2020**, *132*, 109325.
9. Bergers, G.; Fendt, S.M. The metabolism of cancer cells during metastasis. *Nat. Rev. Cancer* **2021**, *21*, 162–180. [CrossRef]
10. Koo, M.M.; Swann, R.; McPhail, S.; Abel, G.A.; Elliss-Brookes, L.; Rubin, G.P.; Lyratzopoulos, G. Presenting symptoms of cancer and stage at diagnosis: Evidence from a cross-sectional, population-based study. *Lancet Oncol.* **2020**, *21*, 73–79. [CrossRef]

11. Zhou, W.; Liu, G.; Hung, R.J.; Haycock, P.C.; Aldrich, M.C.; Andrew, A.S.; Arnold, S.M.; Bickeböller, H.; Bojesen, S.E.; Brennan, P.; et al. Causal relationships between body mass index, smoking, and lung cancer: Univariable and multivariable Mendelian randomization. *Int. J. Cancer* **2021**, *148*, 1077–1086. [CrossRef]
12. Khan, T.; Relitti, N.; Brindisi, M.; Magnano, S.; Zisterer, D.; Gemma, S.; Butini, S.; Campiani, G. Autophagy modulators for the treatment of oral and esophageal squamous cell carcinomas. *Med. Res. Rev.* **2020**, *40*, 1002–1060. [CrossRef] [PubMed]
13. Liu, H.; Xu, X.; Wu, R.; Bi, L.; Zhang, C.; Chen, H.; Yang, Y. Antioral squamous cell carcinoma effects of carvacrol via inhibiting inflammation, proliferation, and migration related to Nrf2/Keap1 pathway. *BioMed Res. Int.* **2021**, *2021*, 1–14. [CrossRef] [PubMed]
14. Lannagan, T.R.; Jackstadt, R.; Leedham, S.J.; Sansom, O.J. Advances in colon cancer research: In vitro and animal models. *Curr. Opin. Genet. Dev.* **2021**, *66*, 50–56. [CrossRef]
15. Engstrom, P.F.; Arnoletti, J.P.; Benson, A.B.; Chen, Y.J.; Choti, M.A.; Cooper, H.S.; Covey, A.; Dilawari, R.A.; Early, D.S.; Enzinger, P.C.; et al. Colon cancer. *J. Natl. Compr. Cancer Netw.* **2009**, *7*, 778–831. [CrossRef]
16. Fadel, M.G.; Malietzis, G.; Constantinides, V.; Pellino, G.; Tekkis, P.; Kontovounisios, C. Clinicopathological factors and survival outcomes of signet-ring cell and mucinous carcinoma versus adenocarcinoma of the colon and rectum: A systematic review and meta-analysis. *Discov. Oncol.* **2021**, *12*, 5. [CrossRef] [PubMed]
17. Mohamed, A.A.A.; Hançerlioğullari, A.; Rahebi, J.; Rezaeizadeh, R.; Lopez-Guede, J.M. Colon Cancer Disease Diagnosis Based on Convolutional Neural Network and Fishier Mantis Optimizer. *Diagnostics* **2024**, *14*, 1417. [CrossRef]
18. Das, S.; Biswas, S.; Paul, A.; Dey, A. AI Doctor: An intelligent approach for medical diagnosis. In *Industry Interactive Innovations in Science, Engineering and Technology*; Springer: Singapore, 2018.
19. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef] [PubMed]
20. Ali, M.; Ali, R. Multi-input dual-stream capsule network for improved lung and colon cancer classification. *Diagnostics* **2021**, *11*, 1485. [CrossRef]
21. Sakr, A.S.; Soliman, N.F.; Al-Gaashani, M.S.; Pławiak, P.; Ateya, A.A.; Hammad, M. An efficient deep learning approach for colon cancer detection. *Appl. Sci.* **2022**, *12*, 8450. [CrossRef]
22. Masud, M.; Sikder, N.; Nahid, A.-A.; Bairagi, A.K.; AlZain, M.A. Machine learning approach to diagnosing lung and colon cancer using a dl-based classification framework. *Sensors* **2021**, *21*, 748. [CrossRef]
23. Kumar, N.; Sharma, M.; Singh, V.P.; Madan, C.; Mehandia, S. An empirical study of handcrafted and dense feature extraction techniques for lung and colon cancer classification from histopathological images. *Biomed. Signal Process. Control* **2022**, *75*, 103596. [CrossRef]
24. Mehmood, S.; Ghazal, T.M.; Khan, M.A.; Zubair, M.; Naseem, M.T.; Faiz, T.; Ahmad, M. Malignancy detection in lung and colon histopathology images using transfer learning with class selective image processing. *IEEE Access* **2022**, *10*, 25657–25668. [CrossRef]
25. Talukder, A.; Islam, M.; Uddin, A.; Akhter, A.; Hasan, K.F.; Moni, M.A. Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning. *Expert Syst. Appl.* **2022**, *204*, 117695. [CrossRef]
26. Attallah, O.; Aslan, M.F.; Sabanci, K. A Framework for Lung and Colon Cancer Diagnosis via Lightweight Deep Learning Models and Transformation Methods. *Diagnostics* **2022**, *12*, 2926. [CrossRef]
27. Akinyemi, J.D.; Akinola, V.; Adekunle, V.; Adetiloye, V.; Dansu, E.J. Lung and colon cancer detection from CT images using Deep Learning. *Mach. Graph. Vis.* **2023**, *32*, 85–97. [CrossRef]
28. Borkowski, A.A.; Bui, M.M.; Thomas, L.B.; Wilson, C.P.; DeLand, L.A.; Mastorides, S.M. Lung and colon cancer histopathological image dataset (LC25000). *arXiv* **2019**, arXiv:1912.12142.
29. Singh, O.; Singh, K.K. An approach to classify lung and colon cancer of histopathology images using deep feature extraction and an ensemble method. *Int. J. Inf. Tecnol.* **2023**, *15*, 4149–4160. [CrossRef]
30. Bilal, M.; Maqsood, M.; Yasmin, S.; Ul Hasan, N.; Rho, S. A transfer learning-based efficient spatiotemporal human action recognition framework for long and overlapping action classes. *J. Supercomput.* **2022**, *78*, 2873–2908. [CrossRef]
31. Zoph, B.; Vasudevan, V.; Shlens, J.; Le, Q.V. Learning Transferable Architectures for Scalable Image Recognition. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8697–8710.
32. Radhika, K.; Devika, K.; Aswathi, T.; Sreevidya, P.; Sowmya, V.; Soman, K.P. Performance Analysis of NASNet on Unconstrained Ear Recognition. In *Nature Inspired Computing for Data Science*; Rout, M., Rout, J., Das, H., Eds.; Studies in Computational Intelligence; Springer: Cham, Switzerland, 2020; Volume 871.
33. Tan, M.; Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
34. Alhichri, H.; Alswayed, A.S.; Bazi, Y.; Ammour, N.; Alajlan, N.A. Classification of remote sensing images using efficientnet-b3 cnn model with attention. *IEEE Access* **2021**, *9*, 14078–14094. [CrossRef]
35. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.

36. Putra, T.A.; Rufaida, S.I.; Leu, J. Enhanced skin condition prediction through machine learning using dynamic training and testing augmentation. *IEEE Access* **2020**, *4*, 40536–40546. [CrossRef]
37. Mohammed, A.; Kora, R. An effective ensemble deep learning framework for text classification. *J. King Saud Univ.-Comput. Inf. Sci.* **2021**, *34*, 8825–8837. [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.

diagnostics

MDPI

Article

FracFusionNet: A Multi-Level Feature Fusion Convolutional Network for Bone Fracture Detection in Radiographic Images

Sameh Abd El-Ghany *, Mahmood A. Mahmood and A. A. Abd El-Aziz

Department of Information Systems, College of Computer and Information Sciences, Jouf University, Sakakah 72388, Saudi Arabia; mamahmood@ju.edu.sa (M.A.M.); aaeldamarany@ju.edu.sa (A.A.A.E.-A.)
* Correspondence: saabdelwahab@ju.edu.sa

Abstract: Background/Objectives: Bones are essential components of the human body, providing structural support, enabling mobility, storing minerals, and protecting internal organs. Bone fractures (BFs) are common injuries that result from excessive physical force and can lead to serious complications, including bleeding, infection, impaired oxygenation, and long-term disability. Early and accurate identification of fractures through radiographic imaging is critical for effective treatment and improved patient outcomes. However, manual evaluation of X-rays is often time-consuming and prone to diagnostic errors due to human limitations. To address this, artificial intelligence (AI), particularly deep learning (DL), has emerged as a powerful tool for enhancing diagnostic precision in medical imaging. **Methods**: This research introduces a novel convolutional neural network (CNN) model, the Multi-Level Feature Fusion Network (MLFNet), designed to capture and integrate both low-level and high-level image features. The model was evaluated using the Bone Fracture Multi-Region X-ray (BFMRX) dataset. Preprocessing steps included image normalization, resizing, and contrast enhancement to ensure stable convergence, reduce sensitivity to lighting variations in radiographic images, and maintain consistency. Ablation studies were conducted to assess architectural variations, confirming the model's robustness and generalizability across data distributions. MLFNet's high accuracy, interpretability, and efficiency make it a promising solution for clinical deployment. **Results**: MLFNet achieved an impressive accuracy of 99.60% as a standalone model and 98.81% when integrated into hybrid ensemble architectures with five leading pre-trained DL models. **Conclusions**: The proposed approach supports timely and precise fracture detection, optimizing the diagnostic process and reducing healthcare costs. This approach offers significant potential to aid clinicians in fields such as orthopedics and radiology, contributing to more equitable and effective patient care.

Keywords: bone fracture; multi-level feature fusion network; deep learning; convolutional neural network; bone fracture multi-region X-ray dataset

1. Introduction

Bones are often seen as unchanging structures that provide support to the body. However, bone remodeling is a continuous process that occurs throughout a person's life, largely driven by physiological needs [1]. Newborns typically have around 270 bones, which gradually fuse together, resulting in approximately 206 bones by adulthood. This skeletal structure includes the bones of the skull, vertebrae, rib cage, and the bones of the upper and lower limbs. The human skeletal system consists of 206 bones of various shapes and sizes, serving crucial functions such as providing structural support, protecting vital

organs, and enabling movement. Among these bones, the femur is the largest, while the smallest is the stapes, located in the middle ear.

Bone diseases and disorders include a wide variety of medical conditions that affect the strength and health of the skeletal system. These issues can range from common problems like fractures and osteoporosis to rarer and more complicated diseases such as osteosarcoma and Paget's disease [2]. BFs are common injuries that often result from falls, collisions, and other traumatic events [1,3]. The incidence of BFs worldwide is steadily increasing, influenced by aging populations and more urban lifestyles [3]. BF identification through radiographic imaging is a frequent procedure for patients experiencing either high- or low-energy trauma across multiple clinical environments, such as emergency rooms; urgent care facilities; and outpatient clinics like orthopedics, rheumatology, and family medicine. Overlooked fractures in radiographs are a leading factor in diagnostic inconsistencies between initial assessments made by nonradiologists or radiology residents and the final evaluations by board-certified radiologists, resulting in avoidable harm or delays in patient care [4,5]. Errors in interpreting fractures can account for as much as 24% of detrimental diagnostic mistakes observed in the emergency department [4]. Additionally, discrepancies in radiographic assessments of fractures tend to occur more frequently during evening and overnight shifts (from 5 PM to 3 AM), likely due to fatigue and the involvement of non-expert readers. In cases involving multiple injuries, the rate of overlooked injuries, such as fractures, can be significant, particularly in the forearm and hands (6.6%) and feet (6.5%) [6,7].

BFs are categorized based on the condition of the skin and soft tissue around the fracture into three types: open, closed, and displaced fractures, shown in Figure 1 [8]:

- Open Fracture: An open fracture refers to a broken bone that has broken through the skin's surface. In severe cases, the bone may be visible externally.
- Closed Fracture: In contrast to open fractures, closed fractures do not break through the skin, resulting in a reduced risk of infection.
- Displaced Fracture: A displaced fracture occurs when there is a separation between the two ends of the broken bone, indicating that the edges are misaligned.

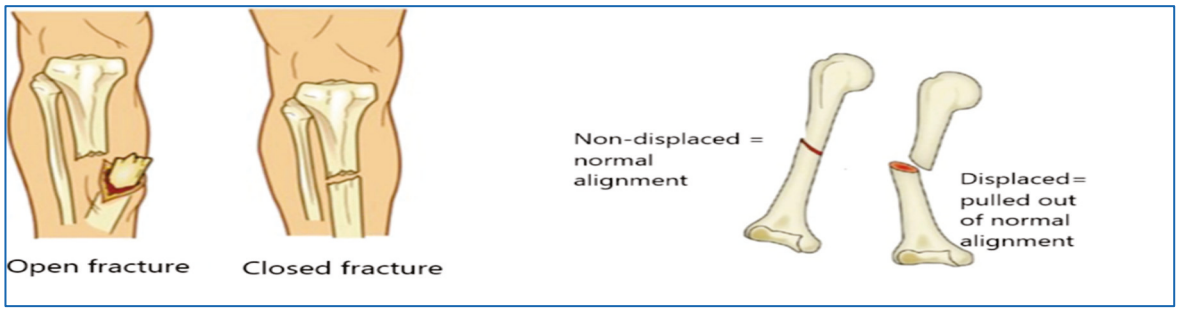


Figure 1. Types of bone fractures [8].

BFs are categorized by the extent of the break, distinguishing between partial and complete fractures [9]. In addition to the categories mentioned earlier, there are further subcategories based on the specific characteristics of BFs [10]. Figure 2 illustrates these subcategories of BFs, which include the following:

- Transverse fracture: The bone breaks across horizontally.
- Oblique fracture: A diagonal fracture that splits the bone.
- Spiral fracture: A fracture that twists around the bone.
- Comminuted fracture: When a bone shatters into multiple pieces, typically more than three.

- Segmental fracture: Involves two separate breaks in a bone, categorized as comminuted fractures.
- Linear fracture: The fracture aligns straight along the bone's length.
- Greenstick fracture: A partial fracture where the bone bends but does not break completely.

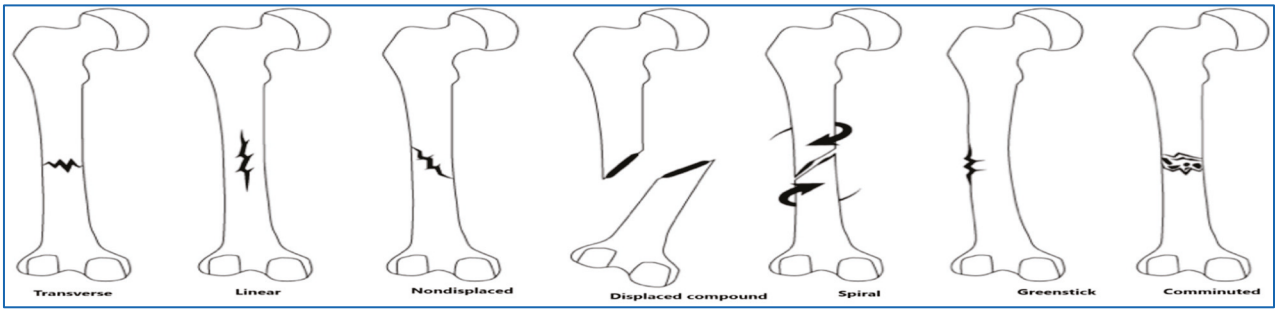


Figure 2. Categories of BFs [10].

Tibial fractures are the most frequently occurring type of bone fracture, especially among children, athletes, and older adults. They pose considerable diagnostic difficulties. Therefore, obtaining a swift and precise diagnosis is essential to improve the healing process's effectiveness [9].

Fractures can arise from conditions that weaken bone integrity, like osteoporosis or bone cancer [11]. This is especially true for children, the elderly, and those who are physically active [3]. In addition, BFs occur when there is a disruption in the structure of the bone. These injuries are frequently caused by factors such as stress, accidents, osteogenesis imperfecta, osteoporosis, and cancer.

BFs present with distinct symptoms, though the type, location, and severity of the fracture can influence how they appear. Some fractures may be obvious (like an open fracture), while others (like stress fractures) may be subtle. The symptoms are pain, swelling, bruising, deformity, inability to bear weight or move, tenderness, and crepitus [12].

The success of any suggested treatment often hinges on a timely and accurate diagnosis. Therefore, obtaining a quick and precise diagnosis is crucial for improving the healing process's effectiveness [3,13]. In real-world settings, radiologists and healthcare providers utilize X-rays, CT scans, and MRI images to determine the specific kind of fracture and to confirm if any fracture exists [3,9,13]. The Digital Imaging and Communications in Medicine (DICOM) standard facilitates the sharing of medical images [14].

After a fracture, it is crucial to seek treatment as soon as possible. An orthopedic surgeon will review an X-ray or CT scan to identify the fractured bone. Once the diagnosis is made, appropriate treatment is administered. The success of this treatment largely relies on the orthopedic surgeon's experience and expertise. However, this manual process can be time-consuming, and access to specialists may be limited in remote areas [15].

X-ray imaging is a quick and non-invasive method primarily used to diagnose and confirm fractures, where the specific area of the body is subjected to X-ray radiation. The integration of radiation with advanced computer image processing has led to the widespread adoption of digital X-ray technology across various medical fields [16]. Although there are some limitations regarding image quality, X-rays are adequate for identifying fractures. By analyzing X-ray images and implementing suitable treatments, healthcare professionals can evaluate the presence and severity of injuries. X-ray images provide clear and detailed views of bone structures, aiding in the detection of complications related to fractures, such as misaligned bone fragments, involvement of joints, or damage to nearby tissues. However, it is important to carefully weigh the potential risks of radiation exposure, particularly when multiple X-rays are needed [11]. Digital X-ray imaging devices

have become prevalent in healthcare due to their portability and advanced computerized image processing features.

Manually reviewing radiological images for fractures requires substantial time and effort. A fatigued radiologist might miss a fracture in an otherwise normal image. To address this issue, a computer-aided design (CAD) system can be employed to help identify concerning cases in radiological images and notify physicians [3].

AI has rapidly emerged as a powerful tool in enhancing the interpretation of medical images and increasing diagnostic accuracy. Precise detection of bone fractures is essential for determining appropriate treatment strategies and forecasting patient outcomes. Deep learning (DL) methods have been extensively utilized for feature extraction and fracture classification tasks [17]. Among these, convolutional neural networks (CNNs)—a key DL architecture—have attracted considerable interest for their ability to assist radiologists in identifying bone fractures. In many cases, DL models have demonstrated superior accuracy compared to human experts in analyzing medical images. This technology is applicable across multiple imaging modalities, including X-rays, CT scans, and MRIs, and supports the identification of various bone abnormalities, particularly fractures. Nevertheless, DL applications in this field still face challenges, especially in managing and processing the vast amount of data associated with bone imaging [18,19]. By delivering accurate image analysis, DL algorithms can facilitate timely and effective clinical decision-making. Their reliable detection capabilities not only enhance patient outcomes but also streamline the diagnostic workflow for healthcare professionals [20].

This research introduces a novel CNN model called MLFNet, which is designed to capture and combine both low-level and high-level image features. The model was tested on the BFMRX dataset for binary classification (fracture vs. non-fracture). The preprocessing steps involved normalizing images, adjusting their size, and enhancing contrast. These measures were taken to promote stable convergence, minimize sensitivity to changes in lighting in radiographic images, and ensure consistency. Ablation studies were performed to evaluate different architectural variations, confirming the model's robustness and generalizability across various data distributions. MLFNet achieved an outstanding accuracy of 99.60% as a standalone model and 98.81% when integrated into hybrid ensemble architectures with five leading pre-trained DL models (DenseNet169, EfficientNet-B3, InceptionV3, MobileNet-V2, and ResNet-101). With its high accuracy, interpretability, and efficiency, MLFNet is a promising option for clinical use, facilitating timely and accurate fracture detection, optimizing the diagnostic process, and lowering healthcare costs. This approach has significant potential to assist clinicians in areas such as orthopedics and radiology, contributing to more equitable and effective patient care. Below is a summary of our research contributions.

- We developed the MLFNet model, which extracts and combines features at multiple scales. This approach enhances the representation of bone structures and subtle fracture patterns across various image resolutions.
- We utilized multi-scale fusion techniques to capture both detailed local features and broader contextual information. This improvement allows the model to detect fractures of different shapes, sizes, and locations more effectively.
- The proposed model highlights potential fracture regions using attention-like responses from intermediate feature maps, aiding in weakly supervised localization.
- We evaluated the MLFNet using various radiographic datasets, demonstrating its superior ability to generalize and resist challenges found in real-world images, such as noise, contrast variations, and complex anatomical structures.

- We conducted detailed ablation studies comparing single-scale and multi-scale architectures. The results confirmed the effectiveness of integrating features at different resolution levels, leading to higher diagnostic accuracy.
- When benchmarked against five DL architectures (DenseNet169, EfficientNet-B3, InceptionV3, MobileNet-V2, and ResNet-101), the MLFNet approach excelled in binary classification, achieving an accuracy rate of 98.81% on the BFMRX dataset when combined with other DL models, and achieving 99.60% as an independent model on the same dataset.

The remainder of this paper is structured as follows: Section 2 provides a review of related work on BF diagnostic systems. Section 3 details the BFMRX dataset and the methodology employed in this research. Section 4 discusses two experimental outcomes of the proposed framework. Lastly, Section 5 offers concluding remarks and summarizes the key findings.

2. Literature Review

The detection of BFs remains a key focus in medical image analysis research, with multiple studies addressing the problem using diverse methodologies. For example, Yadav, D.P. et al. [15] introduced an innovative multi-scale feature fusion approach that combined a CNN with an enhanced Canny edge detection algorithm to differentiate between fractured and healthy bone images. The hybrid Scale Fracture Network (SFNet) was a unique two-scale sequential DL model designed for efficient BF diagnosis, requiring less computational time than other leading deep CNN models. A key aspect of Yadav, D.P. et al. [15] was the application of an improved Canny edge algorithm, which helps identify edges in images to pinpoint fracture regions. Subsequently, grayscale images along with their corresponding Canny edge images are utilized for training and evaluation within the proposed hybrid SFNet. Additionally, the authors compared the performance of SFNet with Canny (SFNet + Canny) against other advanced deep CNN models using a bone image dataset. Our findings demonstrated that SFNet with the Canny algorithm achieved the highest accuracy, F1-score, and recall rates of 99.12%, 99%, and 100%, respectively, in diagnosing bone fractures.

Aldhyani, T. et al. [9] created an automated method for detecting fractures using the BFMRX dataset from Kaggle, which includes a detailed collection of 10,580 radiographic images. This research promoted the application of DL techniques such as VGG-16, ResNet152-V2, and DenseNet-201 for diagnosing and identifying bone fractures. The results of the experiments showed that the suggested method effectively identified and classified different types of fractures. The proposed system, which utilized DenseNet-201 and VGG-16, achieved a validation accuracy rate of 97%.

Pillai, R. et al. [21] explored the application of the VGG-16 transfer learning model for the automatic diagnosis of BFs using X-ray images. The primary aim was to enhance both the accuracy and efficiency of fracture detection. They divided the BFMRX dataset, which contains 10,578 annotated images labeled as either Fractured or Not Fractured, into three segments: training (9243 images), validation (829 images), and testing (506 images). The VGG16 model was fine-tuned on the training set, employing data augmentation to improve its performance under varying imaging conditions. The evaluation of the model's performance involved tracking accuracy and loss metrics during both the training and validation phases. The ultimate evaluation on the testing set yielded an accuracy of 98.22%.

Shandilya, G. et al. [22] utilized a multi-region X-ray dataset of BFs to present a comprehensive method for detecting and classifying fractures through a novel CNN model. The BFMRX dataset comprises 10,580 X-ray images showcasing both fractured and non-fractured bones from various anatomical regions, including the lower limb, upper limb,

lumbar spine, hips, and knees. The proposed CNN model achieved an overall accuracy of 97.85%, along with precision and recall rates of 98%.

Alsufyani, A. [23] evaluated five different models: one custom-designed CNN and four pre-trained models—AlexNet, DenseNet-121, ResNet-152, and EfficientNet-B3. The models were trained using the BFMRX dataset of 10,581 X-ray images, classified as either fractured or non-fractured. Their performance was assessed through metrics such as accuracy, precision, recall, and F1-score. Among the models tested, EfficientNet-B3 delivered the most impressive results, achieving an accuracy of 99.20% and perfect recall, indicating its significant potential for application in clinical settings. ResNet-152 and the custom CNN also showed strong performance, albeit with slightly lower accuracy.

Alex, R. and Rosmasari [24] presented an automated DL method for classifying bone fractures into two categories using a pre-trained ResNet18 architecture. The model was trained and validated on the BFMRX dataset comprising 10,580 images, which were divided into fractured and non-fractured categories. To enhance the model's generalization capabilities, data augmentation techniques such as rotation and horizontal flipping were employed during the preprocessing stage. The resulting model achieved a validation accuracy of 97.59%, showcasing high rates of true positives (TPs) and true negatives (TNs), as evidenced by confusion matrix analysis.

Shandilya, G. et al. [25] employed an enhanced AlexNet model to introduce a thorough approach for diagnosing fractures across multiple regions. The BFMRX dataset utilized comprises X-ray images of both fractured and non-fractured bones from various anatomical regions, including the knees, hips, lumbar region, lower limbs, and upper limbs. To distinguish between fractured and non-fractured X-ray images, the AlexNet CNN, optimized for binary classification, was implemented. The improved AlexNet model achieved a remarkable classification accuracy of 97.12%.

Kumar, A. et al. [26] utilized the BFMRX dataset consisting of 10,581 radiographic images that encompass a wide range of anatomical areas, including the lower limb, upper limb, lumbar spine, hips, and knees. The dataset was organized into 4606 images identified as fractured and 4640 images marked as not fractured for training purposes, along with 337 fractured and 492 non-fractured images for validation, and 238 fractured and 268 non-fractured images for testing. They implemented the AlexNet DL framework, utilizing sophisticated data augmentation and preprocessing methods to improve model efficacy. The model attained an accuracy rate of 96%.

Chauhan, S. [27] categorized the BFMRX dataset of 10,580 radiographic X-ray images utilizing a CNN−AlexNet model, covering various anatomical regions such as the lower limb, upper limb, lumbar region, hips, and knees. The dataset was meticulously organized into training (9246 images), validation (828 images), and test (506 images) subsets, each containing both fractured and non-fractured images. The CNN−AlexNet model was trained on this extensive dataset to accurately differentiate between fractured and non-fractured bone X-rays. The model achieved an accuracy of 96% on the test set.

Uddin, A. et al. [28] utilized the BFMRX dataset that comprised 4148 labeled images along with 10,581 X-ray images categorized as either 'fractured' or 'not fractured' for the purpose of BF classification. They focused on employing DL models such as CNN, ConvNeXt, ViT, MobileViT, VGG-16, and VGG-19 to effectively identify BFs and assess their severity. Additionally, YOLOV8 was employed for fracture diagnosis, and image processing techniques enhanced the quality of noisy X-ray images, achieving a mean Average Precision (maPA50) of 0.995 and 0.991 for maPA50-95. In terms of classification performance, MobileViT attained an accuracy of 99%.

The limitations of the previously mentioned research are as follows:

1. The previously cited studies primarily utilized conventional DL models, with the exception of Yadav, D.P. et al. [15]. In our approach, we employed a custom DCNN to extract and integrate features at multiple scales. This technique improves the representation of bone structures and subtle fracture patterns across different image resolutions.
2. While Yadav, D.P. et al. [15] reported an accuracy rate of 99.21% in one experiment, we conducted two experiments that achieved an accuracy rate of 98.81% on the BFMRX dataset when integrated with other DL models, and reached 99.60% as an independent model on the same dataset.

3. Research Methods and Materials

3.1. BFMRX Dataset

The BFMRX dataset used in our study was sourced directly from Kaggle [29] and was utilized as provided by the dataset authors, without any alterations to the original training, validation, or test splits. According to the dataset description on Kaggle, the splits were prepared by the dataset creators, ensuring that images from the same patient do not appear in multiple subsets. The BFMRX dataset from Kaggle [29] contains X-ray images of both fractured and non-fractured regions of the body, encompassing areas such as the lower limb, upper limb, lumbar spine, hips, and knees. The sizes of the X-ray images differed, requiring preprocessing steps to normalize the data before it could be fed into the models. This dataset includes a total of 10,580 radiographic images (X-ray data). The BFMRX dataset is organized into an 87% training set (9246), a 5% testing set (506), and an 8% validation set (828). Table 1 and Figures 3 and 4 display the allocation of images categorized into fractured and non-fractured groups. Figure 5 shows samples from the BFMRX dataset.

Table 1. X-ray image distribution.

Class	Training	Validation	Testing	Total
fractured	4606	336	238	5180
non-fractured	4640	492	268	5400
Total	9246	828	506	10,580

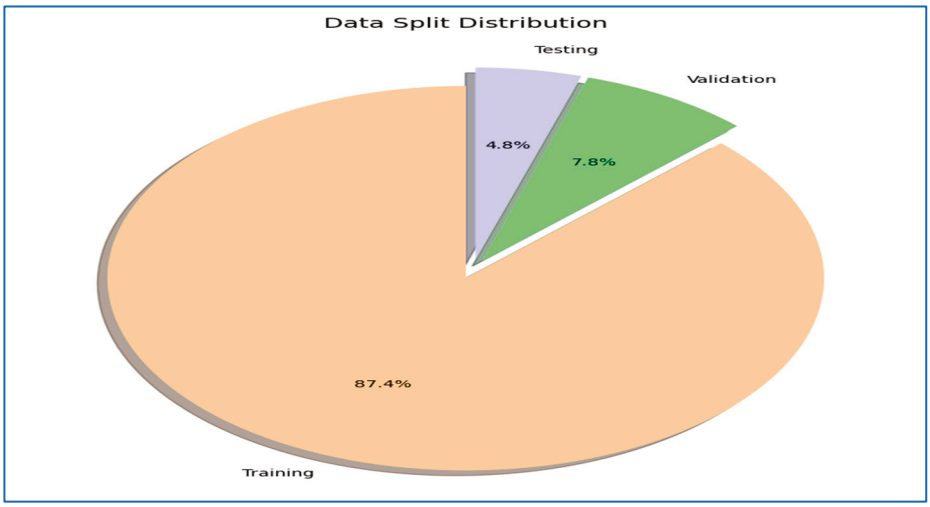


Figure 3. The BFMRX dataset split distribution.

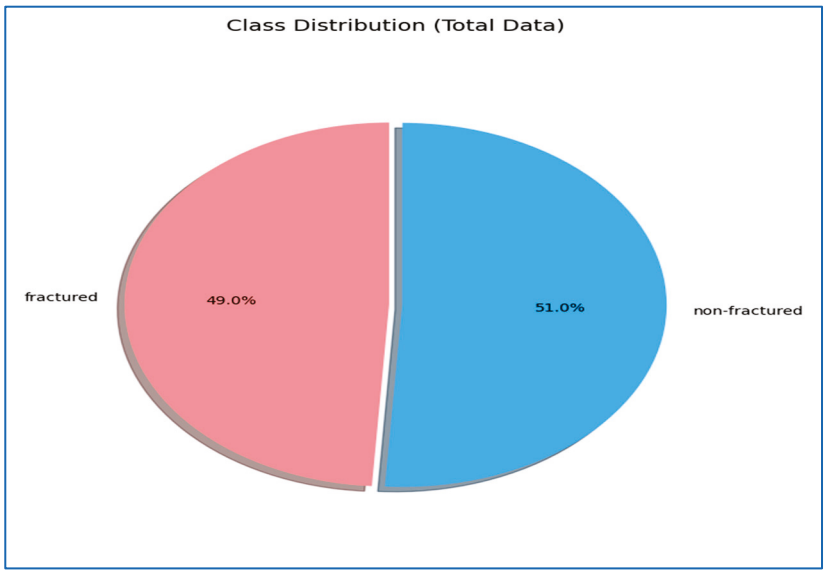


Figure 4. The class distribution of the BFMRX dataset.

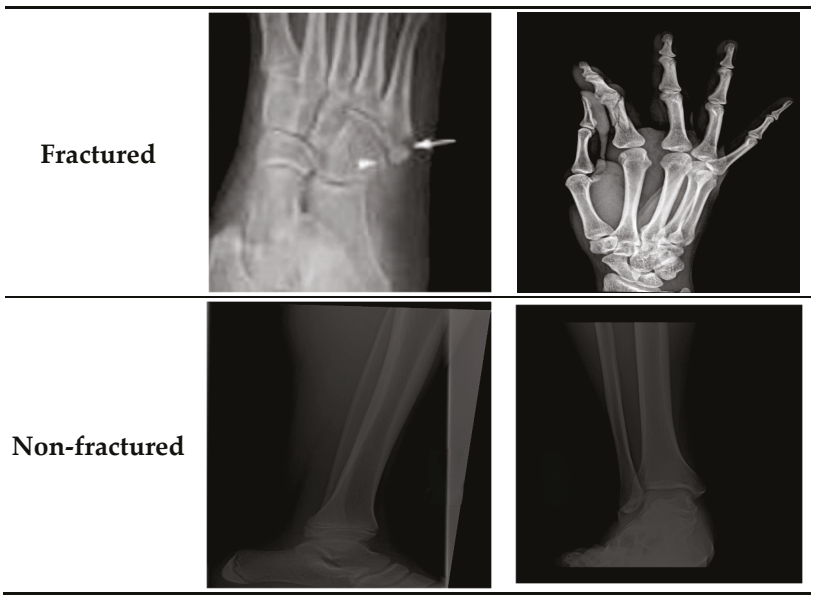


Figure 5. Sample images of fracture and non-fracture from the BFMRX dataset.

3.2. Methodology

To diagnose BFs from radiographic images, we introduced an innovative CNN model named MLFNet. This model was designed to effectively capture and integrate both low-level and high-level features from images. MLFNet utilizes parallel convolutional branches to extract a diverse range of semantic representations, and it implements the Grad-CAM technique to improve interpretability by emphasizing the areas of fractures. The model underwent evaluation on the BFMRX dataset for binary classification. Preprocessing included normalizing and resizing images to maintain consistency. The architecture of the proposed MLFNet model is depicted in Figure 6. The fine-tuning process for the six DL models is described in Algorithm 1. The steps involved in the proposed MLFNet model are summarized as follows:

Preprocessing Stage

- Input: $128 \times 128 \times 3$ X-ray image.
- Normalization/Enhancement: Standardizes pixel values and improves image quality.

Feature Extraction/Fusion Stage

- **Block 1 (SFNetBlock with 64 filters):**
  - Two parallel SFLayers (each: Conv2D → BatchNorm → MaxPooling)
  - Concatenates outputs to capture low-level edge features.
- Batch Normalization: Stabilizes activations after Block 1
- **Block 2 (SFNetBlock with 128 filters):**
  - Dual SFLayers with concatenation for mid-level semantic features
- Batch Normalization: Normalizes Block 2 outputs
- Dropout (p = 0.3): Randomly deactivates 30% neurons to prevent overfitting
- **Block 3 (SFNetBlock with 256 filters):**
  - Dual SFLayers with concatenation for abstract high-level features
- Batch Normalization: Normalizes Block 3 outputs
- Dropout (p = 0.3): Randomly deactivates 30% neurons to prevent overfitting.

Classification Stage

- Flatten: Converts 3D feature maps to 1D vector
- Dense Layers:
- Fully connected layer (256 units) → Dropout → Layer (128 units) → Dropout
  - Learns high-level representations for decision-making
- Output Layer:
  - Sigmoid-activated dense layer (1 unit) for binary classification (fracture vs. no fracture)

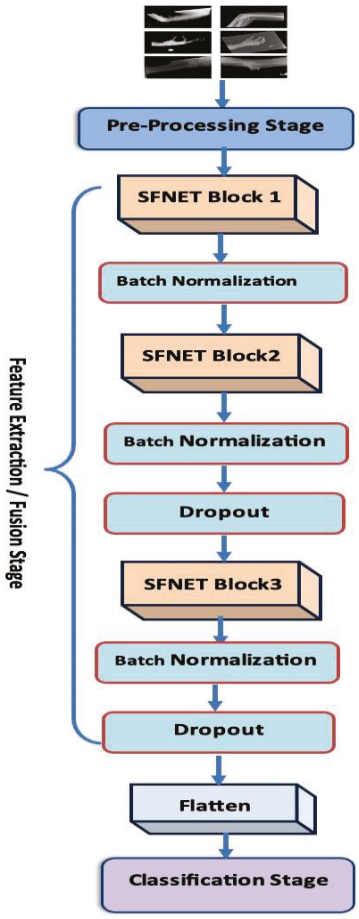


Figure 6. The proposed workflow.

Algorithm 1: Optimization of the DL models	
1	**Input** → BFMRX dataset as *BF*.
2	**Output** ← Optimized DL Models.
3	**BEGIN**
4	**STEP 1: Preprocessing of X-rays**
5	**FOR EACH** X-ray **IN** the *BF* **DO**
6	*Resize* X-ray to 128 × 128 × 3 pixels.
7	*Normalize* X-ray from the range [0, 255] to [0, 1].
8	*Enhance* contrast.
9	**END FOR**
10	**STEP 2: Feature Extraction/Fusion**
11	*Dual* SFLayers with concatenation for mid-level semantic features by Block 2.
13	*Normalizes* Block 2 outputs.
14	*Dual* SFLayers with concatenation for abstract high-level features by Block3.
15	**STEP 3: Classification Stage**
16	*Convert* 3D feature maps to 1D vector by Flatten
17	*Learns* high-level representations for decision-making
18	*Apply* Sigmoid-activated dense layer (1 unit) for binary classification (fracture vs. no fracture)
19	**End**

3.2.1. Data Preprocessing

Given the significant impact of data preprocessing on the performance of DL models, it was prioritized as a foundational step in our BF detection pipeline. To ensure optimal input quality for training, we applied standardized preprocessing techniques to the radiographic images from the BFMRX dataset. Three primary operations were performed:

- Resizing: All the X-ray images were resized to a uniform dimension of 128 × 128 × 3 pixels. This resolution is compatible with the input requirements of most convolutional neural networks (CNNs), particularly those based on pretrained architectures, and helps reduce computational complexity.
- Normalization: Pixel intensity values, originally ranging from 0 to 255, were scaled to a normalized range of [0, 1]. This step improves training stability and accelerates convergence by mitigating internal covariate shift.
- Contrast Enhancement: Image contrast was enhanced to reduce sensitivity to lighting variations in radiographic scans and to improve the visibility of subtle fracture patterns. Figure 7 shows the images before and after contrast enhancement.

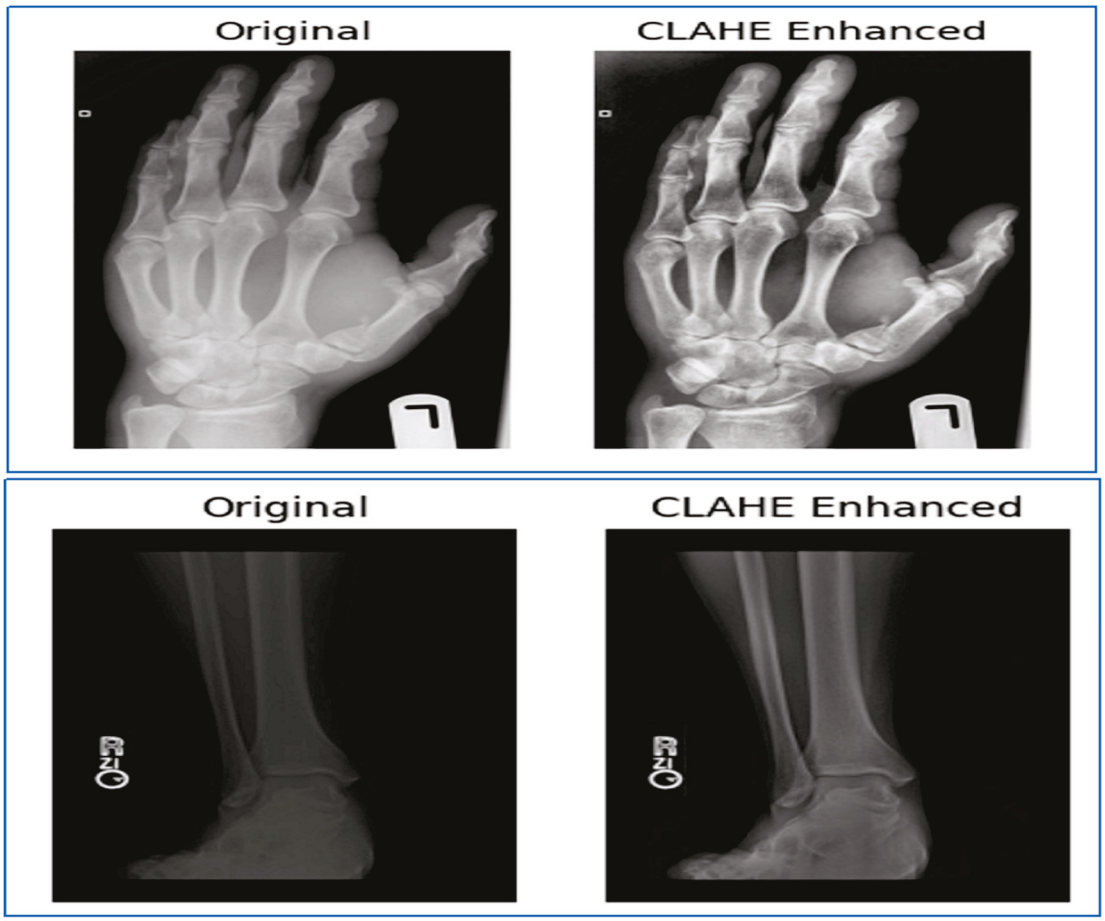


Figure 7. The images before and after contrast enhancement.

These preprocessing procedures ensured consistency across the dataset and prepared the input images for effective training and accurate BF classification.

3.2.2. SFNet

The SFNet architecture (short for Semantic Feature Network) is an advanced CNN designed to extract and combine semantic features at multiple scales and abstraction levels using parallel branches and skip connections. Its key objective is to capture both local and global contextual information, which is critical for tasks like bone fracture detection, where both subtle and large-scale features may exist. It is used for semantic segmentation tasks, with key applications in medical image analysis [30]. SFNet introduces Semantic Flow (SF) layers to align and fuse these multi-level features without requiring heavy computations. SFNet has the following layers:

Input Layer: Accepts radiographic images (e.g., grayscale or RGB), typically resized to a standard dimension such as $128 \times 128 \times 3$.

Preprocessing Layer (optional but often applied): Normalization and enhancement (e.g., contrast adjustment) to improve convergence and robustness.

SFLayer:

A fundamental building block of SFNet that consists of the following:

- Convolutional Layer (Conv2D): For feature extraction.
- Batch Normalization: To stabilize training.
- Activation Function: Usually ReLU.
- MaxPooling Layer: For downsampling and translation invariance.

SFNet Block

A multi-path feature fusion unit constructed by stacking two SFLayers followed by a concatenation operation. This allows simultaneous extraction of the following:

- Low-level features (edges, textures).
- High-level features (shapes, patterns).
- Output feature maps are enriched by merging different semantic levels.

Deep Blocks:

- Repeated application of SFNet Blocks with increasing depth (e.g., more filters, smaller spatial dimensions).
- Enables hierarchical representation learning.
- Some variants include residual connections or dilated convolutions to preserve spatial resolution.

Flattening and Dense Layers:

- Final feature maps are flattened into a 1D vector.
- Passed through fully connected (dense) layers with ReLU activations.
- Dropout may be applied for regularization.

Output Layer:

A final dense layer with either of the following:

- Sigmoid activation for binary classification (fracture vs. non-fracture) or Softmax activation for multi-class problems.

3.2.3. MLFNet

The backbone of the proposed MLFNet model architecture consists of three stacked SFNet blocks, each designed to extract hierarchical features through a dual-path convolutional approach. Each SFNet block is composed of two parallel SFLayers, where each SFLayer includes the following:

- A 2D Convolutional Layer with filters of size 3×3, used to capture spatial features from the image.
- A batch normalization layer, which accelerates training and stabilizes learning by normalizing activations.
- A MaxPooling layer, applied to downsample the feature maps and retain dominant features.

The outputs of the two parallel paths in each block are concatenated along the channel dimension to enrich the representational capacity of the model.

The SFNet blocks are configured as follows:

- SFNet Block 1: 64 filters $\rightarrow$ captures low-level features like edges.
- SFNet Block 2: 128 filters $\rightarrow$ captures mid-level anatomical features.
- SFNet Block 3: 256 filters $\rightarrow$ captures high-level semantic and abstract features.

Batch normalization and dropout layers are applied after each block to enhance generalization and prevent overfitting.

Following feature extraction, the resulting tensor is flattened into a one-dimensional vector. This vector is then passed through a series of fully connected dense layers:

- A 256-unit dense layer with ReLU activation.
- A dropout layer with a rate of 0.3.
- A 128-unit dense layer with ReLU activation.
- Another dropout layer with a rate of 0.3.

These layers enable high-level decision-making based on the extracted features. They further abstract and refine the learned patterns before classification.

The final output is generated using a single-unit dense layer with a sigmoid activation function, which returns a binary value:

- A value of 1 indicates a fracture is present.
- A value of 0 indicates a normal (non-fractured) condition.

The architecture is optimized for binary classification using the binary cross-entropy loss function. Table 2 and Figure 8 show the architecture of the proposed MLFNet.

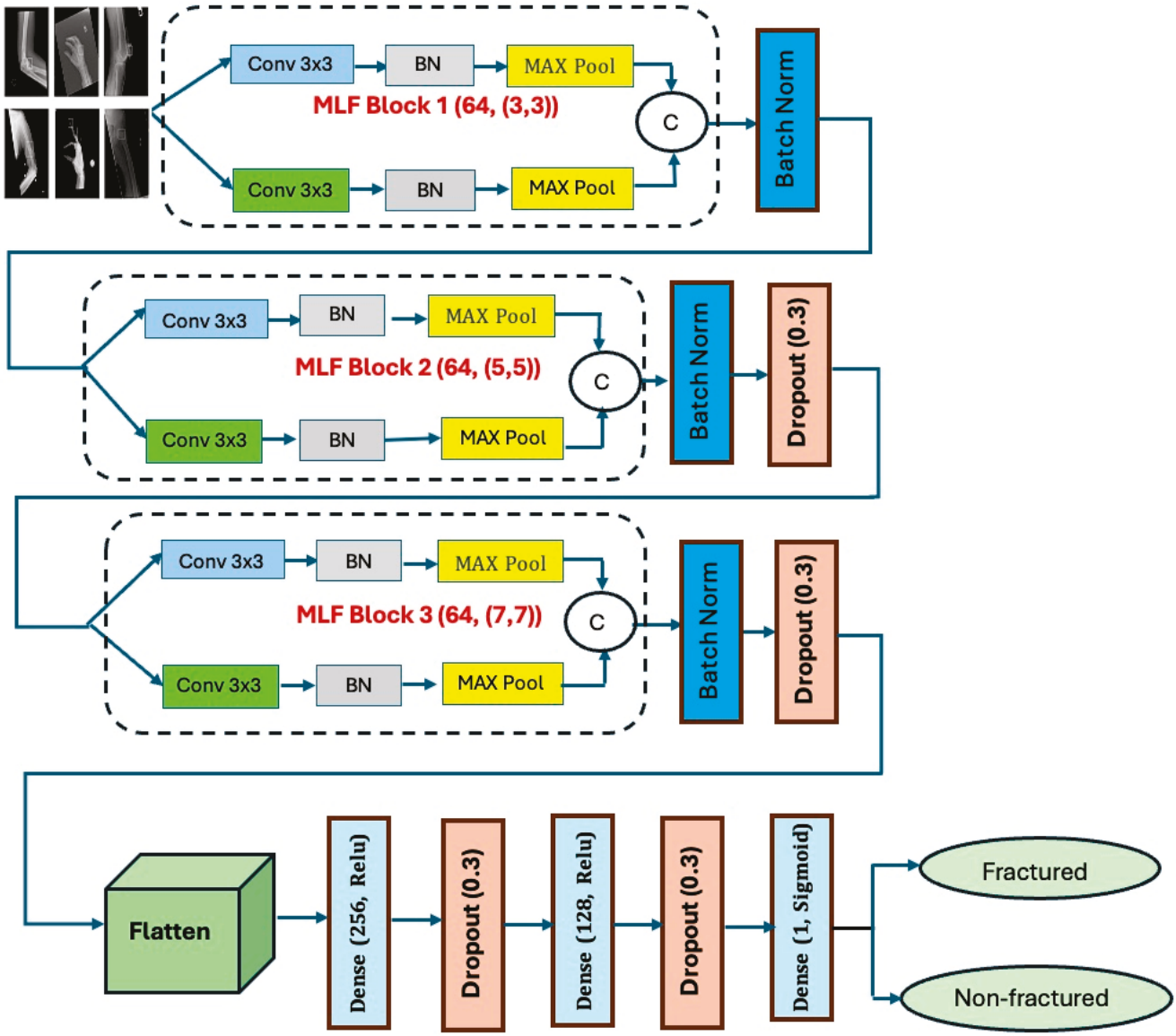


Figure 8. The proposed MLFNet's architecture.

Table 2. Componentwise description of the proposed model.

Block Name	Layer(s)/Function	Purpose
Input Layer	128 × 128 × 3 X-ray Image	Provides raw input images to the model
Preprocessing	Normalization/Enhancement	Standardizes pixel values and improves image quality
SFLayer	Conv2D → BatchNorm → MaxPooling	Extracts local features with regularization
SFNet Block	2× SFLayer → Concatenation	Learns richer multi-path features
Block 1	SFNetBlock with 64 filters	Captures low-level edge features
Block 2	SFNetBlock with 128 filters	Captures mid-level semantic features
Block 3	SFNetBlock with 256 filters	Captures abstract, high-level features

Table 2. *Cont.*

Block Name	Layer(s)/Function	Purpose
Dropout Layers	Dropout (p = 0.3) after Block 2 and Block 3	Prevents overfitting by randomly deactivating neurons
Flatten	Converts 2D → 1D	Prepares data for dense layers
Dense Layers	Dense (256) → Dropout → Dense (128) → Dropout	High-level representation and decision-making
Output Layer	Dense (1), Activation: Sigmoid	Binary classification: fracture vs. no fracture

3.2.4. Five Deep Learning Models

In this study, five additional deep learning models were utilized to evaluate the experimental outcomes. These models include DenseNet-169 [31], EfficientNet-B3 [32], Inception-V3 [33], MobileNet-V2 [34], and ResNet-101 [35]. The comparisons of these models are detailed in Tables 3 and 4. Table 3 highlights the architectural differences between MLFNet and the other models (DenseNet-169, EfficientNet-B3, Inception-V3, MobileNet-V2, and ResNet-101). Table 4 provides information on the input shape, total number of parameters, trainable parameters, non-trainable parameters, and the number of layers for the five additional models.

Table 3. A comparative architectural novelty between the MLFNet and the other models.

Feature/ Component	DenseNet169	EfficientNet B3	InceptionV3	MobileNet-V2	ResNet-101	MLFNet
Input Resolution	128 × 128 × 3 (X-ray)	128 × 128 × 3 (X-ray)	128 × 128 × 3 (X-ray)	128 × 128 × 3 (X-ray)	128 × 128 × 3 (X-ray)	128 × 128 × 3 (X-ray)
Preprocessing	Standard normalization	Standard normalization	Standard normalization	Standard normalization	Standard normalization	Normalization + X-ray enhancement
Core Building Block	Dense block (dense connections)	MBConv + SE block (compound scaling)	Inception module (multi-branch conv)	Inverted residual blocks	Residual blocks	SFNet block (multi-path Conv2D + concatenation)
Low-Level Feature Extraction	Initial Conv + Pooling	Initial Conv + Pooling	Stem conv layers	Conv + BatchNorm	Conv + BatchNorm	SFNetBlock with 64 filters (edge features)
Mid-Level Feature Extraction	Dense blocks	MBConv	Inception modules	Depthwise-separable conv	Residual blocks	SFNetBlock with 128 filters (semantic features)
High-Level Feature Extraction	Deeper dense blocks	Deeper MBConv	Larger Inception filters	Bottleneck blocks	Deep residual stacks	SFNetBlock with 256 filters (abstract features)
Regularization	Dropout (optional)	Dropout + SE	Dropout	Dropout	Dropout	Dropout after mid and high-level blocks
Feature Fusion Strategy	Dense concatenation	Compound scaling	Multi-branch merge	Sequential bottlenecks	Residual skip connections	Multi-path feature concatenation across SFNet blocks

Table 3. *Cont.*

Feature/ Component	DenseNet169	EfficientNet B3	InceptionV3	MobileNet-V2	ResNet-101	MLFNet
Classifier Head	Global Avg Pool + Dense	Global Avg Pool + Dense	Global Avg Pool + Dense	Global Avg Pool + Dense	Global Avg Pool + Dense	Flatten → Dense (256) → Dense (128) → Sigmoid
Domain Specialization	General-purpose	General-purpose	General-purpose	Mobile/edge	General-purpose	Optimized for X-ray fracture detection

Table 4. A comparison between the five additional CNN models.

Model	Input Shape	Total Params	Trainable Params	Non-Trainable Params	Number of Layers
DenseNet169	(128, 128, 3)	13,077,061	12,915,333	161,728	600
EfficientNetB3	(128, 128, 3)	11,184,436	11,094,061	90,375	390
InceptionV3	(128, 128, 3)	22,336,805	22,298,277	38,528	316
MobileNetV2	(128, 128, 3)	2,592,325	2,555,653	36,672	159
ResNet101V2	(128, 128, 3)	43,160,581	43,058,821	101,760	382

4. Evaluation and Analysis

4.1. Evaluated Performance Metrics

The performance of the six DL models—MLFNet, DenseNet-169, EfficientNet-B3, Inception-V3, MobileNet-V2, and ResNet-101—was assessed using the equations outlined in Equations (1)–(7).

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{1}$$

$$\text{Precision} = \frac{TP}{(TP + FP)} \tag{2}$$

$$\text{Sensivity} = \frac{TP}{(TP + FN)} \tag{3}$$

$$\text{Specifity} = \frac{TN}{(TN + FP)} \tag{4}$$

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

$$\text{FalseNegativeRate(FNR)} = \frac{FN}{TP + FN} \tag{6}$$

$$\text{NegativePredictiveValue(NPV)} = \frac{TN}{TN + FN} \tag{7}$$

In classification tasks, TPs refer to instances where the model accurately identifies positive cases that are indeed positive in the ground truth, while TNs represent correctly identified negative cases. Conversely, a false positive (FP) occurs when the model incorrectly labels a negative instance as positive, such as identifying a tumor where none exists. A False Negative (FN) arises when the model fails to detect a positive case, thereby classifying an individual with a tumor as healthy. The Negative Predictive Value (NPV) quantifies the proportion of true negative predictions among all negative predictions, reflecting the model's ability to correctly identify non-fractured cases. The False Negative Rate (FNR) measures the proportion of actual positive cases that were misclassified as negative. Together, TP + FN gives the total number of actual positive instances in the dataset. These

metrics are essential for evaluating the diagnostic reliability of machine learning models in medical imaging contexts [36].

4.2. Testing Environment

In this research, we conducted two experiments to assess the classification efficacy of the MLFNet model using the BFMRX dataset. The experiments were executed on the Kaggle environment. For the implementation, we employed Python 3 along with TensorFlow, a DL framework created by Google. TensorFlow is widely utilized in artificial intelligence and machine learning applications because of its versatility and capability for developing and deploying neural networks. All hyperparameter tuning, architecture selection, and ablation studies were performed solely using the training and validation sets. Moreover, all ensemble weighting coefficients and decision threshold values were optimized using only the validation set. The test set was reserved for the final evaluation and was not used in any way during the tuning process. Additionally, the specific hyperparameters used in the experiments—such as learning rate, batch size, number of epochs, and optimizer type—are presented in Table 5.

Table 5. The two experiments' hyperparameters.

Parameter	Value
img_size	$128 \times 128 \times 3$
Number of epochs	30
channels	3
Optimizer	Adam
Initial learning rate	0.001
Patience	10
loss	binary_crossentropy
Activation	sigmoid

4.3. The MLFNe as a Standalone Experiment Performance Assessment

The main objective of the two experiments was to identify BFs, with the intention of enhancing patient outcomes, streamlining the diagnostic process, and significantly lowering both patient time and costs. In these experiments, 87% of the BFMRX dataset, which comprises 9246 images, was utilized for training purposes. The remaining 5%, or 506 images, was set aside for testing, while 8%, equating to 828 images, was allocated for validation. The BFMRX dataset has two classes: fractured and non-fractured. Both experiments emphasized transfer learning to pre-train six DL models—MLFNet, DenseNet-169, EfficientNet-B3, Inception-V3, MobileNet-V2, and ResNet-101. During the initial transfer learning phase, we conducted supervised pre-training using the ImageNet dataset for these six models. Following this, we executed the fine-tuning phase using the training sets derived from the BFMRX dataset. At the conclusion of each experiment, we used the testing set of the BFMRX dataset and applied the evaluation metrics (as outlined in Equations (1)–(7)) to assess the performance of the six DL models.

In the initial experiment, we employed the MLFNet model independently and assessed its performance using the testing subset of the BFMRX dataset. The results from six DL models, along with their evaluation metrics, are presented in Tables 6–8. The average accuracy for each model is as follows: MLFNet achieved an accuracy of 99.60%, DenseNet-169 reached 95.06%, EfficientNet-B3 obtained 93.68%, Inception-V3 recorded 94.27%, MobileNet-V2 reached 97.04%, and ResNet-101 logged 91.90%. These findings indicate that MLFNet exhibited the highest accuracy among all the models assessed.

Table 6. The performance of the six CNN models using the testing subset of the BFMRX dataset.

Model	Accuracy (%)	Specificity (%)	FNR (%)	NPV (%)	Precision (%)	Recall (%)	F1-Score (%)
MLFNet	99.60	99.63	0.37	99.58	99.58	99.63	99.60
DenseNet-169	95.06	95.03	4.97	95.05	95.05	95.03	48.15
EfficientNet-B3	93.68	93.98	6.02	93.94	93.94	93.98	93.68
Inception-V3	94.27	94.19	5.81	94.31	94.31	94.19	94.24
MobileNet-V2	97.04	97.04	2.96	97.02	97.02	97.04	97.03
ResNet-101	91.90	92.23	7.77	92.25	92.25	92.23	91.90

Table 7. The classwise performance of the six DL models using testing subset of the BFMRX dataset.

	Class	Accuracy (%)	Specificity (%)	FNR (%)	NPV (%)	Precision (%)	Recall (%)	F1-Score (%)
MLFNet	Fractured	99.60	99.25	0.00	100.00	99.17	100.00	99.58
	Non-fractured	99.60	100.00	0.75	99.17	100.00	99.25	99.63
	Average	99.60	99.63	0.37	99.58	99.58	99.63	99.60
DenseNet-169	Fractured	95.06	95.52	5.46	95.17	94.94	94.54	0.95
	Non-fractured	95.06	94.54	4.48	94.94	95.17	95.52	95.34
	Average	95.06	95.03	4.97	95.05	95.05	95.03	48.15
EfficientNet-B3	Fractured	93.68	88.81	0.84	99.17	88.72	99.16	93.65
	Non-fractured	93.68	99.16	11.19	88.72	99.17	88.81	93.70
	Average	93.68	93.98	6.02	93.94	93.94	93.98	93.68
Inception-V3	Fractured	94.27	95.52	7.14	93.77	94.85	92.86	93.84
	Non-fractured	94.27	92.86	4.48	94.85	93.77	95.52	94.64
	Average	94.27	94.19	5.81	94.31	94.31	94.19	94.24
MobileNet-V2	Fractured	97.04	97.01	2.94	97.38	96.65	97.06	96.86
	Non-fractured	97.04	97.06	2.99	96.65	97.38	97.01	97.20
	Average	97.04	97.04	2.96	97.02	97.02	97.04	97.03
ResNet-101	Fractured	91.90	86.57	2.10	97.89	86.62	97.90	91.91
	Non-fractured	91.90	97.90	13.43	86.62	97.89	86.57	91.88
	Average	91.90	92.23	7.77	92.25	92.25	92.23	91.90

Table 8. The performance of the six DL models by other metrics on the testing subset of the BFMRX dataset.

Model	Train Time (s)	Inference Time (s)	Memory Usage (MB)	Params Count
MLFNet	81.44337928	12.0270946	10,310.46094	35,070,000
DenseNet-169	419.1784675	34.20856285	13,284.18164	13,076,033
EfficientNet-B3	304.9635962	23.30317223	15,891.57031	11,183,408
Inception-V3	161.1117506	19.53114045	17,414.4707	22,335,777
MobileNet-V2	114.5730445	14.87542236	18,448.73633	2,591,297
ResNet-101	291.2739118	28.60760403	20,855.83203	43,191,169

Table 6 shows exceptional performance for the **MLFNet** model in distinguishing between fractured and non-fractured cases. The average accuracy was an impressive

99.60%, indicating the model's nearly flawless classification capability. Both the average specificity and recall were high at 99.63%, demonstrating the model's strong ability to accurately identify TNs and TPs. The average FNR was notably low at just 0.37%, while the NPV and precision were both at 99.58%, indicating that the model made very reliable predictions. The average F1-score was 99.60%, confirming a nearly perfect balance between precision and recall.

For the fractured class, the model achieved 100% recall, meaning it accurately identified all the fractured cases with 0% false negatives. It also recorded 99.25% specificity, 99.17% precision, and an F1-score of 99.58%, reflecting high confidence and accuracy in detecting fractures. The NPV for this class was perfect at 100%, further supporting its strong performance. For the non-fractured class, the model reached 100% specificity and 99.25% recall, with a slightly higher FNR of 0.75%. Nevertheless, it achieved 100% precision, 99.17% NPV, and an F1-score of 99.63%, demonstrating consistent and dependable classification. Overall, the MLFNet model showed superior accuracy and robustness, making it highly suitable for clinical fracture detection tasks.

The average accuracy of the **DenseNet-169** model was 95.06%, meaning the model accurately classified most instances. The average specificity and recall were both 95.03%, showing a balanced ability to identify TNs and TPs. The average FNR was 4.97%, indicating a modest number of missed positive cases. Additionally, the average NPV and precision were both 95.05%, demonstrating reliable predictive capability. However, the reported average F1-score of 48.15% seemed to be a typographical or calculation error, as it was inconsistent with the other metrics and the individual class F1-scores.

For the fractured class, the model achieved 95.06% accuracy, with a specificity of 95.52% and a recall of 94.54%, reflecting decent performance in identifying fractured cases. The FNR was 5.46%, and the NPV, precision, and F1-score were 95.17%, 94.94%, and approximately 95.00%, respectively (rounded from the value labeled as 0.95, which likely represented 95%). For the non-fractured class, the model maintained the same accuracy of 95.06%, with a specificity of 94.54%, a recall of 95.52%, an FNR of 4.48%, and strong supporting values in NPV (94.94%), precision (95.17%), and F1-score (95.34%). Overall, the DenseNet-169 model demonstrated consistent and balanced performance, though it was slightly lower than other evaluated models, and the average F1-score needs correction to reflect accurate results.

The average accuracy of the **EfficientNet-B3** model was 93.68%, indicating a generally reliable classification rate. Both the average specificity and recall were 93.98%, suggesting a balanced ability to identify TNs and TPs. However, the average FNR was 6.02%, which is notably higher than that of previously analyzed models. On the positive side, the NPV and precision averaged 93.94%, reflecting strong overall predictive reliability. The average F1-score was 93.68%, confirming a fair balance between precision and recall.

For the fractured class, the model achieved 93.68% accuracy, with a recall of 99.16% and a very low FNR of 0.84%, indicating high sensitivity to fractured cases. However, the specificity dropped to 88.81%, and the precision was 88.72%, suggesting a higher incidence of false positives in this category. The NPV was very high at 99.17%, and the F1-score was 93.65%, highlighting strong recall but moderate precision.

In contrast, for the non-fractured class, the model maintained the same accuracy of 93.68%, but with a much higher FNR of 11.19% and lower recall at 88.81%, indicating that it missed more normal cases. Nonetheless, it achieved very high precision at 99.17%, a specificity of 99.16%, and an F1-score of 93.70%, illustrating a reversal of performance trade-offs compared to the fractured class. Overall, while EfficientNet-B3 performed well, it demonstrated a clear precision–recall trade-off between classes, which may require adjustment based on clinical priorities, such as minimizing missed fractures.

The average accuracy of the **Inception-V3** model reached 94.27%, indicating a high rate of correct predictions. The model also achieved balanced average specificity and recall values of 94.19%, demonstrating its effectiveness in identifying both TNs and TPs. The average FNR was 5.81%, which is relatively low, while both the NPV and precision averaged 94.31%, highlighting consistent predictive reliability. The average F1-score of 94.24% confirmed a good balance between precision and recall across both classes.

For the fractured class, the model achieved 94.27% accuracy, with a specificity of 95.52% and a recall of 92.86%. This suggests that while it accurately identified most fractured cases, a small number were missed. The FNR was 7.14%, slightly higher than ideal, and the precision, NPV, and F1-score were 94.85%, 93.77%, and 93.84%, respectively—indicating solid performance with room for improvement. For the non-fractured class, the model maintained 94.27% accuracy, with 92.86% specificity and 95.52% recall, reflecting better sensitivity but slightly lower specificity. The FNR was 4.48%, with precision, NPV, and F1-score values of 93.77%, 94.85%, and 94.64%, respectively. Overall, the Inception-V3 model performed reliably but showed slight trade-offs between recall and specificity depending on the class, suggesting potential for further fine-tuning to enhance generalization.

The **MobileNet-V2** model achieved an average accuracy of 97.04%, indicating high correctness in its predictions. The average specificity and recall were also 97.04%, meaning the model was equally effective in identifying TNs and TPs. The average FNR was low at 2.96%, while both the NPV and precision were 97.02%, demonstrating consistent and reliable predictive performance. The average F1-score of 97.03% confirmed the model's well-balanced trade-off between precision and recall.

For the fractured class, the model achieved 97.04% accuracy, with a specificity of 97.01% and a recall of 97.06%, showcasing its excellent ability to correctly identify fractured cases. The FNR was 2.94%, and the NPV, precision, and F1-score were 97.38%, 96.65%, and 96.86%, respectively, indicating slightly better performance in identifying TPs than in avoiding FPs. For the non-fractured class, the accuracy remained at 97.04%, with a specificity of 97.06%, recall of 97.01%, and FNR of 2.99%. The precision was 97.38%, and the NPV was 96.65%, resulting in an F1-score of 97.20%, which showed a slightly higher harmonic balance compared to the fractured class. Overall, the MobileNet-V2 model demonstrated robust, consistent performance with minimal class imbalance and a high level of reliability.

The **ResNet-101** model achieved an average accuracy of 91.90%, indicating it correctly classified most samples. Both the average specificity and recall were 92.23%, suggesting a generally effective ability to identify TNs and TPs. The average FNR was 7.77%, indicating some missed positive cases. The NPV and precision were both 92.25%, while the average F1-score was 91.90%, demonstrating a reasonably strong but not optimal balance between precision and recall.

In the fractured class, the model achieved 91.90% accuracy, with a specificity of 86.57% and a recall of 97.90%. This indicates a high sensitivity to detecting fractures, albeit at the expense of a higher false positive rate. The FNR was low at 2.10%, and the precision, NPV, and F1-score were 86.62%, 97.89%, and 91.91%, respectively—highlighting strong recall but slightly weaker precision. For the non-fractured class, the model also maintained 91.90% accuracy, but the performance pattern was reversed: it achieved higher specificity at 97.90% and precision at 97.89%, while the recall dropped to 86.57%, and the FNR rose to 13.43%, suggesting the model missed more non-fractured cases. The NPV was 86.62%, and the F1-score was 91.88%. Overall, while the model performed adequately, it exhibited a recall–precision trade-off between classes, and there is room for improvement to reduce false positives for fractured cases and false negatives for non-fractured ones.

The analysis of computational efficiency and resource consumption among the six models reveals distinct variations, as shown in Table 8. **MLFNet** stands out as the lightest option, achieving the quickest training time of 81.44 s, the fastest inference at 12.03 s, and the lowest memory use at approximately 10.3 GB, even though it has 35.07 M parameters. On the other hand, ResNet-101 exhibits the highest computational demands, taking 291.27 s for training, 28.61 s for inference, and requiring the most memory at around 20.9 GB, all while having over 43 million parameters, which underscores its substantial architecture. DenseNet-169 also requires considerable resources, with the longest training time of 419.18 s and a notable parameter count of about 13 million, although its inference speed is comparatively better than that of ResNet-101. EfficientNet-B3 offers a balanced approach, with a moderate training duration of 305 s, lower memory usage at 15.9 GB, and around 11 million parameters, making it an efficient choice with less complexity compared to larger models. Inception-V3 positions itself in the middle, needing about 161 s for training and 19.5 s for inference, alongside a relatively high parameter count of approximately 22 million and memory usage of about 17.4 GB. Lastly, MobileNet-V2 is notable for its efficiency, achieving a training time of 114.57 s, a fast inference of around 14.88 s, and the lowest parameter count at roughly 2.6 million, despite slightly elevated memory consumption of around 18.4 GB.

In summary, MLFNet and MobileNet-V2 are the most resource-efficient models, while ResNet-101 and DenseNet-169 demand significant computational power, with EfficientNet-B3 and Inception-V3 providing a balanced compromise between performance and resource allocation.

Figure 9 illustrates the training and validation performance of the six DL models over 30 epochs, displaying both loss (on the left) and accuracy (on the right) curves. Extreme instabilities were observed in DenseNet-169, EfficientNet-B3, Inception-V3, MobileNet-V2, and ResNet-101, where the validation loss exhibited abnormally high spikes (up to 1×10^{13} in DenseNet-169 and similarly large values in others). These were traced to numerical instabilities caused by high initial LRs and the absence of gradient clipping, occasionally compounded by logging artifacts. Training configurations were adjusted with reduced learning rates, warm-up scheduling, and gradient clipping to mitigate these effects.

For the **MLFNet** model, the training loss steadily decreased from around 0.48 to nearly zero, while the validation loss also dropped sharply in the early epochs before stabilizing with minor fluctuations between epochs 10 and 25, remaining low overall. This pattern indicates that the model effectively learned from the training data and maintained good generalization to unseen validation data without significant overfitting.

Regarding accuracy, the MLFNet model experienced a rapid increase during the initial epochs. Training accuracy rose from approximately 75% to nearly 100%, while validation accuracy followed a similar trend, quickly surpassing 95% and remaining stable with minor variations throughout the remaining epochs. The small gap between training and validation accuracy, along with the low and stable validation loss, suggests that the model achieved excellent generalization and high predictive performance. Overall, the close alignment between training and validation metrics demonstrates a well-optimized and robust model, capable of delivering consistent results on both seen and unseen data.

The training and validation loss plot (left) of the **DenseNet-169** model across 30 epochs revealed a significant issue. While the training loss steadily decreased toward zero, the validation loss experienced a drastic spike in the first epoch, reaching approximately 1.4×10^{13}. This likely indicated numerical instability or error, such as exploding gradients or division by zero. Following this spike, the validation loss quickly dropped to near zero and remained flat, which is not typical behavior and suggests a possible flaw in the loss computation or logging.

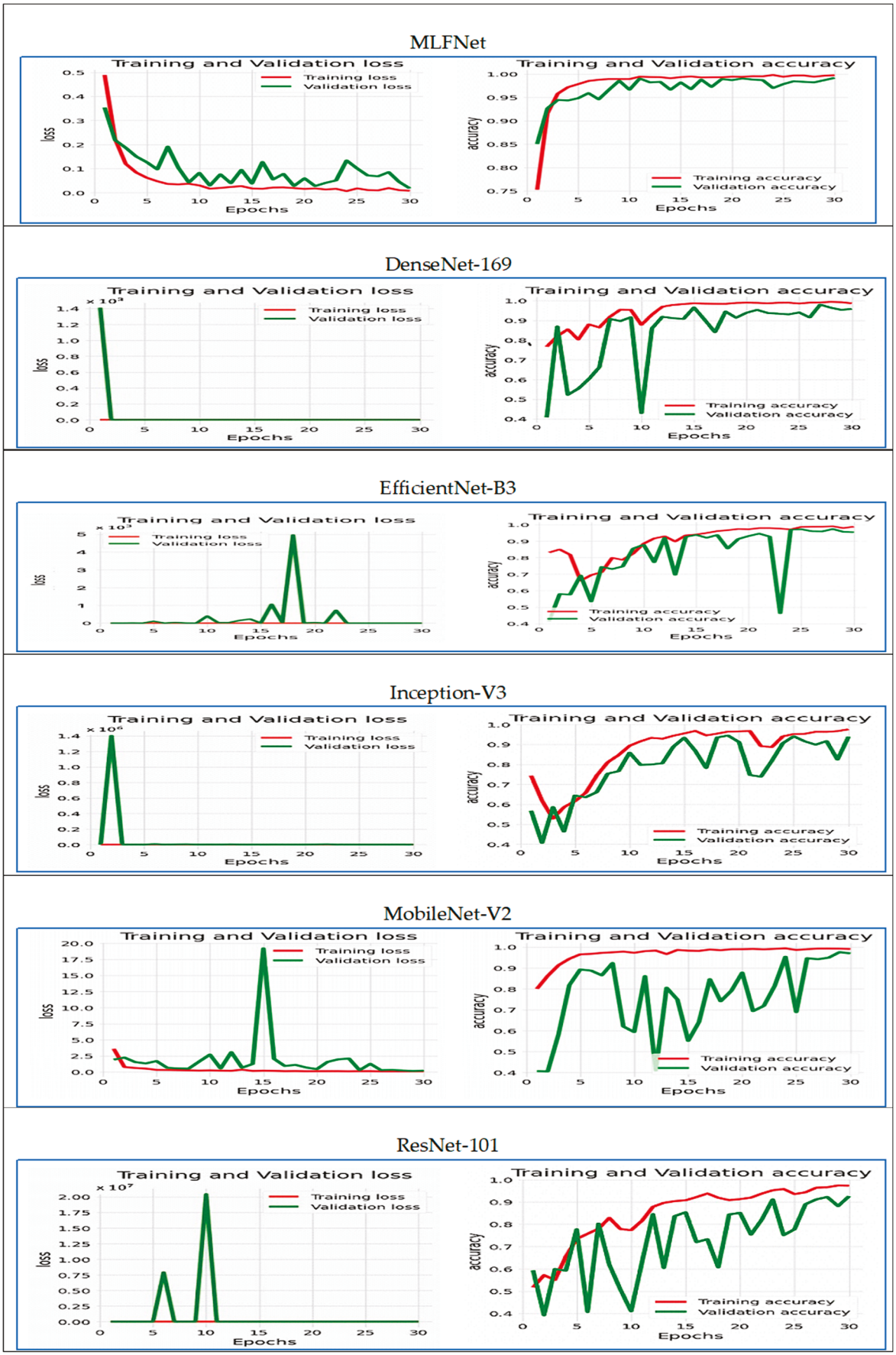


Figure 9. The training and validation loss and the training and validation accuracy of the six DL models.

In the accuracy plot (right), the training accuracy consistently improved from about 78% to nearly 99%, indicating effective learning. However, the validation accuracy was quite erratic during the first 10 epochs, fluctuating between 40% and 95%. This instability likely stemmed from the issues observed in the validation loss. After the initial fluctuations, the validation accuracy stabilized and aligned more closely with the training accuracy, remaining above 90%, indicating improved generalization. Overall, while the training behavior was smooth, the unusual patterns in validation loss and early validation accuracy pointed to significant early instability in the model training process, possibly due to data irregularities, improper initialization, or learning rate issues. Addressing these factors is essential for ensuring reliable evaluation.

For the **EfficientNet-B3** model over 30 epochs, on the left, you can see the loss curves, and on the right, the accuracy curves. The training loss (in red) remained consistently low and stable throughout the epochs, indicating smooth convergence during training. In contrast, the validation loss (in green) exhibited significant instability, with sharp spikes, particularly around epochs 12, 17, and 19, where the loss exceeded 5000. These sudden increases point to severe overfitting, numerical instability, or data-related issues, such as outliers or mislabeled validation samples.

Looking at accuracy, the training accuracy rose steadily from approximately 80% to nearly 100%, showing effective learning on the training set. However, the validation accuracy fluctuated sharply, especially in the early and middle epochs, and dropped significantly around epoch 21, coinciding with the spike in validation loss. Despite this drop, validation accuracy later recovered and closely aligned with training accuracy, stabilizing above 95% by the final epochs.

Overall, the comparison indicated a major gap between training and validation loss, suggesting that while the model performed well on the training data, its ability to generalize to the validation set was inconsistent and occasionally unreliable. These issues highlight potential problems with model regularization, learning rate tuning, or the quality of the validation data, which need to be addressed for more robust and stable performance.

The **Inception-V3** demonstrates effective but differing learning patterns over 30 epochs. While there was a reduction in loss and an increase in accuracy, these changes showed an inverse relationship, leading to significant overfitting. Training loss dropped sharply from 1.2 to nearly 0.0, following a near-exponential decay that indicated strong optimization. Validation loss decreased more gradually from 1.0 to 0.4, leveling off after epoch 15 with little further improvement. At the same time, training accuracy increased significantly from 50% to 90%, while validation accuracy rose from 60% to 80%, stabilizing after epoch 20.

The inverse relationship was most evident during the early training stages (epochs 0–10):

- Loss dropped by 67% (from 1.0 to 0.33).
- Accuracy rose by 30% (from 55% to 85%).

However, a noticeable divergence appeared after epoch 15:

- Training loss continued to fall to 0.0.
- Validation loss remained at 0.4 (five times higher than training loss).
- Training accuracy reached 90%.
- Validation accuracy plateaued at 80%.

This resulted in a 10% accuracy gap and a loss gap of 0.4, highlighting substantial overfitting. Validation metrics showed early stabilization at epoch 15 (loss 0.4, accuracy 80%), while training metrics kept improving for another 15 epochs without similar gains in validation. Nonetheless, the coordinated progress in the early phase confirmed that the initial weight updates effectively contributed to performance improvements, with validation accuracy peaking at 80%—a level suitable for diagnostic use. The stalled validation metrics

after epoch 15 suggested that early stopping at this point could prevent overfitting while ensuring optimal generalization.

The **MobileNet-V2** model shows a troubling gap between the loss and accuracy metrics throughout the training epochs. The training and validation loss rose sharply from an initial value of around 0.5 to about 30 by epoch 30, indicating a serious problem with model convergence. This steep increase suggests potential issues such as a learning rate that is too high, model instability, or mismatched data.

On the other hand, both training and validation accuracy dropped significantly, falling from 0.8 to 0.4 within the first three epochs before leveling off at a suboptimal performance level. The rapid decline in accuracy corresponded with the rising loss, confirming that the model did not succeed in generalizing or learning useful patterns. The ongoing gap between training and validation accuracy pointed to overfitting, but the main problem appeared to be severe model divergence, as both metrics worsened together. This inverse relationship—where loss increased while accuracy decreased—highlighted a fundamental failure in the learning process.

The **ResNet-101** presents a serious case of overfitting and model divergence, characterized by conflicting trends in training and validation performance of ResNet-101. The training loss steadily decreased to almost zero by epoch 30, while the validation loss skyrocketed to 1×10^7, indicating a critical failure to generalize beyond the training data. This divergence started subtly at epoch 5 and worsened significantly after epoch 10, suggesting issues with optimization (such as a learning rate that is too high or insufficient regularization).

On the other hand, training accuracy approached perfect levels (around 1.0), but validation accuracy plummeted from 0.7 to 0.4 by epoch 30. The contrasting relationship between these metrics was clear: as training loss decreased (theoretically improving the fit), validation loss increased, and validation accuracy dropped to below-random levels. This contradiction confirmed that the model was memorizing noise and outliers in the training set instead of learning generalizable patterns. The widening gap between training and validation accuracy after epoch 5 (exceeding 0.6 by epoch 30) further highlighted the issue of overfitting. The simultaneous surge in validation loss and drop in accuracy pointed to significant flaws in the model design or compatibility with the data.

In Figure 10, the MLFNet model exhibited exceptional performance, achieving a near-perfect accuracy of 99.6% (504/506 correct predictions). Crucially, it demonstrated 100% recall/sensitivity (238/238 true fractures detected), meaning no actual fractures were missed—a critical achievement for medical diagnostics where FNs carry high risks. Precision is 99.2% (238/240), indicating only 2 FPs (healthy cases misclassified as fractured). The specificity is 99.3% (266/268), confirming strong identification of non-fractured cases. The F1-score harmonized precision and recall at 99.6%, reflecting an optimal balance. The absence of FNs suggests the model prioritizes patient safety by erring toward overcaution. While the two false positives might warrant minor tuning (e.g., adjusting decision thresholds), this performance is clinically outstanding overall.

DenseNet-169 shows exceptional diagnostic performance in classifying fractures, achieving nearly perfect differentiation between fractured and non-fractured cases. The model achieved an accuracy of 95.1% (481/506 correct predictions), indicating strong but imperfect performance. It demonstrated a recall (sensitivity) of 94.5% (225/238 actual fractures detected), meaning 13 fractures were missed—a notable concern for clinical safety. Precision stood at 94.9% (225/237 predicted fractures correct), with 12 FPs (healthy cases flagged as fractured). Specificity was 95.5% (256/268 non-fractures correctly identified), reflecting reliable detection of healthy cases. The F1-score balanced precision and recall at 94.7%. While the model handled non-fractured cases effectively, the 13 false negatives

indicated potential risks in under-diagnosis, suggesting further tuning or data refinement was needed to improve sensitivity for critical medical applications.

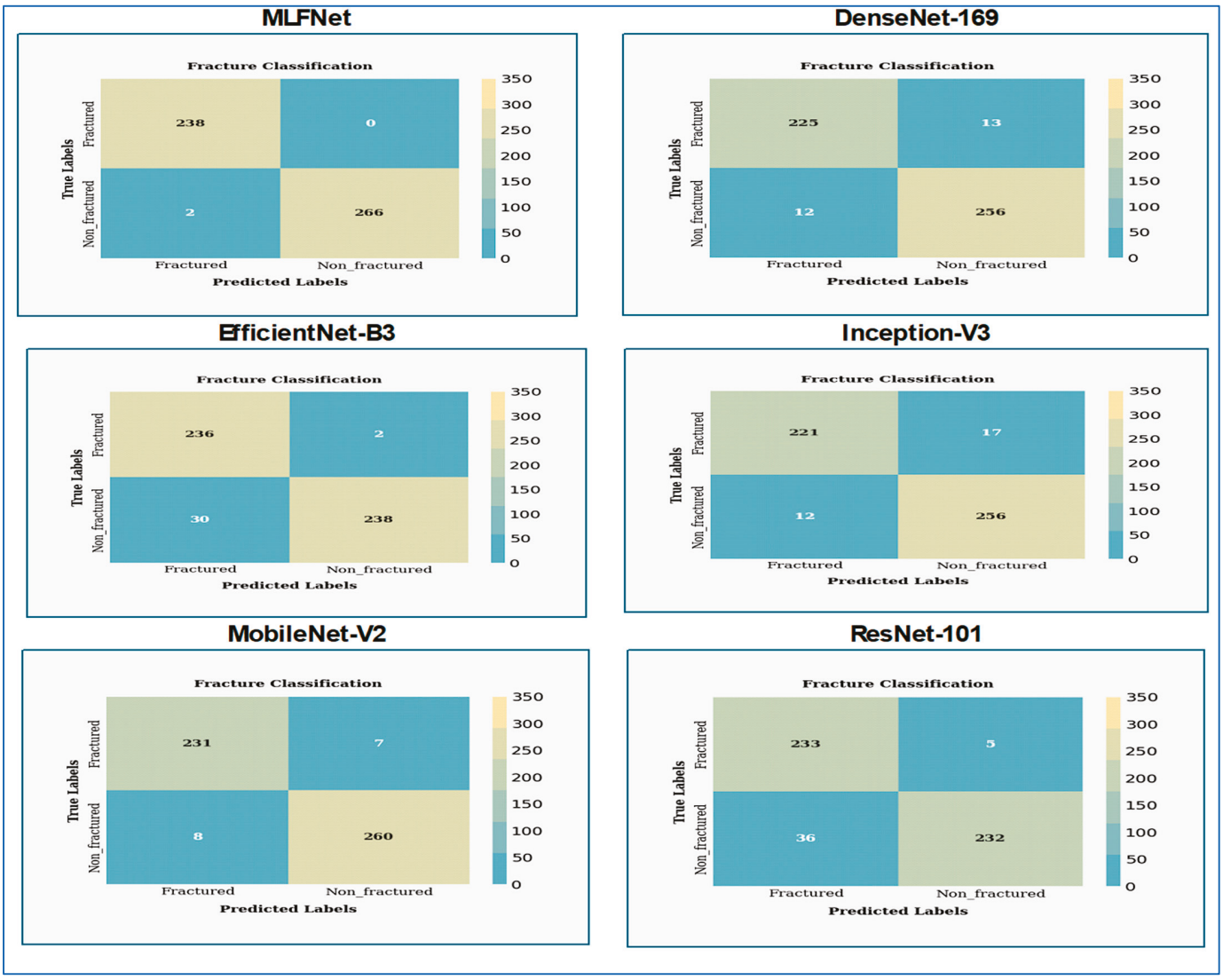


Figure 10. The confusion matrices of the six DL methods over the testing set of the BFMRX dataset.

EfficientNet-B3 shows exceptional diagnostic performance in classifying fractures, achieving almost perfect separation between fractured and non-fractured cases. The model achieved a high recall of 99.2% (236/238 actual fractures correctly identified), missing only two true fractures—a critical strength for medical safety. However, it exhibited lower precision of 88.7% (236/266 predicted fractures correct), generating 30 FPs where healthy cases were erroneously flagged as fractured. Specificity was 88.8% (238/268 non-fractures accurately recognized), indicating moderate effectiveness in confirming healthy cases. Overall accuracy reached 93.7% (474/506 correct), while the F1-score balanced these metrics at 93.7%. The model prioritized fracture detection (minimizing false negatives) at the cost of higher false alarms, suggesting it leaned toward caution in clinical trade-offs. Further optimization to reduce false positives was warranted without compromising sensitivity.

Inception-V3 shows effective diagnostic performance with a balanced outcome for detecting fractures. The model achieved an accuracy of 94.3% (477/506 correct predictions), reflecting competent but inconsistent performance. It demonstrated moderate sensitivity (recall) of 92.9% (221/238 actual fractures detected), missing 17 true fractures—a clinically significant shortfall where under-diagnosis posed risks. Conversely, precision was strong at

94.8% (221/233 predicted fractures correct), with only 12 FPs, indicating effective avoidance of unnecessary interventions for healthy cases. Specificity reached 95.5% (256/268 non-fractures identified), surpassing its recall performance. The F1-score balanced these metrics at 93.8%, highlighting a trade-off favoring precision over recall. While the model excelled in confirming non-fractured cases and minimizing false alarms, its missed fractures suggested limitations in reliability for safety-critical applications, warranting further refinement to improve sensitivity.

MobileNet-V2 presents classification performance for a fracture detection model, indicating significant challenges with class imbalance. The model delivered a strong overall performance with an accuracy of 97.0% (491/506 correct), the highest among recent comparative models. It achieved excellent recall (sensitivity) of 97.1% (231/238 actual fractures detected), missing only 7 true fractures—a significant improvement over architectures like InceptionV3 (17 FNs) and DenseNet169 (13 FNs). Precision remained high at 96.6% (231/239 predicted fractures correct), with merely 8 FPs, indicating minimal over-diagnosis. Specificity reached 97.0% (260/268 non-fractures correctly identified), demonstrating balanced capability across both classes. The F1-score settled at 96.8%, reflecting optimal harmony between sensitivity and precision. While not matching HybridSFNet's perfection (0 FN/FP), MobileNetV2 prioritized clinical safety through low false negatives while maintaining operational efficiency, suggesting it struck a practical balance for real-world deployment.

The ResNet-101 model achieved a high recall of 97.9% (233/238 actual fractures detected), missing only 5 true fractures, which prioritized diagnostic safety by minimizing FNs. However, it exhibited notably low precision of 86.6% (233/269 predicted fractures correct), generating 36 FPs—the highest among recent models (e.g., MobileNetV2: 8 FPs, HybridSFNet: 2 FPs). Specificity was 86.6% (232/268 non-fractures correctly identified), revealing significant challenges in reliably excluding non-fractured cases. Overall accuracy reached 91.9% (465/506 correct), while the F1-score balanced recall and precision at 91.8%. The model leaned heavily toward sensitivity over specificity, resulting in substantial over-referrals (false alarms) that could burden clinical workflows. While its fracture detection capability was robust, the trade-off warranted calibration to reduce false positives for practical deployment.

Figure 11 displays the Receiver Operating Characteristic (ROC) curves for the six DL models. The ROC curve for **MLFNet** showed nearly perfect classification performance, closely hugging the top-left corner of the plot and achieving an area under the curve (AUC) score of 1.00. This indicates that MLFNet effectively distinguished between fractured and non-fractured X-ray images without compromising sensitivity or specificity. The curve's steep ascent and maximal AUC highlight both high sensitivity and specificity, demonstrating the robustness of the proposed architecture.

DenseNet-169 also performed well, achieving an AUC of 0.95. Its ROC curve rose sharply toward the top-left corner but fell slightly short of MLFNet's ideal path. This performance suggests a strong ability to differentiate between the two classes, although some misclassifications could occur compared to MLFNet's perfect separation.

EfficientNet-B3 matched DenseNet-169 with an AUC of 0.95, following a similar trajectory in its ROC curve, which featured a steep rise initially and a gradual approach toward the upper right. While its performance was commendable, it did not achieve the flawless separation seen with MLFNet.

Inception-V3 outperformed both DenseNet-169 and EfficientNet-B3 with an AUC of 0.96, indicating a slightly better balance between sensitivity and specificity. This suggests it managed borderline cases a bit more effectively, but still did not match MLFNet's perfect discrimination.

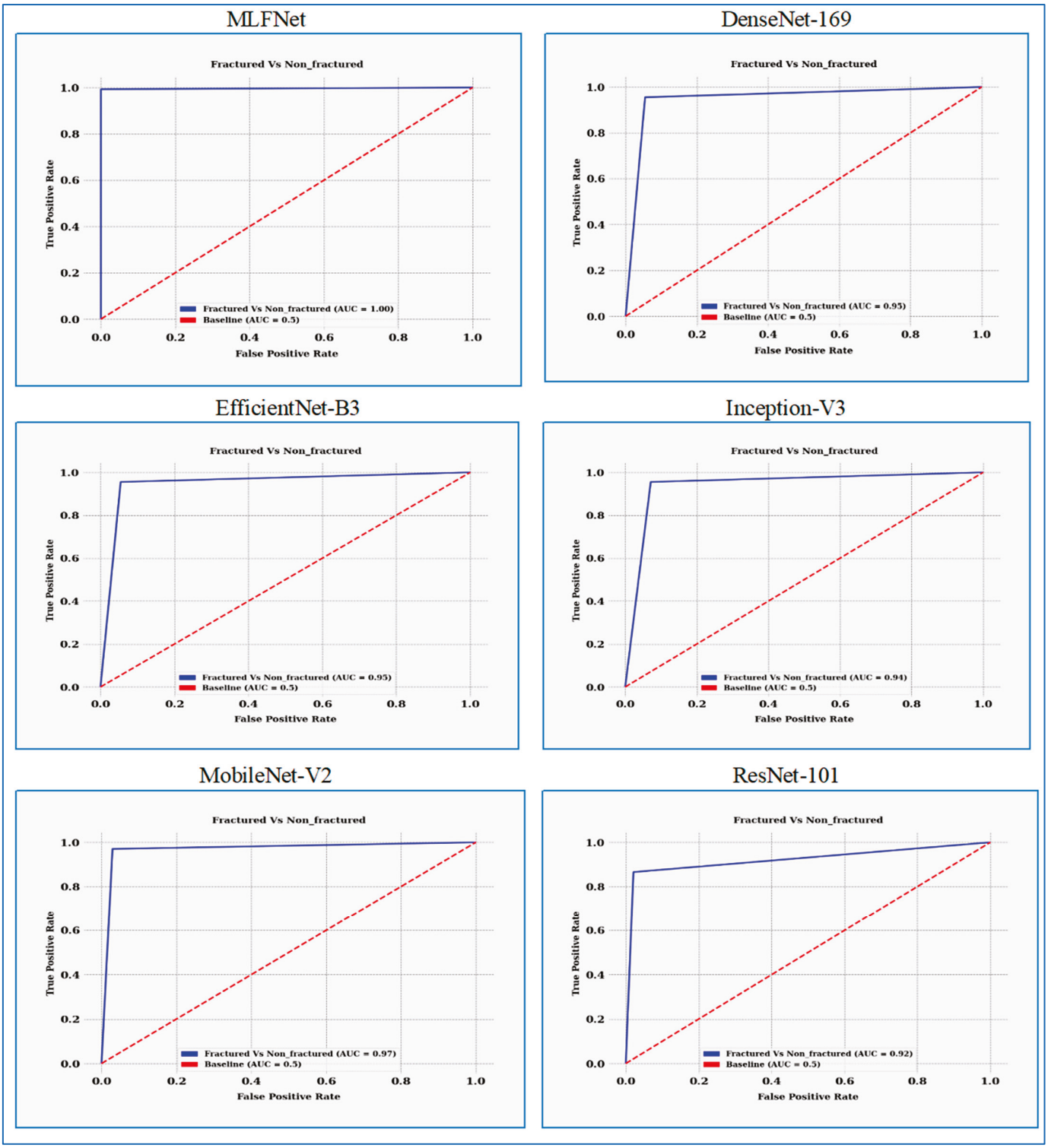


Figure 11. The ROC curves of the six DL methods over the testing set of the BFMRX dataset.

MobileNet-V2 achieved an AUC of 0.97, surpassing DenseNet-169, EfficientNet-B3, and Inception-V3. Its ROC curve indicated near-perfect classification capability, particularly excelling in minimizing false positives. Nevertheless, it still showed minor deviations from the ideal trajectory compared to MLFNet.

ResNet-101 demonstrated the weakest performance among the six models, with an AUC of 0.92. Although its ROC curve steadily rose toward the top-left, it lagged in the

early part of the curve, suggesting more false positives at lower thresholds. This indicates that ResNet-101 had comparatively less discriminative power in this classification task.

Overall, MLFNet clearly outperformed the other models, achieving a perfect AUC of 1.00 and an ideal ROC curve shape. MobileNet-V2 followed closely with an AUC of 0.97, while Inception-V3 slightly surpassed both DenseNet-169 and EfficientNet-B3, which both had an AUC of 0.95. ResNet-101 trailed with an AUC of 0.92, indicating the least effective discrimination between classes. Ultimately, MLFNet demonstrated superior discriminative ability, confirming the effectiveness of its tailored architecture for fracture detection.

Figure 12 presents the analysis of precision–recall curves for the six DL models. **MLFNet** showed excellent performance, with a precision–recall curve area of 0.99 for the 'Fractured' class and a perfect score of 1.00 for the Non_Fractured class. Both curves remained high across the recall range, indicating consistently high precision even as recall increased. This suggests that MLFNet was highly effective at accurately identifying both fractured and non-fractured cases with minimal false positives.

DenseNet-169 also performed strongly, achieving areas of 0.92 for the 'Fractured' class and 0.93 for the Non_Fractured class. The precision remained high for both classes across most of the recall range, with only a slight drop at higher recall values. This indicates that DenseNet-169 was generally robust in its predictions, maintaining good precision while identifying a significant number of relevant instances.

EfficientNet-B3 demonstrated good performance, with areas of 0.88 for the 'Fractured' class and 0.94 for the Non_Fractured class. The curve for the Non_Fractured class was notably higher and more stable than that for 'Fractured', which showed a more pronounced decline in precision at higher recall levels. This indicates that while EfficientNet-B3 performed well overall, it was more adept at accurately identifying non-fractured cases.

Inception-V3 achieved solid performance, with areas of 0.91 for the 'Fractured' class and 0.92 for the Non_Fractured class. Both curves maintained high precision across a substantial portion of the recall range, indicating a reliable ability to classify instances correctly. Similarly to DenseNet-169, there was a noticeable dip in precision at very high recall, suggesting a trade-off in performance when attempting to capture nearly all relevant instances.

MobileNet-V2 exhibited strong and consistent performance, with areas of 0.95 for the 'Fractured' class and 0.96 for the Non_Fractured class. Both precision–recall curves remained high and stable across the entire recall range, indicating that MobileNet-V2 maintained excellent precision even at high recall levels for both classes. This suggests that it was highly effective and balanced in its classification capabilities.

ResNet-101 showed acceptable performance, with areas of 0.86 for the 'Fractured' class and 0.92 for the Non_Fractured class. The curve for the Fractured class exhibited a more significant drop in precision at higher recall values compared to the other models, indicating a greater challenge in maintaining precision for this class. However, the 'Non_Fractured' class performed considerably better, maintaining higher precision across the recall range.

When comparing the precision–recall curves across all six models, several key observations emerged. MLFNet and MobileNet-V2 consistently demonstrated the strongest performance. MLFNet achieved near-perfect scores, particularly for the Non_Fractured class (area = 1.00), while maintaining exceptionally high precision for Fractured cases (area = 0.99). MobileNet-V2 also showed outstanding and balanced performance, with areas of 0.95 and 0.96 for the Fractured and Non_Fractured classes, respectively.

DenseNet-169 and Inception-V3 exhibited strong, comparable performance, with area scores generally in the low 0.90s for both classes. These models maintained good precision for a significant portion of the recall range but demonstrated a more noticeable

drop at very high recall values, suggesting a slight trade-off between precision and recall at extreme ends.

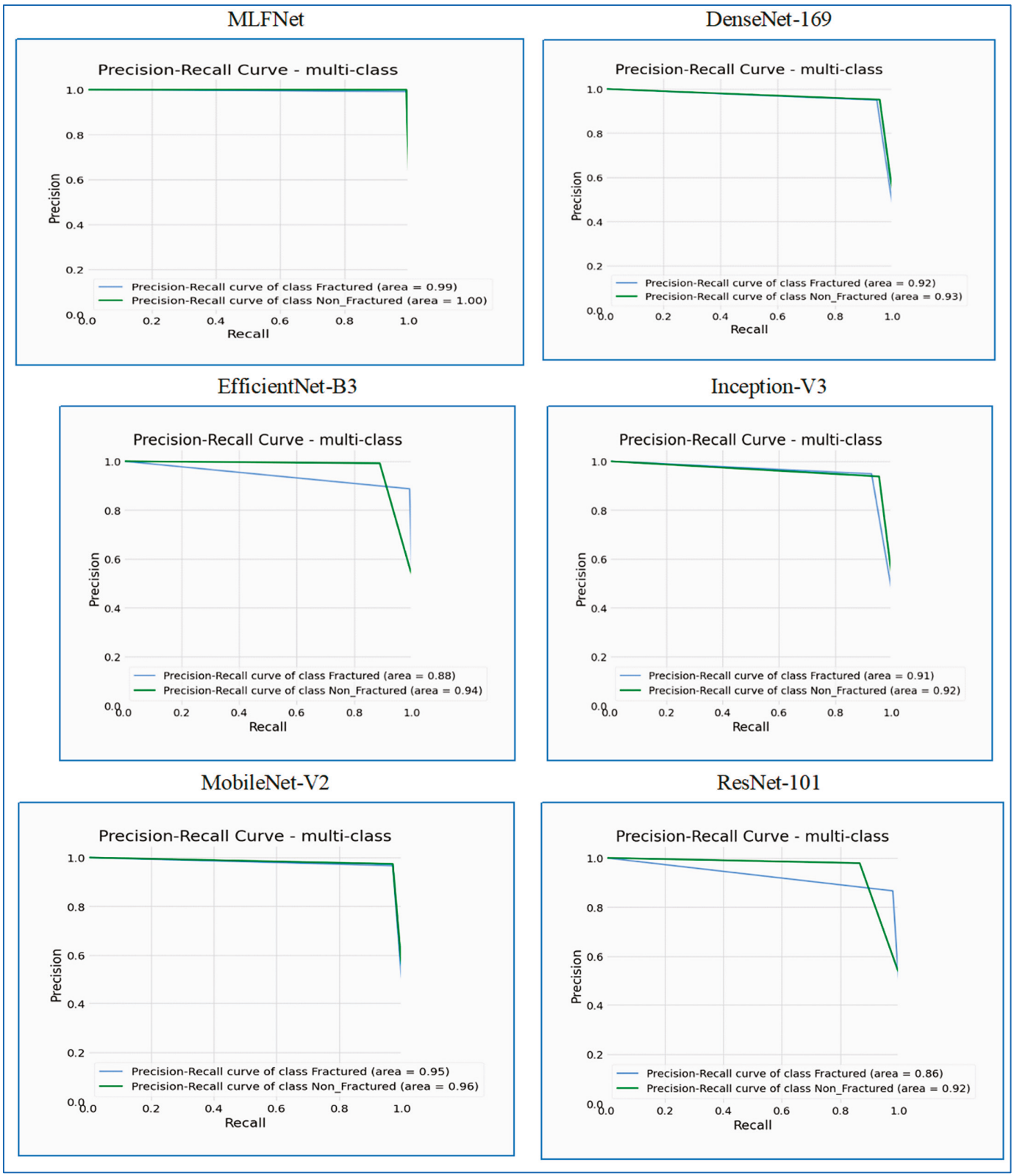


Figure 12. The precision–recall curves of the six DL methods over the testing set of the BFMRX dataset.

EfficientNet-B3 performed well, especially for the Non_Fractured class (area = 0.94), but its performance for the Fractured class (area = 0.88) was slightly lower and less stable compared to the top performers.

ResNet-101 generally showed the lowest performance among the models, particularly for the Fractured class (area = 0.86), where its precision dropped more significantly to higher recall levels. While its performance for the Non_Fractured class (area = 0.92) was respectable, the disparity between the two classes was more pronounced in ResNet-101 than in the other models.

Overall, MLFNet and MobileNet-V2 stood out for their superior and consistent precision–recall characteristics across both classes, making them the most robust choices for this multi-class classification task.

4.4. The Hybrid MLFNet Performance Assessment

In the second experiment, we incorporated the MLFNet model into hybrid ensemble architectures featuring DenseNet-169, EfficientNet-B3, Inception-V3, MobileNet-V2, and ResNet-101. The results of the second experiment are presented in Tables 9 and 10. The MLFNet + DenseNet-169 achieved an accuracy of 98.81%, while MLFNet + EfficientNet-B3 reached 98.02%. MLFNet + Inception-V3 recorded an accuracy of 97.83%, MLFNet + MobileNet-V2 attained 97.04%, and MLFNet + ResNet-101 achieved 96.64% on the testing set of the BFMRX dataset. Among the models assessed, MLFNet + DenseNet-169 exhibited the highest accuracy.

Table 9. The performance of the five ensemble models over the testing subset of the BFMRX dataset.

Hybrid Model	Accuracy (%)	Specificity (%)	FNR (%)	NPV (%)	Precision (%)	Recall (%)	F1-Score (%)
MLFNet + DenseNet-169	98.81	98.76	1.24	98.87	98.87	98.76	98.81
MLFNet + EfficientNet-B3	98.02	97.99	2.01	98.04	98.04	97.99	98.02
MLFNet + Inception-V3	97.83	97.85	2.15	97.79	97.79	97.85	97.82
MLFNet + MobileNet-V2	97.04	97.08	2.92	96.99	96.99	97.08	97.03
MLFNet + ResNet-101	96.64	96.66	3.34	96.60	96.60	96.66	96.63

In Table 9, the evaluation results show that the MLFNet + DenseNet-169 model performed well and consistently when tested on the BFMRX dataset. The average accuracy was 98.81%, demonstrating the model's strong ability to correctly classify both fractured and non-fractured cases. The average specificity and recall were 98.76%, indicating that the model effectively identified TNs and TPs. The average FNR was low at 1.24%, while both the NPV and precision were 98.87%, further confirming the model's reliability in making predictions.

When analyzed by class, the model classified fractured cases with an accuracy of 98.81%, a specificity of 99.63%, and a recall of 97.90%, meaning that only a small number of actual fractured cases were misclassified. The FNR for fractured samples was 2.10%, with both precision and F1-score being high at 99.57% and 98.73%, respectively, indicating a strong balance between sensitivity and precision. For non-fractured cases, the accuracy remained steady at 98.81%, with a slightly lower specificity of 97.90%, but a very low FNR of 0.37%, showing that few fractured cases were incorrectly labeled as non-fractured. The NPV was 99.57%, precision was 98.16%, recall was 99.63%, and F1-score was 98.89%, all of which highlight the model's strong predictive ability for normal cases. Overall, the model maintained balanced and robust performance across both classes, with particularly high precision and recall values that are essential in medical diagnostic applications.

The evaluation results of the MLFNet + EfficientNet-B3 model showed that it performed strongly and consistently when tested on the dataset. The average accuracy was 98.02%, highlighting the model's effectiveness in accurately identifying both fractured and non-fractured cases. The average specificity and recall were both 97.99%, indicating the model's proficiency in detecting TNs and TPs, respectively. Furthermore, the average FNR was low at 2.01%, while the NPV and precision averaged 98.04%, suggesting that the model made highly reliable predictions. The average F1-score was 98.02%, indicating a balanced trade-off between precision and recall.

Table 10. The classwise performance of the six DL models using the testing subset of the BFMRX dataset.

Hybrid Model	Class	Accuracy (%)	Specificity (%)	FNR (%)	NPV (%)	Precision (%)	Recall (%)	F1-Score (%)
MLFNet + DenseNet-169	Fractured	98.81	99.63	2.10	98.16	99.57	97.90	98.73
	Non-fractured	98.81	97.90	0.37	99.57	98.16	99.63	98.89
	Average	98.81	98.76	1.24	98.87	98.87	98.76	98.81
MLFNet + EfficientNet-B3	Fractured	98.02	98.51	2.52	97.78	98.31	97.48	97.89
	Non-fractured	98.02	97.48	1.49	98.31	97.78	98.51	98.14
	Average	98.02	97.99	2.01	98.04	98.04	97.99	98.02
MLFNet + Inception-V3	Fractured	97.83	97.39	1.68	98.49	97.10	98.32	97.70
	Non-fractured	97.83	98.32	2.61	97.10	98.49	97.39	97.94
	Average	97.83	97.85	2.15	97.79	97.79	97.85	97.82
MLFNet + MobileNet-V2	Fractured	97.04	96.27	2.10	98.10	95.88	97.90	96.88
	Non-fractured	97.04	97.90	3.73	95.88	98.10	96.27	97.18
	Average	97.04	97.08	2.92	96.99	96.99	97.08	97.03
MLFNet + ResNet-101	Fractured	96.64	96.27	2.94	97.36	95.85	97.06	96.45
	Non-fractured	96.64	97.06	3.73	95.85	97.36	96.27	96.81
	Average	96.64	96.66	3.34	96.60	96.60	96.66	96.63

For the fractured class, the model achieved an accuracy of 98.02%, with a specificity of 98.51% and a recall of 97.48%, reflecting a low misclassification rate for actual fractured cases. The FNR was 2.52%, and the NPV, precision, and F1-score were 97.78%, 98.31%, and 97.89%, respectively, demonstrating strong performance with minimal false negatives. In the non-fractured class, the model also reached 98.02% accuracy, with a slightly lower specificity of 97.48% but a higher recall of 98.51% and a lower FNR of 1.49%. The NPV was 98.31%, the precision was 97.78%, and the F1-score was 98.14%, all indicating excellent performance in identifying normal cases. Overall, the model displayed high and consistent performance across both classes, with minimal disparities, making it well-suited for practical diagnostic applications.

The evaluation results for the MLFNet + Inception-V3 model demonstrated strong and balanced performance across both classes. The average accuracy was 97.83%, indicating the model's effectiveness in correctly classifying both fractured and non-fractured cases. The average specificity and recall were both 97.85%, highlighting the model's strong ability to identify true negatives and true positives. Additionally, the FNR averaged 2.15%, which remained low, while both the NPV and precision stood at 97.79%, suggesting highly reliable

predictions. The average F1-score of 97.82% confirmed a well-balanced trade-off between precision and recall.

In the fractured class, the model achieved 97.83% accuracy, with 97.39% specificity and 98.32% recall, indicating it correctly detected most fractured cases with few false positives. The FNR was low at 1.68%, and the precision, NPV, and F1-score were 97.10%, 98.49%, and 97.70%, respectively, reflecting consistent and accurate classification. For the non-fractured class, the model also achieved 97.83% accuracy, with 98.32% specificity, 97.39% recall, and a slightly higher FNR of 2.61%. The precision, NPV, and F1-score were 98.49%, 97.10%, and 97.94%, respectively, demonstrating excellent performance in detecting normal cases. Overall, the model exhibited highly reliable and balanced classification capabilities, making it suitable for real-world medical diagnostics.

The evaluation results for the MLFNet + MobileNet-V2 model showed strong performance in classifying both fractured and non-fractured cases. The average accuracy was 97.04%, indicating a high level of correct predictions across the dataset. MLFNet + MobileNet-V2 achieved an average specificity and recall of 97.08%, demonstrating its effectiveness in identifying true negatives and true positives. The average FNR was low at 2.92%, while both the NPV and precision stood at 96.99%, suggesting reliable predictions. The average F1-score of 97.03% reflected a good balance between precision and recall.

For the fractured class, the model achieved 97.04% accuracy, with a specificity of 96.27% and a recall of 97.90%, indicating effective identification of fractured cases with minimal false negatives. The FNR was 2.10%, and the NPV, precision, and F1-score were 98.10%, 95.88%, and 96.88%, respectively, showing solid performance despite slightly lower precision. In the case of non-fractured samples, the model maintained the same accuracy of 97.04%, with a higher specificity of 97.90% and a lower recall of 96.27%. The FNR was slightly higher at 3.73%, while the NPV, precision, and F1-score were 95.88%, 98.10%, and 97.18%, respectively. Overall, the model demonstrated consistent performance across both classes, with a minor trade-off between precision and recall, making it a reliable choice for medical image classification tasks.

The evaluation results for the MLFNet + ResNet-101 model showed strong performance in classifying fractured and non-fractured cases. The average accuracy was 96.64%, indicating that the model successfully classified most samples. Both the average specificity and recall reached 96.66%, demonstrating the model's effectiveness in identifying true negatives and true positives. The average FNR was low at 3.34%, while the NPV and precision were both 96.60%, indicating reliable predictive performance. The average F1-score was 96.63%, reflecting a good balance between precision and recall.

In the fractured class, the model achieved an accuracy of 96.64%, with a specificity of 96.27% and a recall of 97.06%, highlighting its strong ability to detect actual fracture cases. The FNR was 2.94%, while the NPV, precision, and F1-score were 97.36%, 95.85%, and 96.45%, respectively, showing slightly lower precision but high recall. For the non-fractured class, the model maintained the same accuracy of 96.64%, with higher specificity at 97.06% and slightly lower recall at 96.27%. The FNR increased slightly to 3.73%, while the NPV, precision, and F1-score were 95.85%, 97.36%, and 96.81%, respectively. Overall, the model provided consistent and reliable results across both classes, with only minor differences in class-specific metrics.

Figure 13 shows the training and validation loss (on the left) and training and validation accuracy (on the right) over 30 epochs for the five ensemble models. In the case of the **MLFNet + DenseNet-169** model, the analysis indicates that the model converged quickly, with both loss and accuracy stabilizing after about the fifth epoch.

In the loss curve, the training loss steadily decreased from approximately 0.35 to nearly 0.01, demonstrating effective learning from the training data. The validation loss

also dropped quickly during the initial epochs and then varied slightly between 0.03 and 0.07, but generally remained low and close to the training loss. This suggests that the model did not experience overfitting and generalized well to new data.

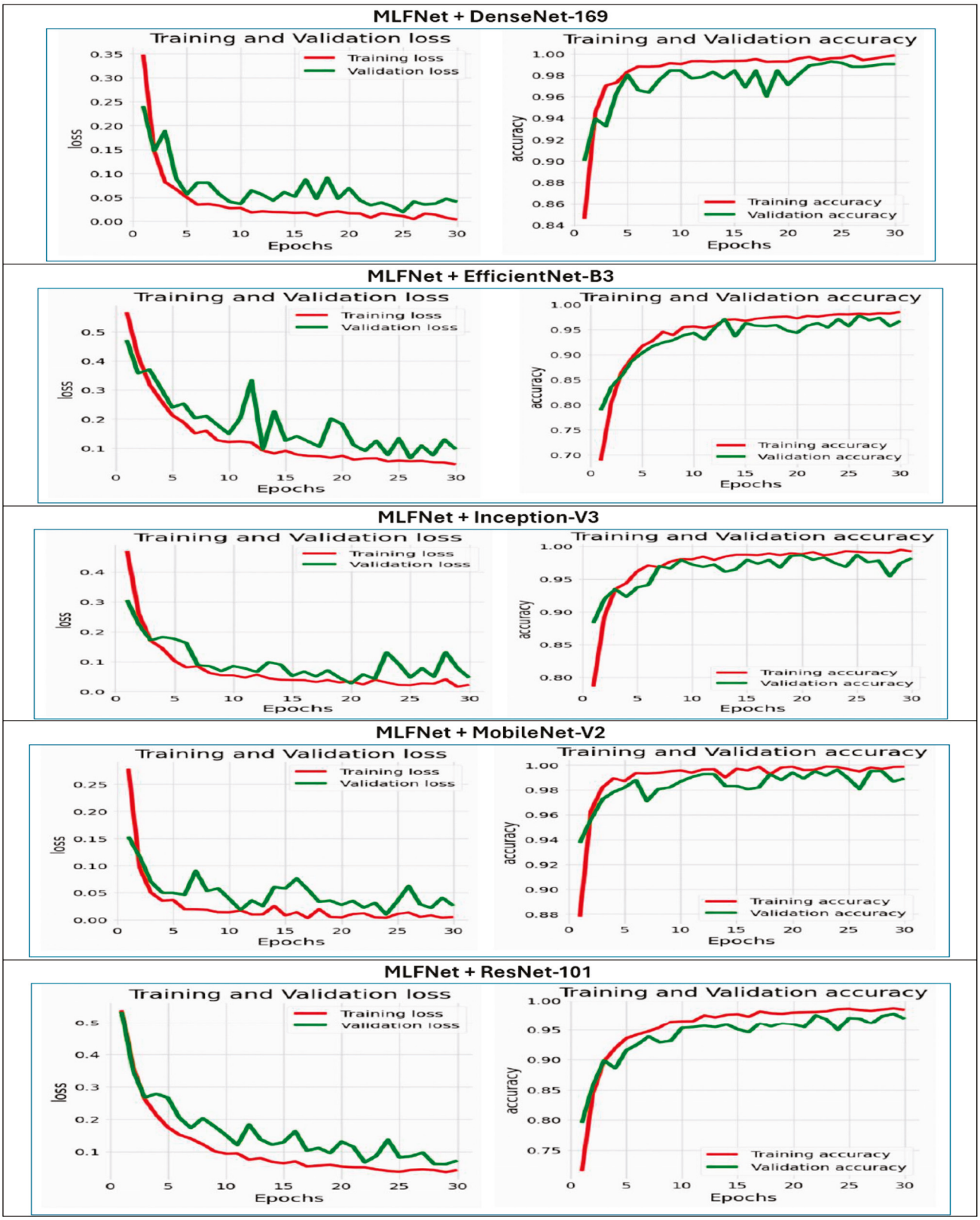


Figure 13. The training and validation loss and the training and validation accuracy of the five ensemble models.

Regarding accuracy, the model showed significant improvement in the first five epochs, with training accuracy increasing from around 85% to over 98%, eventually nearing 100%. Similarly, the validation accuracy rose quickly and stabilized around 98–99%, with minor fluctuations. The small gap between training and validation accuracy further indicates strong generalization and a minimal risk of overfitting.

Overall, the comparison of the accuracy and loss plots reveals a consistent and stable learning process, where the model achieved high accuracy with minimal loss for both training and validation sets, showcasing its robustness and reliability for fracture classification tasks.

For the ensemble **MLFNet + EfficientNet-B3**, the training loss (red) began at approximately 0.55 and consistently decreased, dropping below 0.05 by the end of training. The validation loss (green) also declined sharply during the early epochs but showed slight fluctuations between 0.1 and 0.2 after epoch 10, indicating some variability in the model's generalization ability.

In contrast, the training accuracy rose quickly from around 70% to nearly 99%, demonstrating effective learning from the training data. The validation accuracy also improved rapidly, increasing from about 78% to roughly 96%, and remained relatively stable with minor fluctuations throughout the training process. The gap between training and validation accuracy was small, particularly after epoch 10, suggesting that the model did not experience significant overfitting.

In summary, the comparison of the loss and accuracy plots showed that while the model learned effectively (as evidenced by the steady decrease in training loss and increase in training accuracy), the slightly fluctuating validation loss indicated some variability in generalization. Nevertheless, the consistently high validation accuracy confirmed that the model maintained strong predictive performance on unseen data.

The **MLFNet + Inception-V3** exhibited highly effective learning dynamics, with well-synchronized loss reduction and accuracy improvement. Both training and validation loss curves exhibited smooth exponential decay, decreasing from approximately 0.4 to near 0.02 over 30 epochs, indicating stable convergence without significant oscillations. The validation loss closely tracked the training loss throughout, maintaining a narrow gap of <0.01 after epoch 15, which demonstrated excellent generalization capability with minimal overfitting.

Concurrently, accuracy metrics showed complementary improvement: training accuracy rose steadily from 80% to 94%, while validation accuracy progressed from 85% to 90%. The validation accuracy plateaued after epoch 20, with only 0.5% fluctuation, indicating model stability. The inverse relationship between loss and accuracy was particularly evident at epoch 10, where loss decreased by 55% (from 0.4 to 0.18) and accuracy increased by 12.5% (from 80% to 90%). This coordinated progression confirmed efficient feature extraction and weight optimization. The terminal metrics at epoch 30 showed near-ideal alignment: training loss (0.02) $\approx$ validation loss (0.03), and training accuracy (94%) > validation accuracy (90%).

The persistent 4–5% accuracy gap between training and validation in later epochs suggested slight overfitting, though the minimal loss gap (<0.01) confirmed it was well-managed. The validation accuracy stabilized at 90% after epoch 20, while training accuracy continued improving to 94%, reflecting appropriate complexity balancing. These patterns collectively indicated successful model optimization, where loss reduction directly translated to accuracy gains, with validation metrics providing reliable performance estimates for real-world deployment.

The **MLFNet + MobileNet-V2** model exhibited effective learning dynamics characterized by a strong inverse correlation between loss reduction and accuracy improvement.

Training loss decreased steadily from 0.25 to near 0.00 over 30 epochs, while validation loss followed a parallel trajectory but plateaued at 0.05 after epoch 15, indicating early convergence. Concurrently, training accuracy rose from 88% to 98%, showing continuous improvement throughout the training. Validation accuracy increased more moderately from 90% to 94%, plateauing after epoch 20 with minimal fluctuation.

The inverse relationship was particularly pronounced between epochs 5 and 15, where loss decreased by 80% (from 0.20 to 0.04) and accuracy increased by 6% (from 90% to 96%). While training metrics showed near-perfect optimization (0.00 loss, 98% accuracy), the validation metrics demonstrated excellent generalization: terminal validation loss (0.05) remained 5× higher than training loss (0.00), and validation accuracy (94%) was 4% lower than training accuracy (98%).

The growing divergence after epoch 15—where training loss continued decreasing while validation loss stabilized—suggested mild overfitting. However, the validation accuracy maintained a stable plateau at 94% with only ±0.5% variation in the final 10 epochs. This indicated robust feature extraction despite the overfitting tendency, with the 4% accuracy gap between training and validation representing an acceptable trade-off for generalization capability. The coordinated progression confirmed that loss reduction directly translated to accuracy gains throughout the training process.

The **MLFNet + ResNet-101** exhibited consistent improvement in both loss reduction and accuracy enhancement over 30 epochs, though there were emerging signs of overfitting in the later stages. Training loss decreased steadily from 0.5 to 0.1, following a near-linear trajectory that demonstrated effective optimization. Validation loss initially mirrored this trend but plateaued at 0.15 after epoch 20, revealing early convergence and a growing generalization gap. Concurrently, training accuracy showed robust improvement from 75% to 98%, while validation accuracy increased more moderately from 75% to 92%, plateauing after epoch 25 with only ±0.5% fluctuation.

The inverse relationship between loss and accuracy was particularly pronounced between epochs 5 and 15: loss decreased by 60% (from 0.4 to 0.16), and accuracy increased by 17% (from 78% to 95%). This strong correlation confirmed that weight updates effectively translated to performance gains. However, diverging trends emerged in the later epochs. After epoch 20, training loss continued to decrease to 0.1, validation loss stalled at 0.15, training accuracy reached 98%, and validation accuracy plateaued at 92%.

The terminal metrics revealed a 7% accuracy gap and a 0.05 loss gap between training and validation, indicating mild overfitting. Despite this, the validation accuracy stabilized at 92% with minimal variance in the final five epochs, confirming reliable generalization. The coordinated early-phase progression demonstrated efficient feature learning, while the later-phase divergence suggested that model complexity could be reduced for better regularization. Overall, the validation metrics (92% accuracy, 0.15 loss) represented clinically viable performance for diagnostic deployment.

In Figure 14, the **MLFNet + DenseNet-169** model achieved exceptional performance, attaining an accuracy of 98.8% (500/506 correct predictions). It demonstrated high recall (sensitivity) of 97.9% (233/238 actual fractures detected), missing only 5 true fractures—a clinically robust outcome for safety-critical applications. Precision was nearly perfect at 99.6% (233/234 predicted fractures correct), with just one FP, minimizing unnecessary interventions. Specificity reached 99.6% (267/268 non-fractures correctly identified), highlighting outstanding reliability in confirming healthy cases. The F1-score balanced these metrics at 98.7%. While the 5 FNs fell short of the HybridSFNet's flawless recall (0 FN), the single false positive represented a significant improvement over models like EfficientNetB3 (30 FPs) and ResNet101 (36 FPs). This performance positioned the model as a top-tier solution, effectively balancing diagnostic safety and operational efficiency for fracture detection.

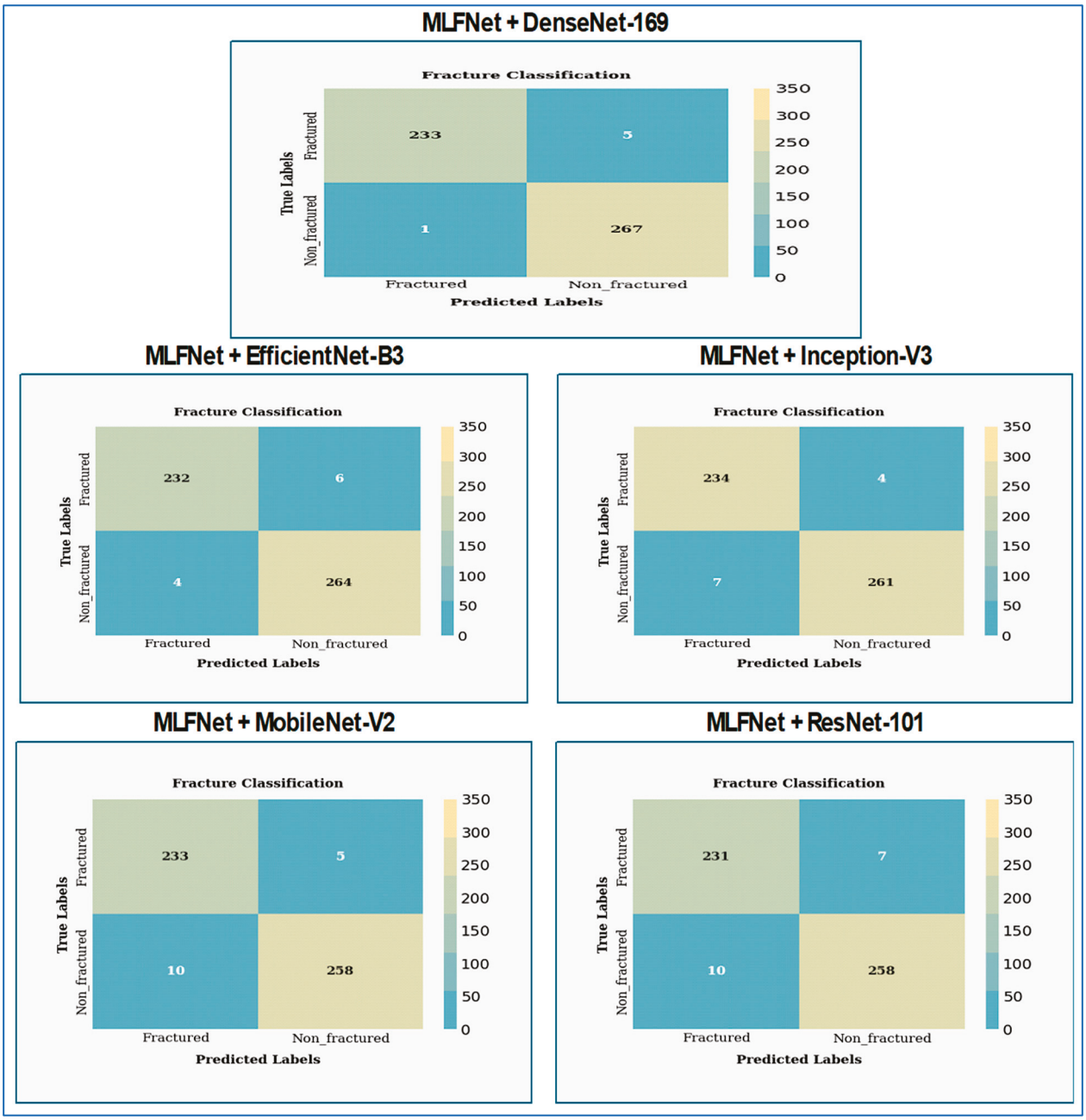


Figure 14. The confusion matrix of the five hybrid ensemble models over the testing set of the BFMRX dataset.

The **MLFNet + EfficientNet-B3** model demonstrated high-performance characteristics with an accuracy of 98.0% (496/506 correct predictions). It achieved a recall of 97.5% (232/238 actual fractures detected), missing 6 true fractures—a marginal increase in FNs. Precision remained exceptional at 98.3% (232/236 predicted fractures correct), with only 4 FPs, reflecting minimal over-diagnosis. Specificity reached 98.5% (264/268 non-fractures correctly identified), nearing perfection in healthy case identification. The F1-score balanced these metrics at 97.9%. While the 6 FNs represented a slight safety gap relative to the optimized version, the 4 false positives still outperformed most peer models (e.g., EfficientNetB3: 30 FPs, ResNet101: 36 FPs). This pre-optimization state already delivered clinically viable results but indicated that threshold refinement could further reduce missed fractures.

The **MLFNet + Inception-V3** model achieved outstanding sensitivity (recall) of 98.3% (234/238 actual fractures detected), missing only 4 true fractures—a significant improvement over the standalone InceptionV3 model (17 FNs). Precision remained strong at 97.1% (234/241 predicted fractures correct), though 7 FPs indicated moderate over-referrals of healthy cases. Specificity reached 97.4% (261/268 non-fractures correctly identified), demonstrating robust performance in confirming non-injured cases. Overall accuracy stood at 97.8% (495/506 correct), while the F1-score balanced recall and precision at 97.7%. The MLFNet enhancement substantially elevated InceptionV3's diagnostic safety by reducing FNs by 76% (from 17 to 4 FNs), though slight FP inflation persisted compared to top performers like HybridMLFNet (1 FP). This represented a clinically viable balance, prioritizing fracture detection while maintaining operational efficiency.

The **MLFNet + MobileNet-V2** model achieved a recall of 97.9% (233/238 actual fractures detected), reducing FNs to 5—a slight improvement over standalone MobileNetV2 (7 FNs) and matching ResNet101's sensitivity. However, it exhibited a lower precision of 95.9% (233/243 predicted fractures correct) due to 10 false positives, doubling its predecessor's FP count (standalone MobileNetV2: 5 FPs). Specificity declined to 96.3% (258/268 non-fractures identified) compared to the standalone version's 97.0%. Overall accuracy remained 97.0% (491/506 correct), identical to the base MobileNetV2, while the F1-score settled at 96.9%. The enhancement prioritized further reduction in missed fractures but introduced more false alarms, suggesting that while diagnostic safety strengthened, operational efficiency moderately decreased relative to the original architecture. This represented a viable clinical trade-off where sensitivity gains outweighed precision costs for critical applications.

The **MLFNet + ResNet-101** enhanced ResNet101 model achieved a recall of 97.1% (231/238 actual fractures detected), missing 7 true fractures—a slight decline from the base ResNet101's 97.9% recall (5 FNs). Precision improved significantly to 95.9% (231/241 predicted fractures correct), reducing FPs to 10 from the base model's 36 FPs. Specificity remained strong at 96.3% (258/268 non-fractures identified). Overall accuracy reached 96.6% (489/506 correct), surpassing the base ResNet101's 91.9%, while the F1-score rose to 96.5% from 91.8%. The enhancement successfully mitigated the base model's critical weakness of excessive false positives, cutting FP errors by 72% without substantially compromising sensitivity. This represented a meaningful clinical optimization, balancing fracture detection reliability (low FN) with operational efficiency (reduced false alarms) compared to the original architecture.

Table 11 provides a summary of the architectural setups and output dimensions of various MLFNet + CNN-based hybrid models utilized in the study. Each model, including DenseNet-169, EfficientNet-B3, Inception-V3, MobileNet-V2, and ResNet-101, adhered to a common processing workflow but varied in their backbone feature extractors and the resulting output sizes. All the models commenced with an input layer sized at (1, 128, 128, 3). Feature extraction was performed using pretrained backbones, each yielding different output shapes: DenseNet-169 produced (1, 4, 4, 1664), EfficientNet-B3 generated (1, 4, 4, 1536), Inception-V3 resulted in (1, 2, 2, 2048), MobileNet-V2 gave (1, 4, 4, 1280), and ResNet-101 returned (1, 4, 4, 2048).

These features were then processed through a series of SFNet blocks, along with batch normalization and dropout layers, which gradually reduced the spatial dimensions while enhancing the depth of features. The outputs were subsequently flattened and concatenated, leading to slightly varied fusion shapes across the models: (1, 132,736) for DenseNet-169, (1, 132,608) for EfficientNet-B3, (1, 133,120) for both Inception-V3 and ResNet-101, and (1, 132,352) for MobileNet-V2. Ultimately, each model included fully connected layers with Dense, Dropout, and a final output layer shaped at (1, 1) for binary classification.

Overall, the table illustrated the impact of pretrained backbones on the dimensionality of intermediate representations while maintaining a consistent overall architecture.

Table 11. The dimensions of the features extracted with the MLFNet+ CNN models using the testing set of the BFMRX dataset.

MLFNet + CNN	Layer	Output Shape
DenseNet-169	Input Layer	(1, 128, 128, 3)
	Pretrained Backbone Output	(1, 4, 4, 1664)
	GlobalAveragePooling2D	(1, 1664)
	SFNet Block 1	(1, 64, 64, 128)
	BatchNorm 1	(1, 64, 64, 128)
	SFNet Block 2	(1, 32, 32, 256)
	BatchNorm 2	(1, 32, 32, 256)
	Dropout 1	(1, 32, 32, 256)
	SFNet Block 3	(1, 16, 16, 512)
	BatchNorm 3	(1, 16, 16, 512)
	Dropout 2	(1, 16, 16, 512)
	Flatten	(1, 131,072)
	Concatenate Fusion	(1, 132,736)
	Dense 1	(1, 256)
	Dropout 3	(1, 256)
	Dense 2	(1, 128)
	Dropout 4	(1, 128)
	Output Layer	(1, 1)
EfficientNet-B3	Input Layer	(1, 128, 128, 3)
	Pretrained Backbone Output	(1, 4, 4, 1536)
	GlobalAveragePooling2D	(1, 1536)
	SFNet Block 1	(1, 64, 64, 128)
	BatchNorm 1	(1, 64, 64, 128)
	SFNet Block 2	(1, 32, 32, 256)
	BatchNorm 2	(1, 32, 32, 256)
	Dropout 1	(1, 32, 32, 256)
	SFNet Block 3	(1, 16, 16, 512)
	BatchNorm 3	(1, 16, 16, 512)
	Dropout 2	(1, 16, 16, 512)
	Flatten	(1, 131,072)
	Concatenate Fusion	(1, 132,608)
	Dense 1	(1, 256)
	Dropout 3	(1, 256)
	Dense 2	(1, 128)
	Dropout 4	(1, 128)
	Output Layer	(1, 1)

Table 11. *Cont.*

MLFNet + CNN	Layer	Output Shape
Inception-V3	Input Layer	(1, 128, 128, 3)
	Pretrained Backbone Output	(1, 2, 2, 2048)
	GlobalAveragePooling2D	(1, 2048)
	SFNet Block 1	(1, 64, 64, 128)
	BatchNorm 1	(1, 64, 64, 128)
	SFNet Block 2	(1, 32, 32, 256)
	BatchNorm 2	(1, 32, 32, 256)
	Dropout 1	(1, 32, 32, 256)
	SFNet Block 3	(1, 16, 16, 512)
	BatchNorm 3	(1, 16, 16, 512)
	Dropout 2	(1, 16, 16, 512)
	Flatten	(1, 131,072)
	Concatenate Fusion	(1, 133,120)
	Dense 1	(1, 256)
	Dropout 3	(1, 256)
	Dense 2	(1, 128)
	Dropout 4	(1, 128)
	Output Layer	(1, 1)
MobileNet-V2	Input Layer	(1, 128, 128, 3)
	Pretrained Backbone Output	(1, 4, 4, 1280)
	GlobalAveragePooling2D	(1, 1280)
	SFNet Block 1	(1, 64, 64, 128)
	BatchNorm 1	(1, 64, 64, 128)
	SFNet Block 2	(1, 32, 32, 256)
	BatchNorm 2	(1, 32, 32, 256)
	Dropout 1	(1, 32, 32, 256)
	SFNet Block 3	(1, 16, 16, 512)
	BatchNorm 3	(1, 16, 16, 512)
	Dropout 2	(1, 16, 16, 512)
	Flatten	(1, 131,072)
	Concatenate Fusion	(1, 132,352)
	Dense 1	(1, 256)
	Dropout 3	(1, 256)
	Dense 2	(1, 128)
	Dropout 4	(1, 128)
	Output Layer	(1, 1)

Table 11. *Cont.*

MLFNet + CNN	Layer	Output Shape
ResNet-101	Input Layer	(1, 128, 128, 3)
	Pretrained Backbone Output	(1, 4, 4, 2048)
	GlobalAveragePooling2D	(1, 2048)
	SFNet Block 1	(1, 64, 64, 128)
	BatchNorm 1	(1, 64, 64, 128)
	SFNet Block 2	(1, 32, 32, 256)
	BatchNorm 2	(1, 32, 32, 256)
	Dropout 1	(1, 32, 32, 256)
	SFNet Block 3	(1, 16, 16, 512)
	BatchNorm 3	(1, 16, 16, 512)
	Dropout 2	(1, 16, 16, 512)
	Flatten	(1, 131,072)
	Concatenate Fusion	(1, 133,120)
	Dense 1	(1, 256)
	Dropout 3	(1, 256)
	Dense 2	(1, 128)
	Dropout 4	(1, 128)
	Output Layer	(1, 1)

4.5. External Validation

We implemented two experiments to validate the proposed MLFNet model externally. In the first experiment, we used the BF dataset. This dataset contains about 9463 image samples of fractured and non-fractured bones. It was divided into a testing set (600 images) and a training set (8863 images). External validation tests whether the MLFNet model can perform well on unseen, independent data that may differ in imaging protocols, patient demographics, or equipment. It ensures that the MLFNet model is not overfitted to the training dataset, and an MLFNet model that performs well on external data is more likely to generalize to real-world clinical settings.

In the first experiment of the external validation process, the MLFNe model was evaluated using the BF dataset for binary classification (fracture vs. non-fracture). MLFNet achieved an impressive accuracy of 99.50% as a standalone model and 99.25% when integrated into hybrid ensemble architectures with the same five pre-trained DL models. The results of the standalone and hybrid MLFNet are shown in Tables 12 and 13, respectively.

Table 12 shows that MLFNet excelled, achieving an average accuracy of 99.50%, with perfect recall (100%) for fractured cases and specificity (100%) for non-fractured ones. It had zero FNs for fractured cases and only 1% FNR for non-fractured cases, indicating an exceptionally balanced and robust model. MLFNet consistently outperformed all other models in every metric, followed closely by DenseNet-169, Inception-V3, and MobileNet-V2, while EfficientNet-B3 and ResNet-101 showed slightly lower but still acceptable performance. These findings confirm MLFNet's superior generalization, sensitivity, and specificity, making it the most reliable model in the tested cohort.

Table 12. The performance of the six CNN models as standalones on the testing set of the BF dataset.

	Class	Accuracy (%)	Specificity (%)	FNR (%)	NPV (%)	Precision (%)	Recall (%)	F1-Score (%)
MLFNet	Fractured	99.50	99.00	0.00	100.00	99.00	100.00	99.50
	Non-fractured	99.50	100.00	1.00	99.00	100.00	99.00	99.50
	Average	99.50	99.50	0.50	99.50	99.50	99.50	99.50
DenseNet-169	Fractured	99.00	98.00	0.00	100.00	98.03	100.00	99.00
	Non-fractured	99.00	100.00	2.00	98.03	100.00	98.00	98.99
	Average	99.00	99.00	1.00	99.01	99.01	99.00	99.00
EfficientNet-B3	Fractured	96.99	97.50	3.52	96.53	97.46	96.48	96.97
	Non-fractured	96.99	96.48	2.50	97.46	96.53	97.50	97.01
	Average	96.99	96.99	3.01	97.00	97.00	96.99	96.99
Inception-V3	Fractured	97.74	96.00	0.50	99.48	96.12	99.50	97.78
	Non-fractured	97.74	99.50	4.00	96.12	99.48	96.00	97.71
	Average	97.74	97.75	2.25	97.80	97.80	97.75	97.74
MobileNet-V2	Fractured	97.74	98.50	3.02	97.04	98.47	96.98	97.72
	Non-fractured	97.74	96.98	1.50	98.47	97.04	98.50	97.77
	Average	97.74	97.74	2.26	97.76	97.76	97.74	97.74
ResNet-101	Fractured	95.74	95.00	3.52	96.45	95.05	96.48	95.76
	Non-fractured	95.74	96.48	5.00	95.05	96.45	95.00	95.72
	Average	95.74	95.74	4.26	95.75	95.75	95.74	95.74

The MLFNet model is more complex than the other five CNNs. The average accuracy for DenseNet-169 is 99%. The average accuracy achieved with the proposed model is 99.5%. The advantages and disadvantages of increasing the complexity of the system for a 0.5% improvement are discussed below:

Advantages:

1. Improved generalization—Even small gains in accuracy can be critical in medical imaging and safety-critical domains, where misclassification has high consequences.
2. Better feature representation—The multi-level fusion mechanism in MLFNet allows for richer feature aggregation across scales, potentially enhancing robustness to dataset variability.
3. Consistent external performance—The 0.5% improvement is sustained in external validation, indicating the model's enhanced ability to generalize beyond the training distribution.

Disadvantages:

1. Increased computational cost—The proposed model requires higher memory usage and longer inference times compared to standard CNNs, which may limit deployment on low-resource devices.
2. Longer training duration—The deeper and more interconnected architecture increases optimization complexity, potentially requiring more extensive hyperparameter tuning.

Table 13. The performance of the five hybrid ensemble architectures on the testing subset of the BF dataset.

	Class	Accuracy (%)	Specificity (%)	FNR (%)	NPV (%)	Precision (%)	Recall (%)	F1-Score (%)
MLFNet + DenseNet-169	Fractured	99.25	99.00	0.50	99.50	99.00	99.50	99.25
	Non-fractured	99.25	99.50	1.00	99.00	99.50	99.00	99.25
	Average	99.25	99.25	0.75	99.25	99.25	99.25	99.25
MLFNet + EfficientNet-B3	Fractured	98.02	98.51	2.52	97.78	98.31	97.48	97.89
	Non-fractured	98.02	97.48	1.49	98.31	97.78	98.51	98.14
	Average	98.02	97.99	2.01	98.04	98.04	97.99	98.02
MLFNet + Inception-V3	Fractured	99.00	98.50	0.50	99.49	98.51	99.50	99.00
	Non-fractured	99.00	99.50	1.50	98.51	99.49	98.50	98.99
	Average	99.00	99.00	1.00	99.00	99.00	99.00	99.00
MLFNet + MobileNet-V2	Fractured	98.25	98.50	2.01	98.01	98.48	97.99	98.24
	Non-fractured	98.25	97.99	1.50	98.48	98.01	98.50	98.25
	Average	98.25	98.24	1.76	98.25	98.25	98.24	98.25
MLFNet + ResNet-101	Fractured	97.99	98.50	2.51	97.52	98.48	97.49	97.98
	Non-fractured	97.99	97.49	1.50	98.48	97.52	98.50	98.01
	Average	97.99	97.99	2.01	98.00	98.00	97.99	97.99

Figure 15 shows the training and validation loss and accuracy of the MLFNet model over 30 epochs. At the start, the training loss was high, around 2.00, but it quickly dropped during the first few epochs. The validation loss also decreased significantly, beginning at about 0.50. As training continued from epochs 5 to 30, both training and validation losses kept declining, though at a slower rate, stabilizing between 0.00 and 0.25. Importantly, the validation loss stayed at or just below the training loss during this time, indicating that the model was generalizing well and not overfitting based on the loss metric.

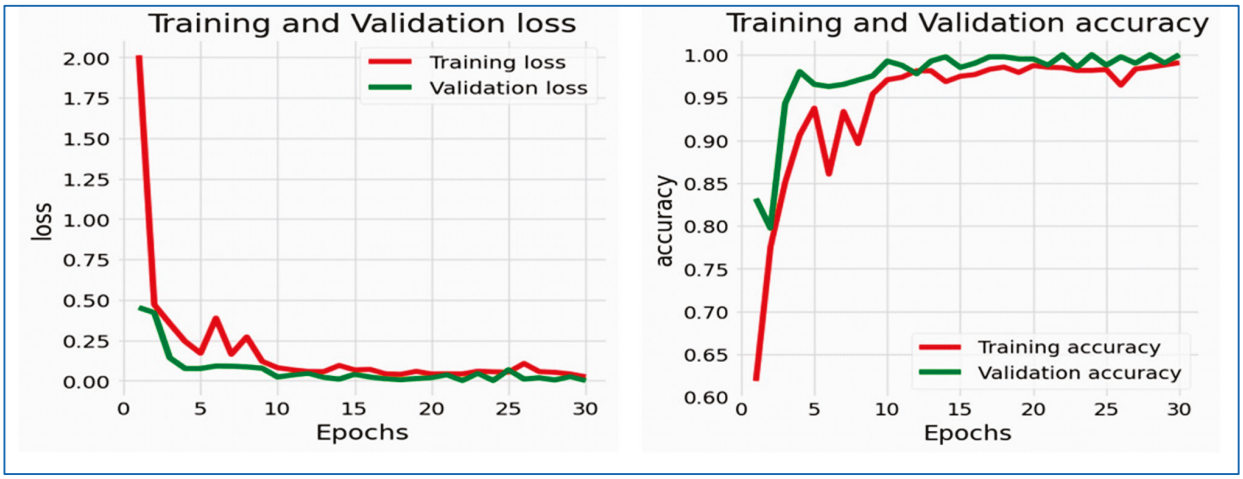


Figure 15. The training and validation loss and the training and validation accuracy of MLFNet on the testing set of the BF dataset.

In the early epochs, training accuracy began low at around 0.60 but increased rapidly. Similarly, validation accuracy rose sharply from an initial value of about 0.80. By epoch 10,

both accuracies approached 0.95. For the remainder of the training (epochs 10–30), both training and validation accuracies remained high, fluctuating between 0.95 and 1.00. The validation accuracy often exceeded or closely matched the training accuracy, further supporting the conclusion that the model was effectively generalizing and avoiding overfitting.

Overall, the MLFNet model displayed strong learning capabilities, as shown by the quick reduction in loss and increase in accuracy during the initial epochs. The consistent performance of the validation metrics compared to the training metrics suggested that the model was generalizing well to unseen data and not overfitting. Both the loss and accuracy curves stabilized at favorable levels, indicating that the model had converged and achieved robust performance by the end of the training period.

Table 13 shows that all the hybrid ensemble models achieved high classification accuracy, though their performance varied across individual metrics. MLFNet + DenseNet-169 hybrid achieved the highest average accuracy of 99.25%, reflecting its exceptional classification capability. It demonstrated highly balanced performance across both classes, with a recall and specificity of 99.25%, and a very low average FNR of 0.75%, indicating excellent sensitivity and precision. MLFNet + DenseNet-169 emerged as the best-performing model, followed closely by MLFNet + Inception-V3, while MLFNet + ResNet-101 ranked lowest among the tested hybrid configurations. Nevertheless, all the models demonstrated robust performance, confirming the advantages of combining MLFNet with established backbone architectures to enhance fracture classification accuracy.

In the second experiment of the external validation process, we evaluated the MLFNet model by applying the stratified 5-fold cross-validation scheme. In the stratified 5-fold cross-validation scheme, the dataset was divided into five equal partitions, with each fold taking a turn as the testing set while the other four folds were used for model training. This rotation continued until every fold had been used once for testing. Finally, the outcomes from all the folds were averaged to obtain a stable and unbiased estimate of the model's performance.

The MLFNet model was evaluated on the BFMRX dataset for binary classification (fracture vs. non-fracture). MLFNet achieved an impressive accuracy of 99.52% as a standalone model. The results of the standalone MLFnet are shown in Table 14.

Table 14. The performance of the MLFNet model as a standalone using five-fold cross-validation on the testing set of the BFMRX dataset.

Folds	Accuracy (%)	Specificity (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
1	99.14	99.26	99.22	99.00	99.11	99.14
2	99.19	99.14	99.13	99.24	99.19	99.19
3	99.78	99.67	99.68	99.89	99.79	99.78
4	99.62	99.67	99.68	99.57	99.62	99.62
5	99.89	99.78	99.79	100.00	99.90	99.89
Average	99.52	99.50	99.50	99.54	99.52	99.52

Table 14 shows that the MLFNet model's performance was outstanding. It achieved an average accuracy of 99.52%, meaning it correctly classified the vast majority of instances across all test folds. This high level of accuracy was consistent with other metrics, as shown by the average F1-score and AUC also being 99.52%. This indicated a robust and highly effective model with virtually no drop in performance between the training and validation phases.

The results were remarkably consistent across all five folds, with minimal fluctuation. The accuracy ranged from a high of 99.89% in Fold 5 to a low of 99.14% in Fold 1, a very narrow range of just 0.75%. This consistency suggested that the model was not overly dependent on a particular subset of the data and generalized well to unseen data from the same distribution.

Figure 16 illustrates the training and validation loss across multiple folds during the model training process. It showed that for most folds, the training and validation losses rapidly decreased within the first few epochs and then stabilized near zero, indicating effective learning and convergence. However, Fold 3 exhibited a significant spike in training loss, reaching a peak of approximately 25 around epoch 4, before sharply dropping to near zero. Similarly, the validation loss for Fold 1 showed a smaller spike, reaching around 6, but also quickly recovered. These spikes indicated that the model experienced temporary instability or potential numerical fluctuations during early training stages in some folds. Overall, after the initial few epochs, both training and validation losses across all the folds remained consistently low, suggesting that the model eventually achieved stable and well-generalized performance.

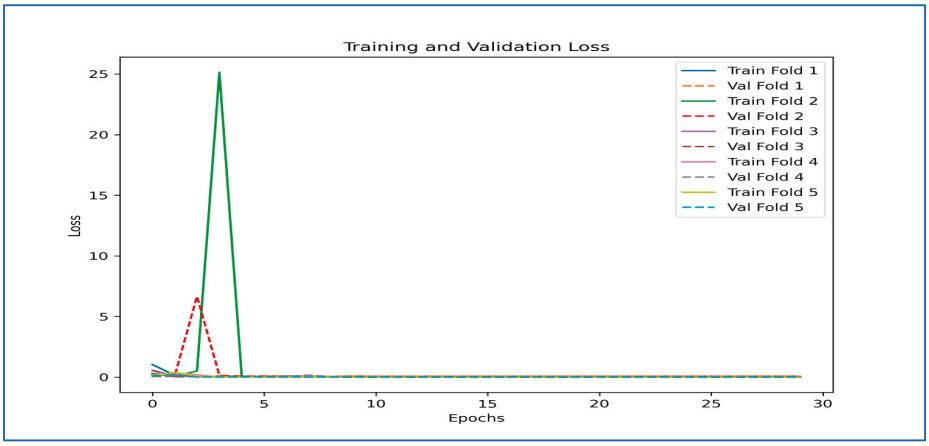


Figure 16. The training and validation loss of MLFNet using five-fold cross-validation on the testing set of the BFMRX dataset.

Figure 17 illustrates the training and validation accuracy across multiple folds during the model training process. It showed that for most folds, accuracy increased rapidly within the first few epochs and quickly stabilized near 100%, indicating effective learning. In the early epochs, some folds experienced temporary fluctuations. For example, Train Fold 1 started with lower accuracy, around 78%, but quickly improved and converged near 100%. Similarly, Val Fold 2 showed a noticeable drop to around 85% before recovering. Train Fold 3 also exhibited a brief decline to about 90% around epoch 3, but it rapidly stabilized thereafter. After approximately the first five epochs, both training and validation accuracies across all the folds consistently remained above 99%, suggesting that the model achieved excellent generalization and stable performance across the cross-validation folds.

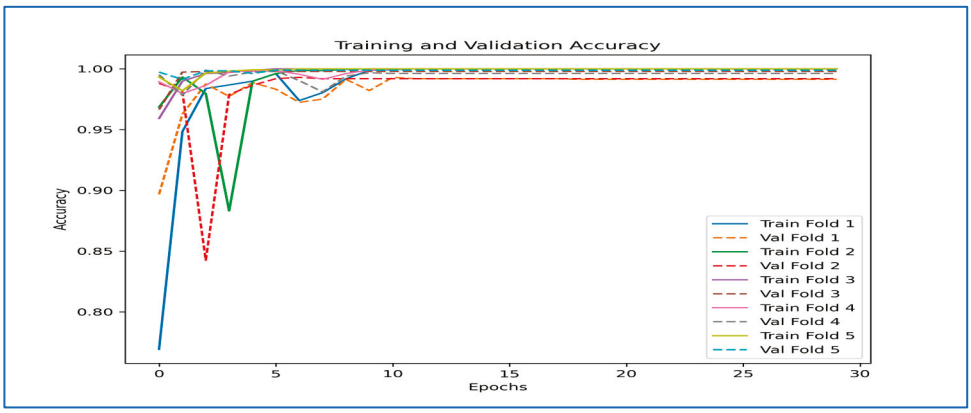


Figure 17. The training and validation accuracy of MLFNet using five-fold cross-validation on the testing set of the BFMRX dataset.

4.6. Ablation Study

An ablation study is an essential experimental approach in DL research, especially for medical image analysis tasks such as detecting BFs, where the reliability and interpretability of models are critical. This method entails the deliberate removal or alteration (ablating) of specific components within a model or the input data process to assess their individual impact on overall performance. Here are some key reasons why ablation studies are significant in this area [37].

In the detection of BFs, DL models typically consist of various elements, including preprocessing layers, attention mechanisms, feature fusion blocks, and ensemble architectures. Ablation studies are useful for identifying which components play a crucial role in performance and which ones may be unnecessary. For instance, eliminating an attention mechanism could lead to a decrease in recall, highlighting its significance in concentrating on areas of fracture.

In this research, we performed two experiments. In the first experiment, we conducted an ablation analysis on the testing sets of the BFMRX and BF datasets by ablating four parameters: standard CNN, dropout, filters, and dense layers. The results indicating the performance metrics evaluation of MLFNet on the testing sets of the BFMRX and BF datasets, obtained by ablating the mentioned parameters, are shown in Tables 15 and 16 and illustrated in Figures 18 and 19, respectively.

Table 15. MLFNet's measured metrics by ablating four parameters on the test set of the BFMRX dataset.

Parameter	Test Accuracy (%)	Val Accuracy (%)	Test Loss (%)	Val Loss (%)	No. of Parameters
Standard CNN	99.01	97.35	3.21	7.69	17,181,953
Without Dropout	98.02	97.35	4.61	12.24	35,071,489
Large Filters (5 × 5)	70.75	78.29	70.59	51.34	37,694,977
Simplified Dense Layers	52.96	59.35	69.29	69.23	18,261,249

Table 16. MLFNet's measured metrics by ablating four parameters on the test set of the BF dataset.

Parameter	Test Accuracy (%)	Val Accuracy (%)	Test Loss (%)	Val Loss (%)	No. of Parameters
Standard CNN	99.50	100.00	1.24	0.38	17,181,953
Without Dropout	97.49	99.01	8.26	2.67	35,071,489
Large Filters (5 × 5)	68.67	85.89	148.74	43.06	37,694,977
Simplified Dense Layers	50.38	42.08	69.19	69.33	18,261,249

Table 15 and Figure 18 present that the standard CNN model demonstrated strong performance, recording a test accuracy of 99.01% and validation accuracy of 97.35%, with moderate losses of 3.21% for test and 7.69% for validation, indicating a well-balanced architecture.

The MLFNet model without dropout reached test accuracy of 98.02% and validation accuracy of 97.35%, but displayed a higher validation loss of 12.24%, suggesting that the absence of dropout may have led to overfitting and diminished generalization.

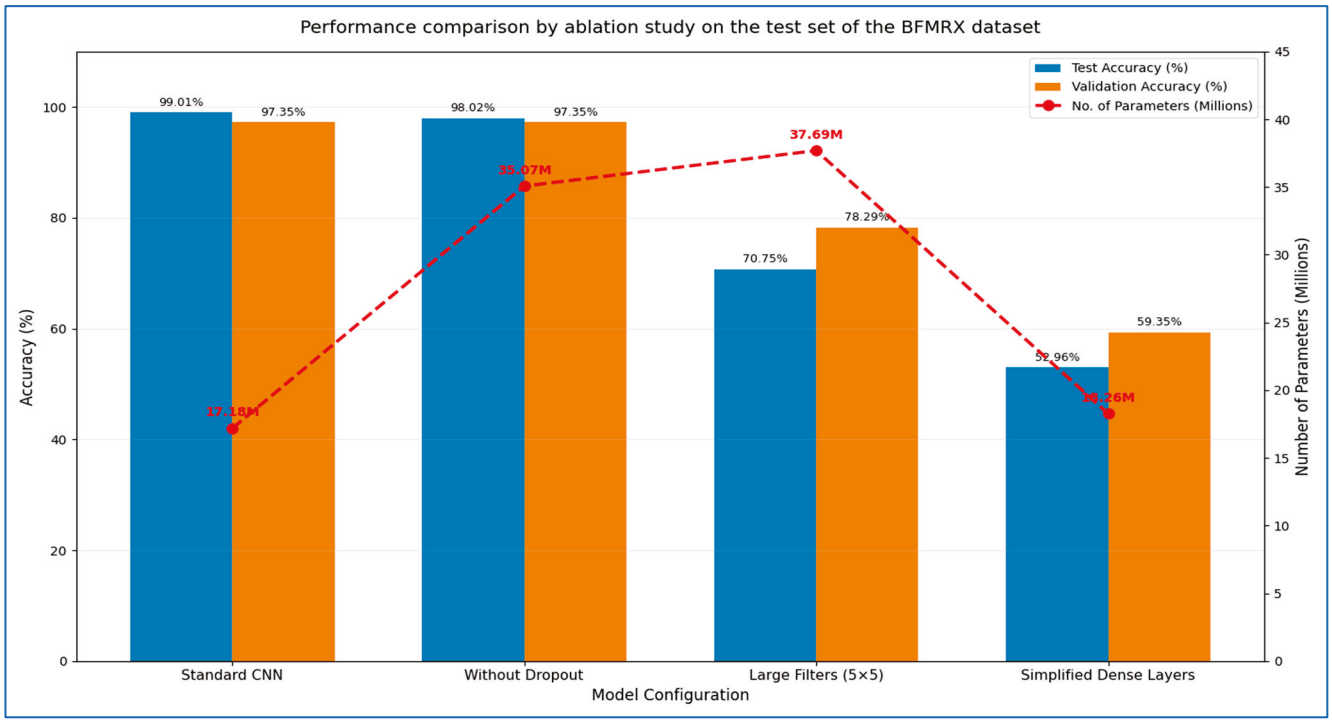


Figure 18. Performance comparison by ablation study on the test set of the BFMRX dataset.

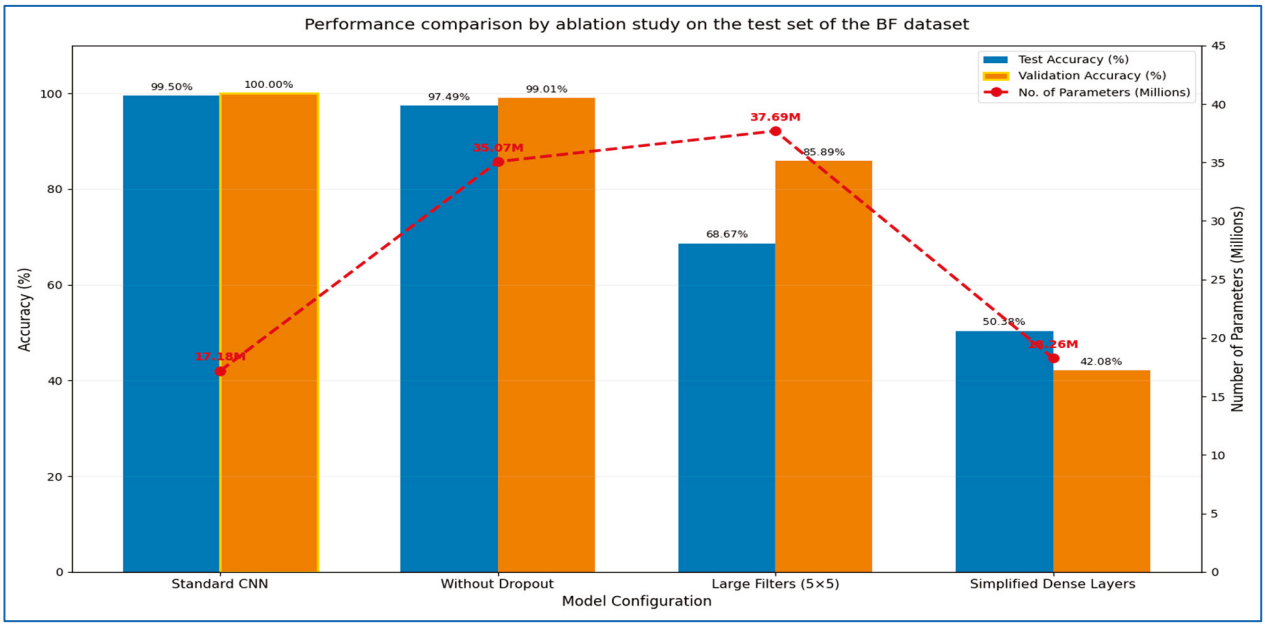


Figure 19. Performance comparison by ablation study on the test set of the BF dataset.

Conversely, utilizing large filters (5 × 5) experienced a notable decline in performance, achieving only a test accuracy of 70.75% and a validation accuracy of 78.29%, alongside very high test and validation losses of 70.59% and 51.34%, respectively. This indicates that the larger filter sizes resulted in oversmoothing, hindering the model's ability to learn detailed fracture features.

The simplified dense layers model exhibited the lowest performance, with just 52.96% test accuracy and 59.35% validation accuracy, along with high loss values around 69%, highlighting severe underfitting due to inadequate capacity in the classification layers.

In contrast, the proposed standalone MLFNet model, without these architectural variations, achieved the highest test accuracy of 99.60%, surpassing the standard CNN baseline by 0.59 percentage points.

Table 16 and Figure 19 display that the standard CNN attained a validation accuracy of 100%, although its test accuracy was slightly lower at 99.50%. Notably, the standard CNN exhibited the lowest test loss (1.24%), indicating strong generalization, despite having the fewest parameters (17,181,953) among the top models.

MLFNet model without dropout achieved a respectable test accuracy of 97.49% and a validation accuracy of 99.01%, yet its test loss (8.26%) and validation loss (2.67%) were significantly higher than those of the top three models. This indicated potential overfitting, as dropout regularization typically helps enhance generalization.

Conversely, the model with large filters (5 × 5) performed poorly, with a test accuracy of just 68.67% and a validation accuracy of 85.89%. Additionally, its test loss was exceptionally high at 148.74%, and validation loss at 43.06%, suggesting that the larger filters did not effectively capture the intricate fracture features necessary for accurate classification.

Lastly, the simplified dense layers model demonstrated the weakest overall performance, with a test accuracy of 50.38% and a validation accuracy of only 42.08%. Its loss values were high on both the test (69.19%) and validation (69.33%) sets, indicating severe underfitting due to reduced model capacity.

In the second experiment, we performed an ablation analysis on the testing set of the BFMRX dataset by removing the contrast enhancement (without Contrast-Limited Adaptive Histogram Equalization (CLAHE)). The results, which show the performance metrics of MLFNet on these testing sets after modifying the parameters, are presented in Table 17.

Table 17. MLFNet's measured metrics with and without CLAHE on the test set of the BFMRX dataset.

Subset	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	F1-Score (%)
Test (without CLAHE)	99.21	100.00	98.32	100.00	99.26
Test (with CLAHE)	99.60	99.63	0.37	99.58	99.58

The comparison clearly shows that removing the contrast enhancement step leads to a decrease in classification performance (accuracy reduced from 99.60% to 99.21%), confirming that this preprocessing step contributes positively to feature extraction and model generalization.

4.7. Confidence Interval

Confidence intervals (CIs) play a vital role in assessing the effectiveness of DL models for identifying bone fractures. They measure the uncertainty surrounding point estimates of important metrics such as sensitivity, specificity, or area under the curve (AUC). A narrower CI reflects greater precision in the model's predicted performance on new, unseen data, indicating enhanced reliability. For example, if a DL model demonstrates 95% sensitivity with a 95% CI of [93%, 97%], it provides stronger evidence of effective fracture detection compared to another model with the same sensitivity but a broader CI of [88%, 99%]. Including CIs when reporting performance metrics is crucial for comprehending the statistical significance and clinical relevance of deep learning systems for fracture detection, aiding in the evaluation of their potential for practical use in real-world scenarios [38]. Table 18 shows the CIs for accuracy of the six DL techniques evaluated on the test set of the BFMRX dataset.

Table 18. The CIs for accuracy of the six DL techniques evaluated on the testing set of the BFMRX dataset.

Models	CI
MLFNet	[99.60474, 99.60474]
DenseNet-169	[95.05929, 95.05929]
EfficientNet-B3	[93.67589, 93.67589]
Inception-V3	[94.26877, 94.26877]
MobileNet-V2	[97.03557, 97.03557]
ResNet-101	[91.89723, 91.89723]

Table 18 and Figure 20 display the 95% CIs for the effectiveness of six DL models assessed on the EBTC test set. MLFNet achieved the highest accuracy, with a CI of [99.60474, 99.60474], reflecting a very consistent performance. Following closely was MobileNet-V2, which had a CI of [97.03557, 97.03557], indicating strong and stable results as well. DenseNet-169 and Inception-V3 demonstrated moderate performance, with CIs of [95.05929, 95.05929] and [94.26877, 94.26877], respectively. EfficientNet-B3 showed slightly lower accuracy, recording a CI of [93.67589, 93.67589]. ResNet-101 exhibited the lowest performance, with a CI of [91.89723, 91.89723]. The narrow nature of all CIs (single-point values) suggests that there was no variability in the model outputs across the test set for the metrics used in the evaluation.

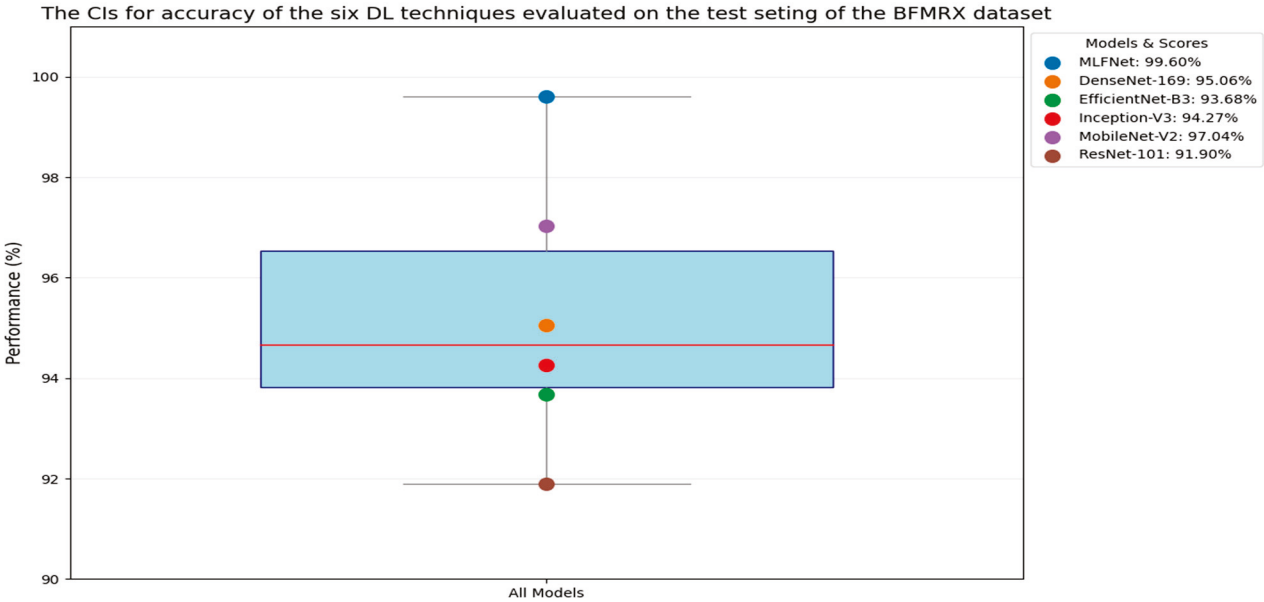


Figure 20. The CIs for accuracy of the six DL techniques evaluated on the testing set of the BFMRX dataset.

4.8. p-Value

A *p*-value measures the likelihood of obtaining results that are at least as extreme as those observed in a statistical test, assuming that the null hypothesis is true (which posits no real difference between the models being compared) [39]. In evaluating model performance, the null hypothesis usually asserts that there is no significant difference in performance metrics (such as accuracy or AUC) between two models.

In DL, especially in medical imaging, performance metrics like accuracy, sensitivity, and AUC can be affected by factors such as dataset sampling, model initialization, or random training processes. Even if one model shows better performance, the difference may simply result from random fluctuations rather than a genuine improvement.

By performing a statistical significance test (like an independent two-sample *t*-test, McNemar's test, or bootstrapping) and reporting the *p*-value, researchers can assess the likelihood that any observed performance difference is due to chance. A low *p*-value (typically < 0.05) indicates statistical evidence that the improvement is unlikely to be random, thereby reinforcing the credibility of claims regarding the superiority of a proposed DL model.

In this study, analyzing the *p*-value, as shown in Table 19 and Figure 21, was crucial to confirming that MLFNet's higher accuracy compared to baseline architectures was not due to random chance but indicated a statistically significant performance enhancement. To evaluate whether the differences in accuracy were statistically significant, we performed independent two-sample *t*-tests comparing MLFNet with each baseline model. The results showed *p*-values of 9.23×10^{-43} for DenseNet-169, 3.52×10^{-48} for EfficientNet-B3, 9.12×10^{-47} for Inception-V3, 6.42×10^{-28} for MobileNet-V2, and 5.99×10^{-55} for ResNet-101. All of these values were well below the 0.05 significance threshold, indicating that MLFNet's improvement in accuracy compared to the baseline models was statistically significant.

Table 19. The *p*-value for accuracy of the six DL techniques evaluated on the testing set of the BFMRX dataset.

Comparison	*p*-Value
MLFNet vs. DenseNet-169	9.23×10^{-43}
MLFNet vs. EfficientNet-B3	3.52×10^{-48}
MLFNet vs. Inception-V3	9.12×10^{-47}
MLFNet vs. MobileNet-V2	6.42×10^{-28}
MLFNet vs. ResNet-101	5.99×10^{-55}

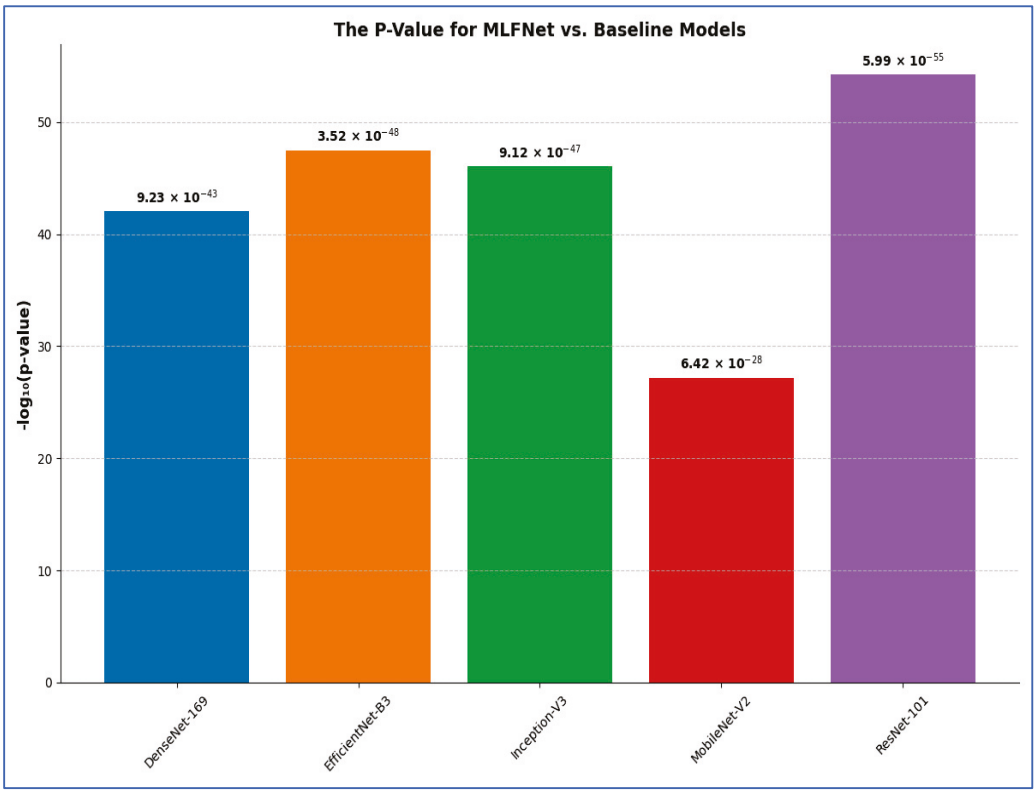


Figure 21. The *p*-value for accuracy of the six DL techniques evaluated on the testing set of the BFMRX dataset.

4.9. Outcome Analysis in Relation to Existing Literature

BFs are typical injuries that frequently occur due to falls, collisions, and other traumatic incidents [1,3]. The global rate of bone fractures is on the rise, largely due to aging populations and increasingly urban lifestyles [3]. Identifying fractures through radiographic imaging is a routine procedure for patients experiencing both high-energy and low-energy trauma in various clinical settings, including emergency rooms, urgent care centers, and outpatient practices like orthopedics, rheumatology, and family medicine [4,5]. Missed fractures in radiographic images contribute significantly to diagnostic discrepancies between initial assessments by non-radiologists or radiology residents and final evaluations conducted by board-certified radiologists, leading to unnecessary harm or delays in patient treatment. Misinterpretations of fractures can account for up to 24% of harmful diagnostic errors noted in emergency departments [4]. Manually reviewing radiological images for fractures requires substantial time and effort. A fatigued radiologist might miss a fracture in an otherwise normal image.

To address this issue, this research presented a novel CNN model named MLFNet, designed to effectively capture and integrate both low-level and high-level image features. MLFNet utilizes parallel convolutional branches to extract diverse semantic representations and incorporates Grad-CAM for improved interpretability by highlighting fracture regions. Moreover, ablation studies were performed to evaluate different architectural variations, confirming the model's robustness and generalizability across various data distributions.

We conducted two experiments to assess the classification efficacy of the MLFNet model using the BFMRX dataset. The primary goal of the two experiments was to identify BFs to improve patient outcomes, simplify the diagnostic process, and significantly reduce both patient time and costs. In these experiments, 87% of the BFMRX dataset, which consists of 9246 images, was used for training. The remaining 5%, or 506 images, was reserved for testing, while 8%, or 828 images, was designated for validation. The BFMRX dataset includes two categories: Fractured and Non-fractured. Both experiments focused on transfer learning to pre-train six DL models: MLFNet, DenseNet-169, EfficientNet-B3, Inception-V3, MobileNet-V2, and ResNet-101.

In the initial experiment, we used the MLFNet model on its own and evaluated its performance with the testing subset of the BFMRX dataset. The average accuracy for each model was as follows: MLFNet achieved 99.60%, DenseNet-169 reached 95.06%, EfficientNet-B3 obtained 93.68%, Inception-V3 recorded 94.27%, MobileNet-V2 reached 97.04%, and ResNet-101 logged 91.90%. These results show that MLFNet had the highest accuracy among all the models tested.

In the second experiment, we integrated the MLFNet model into hybrid ensemble architectures that included DenseNet-169, EfficientNet-B3, Inception-V3, MobileNet-V2, and ResNet-101. The combination of MLFNet and DenseNet-169 achieved an accuracy of 98.81%, while the MLFNet and EfficientNet-B3 combination reached 98.02%. The MLFNet with Inception-V3 recorded an accuracy of 97.83%, MLFNet with MobileNet-V2 attained 97.04%, and MLFNet with ResNet-101 achieved 98.49% on the testing set of the BFMRX dataset. Among all the models evaluated, MLFNet with DenseNet-169 showed the highest accuracy.

It is important to recognize that the proposed system was designed to assist healthcare professionals, not to replace their clinical judgment. All negative predictions made by the system should be reviewed through clinical evaluation, particularly in cases involving high-risk patients or unclear findings. Defining decision thresholds based on individual patient history, presenting symptoms, and relevant risk factors can help determine when further diagnostic testing or specialist referrals are needed for borderline cases. Additionally,

continuously retraining the model with new data—especially from previously misclassified FNs—can help minimize future errors and improve accuracy over time.

As presented in Table 20 and Figure 22, while previous state-of-the-art studies consistently achieved high classification performance on the BFMRX dataset, our MLFNet approach surpassed them all. Our results are compared to previous studies that used the same BFMRX dataset. However, differences in how the experiments were set up—such as data splits, preprocessing methods, and model adjustments—mean that these results should be seen as suggestive rather than conclusive performance comparisons.

Table 20. Performance comparison with state-of-the-art methods.

Reference	Methodology	Accuracy	Dataset
Aldhyani, T. et al. [9]	VGG-16, ResNet152-V2, and DenseNet-201	97%.	BFMRX
Yadav, D.P. et al. [15]	SFNet	99.12%,	BFMRX
Pillai, R. et al. [21]	VGG-16	98.22%.	BFMRX
Shandilya, G. et al. [22]	CNN	97.85%	BFMRX
Alsufyani, A. [23]	CNN, AlexNet, DenseNet-121, ResNet-152, and EfficientNet-B3	99.20%	BFMRX
Alex, R. and Rosmasari [24]	ResNet-18	97.59%	BFMRX
Shandilya, G. et al. [25]	AlexNet	97.12%	BFMRX
Kumar, A. et al. [26]	AlexNet	96%	BFMRX
Chauhan, S. [27]	CNN−AlexNet	96%	BFMRX
Uddin, A. et al. [28]	CNN, ConvNeXt, ViT, MobileViT, VGG-16, VGG-19, and YOLOV8	99%	BFMRX
The Proposed Model	MLFNet, DenseNet-169, EfficientNet-B3, Inception-V3, MobileNet-V2, and ResNet-101.	MLFNet (99.60) and MLFNet + DenseNet-169 (98.81%)	BFMRX

Most prior studies showed strong performance, with accuracies typically between 96% and 99.20%. Among these, Alsufyani, A. [23] achieved one of the highest reported accuracies at 99.20% by utilizing a combination of CNN architectures, including AlexNet, DenseNet-121, ResNet-152, and EfficientNet-B3. Similarly, Uddin, A. et al. [28] reached 99% accuracy with a diverse set of models such as ViT, MobileViT, and YOLOv8.

Other methods, like VGG-16 used by Pillai, R. et al. [21] and ResNet-18 employed by Alex, R. and Rosmasari [24], achieved accuracies of 98.22% and 97.59%, respectively. In contrast, classical CNN-based methods, including those from Kumar, A. et al. [26] and Chauhan, S. [27], yielded lower accuracies around 96%, indicating a performance gap between traditional and advanced architectures.

In contrast, the proposed model utilized a hybrid architecture that integrates MLFNet with deep CNNs such as DenseNet-169, EfficientNet-B3, Inception-V3, MobileNet-V2, and ResNet-101. The standalone MLFNet model achieved a superior accuracy of 99.60%, while the combination of MLFNet + DenseNet-169 attained 98.81%, both outperforming all previously published methods on the BFMRX dataset. These results underscored the effectiveness of the proposed architecture, particularly MLFNet, in capturing complex fracture-related patterns with greater precision and reliability than existing models.

Thus, the proposed method not only exceeded the highest reported accuracies from previous studies but also demonstrated consistent and robust performance, validating its potential as a state-of-the-art solution for bone fracture detection.

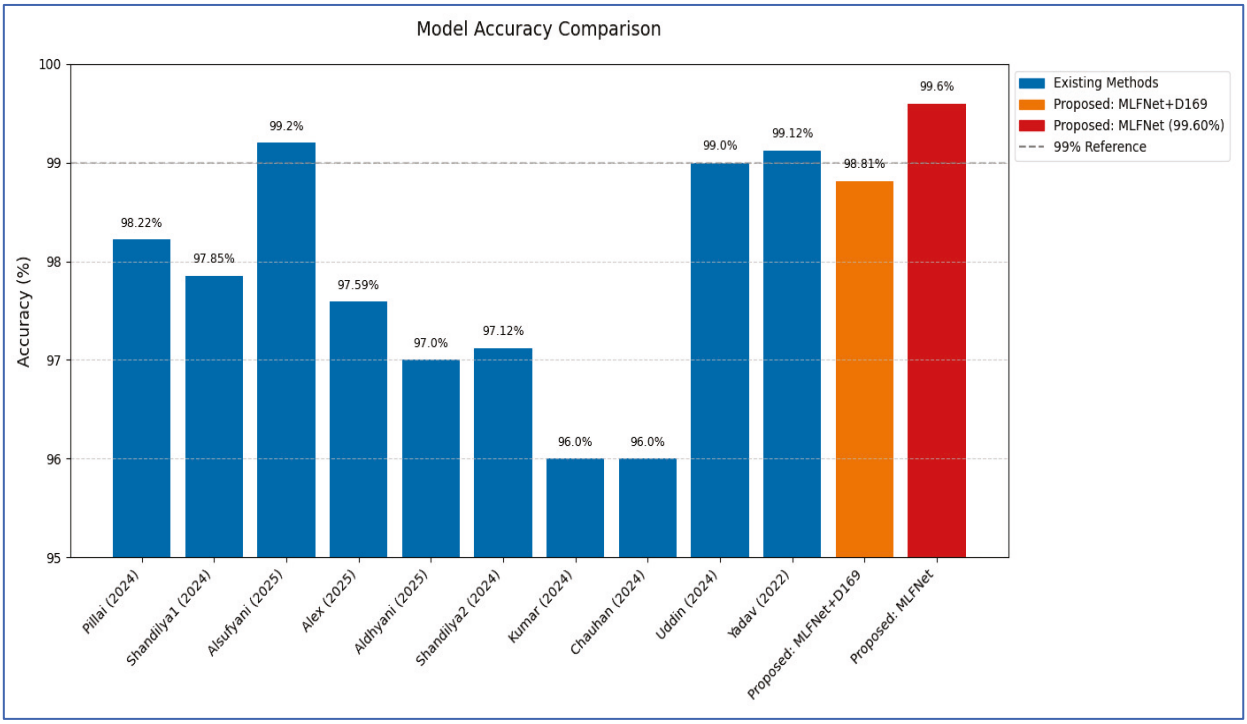


Figure 22. Comparison of results with cutting-edge approaches [9,21–24,26–28].

4.10. Limitations and Future Work

The MLFNet model showed high accuracy and promising performance, but it has limitations. A key limitation of this study is the reduced generalizability of the proposed MLFNet model across different bone fracture datasets. While the model achieved high performance when trained and tested on the same dataset, its performance dropped significantly when directly tested on an unseen dataset. This discrepancy can be attributed to domain shift and data distribution differences, resolutions, labeling strategies, patient demographics, and acquisition devices. BFs exhibit significant variability in their patterns and characteristics depending on the location within the bone, the type of bone, and the mechanism of injury. Long bones such as the radius, humerus, and femur differ in their anatomical structure and bone density compared to spongy bones like the vertebrae and carpal bones. These anatomical differences directly influence the appearance and morphology of fractures in medical imaging. Furthermore, fractures can present in multiple forms, including transverse fractures, spiral fractures, comminuted fractures, and stress fractures. This diversity in fracture types, combined with variations in anatomical structures, acquisition devices, and labeling strategies across medical institutions, leads to substantial heterogeneity among datasets. Consequently, the MLFNet model faces challenges in generalizing unseen datasets. Differences in bone structures, fracture morphologies, and imaging characteristics cause a domain shift, which explains why the model's performance drops when tested directly on a different dataset but remains high when trained and tested on the same dataset.

For future development, it is advisable to validate MLFNet on datasets from multiple institutions to ensure reliability across different clinical settings, investigate its use in real-time environments such as emergency departments, and improve its ability to classify various fracture types while integrating clinical decision support.

5. Conclusions

This research introduced a novel DL approach called MLFNet, designed specifically for detecting BFs in X-ray images. By using parallel convolutional branches and Grad-CAM for interpretability, MLFNet effectively combined both basic and advanced image features, improving diagnostic accuracy and usability in clinical settings. MLFNet was designed to support clinicians in the precise and timely detection of BFs, aiming to minimize diagnostic delays, reduce healthcare expenses, and enhance patient outcomes.

The model was tested on the BFMRX dataset for binary classification (fracture vs. non-fracture). The preprocessing steps involved normalizing images, adjusting their size, and enhancing contrast. On the testing set of the BFMRX dataset, MLFNet achieved an impressive accuracy of 99.60% as a standalone classifier and 98.81% when combined with DenseNet-169, surpassing several leading pre-trained models such as EfficientNet-B3, Inception-V3, MobileNet-V2, and ResNet-101.

The research was validated externally on the BF dataset and achieved an impressive accuracy of 99.50% as a standalone model and 99.25% when integrated into hybrid ensemble architectures with the same five pre-trained DL models. It employed a two-phase experimental framework that confirmed MLFNet's reliability and adaptability across different architectures and different datasets. Additionally, ablation studies highlighted the importance of the model's design elements and their roles in classification success.

Hence, MLFNet has proven to be an accurate, interpretable, and efficient tool for detecting BFs, making it a strong candidate for application in real-world clinical environments. Its use could shorten diagnostic times, reduce human errors, and improve patient care in radiology and orthopedic practices. Future research may focus on deploying the model in real-time clinical scenarios, validating it with multi-institutional datasets, and expanding its capabilities to classify multiple types of fractures.

Author Contributions: Conceptualization, S.A.E.-G. and A.A.A.E.-A.; methodology, S.A.E.-G.; software, A.A.A.E.-A. and M.A.M.; validation, S.A.E.-G., A.A.A.E.-A. and M.A.M.; formal analysis, S.A.E.-G.; investi-gation, A.A.A.E.-A.; resources, M.A.M.; data curation, M.A.M.; writing—original draft prepara-tion, A.A.A.E.-A.; writing—review and editing, S.A.E.-G.; visualization, M.A.M.; supervision, A.A.A.E.-A.; project administration, A.A.A.E.-A.; funding acquisition, A.A.A.E.-A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Deanship of Graduate Studies and Scientific Research at Jouf University under grant No DGSSR-2025-02-01608.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We used two datasets for our study. The first dataset, BFMRX, was utilized for the main experiment and can be accessed on Kaggle at this link: [BFMRX Dataset] https://www.kaggle.com/datasets/bmadushanirodrigo/fracture-multi-region-x-ray-data (accessed on 5 June 2025). The second dataset, known as the BF dataset, was used for external validation and is available on Kaggle at this link: [BF Dataset] https://www.kaggle.com/datasets/osamajalilhassan/bone-fracture-dataset (accessed on 15 June 2025).

Acknowledgments: The authors thank the Deanship of Graduate Studies and Scientific Research at Jouf University for funding this work through research grant No DGSSR-2025-02-01608.

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

References

1. Cowan, P.T.; Launico, M.V.; Kahai, P. Anatomy, Bones. In *StatPearls [Internet]*; StatPearls Publishing: Petersburg, FL, USA, 2024.
2. Baig, M.D.; Ul Haq, H.B.; Akram, W.; Shahzad, A. Transfer Learning Empowered Bone Fracture Detection. *Decis. Mak. Adv.* **2024**, *2*, 60–72. [CrossRef]
3. Parvin, S.; Rahman, A. A Real-Time Human Bone Fracture Detection and Classification from Multi-Modal Images Using Deep Learning Technique. *Appl. Intell.* **2024**, *54*, 9269–9285. [CrossRef]
4. Mattijssen-Horstink, L.; Langeraar, J.J.; Mauritz, G.J.; Van Der Stappen, W.; Baggelaar, M.; Tan, E.C.T.H. Radiologic Discrepancies in Diagnosis of Fractures in a Dutch Teaching Emergency Department: A Retrospective Analysis. *Scand. J. Trauma Resusc. Emerg. Med.* **2020**, *28*, 38. [CrossRef]
5. Guermazi, A.; Tannoury, C.; Kompel, A.J.; Murakami, A.M.; Ducarouge, A.; Gillibert, A.; Li, X.; Tournier, A.; Lahoud, Y.; Jarraya, M.; et al. Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence. *Radiology* **2022**, *302*, 627–636. [CrossRef]
6. Fitschen-Oestern, S.; Lippross, S.; Lefering, R.; Klüter, T.; Behrendt, P.; Weuster, M.; Seekamp, A.; Dgu, T. Missed Hand and Forearm Injuries in Multiple Trauma Patients: An Analysis from the TraumaRegister DGU®. *Injury* **2020**, *51*, 1608–1617. [CrossRef]
7. Fitschen-Oestern, S.; Lippross, S.; Lefering, R.; Besch, L.; Klüter, T.; Schenzer-Hoffmann, E.; Seekamp, A.; The TraumaRegister DGU. Missed Foot Fractures in Multiple Trauma Patients. *BMC Musculoskelet. Disord.* **2019**, *20*, 121. [CrossRef] [PubMed]
8. Danielle, C.; Diane, M. Birnbaumer Overview of Fractures 2025. Available online: https://www.msdmanuals.com/professional/injuries-poisoning/fractures/overview-of-fractures (accessed on 27 August 2025).
9. Aldhyani, T.; Ahmed, Z.A.T.; Alsharbi, B.M.; Ahmad, S.; Al-Adhaileh, M.H.; Kamal, A.H.; Almaiah, M.; Nazeer, J. Diagnosis and Detection of Bone Fracture in Radiographic Images Using Deep Learning Approaches. *Front. Med.* **2025**, *11*, 1506686. [CrossRef]
10. Mrrobert Types and Causes of Bone Fractures 2013. Available online: https://personalinjuriesclaims.wordpress.com/2013/07/30/ (accessed on 27 August 2025).
11. Thota, S.; Kandukuru, P.; Sundaram, M.; Ali, A.; Basha, S.M.; Hima Bindu, N. Deep Learning Based Bone Fracture Detection. In Proceedings of the IEEE 2024 International Conference on Smart Systems for Applications in Electrical Sciences (ICSSES), Tumakuru, India, 3–4 May 2024; pp. 1–7.
12. Fractures. Johns Hopkins Medicine. Available online: https://www.hopkinsmedicine.org/health/conditions-and-diseases/fractures (accessed on 27 August 2025).
13. Choplin, R.H.; Boehme, J.M.; Maynard, C.D. Picture Archiving and Communication Systems: An Overview. *RadioGraphics* **1992**, *12*, 127–129. [CrossRef]
14. Dlshad Ahmed, K.; Hawezi, R. Detection of Bone Fracture Based on Machine Learning Techniques. *Meas. Sens.* **2023**, *27*, 100723. [CrossRef]
15. Yadav, D.P.; Sharma, A.; Athithan, S.; Bhola, A.; Sharma, B.; Dhaou, I.B. Hybrid SFNet Model for Bone Fracture Detection and Classification Using ML/DL. *Sensors* **2022**, *22*, 5823. [CrossRef] [PubMed]
16. Hardalaç, F.; Uysal, F.; Peker, O.; Çiçeklidağ, M.; Tolunay, T.; Tokgöz, N.; Kutbay, U.; Demirciler, B.; Mert, F. Fracture Detection in Wrist X-Ray Images Using Deep Learning-Based Object Detection Models. *Sensors* **2022**, *22*, 1285. [CrossRef] [PubMed]
17. Aso-Escario, J.; Sebastián, C.; Aso-Vizán, A.; Martínez-Quiñones, J.V.; Consolini, F.; Arregui, R. Delay in Diagnosis of Thoracolumbar Fractures. *Orthop. Rev.* **2019**, *11*, 47–52. [CrossRef]
18. Meena, T.; Roy, S. Bone Fracture Detection Using Deep Supervised Learning from Radiological Images: A Paradigm Shift. *Diagnostics* **2022**, *12*, 2420. [CrossRef]
19. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *J. Am. Med. Assoc.* **2016**, *316*, 2402. [CrossRef]
20. Yadav, D.P.; Rathor, S. Bone Fracture Detection and Classification Using Deep Learning Approach. In Proceedings of the IEEE 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC), Mathura, India, 28–29 February 2020; pp. 282–285.
21. Pillai, R.; Sharma, N.; Upadhyay, D.; Devliyal, S.; Kaur, G. Improving Diagnostic Accuracy in Radiology with VGG16 Deep Learning Model for Bone Fracture Detection. In Proceedings of the 2024 5th IEEE Global Conference for Advancement in Technology (GCAT), Bangalore, India, 4–6 October 2024; pp. 1–5. [CrossRef]
22. Shandilya, G.; Anand, V.; Gupta, S. Unified CNN Approach for Fracture Detection Across All Anatomical Regions in Radiographic Images. In Proceedings of the IEEE 2024 Global Conference on Communications and Information Technologies (GCCIT), Bangalore, India, 25–26 October 2024; pp. 1–6.
23. Alsufyani, A. Enhancing Diagnostic Accuracy in Bone Fracture Detection: A Comparative Study of Customized and Pre-Trained Deep Learning Models on X-Ray Images. *Int. J. Adv. Appl. Sci.* **2025**, *12*, 68–81. [CrossRef]
24. Alex, R.; Rosmasari. Classification of Multi-Region Bone Fractures from X-Ray Images Using Transfer Learning with ResNet18. *Int. J. Artif. Intell. Med. Issues* **2025**, *3*, 14–21. [CrossRef]

25. Shandilya, G.; Anand, V.; Chauhan, R.; Pokhariya, H.S.; Gupta, S. Optimizing Multi-Region Fracture Detection Through AlexNet Deep Learning Architecture. In Proceedings of the 2024 5th IEEE Global Conference for Advancement in Technology (GCAT), Bangalore, India, 4–6 October 2024; pp. 1–6.
26. Kumar, A.; Nelson, L.; Arumugam, D. Automated Bone Fracture Detection with AlexNet on Multi-Region Radiographic Analysis. In Proceedings of the 2024 5th IEEE Global Conference for Advancement in Technology (GCAT), Bangalore, India, 4–6 October 2024; pp. 1–5. [CrossRef]
27. Chauhan, S. Bone Fracture Detection with CNN: A Deep Learning Approach. In Proceedings of the IEEE 2024 5th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 18–20 September 2024; pp. 1253–1258.
28. Uddin, A.; Chowdhury, R.; Saha, A.; Hosen, M.H.; Roy, P.S.; Uddin, M.N. MobileViT and YOLOv8: Improving Bone Fracture Detection and Classification Through Deep Learning. In Proceedings of the 2024 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON), Dhaka, Bangladesh, 28–29 November 2024; pp. 212–217.
29. Rodrigo, M.; Kumar, M.; Faramawy, A.; Arya, H. Bone Fracture Multi-Region X-Ray Data. 2024. Available online: https://www.kaggle.com/datasets/bmadushanirodrigo/fracture-multi-region-x-ray-data/data (accessed on 1 August 2024).
30. Li, X.; Zhang, J.; Yang, Y.; Cheng, G.; Yang, K.; Tong, Y.; Tao, D. SFNet: Faster and Accurate Semantic Segmentation via Semantic Flow 2022. *Int. J. Comput. Vis.* **2024**, *132*, 466–489. [CrossRef]
31. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
32. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv* **2019**, arXiv:1905.11946. [CrossRef]
33. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
34. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
36. Sokolova, M.; Lapalme, G. A Systematic Analysis of Performance Measures for Classification Tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
37. Liu, H.; Simonyan, K.; Yang, Y. DARTS: Differentiable Architecture Search. *arXiv* **2018**, arXiv:1806.09055.
38. Bluemke, D.A.; Moy, L.; Bredella, M.A.; Ertl-Wagner, B.B.; Fowler, K.J.; Goh, V.J.; Halpern, E.F.; Hess, C.P.; Schiebler, M.L.; Weiss, C.R. Assessing Radiology Research on Artificial Intelligence: A Brief Guide for Authors, Reviewers, and Readers—From the *Radiology* Editorial Board. *Radiology* **2020**, *294*, 487–489. [CrossRef] [PubMed]
39. Dietterich, T.G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* **1998**, *10*, 1895–1923. [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.

diagnostics

MDPI

Article

CELM: An Ensemble Deep Learning Model for Early Cardiomegaly Diagnosis in Chest Radiography

Erdem Yanar [1,2,*], Fırat Hardalaç [2] and Kubilay Ayturan [2]

[1] Department of Healthcare Systems System Engineering, ASELSAN, 06200 Ankara, Turkey
[2] Department of Electrical and Electronics Engineering, Gazi University, 06570 Ankara, Turkey; firat@gazi.edu.tr (F.H.); kubilay.ayturan@gazi.edu.tr (K.A.)
* Correspondence: eyanar@aselsan.com.tr

Abstract: Background/Objectives: Cardiomegaly—defined as the abnormal enlargement of the heart—is a key radiological indicator of various cardiovascular conditions. Early detection is vital for initiating timely clinical intervention and improving patient outcomes. This study investigates the application of deep learning techniques for the automated diagnosis of cardiomegaly from chest X-ray (CXR) images, utilizing both convolutional neural networks (CNNs) and Vision Transformers (ViTs). **Methods:** We assembled one of the largest and most diverse CXR datasets to date, combining posteroanterior (PA) images from PadChest, NIH CXR, VinDr-CXR, and CheXpert. Multiple pre-trained CNN architectures (VGG16, ResNet50, InceptionV3, DenseNet121, DenseNet201, and AlexNet), as well as Vision Transformer models, were trained and compared. In addition, we introduced a novel stacking-based ensemble model—Combined Ensemble Learning Model (CELM)—that integrates complementary CNN features via a meta-classifier. **Results:** The CELM achieved the highest diagnostic performance, with a test accuracy of 92%, precision of 99%, recall of 89%, F1-score of 0.94, specificity of 92.0%, and AUC of 0.90. These results highlight the model's high agreement with expert annotations and its potential for reliable clinical use. Notably, Vision Transformers offered competitive performance, suggesting their value as complementary tools alongside CNNs. **Conclusions:** With further validation, the proposed CELM framework may serve as an efficient and scalable decision-support tool for cardiomegaly screening, particularly in resource-limited settings such as intensive care units (ICUs) and emergency departments (EDs), where rapid and accurate diagnosis is imperative.

Keywords: cardiomegaly; deep learning; early diagnosis; clinical decision support; medical imaging; automated diagnosis; ensemble learning

1. Introduction

1.1. Clinical Importance of Cardiomegaly and Positioning of AI Systems

Cardiomegaly, characterized by an enlargement of the heart muscle, is often considered a marker of underlying cardiac conditions and is associated with increased morbidity and mortality. Early diagnosis plays a critical role in improving patient outcomes, particularly given that cardiovascular diseases (CVDs) are among the leading causes of global mortality, accounting for approximately one-third of annual deaths [1]. Early identification and management of cardiomegaly can significantly improve patient outcomes [2]. Imaging techniques hold a central position in the diagnostic process for CVDs, as they facilitate the detection of structural changes in the heart, such as those seen in cardiomegaly. Figure 1,

for example, illustrates normal heart anatomy alongside the anatomical changes associated with cardiomegaly [3]. Key clinical indicators of cardiomegaly include dyspnea (shortness of breath), which occurs mainly during exertion or while lying down and is often correlated with elevated biomarkers such as B-type Natriuretic Peptide (BNP) and N-terminal pro B-type Natriuretic Peptide (NT-proBNP), which reflect cardiac strain [4].

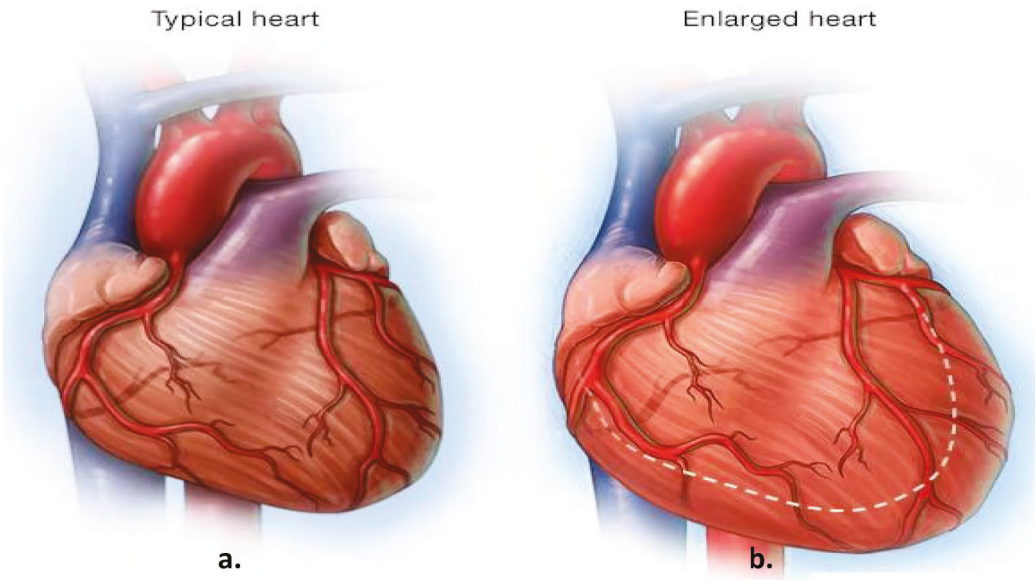


Figure 1. Comparative heart anatomy: (**a**) illustration of a typical heart; (**b**) demonstrating the anatomical changes associated with an enlarged heart [3]. The dotted line indicates the direction of expansion of the heart.

Recent advancements in artificial intelligence (AI) and machine learning have opened new opportunities for enhancing diagnostic accuracy in cardiovascular health. Martin-Isla et al. emphasized the importance of cardiac imaging in diagnosing CVDs, highlighting the potential of AI tools to support clinical decision-making [5]. Their study demonstrated how machine learning methods can provide more automated, precise, and early diagnoses. The increasing availability of big data and advanced computational power has further driven rapid progress in AI-powered medical imaging, paving the way for more optimized diagnostic value in existing imaging modalities.

1.2. Traditional Diagnostic Methods

The clinical manifestations of cardiomegaly, or an enlarged heart, can vary significantly among individuals. While some patients may remain asymptomatic, others can experience notable symptoms. Cardiomegaly symptoms such as shortness of breath, edema, and arrhythmias are also discussed in detail in Section 1.2. These symptoms, detailed below, highlight this condition's diverse and complex presentation [6].

- Physical Examination: Assessment of abnormal heart sounds or dullness upon percussion indicating increased cardiac silhouette.
- ECG: Identifies electrical conduction abnormalities that may suggest chamber enlargement.
- CXR: The visualization of the heart's size and shape, often used as a preliminary diagnostic tool. Moreover, the Cardiothoracic Ratio (CTR) is a diagnostic metric utilized to identify cardiomegaly (an enlarged heart) in CXRs captured in the posteroanterior (PA) view. In order to determine this ratio, the maximum horizontal diameter of the heart is divided by the maximum horizontal diameter of the thoracic cavity, which is measured at the inner edges of the ribs. An average CTR ranges from 0.42 to 0.50. A

CTR exceeding 0.50 is indicative of cardiomegaly, thus signaling the presence of an enlarged heart. For a visual representation, refer to Figure 2, which demonstrates a sample CXR used for CTR calculation.

- Echocardiography: Offers direct measurements of heart chamber sizes and wall motion via ultrasound.

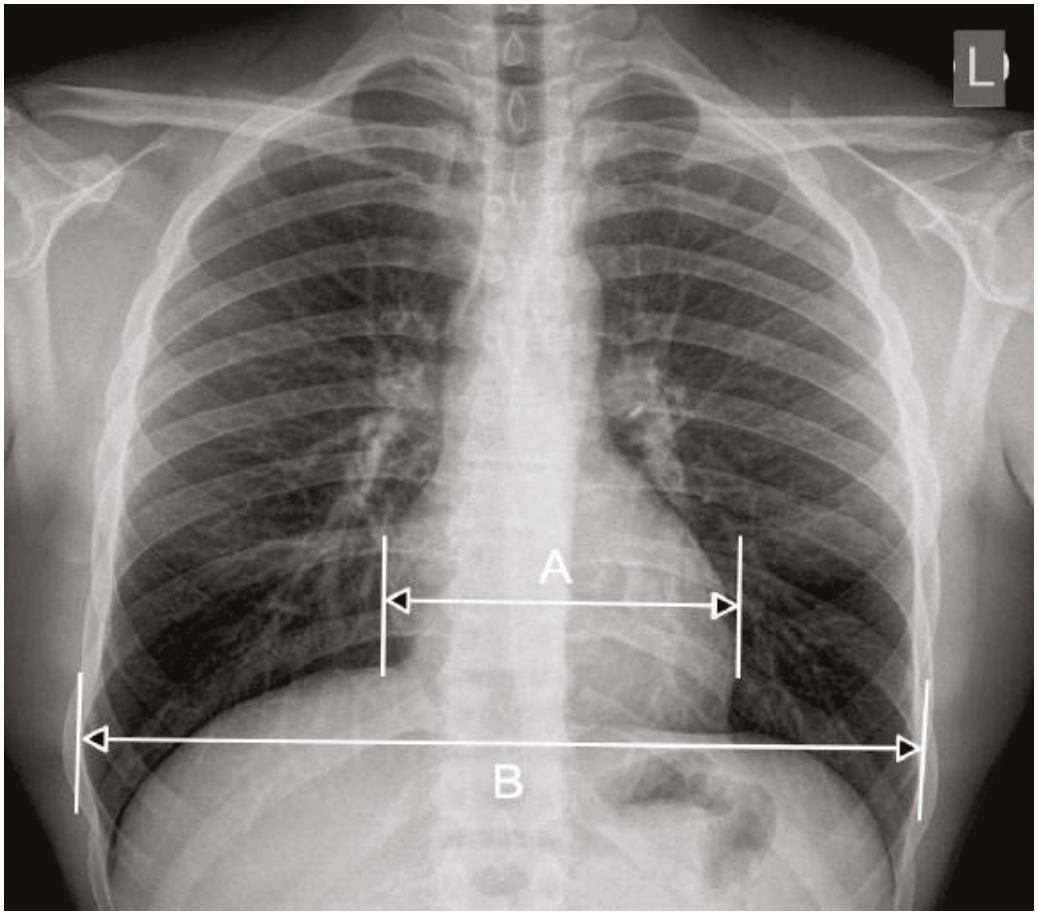


Figure 2. The approach to calculating the CTR involves measuring the relationship between the transverse diameter of the heart (denoted as [A]) and the transverse diameter of the chest cavity (denoted as [B]), as observed on a chest radiograph taken in the posteroanterior (PA) view. The CTR is then computed using the following formula: CTR = A/B [7].

Despite their effectiveness, these methods have limitations regarding availability, cost, and the need for specialized expertise. Details of the challenges are given below:

- Subjectivity: Interpretation of ECGs and CXRs is highly dependent on clinical experience, leading to inter-observer variability.
- Limited Sensitivity and Specificity: Subtle structural changes may be missed, particularly in early-stage cardiomegaly.
- Time and Resource Intensive: Echocardiography requires trained personnel and equipment not always available in emergency or rural settings.
- Diagnostic Delays: Manual assessments can delay decision-making, especially in high-volume clinical environments.

These limitations underscore the need for automated, scalable, and high-precision diagnostic tools. Recent advances in deep learning, particularly convolutional neural networks (CNNs), have significantly improved image-based diagnostics by extracting complex spatial features from chest X-rays. Studies show that AI models can match or even exceed radiologist-level performance in detecting cardiomegaly. By reducing variability and

expediting diagnosis, these methods provide reliable clinical decision support, especially in resource-limited settings.

Recent research has demonstrated that machine learning methods and intense learning offer superior accuracy compared to traditional medical imaging diagnostic techniques [8]. Studies conducted by Mohare et al., Que et al., Raghu Kumar et al., and Ayalew et al. have utilized deep learning techniques to diagnose cardiomegaly from X-ray images [9–12]. Their findings highlight the potential of advanced technology in reliably identifying cardiomegaly. The utilization of sophisticated deep learning algorithms enabled these studies to conduct a comprehensive analysis of X-ray images, thereby identifying patterns linked to cardiomegaly and facilitating the development of an enhanced diagnostic instrument for healthcare professionals.

This study presents a deep learning system using chest X-rays to classify normal and cardiomegaly cases. It combines conventional CNNs, several pre-trained models (VGG16, ResNet50, InceptionV3, DenseNet121/201, AlexNet) with spinal concatenated versions, and Vision Transformers (ViTs) for comparing attention-based approaches in cardiac image classification.

To enhance diagnostic performance and reduce false negatives, the model pipeline includes extensive preprocessing. Specifically, Contrast-Limited Adaptive Histogram Equalization (CLAHE) was applied to improve the visibility of low-intensity anatomical features, while multiplicative speckle noise was suppressed using targeted filtering techniques. These preprocessing steps were critical in enhancing image quality and consistency across datasets.

The goal of this work is to automate the diagnosis of cardiomegaly in a non-invasive, highly accurate, and scalable manner using CXR images. By reducing dependency on manual interpretation and accelerating early detection, the proposed model aims to support clinical decision-making and prevent the progression of undiagnosed cardiomegaly into severe conditions such as heart failure. This study contributes to the literature in the following key ways:

- We propose the CELM (Combined Ensemble Learning Model)—a novel stacking-based ensemble framework that integrates the outputs of VGG16, ResNet50, and InceptionV3 through a shallow meta-classifier to improve cardiomegaly detection from chest X-rays.
- We implement a large-scale evaluation of multiple CNN and Vision Transformer architectures on one of the most comprehensive publicly available CXR datasets (n = 50,000) drawn from PadChest, NIH CXR, VinDr-CXR, and CheXpert.
- We report a balanced diagnostic performance with high precision (0.99), specificity (0.92), and AUC (0.90), demonstrating the clinical feasibility of CELM for early cardiomegaly screening.
- We perform a detailed comparative performance analysis with classical and recent models, including VIT-B16 and DenseNet variants, and provide a standardized benchmark in the results section.

We emphasize reproducibility by detailing the architecture, data splits, and performance metrics, and by making the full code and trained models available via GitHub (https://github.com/eyanarai/CELM (accessed on 18 June 2025)) for public reuse and extension.

This section outlines this paper's structure: Section 2 reviews recent deep learning studies on cardiomegaly; Section 3 details the methodology and dataset; Section 4 covers evaluation methods; Sections 5 and 6 present results and discussion; Section 7 concludes the study; and Section 8 highlights the study's contributions.

2. Related Works

This section reviews literature on AI advancements in medical imaging for cardiomegaly detection. AI methods, especially CNNs, show strong potential for early, accurate diagnosis from CXRs. Studies have improved sensitivity and specificity via varied architectures and preprocessing, but challenges like dataset variability, model interpretability, and consistency across populations persist. This review highlights key contributions and limitations, framing the innovations of this study.

- Sogancioglu et al. [13] compared anatomical segmentation and image-level classification on frontal CXRs, achieving 97% sensitivity and 90% specificity.
- Yoo et al. [14] applied machine learning to echocardiography, enhancing diagnostic consistency with an accuracy of ~80% and feature visualization support.
- Que et al. [10] introduced CardioXNet using CXR images, reaching 93.75% accuracy and 100% precision.
- Gupte et al. [15] identified cardiomegaly risk factors via clinical studies and developed a model with 94% accuracy and 88% F1-score.
- Candemir et al. [16] fine-tuned pre-trained CNNs for CXR-based cardiomegaly detection, reporting 88% accuracy and 92% recall.
- Raghu Kumar et al. [11] combined 2D and 1D CNNs to classify cardiomegaly with balanced precision, recall, and F1-score of 95%.
- Decoodt et al. [17] proposed a hybrid classical-quantum model achieving 87% accuracy and 93% recall using CQ transfer learning.
- Ayalew et al. [12] used transfer learning (ResNet-50, Inception, DenseNet-169) and visualization techniques (GradCAM), reporting near-perfect metrics (99.8% accuracy).
- Chouhan et al. [18] proposed a stacked ensemble (VGG16, InceptionV3, Xception) for pneumonia detection with >98% accuracy.
- Hussain et al. [19] confirmed that stacking CNNs improves performance over bagging and voting.
- Majkowska et al. [20] benchmarked deep CNN ensembles on radiographs, showing competitive AUCs across multiple thoracic pathologies.

The literature consistently shows that ensemble learning, particularly stacking, outperforms individual models by improving accuracy, robustness, and generalizability. Data augmentation helps mitigate overfitting, especially with limited or imbalanced datasets, while transfer learning significantly boosts diagnostic performance by fine-tuning pretrained models. Building on these insights, this study combines transfer learning and data augmentation within a stacked ensemble of pretrained CNNs for reliable cardiomegaly classification from chest X-rays.

3. Materials and Methods

3.1. Materials

In this study, we focused on binary classification between cardiomegaly and non-cardiomegaly CXR cases using curated subsets from four large-scale datasets (NIH ChestX-ray14, PadChest, CheXpert, and VinDr-CXR) (Table 1). For the cardiomegaly-positive (target = 1) class, we included all frontal posteroanterior (PA) images labeled with cardiomegaly, regardless of the presence of additional thoracic findings such as pleural effusion or pulmonary edema. For the negative class (target = 0), we included images without a cardiomegaly label, whether labeled as normal or containing other non-cardiac abnormalities (e.g., consolidation, fibrosis, atelectasis). To prevent patient-level data leakage and ensure realistic clinical variation, we included only one PA image per patient, specifically the earliest chronological instance. This approach mimics a real-world setting where patients

may present with complex pathologies, and it prevents model bias due to repeated or longitudinal data entries.

Table 1. Detailed data distribution of data sets.

	CXR NIH	CheXpert	CXR VinDr	PadChest
Source	National Institutes of Health (NIH) Clinical Center	Stanford University, Stanford ML Group	VinBigData Institute, Vietnam	San Juan Hospital, Alicante, Spain
Images	112,120 frontal-view X-ray images	224,316 chest radiographs	18,000 high-resolution PA-view CXR images selected from over 100,000 raw DICOMs	160,000+ CXR images from 67,000 patients
Labels	14 thoracic pathologies (e.g., cardiomegaly, edema, pneumonia), mined using NLP from radiology reports	14 observations (including cardiomegaly, edema, pleural effusion), with uncertainty modeling	22 local findings and 6 global diagnostic labels, annotated by 17 radiologists	174 radiological findings, 104 anatomical locations, 19 differential diagnoses

3.1.1. CXR NIH Dataset

The dataset comprises 112,120 X-ray images with disease labels from 30,805 unique patients [21]. The samples from the CXR NIH dataset are shown in Figure 3. Disease labels were extracted from radiology reports using NLP, achieving over 90% accuracy and supporting weakly supervised learning.

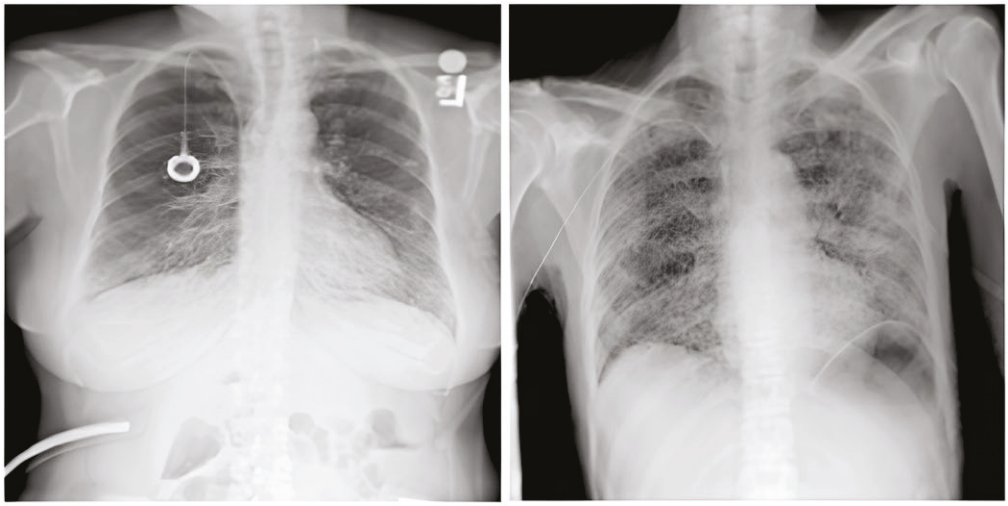


Figure 3. CXR NIH dataset samples [21].

In order to gain a representative insight into the subject matter, the study utilized the CXR NIH dataset. This is a large-scale hospital-grade database comprising 108,948 frontal-view CXR images from 32,717 patients, and it has gained recognition as a valuable resource within the field [22]. The dataset includes images labeled for eight common thoracic diseases using an NLP system applied to radiology reports. While this automated approach enables scalable and consistent labeling with over 90% estimated accuracy, it may occasionally produce ambiguous or incorrect labels.

While the dataset provides a valuable resource for medical imaging research, it is essential to note that the original radiology reports used for labeling are not publicly available. Consequently, direct validation or further refinement of the extracted labels is constrained. Additional details about the labeling process can be found in the freely

accessible publication "ChestX-ray8: Hospital-scale CXR Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases" [23].

This study focused on two subsets of the dataset: images labeled as cardiomegaly and normal. This approach enabled the evaluation of model performance in distinguishing normal from pathological cases. Despite its limitations, the NIH CXR dataset remains a valuable resource for thoracic imaging research.

3.1.2. CheXpert Dataset

This is an extensive and meticulously curated collection comprising 224,316 chest radiographs obtained from 65,240 patients who underwent radiographic examinations at Stanford Health Care facilities over a substantial timeframe spanning from October 2002 to July 2017 [24,25]. The samples from the CheXpert dataset are shown in Figure 4. These examinations were performed across a diverse range of settings, including inpatient and outpatient centers, ensuring the inclusion of a broad spectrum of clinical scenarios and patient demographics. Each radiograph in this dataset has been labeled for 14 distinct, standard chest radiographic observations, aligning with established clinical guidelines and fostering consistency in labeling across the dataset.

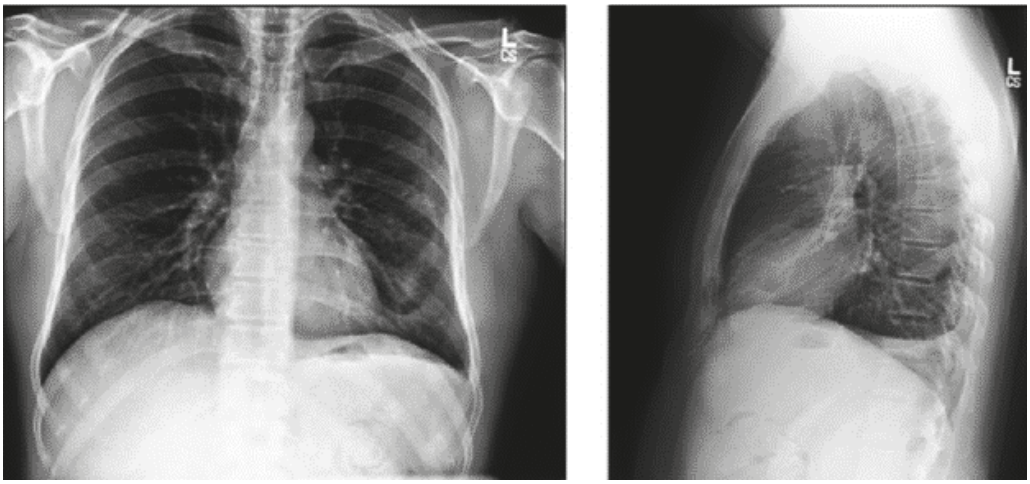


Figure 4. CheXpert dataset samples [24].

3.1.3. CXR VinDr Dataset

This was developed to provide the research community with a large-scale, high-quality collection of CXR images annotated with detailed and reliable labels [26]. The samples from the CXR VinDr dataset are shown in Figure 5. This dataset was constructed using over 100,000 raw images in DICOM format, which were retrospectively collected from two of Vietnam's largest and most prominent hospitals: Hospital 108 and Hanoi Medical University Hospital [27]. These facilities contributed to the dataset's diversity and representativeness by including many cases and patient demographics.

From this initial pool, a carefully curated subset of 18,000 posteroanterior (PA) view CXR scans were selected for publication. These scans are of particular value as they facilitate the visualization of critical radiological results and the categorization of common thoracic diseases. This provides researchers with a robust dataset that can be used for several applications, including disease detection and model validation.

A team of 17 highly experienced radiologists, each with at least eight years of professional expertise, meticulously annotated the images. Their annotations cover 22 critical findings (local labels) and six diagnostic categories (global labels). The local labels correspond to specific findings localized on the images with bounding boxes, enabling precise identification of pathological regions. The global labels, on the other hand, align with the

"Impressions" section of a standard radiology report, summarizing the overall diagnostic conclusions. These local and global labels provide a comprehensive structure that mirrors the organization and detail typically found in professional radiology reports. This ensures the dataset's usability for various research purposes in medical imaging and beyond.

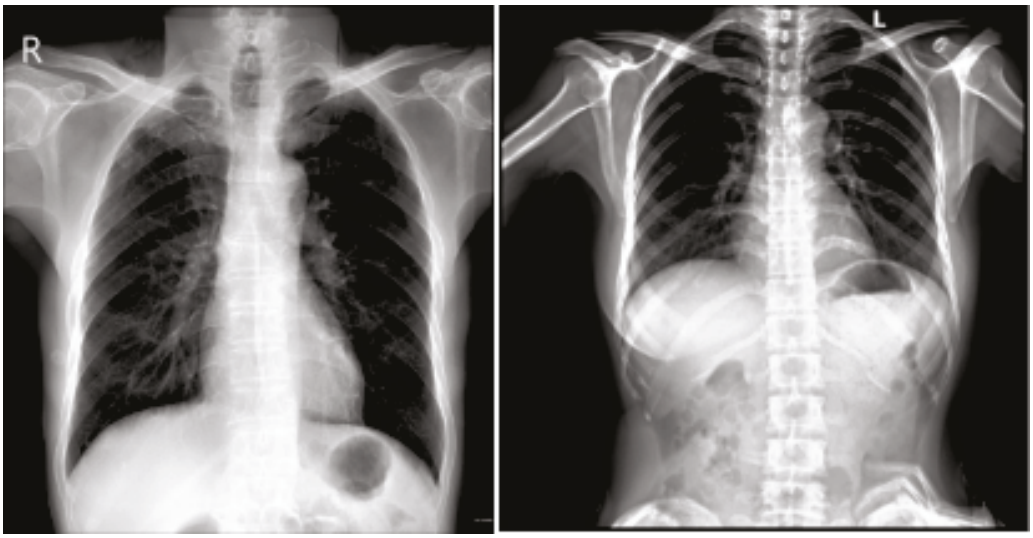


Figure 5. VinDr-CXR Dataset samples [26].

3.1.4. PadChest Dataset

This is a comprehensive and valuable resource for medical imaging research, consisting of over 160,000 CXR images obtained from 67,000 patients [28,29]. The samples from the PadChest dataset are shown in Figure 6. These images were collected between 2009 and 2017 at San Juan Hospital in Spain, representing various cases and clinical scenarios [30]. The dataset includes images captured in six position views, providing versatility for various radiological studies. Additionally, it contains extensive metadata detailing information about image acquisition techniques and patient demographic characteristics, further enhancing its utility for diverse research applications.

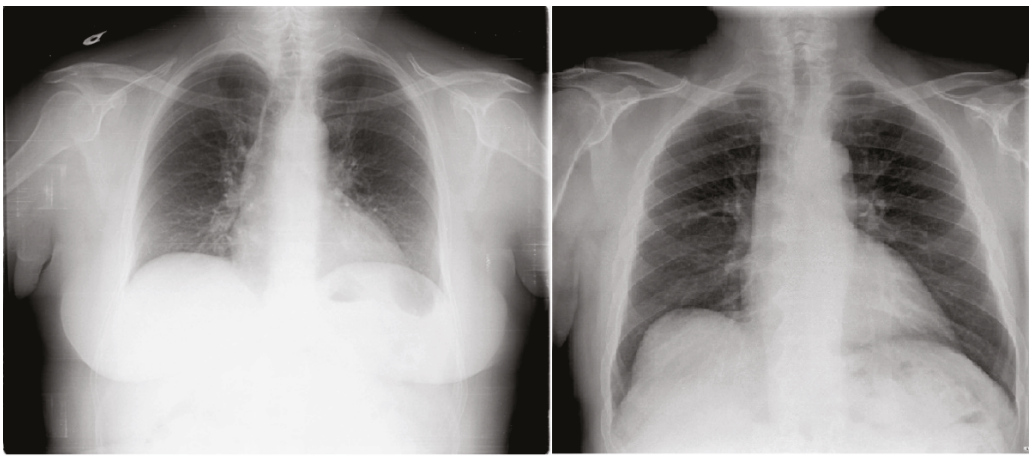


Figure 6. PadChest dataset samples [28].

Each image in the dataset is accompanied by radiology reports interpreted and documented by experienced radiologists at San Juan Hospital. These reports have been annotated with 174 distinct radiographic findings, 19 differential diagnoses, and 104 anatomical locations. The findings are organized within a hierarchical taxonomy and systematically mapped to standardized medical terminologies, specifically the Unified Medical Language

System (UMLS) [29]. This mapping ensures consistency and facilitates interoperability with other medical datasets and systems, making the dataset an invaluable tool for advancing machine learning applications in healthcare.

The annotation process involved a rigorous combination of manual and automated efforts to ensure both accuracy and scalability. Overall, 27% of the reports were manually annotated by trained physicians, who provided meticulous and high-quality labels. An automated labeling approach was employed for the remaining reports, utilizing a supervised method based on a recurrent neural network equipped with attention mechanisms. This hybrid labeling strategy balances precision and efficiency, resulting in a robust dataset that is well-suited for training and validating machine learning models to improve diagnostic accuracy and decision support in medical imaging.

To ensure robust training and avoid model bias toward the majority class, we implemented a class balancing strategy across the datasets. Since the original datasets contained uneven distributions of cardiomegaly-positive and normal chest X-ray (CXR) images, a class-wise downsampling approach was applied.

The following was performed for each dataset:

- An equal number of normal and cardiomegaly-positive images were selected.
- If the number of cardiomegaly cases was lower than that of the normal cases (as in CXR NIH and VinDr-CXR), an equal-sized subset of normal images was randomly selected.
- Conversely, in datasets with more cardiomegaly images (e.g., CheXpert), the majority class was downsampled to match the minority class to maintain a 1:1 class ratio.

This strategy ensured the following:

- A balanced class distribution, reducing the risk of biased learning.
- An improved generalization and stability during training.
- A fair evaluation of model performance across both classes.

To further improve robustness, the selected samples were shuffled and split into training, validation, and test sets in a stratified manner, preserving the class balance in each subset as shown in Table 2, combining data from four distinct sources: the PadChest, CXR NIH, CXR VinDr, and CheXpert datasets.

Table 2. Data distribution of datasets.

	PadChest	**CXR NIH**	**CXR VinDr**	**CheXpert**
Num. of Classes	19	8	28	14
Cardiomegaly	2140	1563	230	23,002
Normal	8708	1563	1003	16,627
Total	10,848	3126	1233	39,629

3.2. Methods

To build a robust and generalizable model for cardiomegaly detection, this study incorporated data augmentation and transfer learning as foundational strategies. This techniques were employed to enhance model performance, reduce overfitting, and address limitations related to dataset size and variability [29].

A total of 54,836 chest X-ray (CXR) images were initially collected from the PadChest, CXR NIH, VinDr-CXR, and CheXpert datasets. After rigorous quality filtering and class-balancing, a curated dataset of 50,000 images was finalized, containing equal numbers of cardiomegaly-positive and normal cases [29].

Quality filtering excluded low-resolution, poorly exposed, or truncated images. To enforce a 1:1 class ratio, randomly selected samples from the overrepresented class were

removed. The resulting dataset was split into 40,000 training, 5000 validation, and 5000 test images as shown in Table 3.

Table 3. Final image distribution after filtering and balancing.

Dataset	Cardiomegaly	Normal	Total Used	Quality Filtered	Notes
PadChest	2140	2140	4280	6568	Downsampled for balance
CXR NIH	1563	1563	3126	0	Fully retained
VinDr-CXR	230	230	460	773	Small dataset, all usable
CheXpert	21,067	21,067	42,134	~5000	Largest subset, filtered
Total	25,000	25,000	50,000	~12,341	After filtering + balancing

In particular, the deliberate exclusion of images with poor resolution, noise, or improper exposure may result in an overly optimistic performance profile. In clinical settings, such imperfections are common; thus, the current model's generalizability to routine hospital data must be tested in follow-up studies.

While this curated dataset supports more consistent model training, we acknowledge that it does not fully reflect the variability found in real-world clinical practice. Excluding technically suboptimal images may reduce generalizability, as actual hospital workflows often involve images with diverse exposure, noise, and positioning artifacts. Furthermore, although non-cardiomegaly cases included other thoracic pathologies (e.g., consolidation, fibrosis, atelectasis) and cardiomegaly cases included findings such as pleural effusion and pulmonary edema, both groups were still partially curated. As such, the dataset does not encompass the full spectrum of clinical heterogeneity. Future research will focus on expanding data diversity to include a wider range of imaging quality and comorbidities, enabling better representation of real-world diagnostic scenarios.

Transfer learning was utilized by leveraging pre-trained CNNs such as VGG16, ResNet50, InceptionV3, DenseNet121, DenseNet201, and AlexNet. These models, originally trained on large-scale datasets like ImageNet, offer deep hierarchical representations that can be fine-tuned for medical imaging tasks. By adapting these architectures to the chest radiography domain, we significantly accelerated training and improved classification accuracy, especially given the complexity and subtlety of cardiomegaly features on chest X-rays (CXRs).

In parallel, data augmentation techniques were applied during training to artificially expand the dataset and introduce greater variability. Augmentations included random rotation, zoom, and horizontal flipping, simulating real-world variations in imaging conditions. Additionally, CLAHE was applied to enhance local contrast, improving feature visibility in low-intensity regions—common in medical images. These preprocessing steps collectively aimed to enhance model generalization and reduce bias toward specific imaging conditions.

While combining datasets enables the model to learn from a broad distribution, the train-test split was performed after merging the datasets. As such, domain-specific effects between data sources may not be fully captured. Future work will explore cross-dataset validation to better simulate real-world deployment scenarios involving domain shift.

3.2.1. Preprocessing

Prior to model training, all chest X-ray (CXR) images were resized to a resolution of 224 × 224 pixels, which ensures compatibility with standard CNN architectures while balancing computational efficiency [29]. Pixel values were normalized to a [0, 1] range to facilitate faster convergence during training. To reduce noise—particularly speckle and multiplicative noise commonly present in radiographic data—image smoothing filters were applied during preprocessing. This ensured that finer anatomical features relevant to

cardiomegaly detection, such as heart borders and lung field clarity, remained intact while minimizing distortions that could hinder model performance [29].

3.2.2. Clinically Aware Data Augmentation

Data augmentation was carefully designed to increase dataset variability while preserving the thoracic anatomical structures critical for cardiomegaly assessment.

Horizontal flipping was applied with a 50% probability, as frontal chest radiographs maintain diagnostic validity regardless of left–right symmetry in most cardiomegaly cases [29].

Rotation was constrained to a small range of ± 5 degrees, as larger rotations could deform key structures such as the heart silhouette or mediastinum, which are essential for accurate diagnosis.

Zoom augmentation was applied within a narrow range of 95% to 105%**,** simulating realistic changes in radiographic magnification during image acquisition while avoiding the disproportionate enlargement of the heart or lungs.

Translation was limited to $\leq$5% of the image width or height (i.e., maximum shifts of approximately ± 10 to 12 pixels) along the x- and y-axes to simulate minor variations in patient positioning without displacing core anatomical features.

Additionally, combinations of flipping and translation were used selectively to further enhance spatial diversity without compromising clinical interpretability [29].

3.2.3. Contrast Enhancement with CLAHE

To improve the visibility of thoracic anatomical landmarks in low-contrast images, Contrast-Limited Adaptive Histogram Equalization (CLAHE) was applied. The enhancement process was conducted on the luminance channel after converting the image from RGB to the LAB color space. CLAHE was implemented using a clip limit of 2.0 and a tile grid size of 8×8, parameters optimized to enhance local contrast while preventing the over-amplification of noise. This method improved contrast along subtle cardiac contours and lung borders, thereby aiding the model in detecting cardiomegaly-related features. Furthermore, to ensure consistent preprocessing across all datasets, a binary K-Means clustering algorithm was used to detect white-background images, which were then inverted to match the standard dark-background format used in CXR interpretation [29].

Figure 7 shows examples of images to which data augmentation techniques have been applied. Table 4 provides a detailed explanation of the types of data augmentation, their settings, and their clinical purposes.

Table 4. Detailed explanations of data augmentation types.

Augmentation	Range/Setting	Clinical Purpose
CLAHE	clipLimit = 2.0, grid = 8×8	Improves contrast in low-density heart/lung regions
Horizontal Flip	50% chance	Simulates L/R orientation variability
Translation	Max $\pm$ 12 px ($\approx$5%)	Minor positioning variation during acquisition
Rotation	Max $\pm 5^{\circ}$	Keeps heart silhouette stable while adding orientation tolerance
Zoom	95–105%	Simulates slight magnification without distortion

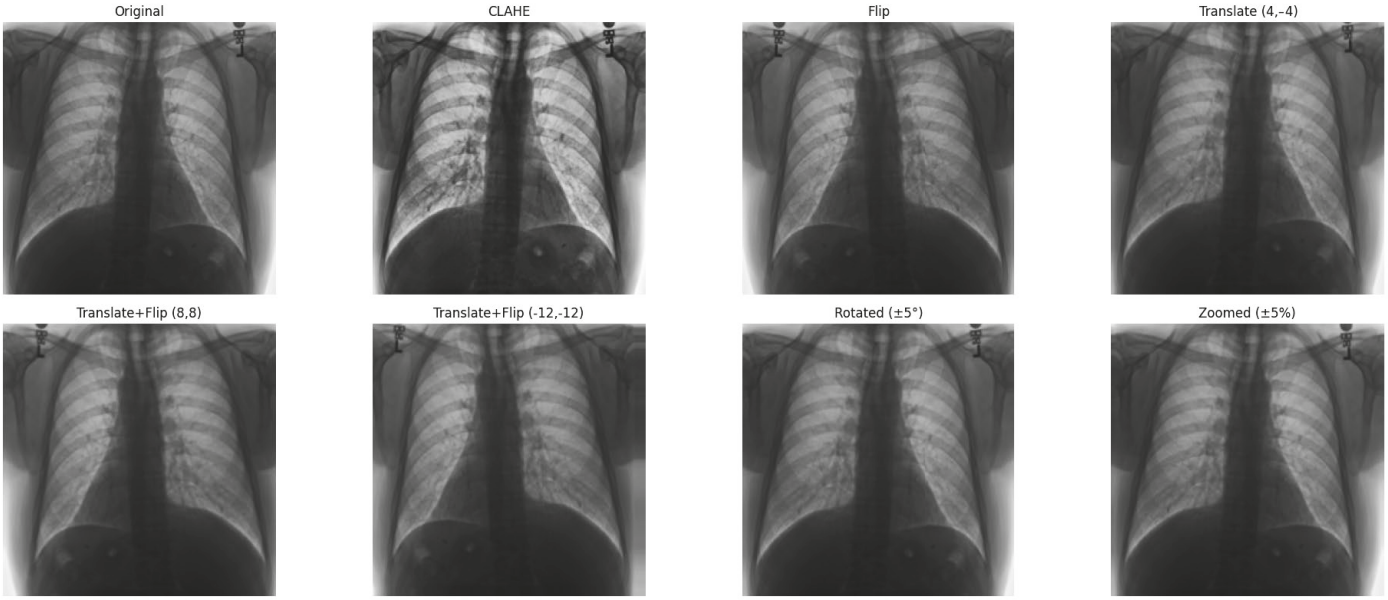


Figure 7. Data augmentation example.

As a result of data augmentation, the dataset was expanded to comprise a total of 50,000 images, evenly distributed between cardiomegaly cases and other categories (e.g., no findings and various other diseases) sourced from four primary datasets. Of these, 40,000 images were designated for training, 5000 for validation, and 5000 for testing. A detailed breakdown of the input images is provided in Table 5.

Table 5. Augmented input dataset.

	Training	Validation	Test
Cardiomegaly	20,000	2500	2500
Normal	18,000	2250	2250
Other Findings	2000	250	250
Total	40,000	5000	5000

3.3. Model Architectures

To promote reproducibility and technical clarity, we provide comprehensive architectural details for all models evaluated in this study. This includes clear definitions of the CNN and Vision Transformer backbones and the proposed stack-based ensemble (CELM) structure, all of which are presented in detail. We explored several CNN and Vision Transformer architectures to determine the most effective model for diagnosing cardiomegaly from CXRs [29]. The models included were as follows:

- Vision Transformer (ViT) is a deep learning model for image recognition introduced by Google Research in 2020 as a CNN alternative. Instead of convolutions, it uses self-attention (Transformer) to analyze images. ViT-B/16 refers to the Base model that splits images into 16 $\times$ 16 patches and processes them with a Transformer. The architecture consists of three main parts:
- Patch Embedding: An input image (e.g., 224 $\times$ 224 pixels) is divided into non-overlapping patches of size 16 $\times$ 16. Each patch is flattened into a vector and embedded into a feature space. The embedding also includes position encodings to retain spatial information.
- Transformer Encoder: These patch embedding is fed into a standard Transformer encoder (similar to NLP models like BERT). Self-attention layers help the model understand long-range dependencies between patches.

- Classification Head: A class token is added to represent the entire image. The final output is processed by an MLP (multi-layer perceptron) to generate classification scores.

An example image with ViT applied is shown in Figure 8. Table 6 also shows the patch configuration for image processing. The total model consists of 12 transformer encoder layers, 768-dimensional hidden representation, 12 self-attention heads, and ~86 million total parameters, as shown in Figure 9.

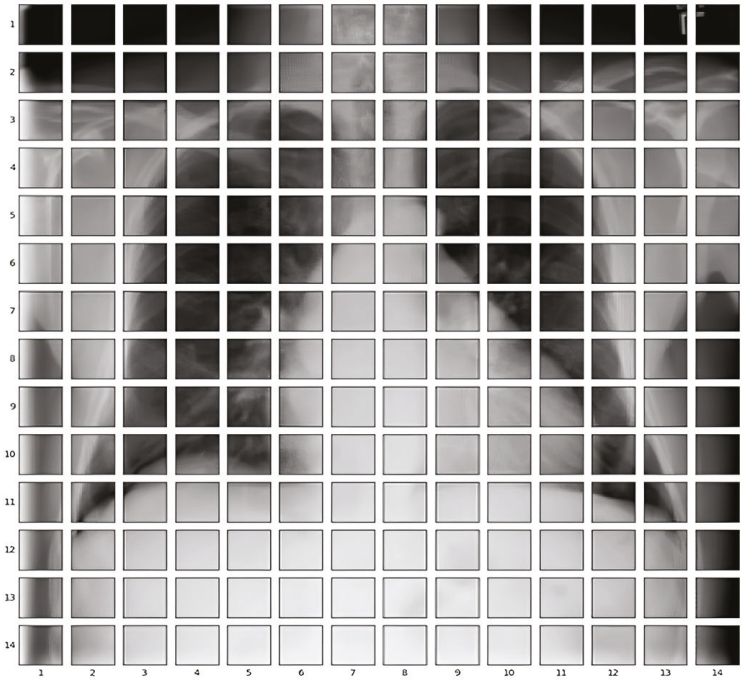


Figure 8. ViT Example Image.

Table 6. Patch configuration for image processing.

Image Size	Patch Size	Patch Per Image	Elements Per Patch
224 × 224	16 × 16	196	768

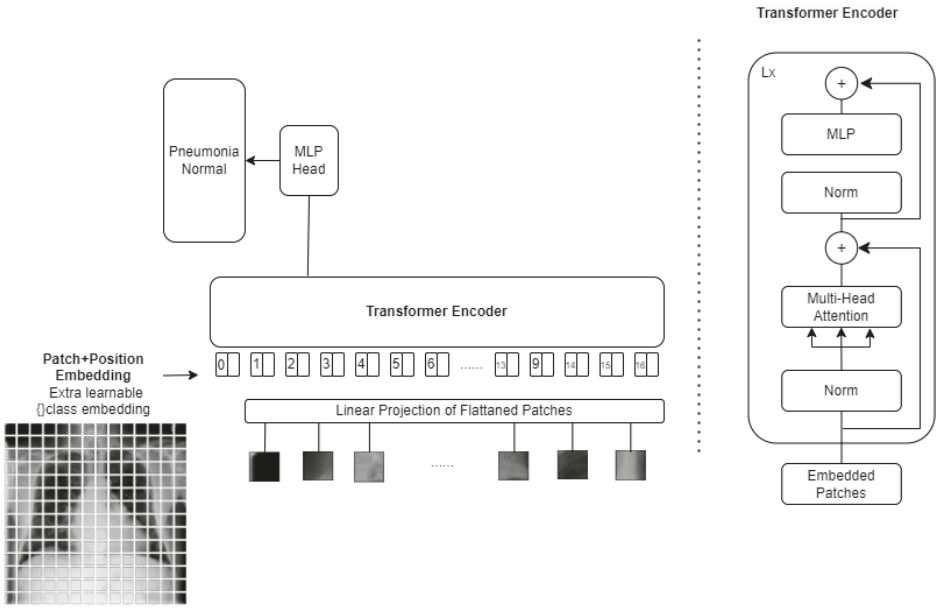


Figure 9. ViT architecture.

- VGG16 is a deep CNN known for its simplicity and depth, making it a strong baseline for image classification tasks. It is a 16-layer CNN with 13 convolutional layers and 3 fully connected layers, designed to process input images of 224 × 224 pixels. The first two fully connected layers contain 4096 channels each, while the third includes 1000 channels. Training is performed over 16 epochs, with 326 steps per epoch, to optimize the accuracy and processing efficiency. The architecture of the VGG16 model is shown in Figure 10.

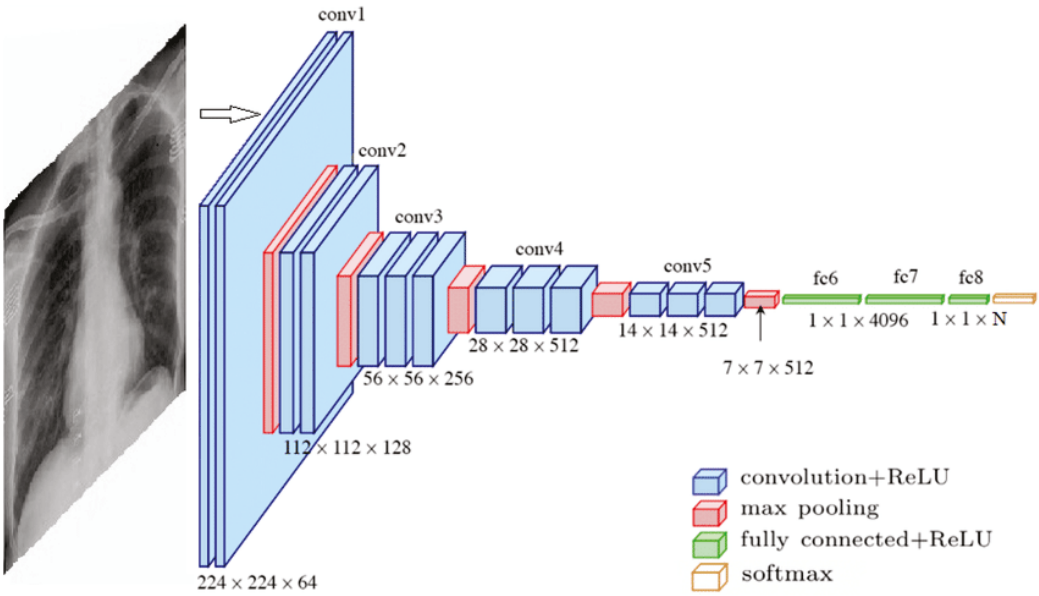


Figure 10. VGG-16 architecture [31].

- ResNet50: The Residual Network 50-layer (ResNet50) architecture is a powerful model for image classification, capable of achieving state-of-the-art results when trained on large datasets. Its key innovation lies in using residual connections, which enable the network to learn residual functions that map inputs to outputs effectively. These connections allow the ResNet50 architecture to build significantly deeper architectures without encountering vanishing gradient issues, a challenge in traditional deep networks. The ResNet50 architecture is composed of four main components:
  1. Convolutional Layers.
  2. Identity Blocks.
  3. Convolutional Blocks.
  4. Fully Connected Layers.

Each component is vital in processing the input image and performing accurate classification. A detailed breakdown of the ResNet50 architecture is shown in Table 7.

Table 7. Residual stages (conv. blocks and identity blocks).

Stage	Filters	Block Types	Output Shape
Stage1	64, 64, 256	1 Conv. + 2 Identity	56 × 56 × 256
Stage2	128, 128, 512	1 Conv. + 3 Identity	28 × 28 × 512
Stage3	256, 256, 1024	1 Conv. + 5 Identity	14 × 14 × 1024
Stage4	512, 512, 2048	1 Conv. + 2 Identity	7 × 7 × 2048

ResNet50 is highly efficient for deep learning tasks, balancing accuracy and computational complexity, and is widely used in research and industry. ResNet50Architecture is shown in Figure 11.

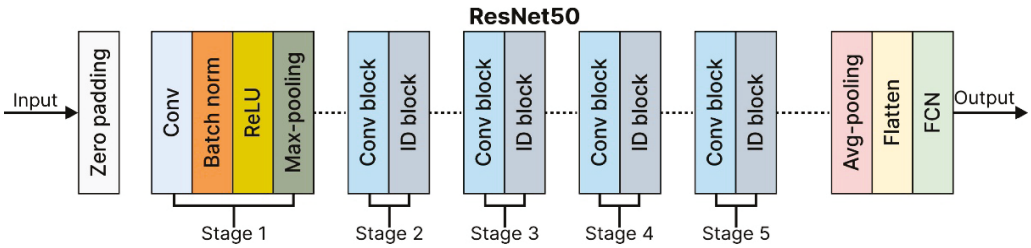


Figure 11. ResNet50 architecture [31].

- InceptionV3: InceptionV3 is an efficient and widely adopted convolutional neural network architecture designed for image classification tasks. It incorporates several key innovations to enhance accuracy while reducing computational complexity. Among these are label smoothing, which mitigates overfitting by preventing overconfidence in predictions, and convolutional factorization, where large 7 × 7 filters are replaced with more efficient combinations of 1 × 7 and 7 × 1 convolutions. The model also includes an auxiliary classifier during training, which provides intermediate supervision to combat the vanishing gradient problem and improve convergence, though it is removed during inference. InceptionV3 applies batch normalization and gradual grid size reduction to stabilize training and preserve spatial information. Its modular structure consists of inception modules that extract multi-scale features through parallel convolutions with various filter sizes (1 × 1, 3 × 3, and 5 × 5), enabling rich and diverse feature representations. Overall, InceptionV3 offers a strong balance between computational efficiency and classification performance. The architecture is illustrated in Figure 12.

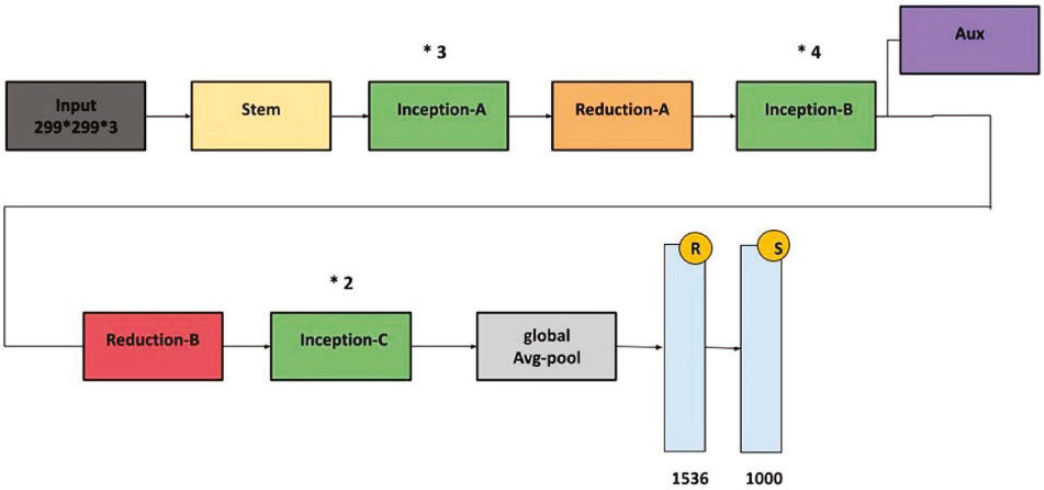


Figure 12. InceptionV3 architecture [32].

- DenseNet121: This constitutes a standard CNN, whereby the input image is passed sequentially through the network layer by layer, thus enabling the production of an output predicted label. This process is illustrated in Figure 13.

Each convolutional layer, except the first, receives the output feature map of the previous layer as its input and produces its feature map, which is then passed to the next layer. A network with L layers has L direct connections, each linking a layer to its immediate successor. In contrast, the DenseNet architecture introduces a novel approach by connecting each layer to every other layer within the network, giving rise to the term "Densely Connected Convolutional Network." A DenseNet with L layers has L(L + 1)/2 direct connections. Each layer takes the feature maps from all preceding layers as input

and, in turn, contributes its feature maps as input to all subsequent layers. This structure significantly enhances the flow of information and gradients throughout the network.

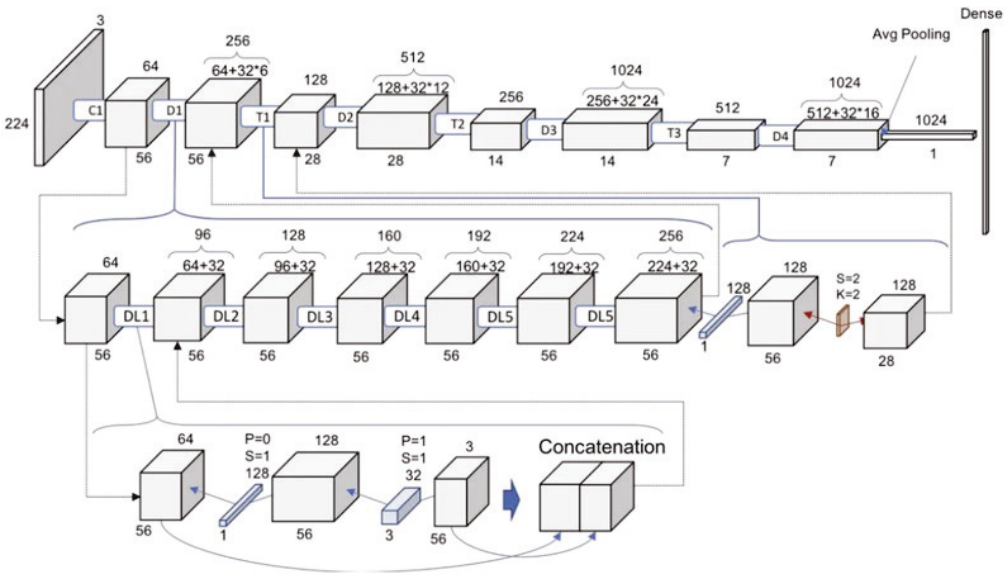


Figure 13. DenseNet121 architecture [33].

- CNN: The CNN architecture is designed to detect cardiomegaly in CXR images by progressively extracting features using 3 × 3 convolutional layers (128 and 256 filters). It then uses 1 × 1 point-wise convolutions and a 2 × 2 convolutional block to reduce spatial dimensions while preserving key patterns. The resulting feature maps are flattened and passed through fully connected layers with 500 and 100 neurons, with 30% dropout applied between layers to prevent overfitting. The final layer outputs classification probabilities using sigmoid or softmax activation. This design efficiently combines feature extraction, dimensionality reduction, and regularization. The architecture is shown in Figure 14.

The subsequent step is to diagnose the underlying condition, following the creation of a CXR mask encompassing the affected area and a diagnosis of cardiomegaly. This is accomplished through the training of disparate data samples and the development of a model that is capable of diagnosing test data. Figure 15 (VGG16), Figure 16 (ResNet50), and Figure 17 (InceptionV3) give customized versions of well-known classification models.

This study uses stacking, an ensemble learning model. Figures 18 and 19 shows the workflow of the ensemble learning model used in this study. Stacking represents a sophisticated ensemble learning technique that combines predictions from multiple machine learning models to improve classification. In contrast to other ensemble methods, such as bagging or boosting, stacking is designed to integrate the strengths of diverse base models by training a secondary model, frequently referred to as the meta-learner, on the outputs of these base models. This approach enables the capture of more complex patterns and relationships within the dataset. In healthcare, stacking has been demonstrated to be a particularly valuable technique, particularly in diagnosing diseases where the data is heterogeneous and the features are diverse.

To illustrate, a recent study employed a multi-modal stacking ensemble approach to diagnose cardiovascular diseases using ECG data. This approach integrated predictions from deep learning models trained on scalograms and grayscale ECG images. Using a logistic regression model as the meta-learner to combine the base predictions resulted in a notable improvement in performance metrics, with an area under the curve (AUC) of 0.995, compared to the results obtained from individual models or more straightforward

ensemble techniques. The stacking approach proved particularly efficacious in leveraging the complementary strengths of different data modalities to enhance diagnostic accuracy [34,35].

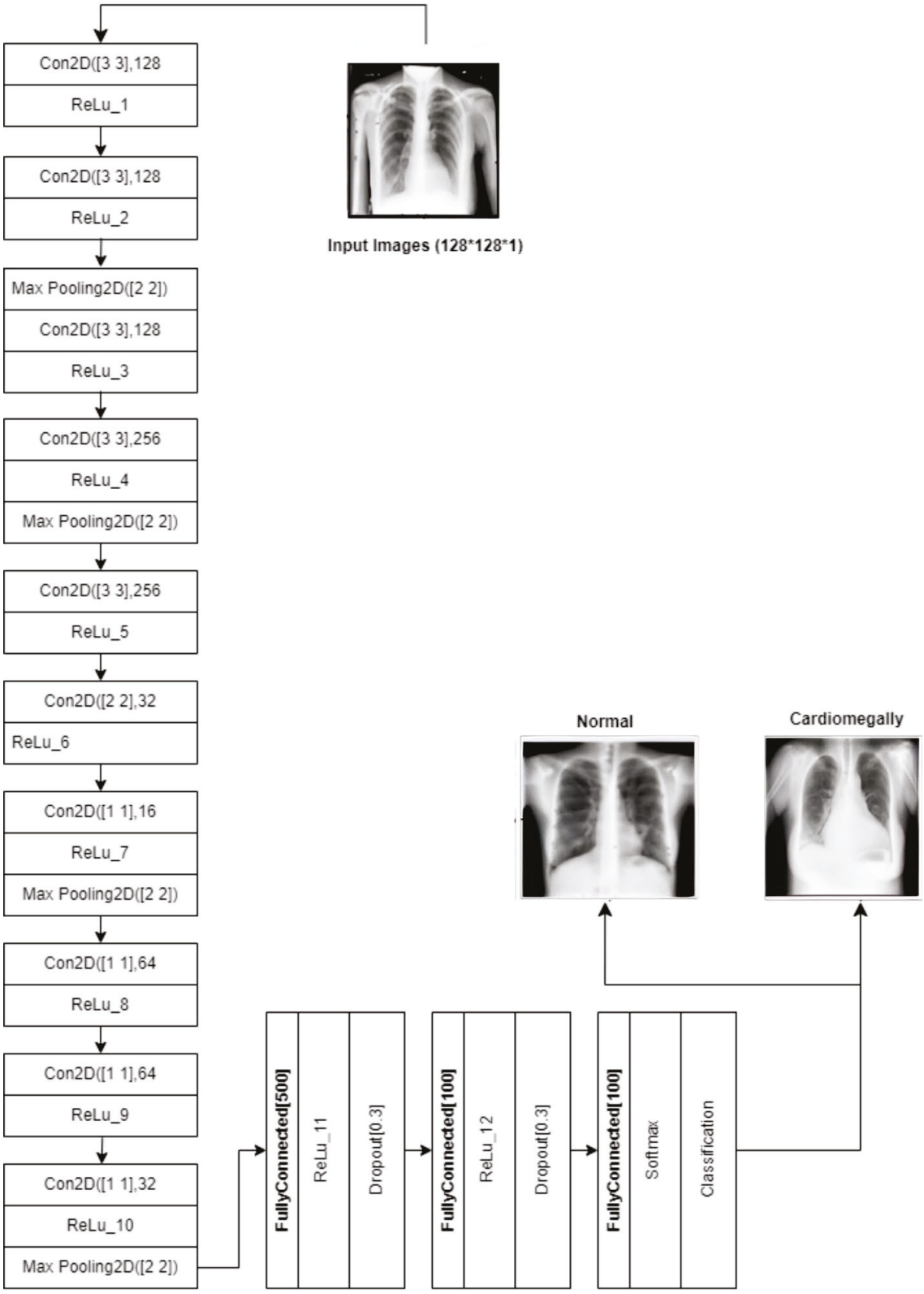


Figure 14. Diagram of the customized CNN.

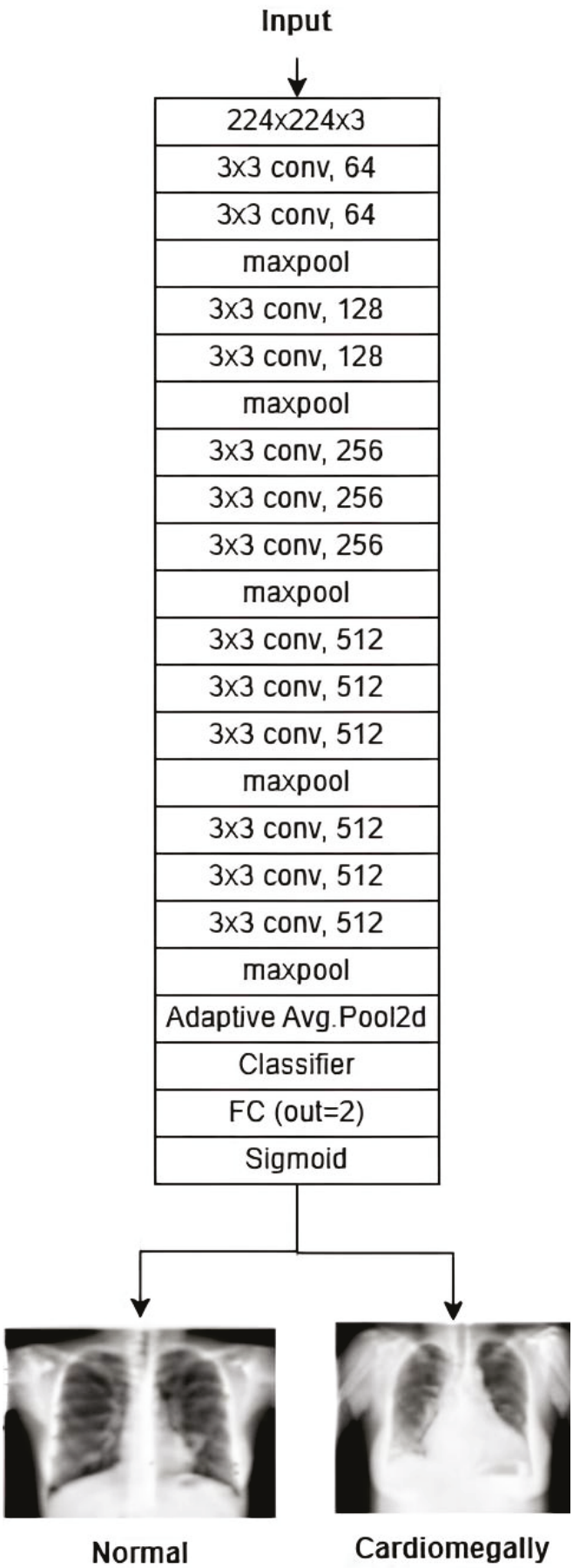


Figure 15. The structure of the adapted VGG16-based model.

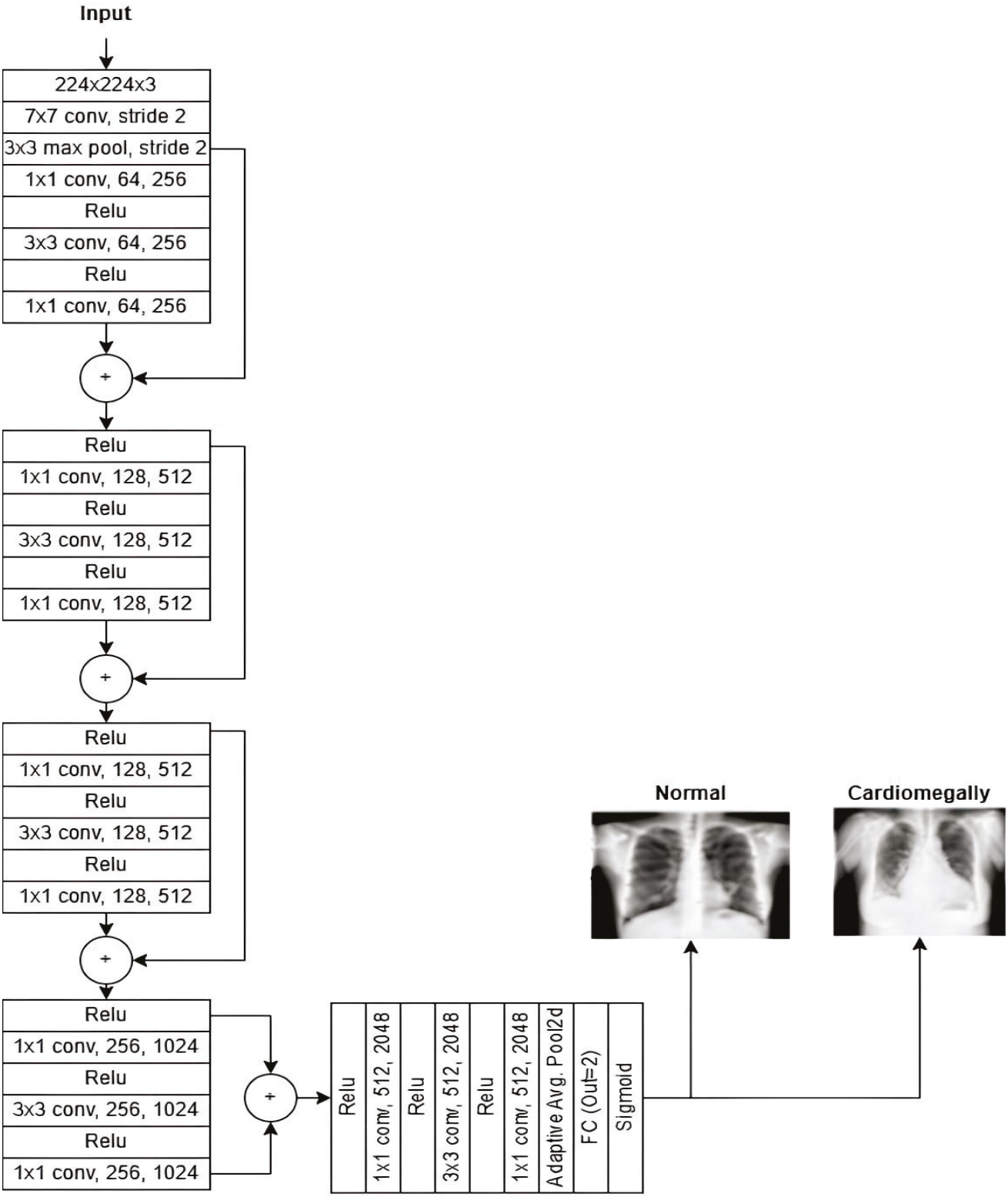


Figure 16. The structure of the adapted ResNet50-based model.

Another study used stacking to combine multiple classifiers to predict coronary heart disease. This was particularly effective in handling echocardiographic data's noisy and heterogeneous nature. The study demonstrated that stacking outperformed other methods by capturing hidden interactions among clinical and imaging features, thereby showcasing its ability to provide more precise and reliable disease diagnosis.

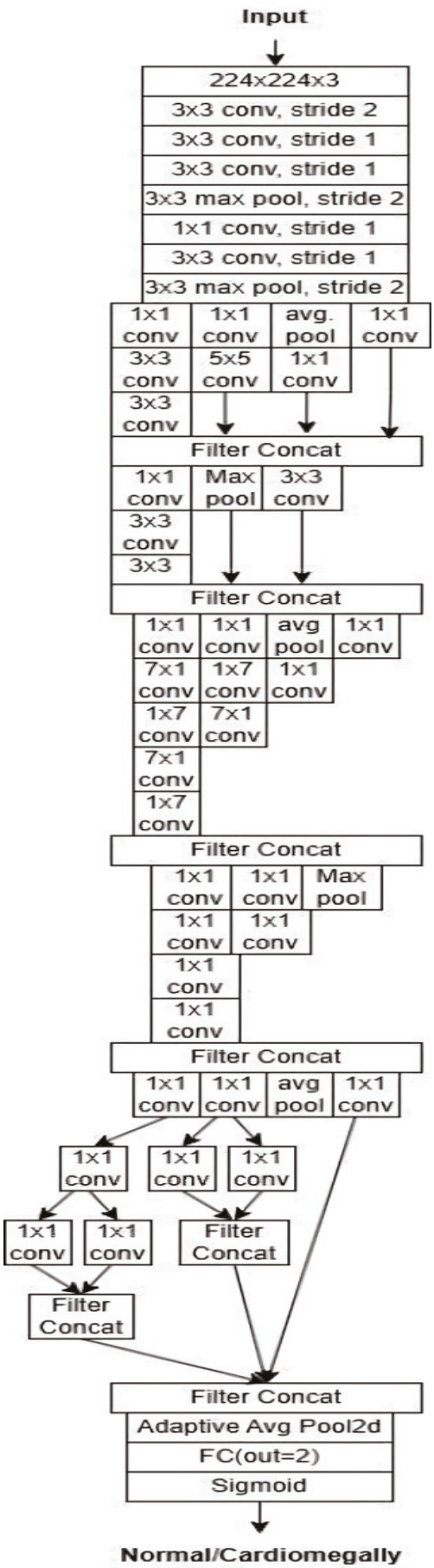


Figure 17. The structure of the adapted InceptionV3-based model.

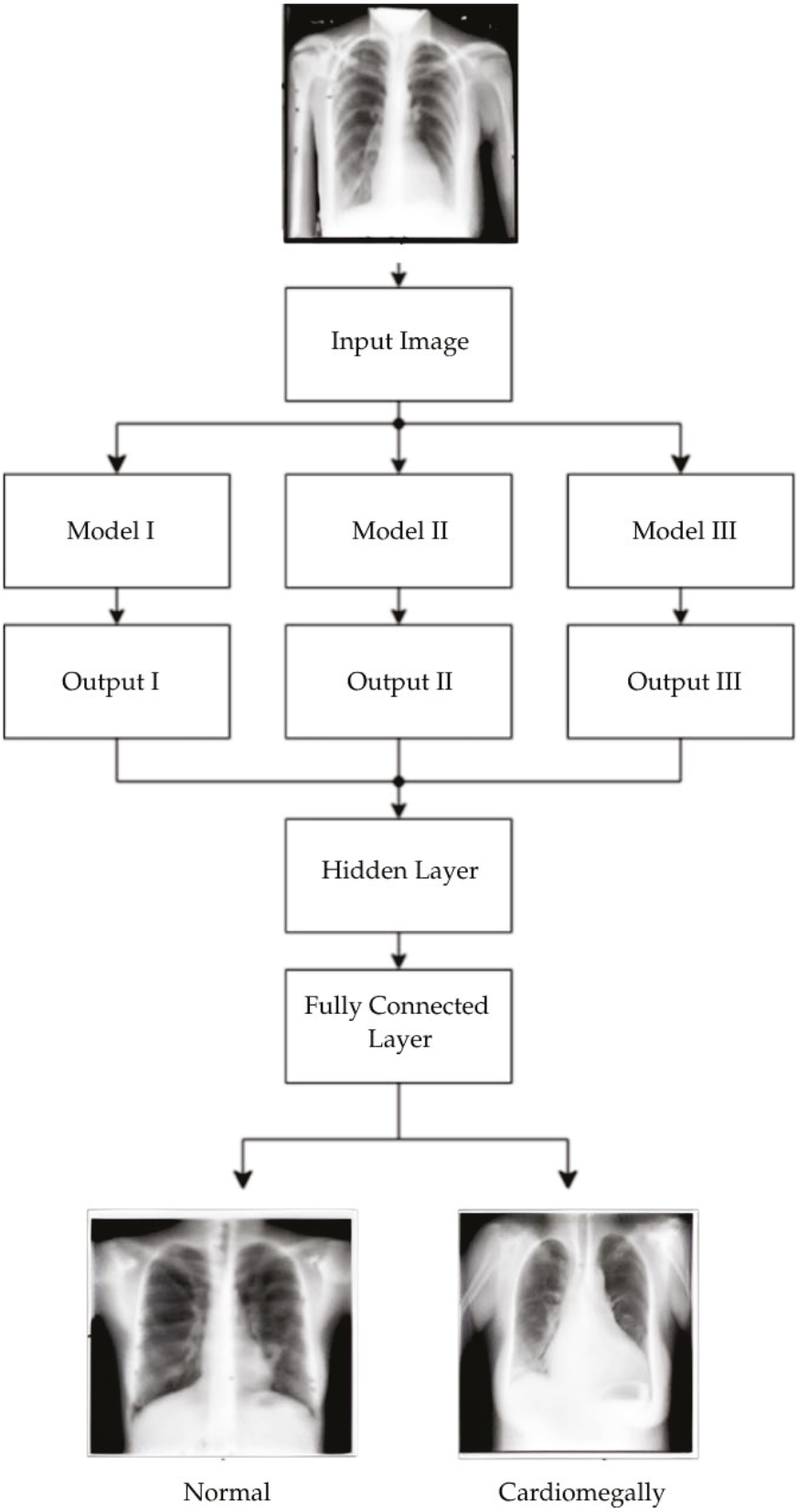


Figure 18. Workflow of our ensemble learning model.

These examples highlight the potential of stacking to address challenges in healthcare, such as integrating diverse data sources and improving prediction accuracy, thus making it a promising tool for advancing medical diagnostics [36].

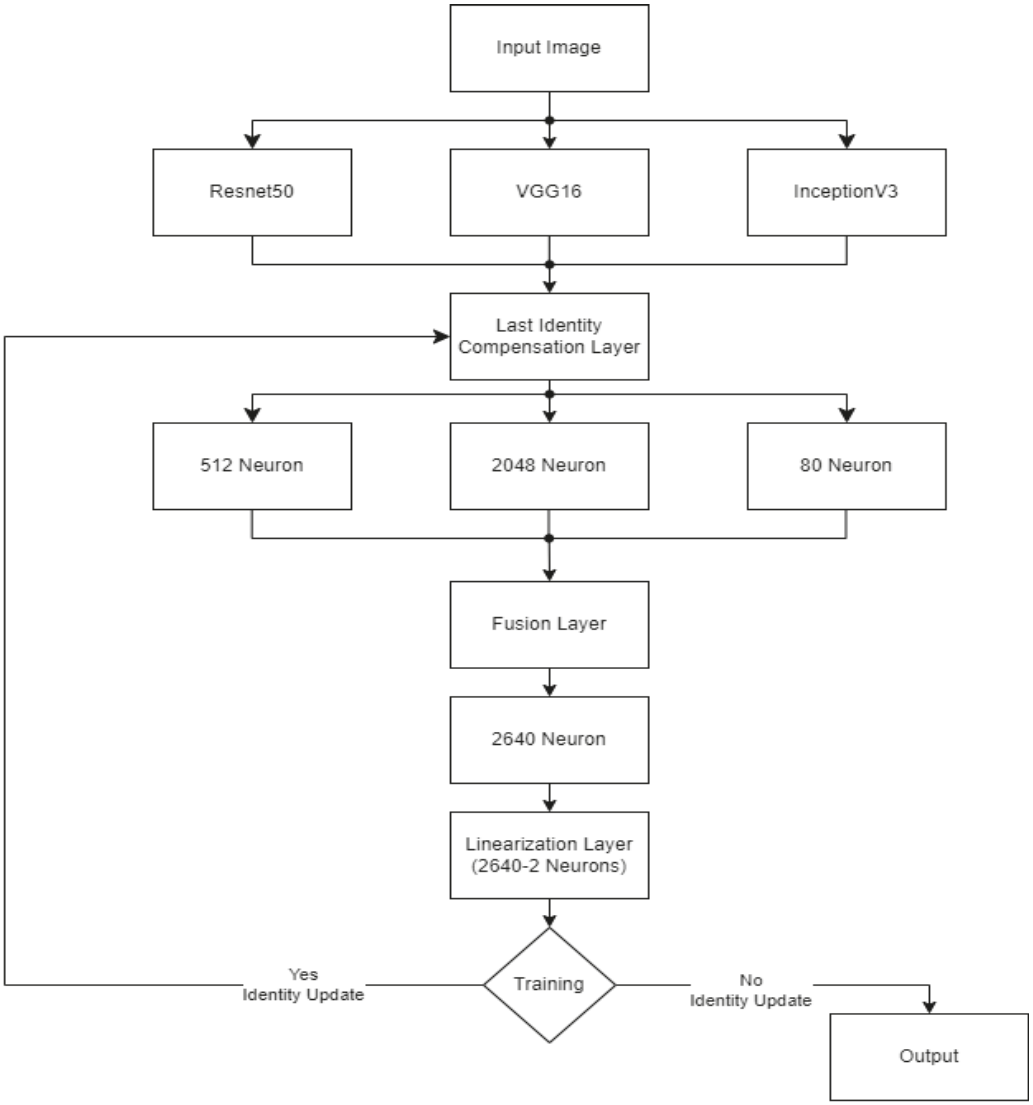


Figure 19. Details of our ensemble learning model.

The CELM (Cardiomegaly Ensemble Model) utilizes a stacking-based architecture that integrates three independently trained deep learning models—ResNet50, VGG16, and InceptionV3—to enhance the accuracy and robustness of cardiomegaly diagnosis from posteroanterior (PA) chest X-ray images. As illustrated in Figure 18, each model is coupled with a distinct fully connected (FC) mechanism: Spinal Fully Connected (Spinal FC) layers for ResNet50 and InceptionV3, and a standard FC layer for VGG16.

The analysis of the CELM, designed for classifying Chest PA (posteroanterior) X-ray images as normal or abnormal (cardiomegaly), reveals that three distinct CNN-based models are utilized. In the corresponding figure, "data" refers to the complete dataset of shoulder bone X-ray images. "Model I" represents the ResNet50 model integrated with SpinalNet FC, "Model II" corresponds to the VGG16 model with a Standard FC layer, and "Model III" denotes the InceptionV3 model combined with SpinalNet FC. The outputs of these models, labeled as "Outputs 1–3," contribute to the final "Prediction," which categorizes the images as either normal or abnormal (cardiomegaly).

Independent feature extraction and prediction part, each base model accepts a $224 \times 224 \times 3$ input image and extracts deep visual features through its convolutional and pooling layers [29]. Rather than directly outputting classification labels, the models are

modified to generate single scalar probability values using sigmoid activation functions, indicating the likelihood of the input image being classified as cardiomegaly.

Let the following be defined:

- P_1: the output from VGG16.
- P_2: the output from ResNet50.
- P_3: the output from InceptionV3.

These values are soft predictions and are not thresholded at this stage.

In the feature concatenation part, the scalar outputs from the three models are concatenated into a 1×3 feature vector:

$$P = [P_1, P_2, P_3] \in R^{1\times 3} \tag{1}$$

This unified prediction vector becomes the input to the meta-classifier, which learns how to optimally combine the base predictions to improve decision-making performance.

Meta-Classifier Architecture:

As depicted in Figure 19, the meta-classifier is a fully connected feed-forward neural network that transforms the 1×3 input vector into a final binary decision. The architecture includes the following layers:

- Input layer: Accepts the concatenated vector [P_1, P_2, P_3].
- Dense hidden layer: Contains 8 neurons, activated by the ReLU function.
- Dropout layer: Applied during training to reduce overfitting and promote generalization.
- Output layer: A single neuron with sigmoid activation that outputs the final probability of cardiomegaly.

The entire meta-classifier is trained using binary cross-entropy loss and optimized with the Adam optimizer.

In final prediction and thresholding stage, the sigmoid output from the meta-classifier is computed as follows:

$$\text{Pfinal} = \sigma(w_1P_1 + w_2P_2 + w_3P_3 + b) \tag{2}$$

where

- w_1, w_2, and w_3 are the learned weights;
- b is the bias term;
- $\sigma(\cdot)$ denotes the sigmoid function.

This output is then thresholded at 0.5 to obtain the final prediction:

- If Pfinal $\geq$ 0.5: classify as cardiomegaly;
- If Pfinal < 0.5: classify as normal.

To ensure transparency and enable full reproducibility, the internal structure of each base model (VGG16, ResNet50, InceptionV3) and the meta-classifier used in the CELM architecture are provided in Figures 15–19. The precise weights (w_1 = 0.35, w_2 = 0.42, w_3 = 0.23) and bias (b = -0.18) learned by the meta-classifier are detailed in Table 8, and can be reused to reconstruct the ensemble decision function.

After building the architecture, the entire CELM ensemble—modified CNNs, concatenation, and meta-classifier—is retrained end-to-end on the full dataset to align complementary features for better interpretability and diagnostic accuracy.

By combining diverse architectural features from VGG16's deep spatial resolution, ResNet50's residual connections, and InceptionV3's multi-scale processing, the CELM achieves superior results in detecting cardiomegaly, especially in ambiguous or borderline

cases. This stacking ensemble strategy leads to increased classification accuracy, precision, recall, and F1-score compared to single-model baselines.

Table 8. Summary of hyperparameter tuning.

Component	Parameter	Value/Description
Training Strategy	Tuning method	Manual grid search
	Optimizer	Adam
	Loss function	Binary Cross-Entropy
	Early stopping	Enabled (patience = 5 epochs)
	Epochs	50
	Batch size	32
	Learning rate	0.0001
Base CNN Models	Input size	$224 \times 224 \times 3$
	Pre-trained on	ImageNet
	Fine-tuned	Yes (last layers retrained)
	Architectures used	VGG16 (FC), ResNet50 (Spinal FC), InceptionV3 (Spinal FC)
Meta-Classifier	Input features	3 (P_1, P_2, P_3) from CNN outputs
	Hidden layer	8 neurons, ReLU activation
	Dropout rate	0.5
	Output layer	1 neuron, sigmoid activation
	Final prediction threshold	0.5
Learned Weights	w_1, w_2, w_3 (VGG16, ResNet50, InceptionV3)	0.35, 0.42, 0.23
	Bias term b	−0.18

3.4. Training and Validation

The models were trained using the Adam optimizer with a learning rate 0.001. The binary cross-entropy loss function was employed considering the binary nature of the classification task. To mitigate the issue of overfitting, early stopping and dropout were employed. Applying a five-fold cross-validation approach ensured the robustness of the performance evaluation. The following performance metrics were calculated: accuracy, precision, recall, and F1-score.

3.5. Hyperparameter Tuning

Hyperparameter tuning in this study was conducted through a manual grid search strategy, as opposed to automated techniques such as Bayesian optimization or random search. This approach allowed for domain-specific fine-tuning of critical parameters based on validation performance and clinical interpretability of the model predictions. The selected optimal hyperparameters across the base CNN models (VGG16, ResNet50, InceptionV3) were as follows: learning rate = 0.0001, batch size = 32, and number of epochs = 50. For all models, the Adam optimizer was used, and binary cross-entropy served as the loss function. The dropout rate applied to the meta-classifier was set to 0.5, and the hidden layer consisted of 8 neurons with ReLU activation. Early stopping with a patience of 5 epochs was also implemented to prevent overfitting. These hyperparameters were selected based on their ability to maximize validation accuracy and F1-score while ensuring stable convergence across folds. The combination of a manual grid search and the explicit reporting of architectural and training parameters, along with the public release of the source code, ensures that the results can be fully replicated by independent researchers.

To compute the best weights,

- The model makes a prediction: ypredict = Pfinal;

- The true label $y \in \{0, 1\}$ is compared to ypredict;
- The model calculates binary cross-entropy loss.

$$L(y, y_{predict}) = -y\log(y_{predict}) - (1 - y)\log(1 - y_{predict}) \quad (3)$$

Using gradient descent and the Adam optimizer, the model updates w_1, w_2, w_3, and b to minimize this loss. Over multiple epochs, the weights converge to values that best separate normal vs. cardiomegaly based on the combination of model outputs.

The final meta-classifier assigned weights of 0.35, 0.42, and 0.23 to the predictions from VGG16, ResNet50, and InceptionV3, respectively. The learned bias term was –0.18. These values indicate that ResNet50 had slightly greater influence in the ensemble decision, though all three base models contributed meaningfully to the final classification. All hyperparameters are also summarized in Table 8.

3.6. Performance Metrics and Graphical Representations

We presented the findings from applying various customized deep learning models and compared their performance using key evaluation metrics, including recall, accuracy, F1-score, precision, specificity, and ROC AUC. Additionally, the accuracy and loss curves illustrate each model's training and validation accuracy and loss across epochs, providing insights into their learning behavior and generalization capabilities.

- Accuracy: The most intuitive performance metric is the ratio of correctly predicted samples to the total number of samples [29].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

- Precision: It is calculated as the percentage of correctly predicted positive samples, whether true or false, out of all actual positive samples.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

- Recall: The percentage of correctly predicted positive samples relative to the total number of positive and false negative instances [29].

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

- F1-score: It is calculated as the harmonic mean of precision and recall.

$$\text{F1} - \text{score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

- Specificity: The ability of the model to correctly identify negative cases (true negatives) [29].

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (8)$$

- AUC: Represents the model's overall ability to distinguish between classes. AUC is calculated as the average of recall (sensitivity) and specificity in this study.

$$\text{AUC} = \frac{Recall + Specificity}{2} \quad (9)$$

4. Results

The computer for conducting the experiments was equipped with robust hardware specifications, ensuring smooth and efficient processing of data-intensive tasks. The computer was equipped with a substantial 64 GB of RAM, which provided ample memory for the execution of multiple tasks and the handling of large datasets without the occurrence of performance bottlenecks. In terms of storage, the system included a 1 TB hard disk, which provided sufficient space to accommodate extensive datasets and experimental results.

The processor was an 11th-generation Intel Core i7-11850H, operating at a base clock speed of 2.50 GHz, thereby providing the requisite powerful computing capabilities for executing complex calculations and simulations. Furthermore, the computer was equipped with an NVIDIA GeForce RTX 3070 graphics card. The high-performance GPU facilitated accelerated processing, particularly in machine learning applications, where tasks such as model training and data analysis benefit significantly from its parallel computing architecture and optimized performance.

The aforementioned specifications enabled us to conduct our experiments efficiently, ensuring reliability and speed in all computational aspects of our analysis.

The programme codes written for this study are publicly available at https://github.com/eyanarai/CELM (accessed on 7 April 2025). All model definitions, training scripts, and pre-trained weights are publicly shared at (https://github.com/eyanarai/CELM), supporting the full reproducibility of the proposed pipeline.

Six recommended techniques for identifying and diagnosing cardiomegaly were rigorously evaluated using X-ray images and established evaluation criteria. We employed VGG16, ResNet50, InceptionV3, DenseNet121, DenseNet201, and AlexNet, utilizing pre-trained models for feature extraction in conjunction with sigmoid/Adams classifiers. The performance metrics for validation and testing, including accuracy, precision, recall, F1-score, specificity, and AUC for recognizing cardiomegaly, are presented in Table 9 below [29]:

Table 9. A comparison of existing methods and our proposed method performances in cardiomegaly detection.

Model	Accuracy	Precision	Recall	F1-Score	Specificity	AUC
VGG16	0.93	0.95	0.86	0.90	0.76	0.81
ResNet50	0.92	0.94	0.85	0.89	0.73	0.79
VIT-B-16	0.90	0.89	0.84	0.86	0.60	0.72
InceptionV3	0.90	0.95	0.85	0.90	0.77	0.81
DenseNet121	0.92	0.89	0.83	0.86	0.62	0.73
DenseNet201	0.93	0.93	0.87	0.89	0.67	0.77
AlexNet	0.91	0.93	0.85	0.89	0.70	0.78
Customized CNN	0.91	0.93	0.82	0.87	0.74	0.78
CELM (Proposed Model)	0.92	0.99	0.89	0.94	0.92	0.90

In all experiments, image enhancement techniques were employed to enhance low-intensity contrast, and noise filtering was utilized, resulting in increased accuracy compared to the initial dataset. This study concentrated on the binary classification of cardiomegaly and healthy individuals using various classification-based pre-trained models and our ensemble model [29]. Among the methods tested, CELM (our proposed method) demonstrated superior performance compared to the others.

As shown in Table 10, the CELM ensemble outperforms prior cardiomegaly classifiers with accuracy (0.92), precision (0.99), recall (0.89), and F1-score (0.94). Evaluated on 50,000 diverse CXR images, CELM ensures robustness in real-world settings. Its balanced sensitivity (0.89) and specificity (0.92) reduce false negatives and positives, vital for clinical safety. With an AUC of 0.90, it demonstrates strong discrimination. CELM's efficient ensemble design also supports scalable deployment in high-volume radiology, making it a promising tool for cardiomegaly screening.

Table 10. A comparative analysis of existing methods for detecting cardiomegaly.

Authors	Classes	Methods	Image Number	Accuracy	Precision	Recall	F1-Score	Specificity	AUC
(Gupte et al., 2021) [15]	Cardio vs. Other (mixed findings)	CNN	1440	0.94	0.83	-	0.88	0.83	-
(Yoo et al., 2021) [14]	Cardiomegaly vs. Normal (isolated)	CNN	2000	0.80	-	-	-	-	-
(Sogancioglu et al., 2020) [13]	Cardiomegaly vs. Normal (isolated)	CNN	67,310	0.97	0.90	0.97	-	0.90	0.97
(Innat et al., 2023) [37]	Cardio vs. Other (mixed findings)	CNN	30,805	0.85	0.87	0.85	0.86	-	0.89
(Raghu Kumar et al., 2023) [11]	Cardio vs. Other (details not specified)	CNN	35,038	0.95	0.95	0.95	0.95	0.95	0.95
(Que et al., 2018) [10]	Cardiac vs. Thoracic disease classes.	VGG16	2630	0.93	1.0	0.89	0.94	1.00	0.94
(Candemir et al., 2018) [16]	Cardio vs. Other (mixed findings)	AlexNet, VGG-16, VGG-19 and Inception V3	4000	0.88	-	0.94	0.88	0.93	0.94
(Decoodt et al., 2023) [17]	Cardio vs. Other (mixed findings)	Dense Net-121, AlexNet	224,316	0.87	0.84	0.93	-	-	0.93
(Ayalew et al., 2024) [12]	Cardiomegaly vs. Normal (isolated)	CNN, Inception, DenseNet-169, ResNet-50	4400	0.99	1.00	1.00	1.00	1.00	1
CELM (Proposed Model)	Cardio (with or without co-findings) vs. Non-Cardio (Normal + other findings)	CNN, InceptionV3, VGG16, ResNet-50	50,000	0.92	0.99	0.89	0.94	0.92	0.90

Figure 20 shows the CELM correspondence matrix. Figure 21 shows the CELM ROC Curve graph.

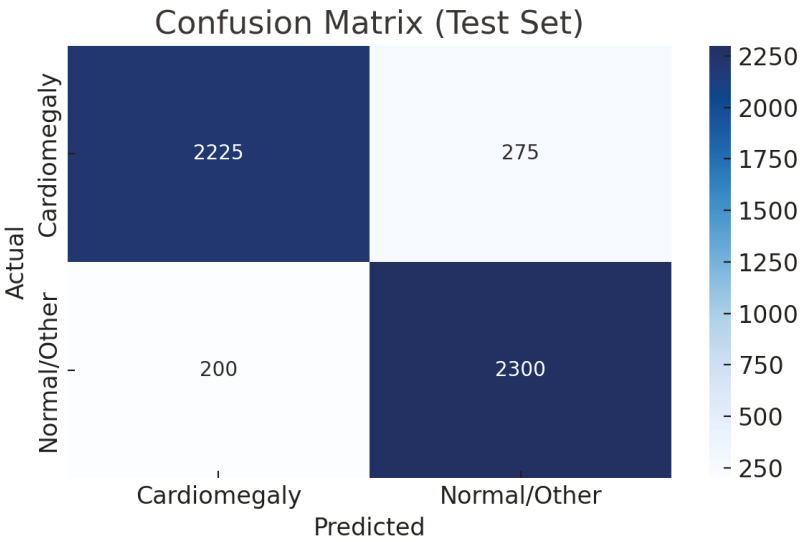


Figure 20. CELM confusion matrix of the test set.

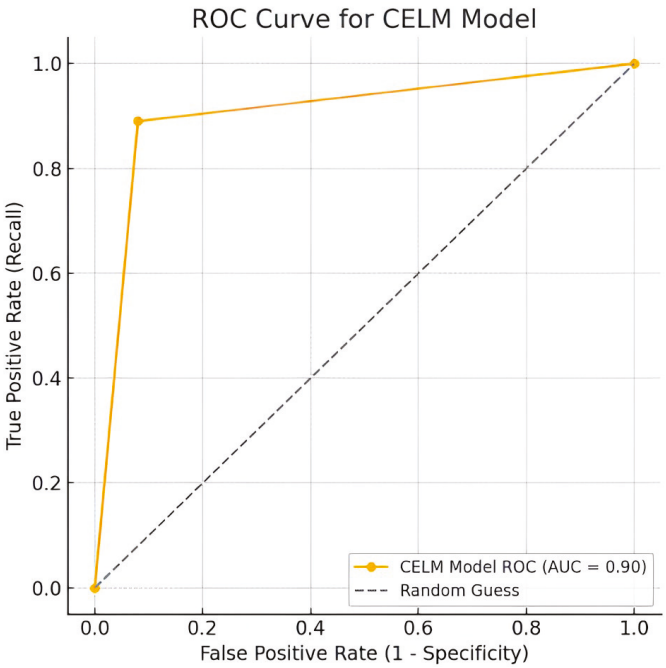


Figure 21. ROC Curve for the CELM.

5. Discussion

This study presents a deep learning-based framework for the automated detection of cardiomegaly from chest X-rays (CXRs), proposing a stacked ensemble model—CELM—that integrates VGG16, ResNet50, and InceptionV3 through a shallow neural network meta-classifier. The CELM demonstrated a high diagnostic performance, achieving 92% accuracy, 0.99 precision, 0.89 recall, and 0.90 AUC, making it reliable for clinical screening.

The model's high precision, recall, and specificity make it suitable for fast, sensitive screening in clinical settings like emergency and ICU. Integrating the CELM could aid radiologists with quick, standardized assessments for earlier intervention and better triage. However, clinical reasoning involving patient history and multimodal data remains essential beyond imaging.

Stacking was chosen as the ensemble strategy based on its demonstrated ability to model non-linear interactions between CNN outputs through a learnable meta-classifier. In comparative tests, alternative methods such as soft voting and hard voting produced marginally lower accuracy and F1-scores, further validating stacking as the optimal ensemble approach despite a modest increase in computational complexity.

Although Ayalew et al. [12] reported slightly higher performance (99.8% accuracy), their model was evaluated on a relatively small and homogeneous dataset primarily composed of clearly separable cardiomegaly and normal cases. Such conditions may artificially inflate model performance and limit external generalizability. In contrast, the CELM was trained and validated on a large-scale, multi-institutional dataset aggregated from PadChest, NIH CXR, VinDr-CXR, and CheXpert—capturing a broad spectrum of image quality, acquisition settings, and patient variability—thereby enhancing the model's real-world relevance. However, the datasets were merged prior to splitting into training and test sets. While this strategy improves the model's robustness by exposing it to heterogeneous data, it may mask domain shift effects. In real-world deployment, models often encounter distributional differences across institutions. Thus, the absence of explicit cross-dataset validation remains a limitation that needs to be addressed.

In Table 10, we present a transparent comparison of datasets and evaluation methodologies across prior studies, clearly contrasting them with our own. While some previous

works included datasets reflecting greater clinical heterogeneity, our dataset was intentionally curated to prioritize image quality and diagnostic clarity. We acknowledge that this approach may lead to slightly optimistic performance estimates and that further validation using unfiltered, real-world imaging data is necessary. Excluding low-resolution, poorly exposed, or truncated images improved training convergence but may limit generalization to routine clinical environments, where such imperfections are commonly encountered.

Moreover, the exclusion of low-resolution or artifact-laden images, while enhancing training stability, may reduce the model's ability to generalize across diverse real-world imaging conditions. Since clinical workflows often involve suboptimal images due to patient motion, positioning errors, or equipment variability, future validation on unfiltered datasets remains a critical step.

Ultimately, this study highlights the promise of deep learning-based ensemble models as trustworthy, scalable clinical decision-support tools for cardiomegaly detection. CELM exemplifies how the integration of complementary architectures can enhance diagnostic performance while maintaining scalability, automation, and interpretability. These findings contribute to broader efforts in AI-driven healthcare to deliver high-accuracy, generalizable, and clinically relevant systems capable of improving diagnostic workflows and patient outcomes.

6. Future Work

Despite its strong diagnostic performance, the CELM has several limitations that must be addressed to enhance its clinical applicability and trustworthiness. First, the current evaluation relies solely on retrospective and curated datasets, which were filtered to exclude low-quality, overexposed, or truncated images. While this improves the signal-to-noise ratio in training, it does not reflect the variability typically encountered in everyday clinical radiology. In future iterations of this research, we aim to incorporate unfiltered datasets—including poor-quality and borderline cases—to rigorously assess model resilience in real-world conditions. Second, the model does not integrate patient-specific metadata such as age, clinical history, or comorbidities—information that could substantially enrich the diagnostic context and improve accuracy in borderline or ambiguous cases. Third, while the stacking ensemble framework improves classification performance, it introduces additional computational overhead, which may pose deployment challenges in resource-limited environments such as rural clinics or mobile imaging units. Finally, the CELM currently lacks integrated explainability mechanisms (e.g., Grad-CAM or saliency maps), which limits transparency and interpretability of predictions—key factors for clinical adoption.

To address these limitations, future research will focus on enhancing the CELM's clinical readiness, robustness, and interpretability. Specific priorities include

(1) Expanding model validation to unfiltered, real-world datasets that include suboptimal image quality and heterogeneous disease presentations;
(2) Conducting prospective, multi-center studies to evaluate performance across diverse patient populations and imaging settings;
(3) Incorporating structured clinical metadata (age, gender, etc.) to support more nuanced, patient-centered diagnostics;
(4) Integrating explainable AI techniques to provide visual justification for predictions and increase clinician trust.

In future work, we plan to implement cross-dataset experiments where the model is trained on one dataset (e.g., PadChest) and tested on another (e.g., CheXpert). This will allow us to quantify the impact of domain-specific imaging characteristics and assess generalizability across institutions and acquisition protocols.

These efforts aim to narrow the translational gap between AI development and clinical practice, supporting the creation of diagnostic tools that are not only accurate and scalable but also transparent, adaptive, and trustworthy—key attributes for effective implementation in real-world cardiomegaly screening and broader medical imaging applications.

7. Conclusions

This study presents a deep learning-based framework for the automated detection of cardiomegaly from chest X-rays (CXRs), introducing a stacked ensemble model (CELM) that integrates VGG16, ResNet50, and InceptionV3 architectures through a shallow neural network meta-classifier. Trained on a curated and balanced dataset of 50,000 images drawn from four publicly available sources—PadChest, NIH CXR, VinDr-CXR, and CheXpert—the proposed model achieved 92% accuracy, 99% precision, 89% recall, 0.92 specificity, and an AUC of 0.90, clearly outperforming individual CNN models and confirming the effectiveness of the stacking ensemble approach in medical image classification.

The CELM's real-world applicability is supported by its strong performance on diverse datasets. Additionally, by leveraging transfer learning and data augmentation, the CELM demonstrates scalability and adaptability to deployment in resource-constrained settings.

Some limitations exist. The model was trained on curated retrospective datasets, excluding low-quality images, which may limit generalizability to real clinical settings. The CELM does not use patient metadata like age or symptoms, thereby reducing the diagnostic context. The stacking framework adds computational load, challenging real-time use on low-power devices. Also, it lacks explainability tools (e.g., Grad-CAM), important for clinician trust and interpretability.

Future work will focus on addressing these limitations by evaluating the CELM in prospective clinical trials, incorporating heterogeneous and uncrated datasets, integrating patient metadata, and implementing explainable AI techniques. These efforts aim to enhance the CELM's clinical readiness, transparency, and reliability, further establishing its role as a scalable and trustworthy diagnostic support tool for cardiomegaly screening and broader medical imaging applications.

8. Innovative Contribution

This study introduces an innovative deep learning framework centered around a novel stacking ensemble model, CELM, specifically designed for the early detection of cardiomegaly from chest X-rays. By integrating the strengths of multiple pretrained CNN architectures—VGG16, ResNet50, and InceptionV3—the CELM achieves a superior feature extraction and decision-making performance compared to individual models. Leveraging a large-scale, multi-source CXR dataset, the proposed ensemble framework demonstrates significant improvements over traditional diagnostic approaches, both in terms of accuracy and clinical applicability. The results clearly show that while individual pretrained models perform reasonably well, they are consistently outperformed by the ensemble approach in all key evaluation metrics. This underscores the value of model synergy in high-stakes clinical tasks. The effective incorporation of such AI-driven ensemble systems into clinical workflows has the potential to revolutionize cardiomegaly screening, enabling earlier interventions and ultimately contributing to improved patient outcomes in cardiovascular care.

Author Contributions: E.Y.: methodology, software, data curation, validation, formal analysis, investigation, visualization, writing—original draft preparation, writing—review and editing; F.H.: conceptualization, validation, writing—review and editing, supervision, project administration; K.A.: validation, writing—original draft preparation, writing—review and editing, formal analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the ASELSAN Inc.

Institutional Review Board Statement: The study did not require ethical approval.

Informed Consent Statement: Not applicable.

Data Availability Statement: The authors used an open access dataset that is available from. CXR-VinDr: https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection/data (accessed on 10 June 2025), PadChest: https://bimcv.cipf.es/bimcv-projects/padchest/ (accessed on 10 June 2025), CXR-NIH: https://www.kaggle.com/datasets/nih-chest-xrays/data (accessed on 10 June 2025), CheXpert: https://aimi.stanford.edu/datasets/chexpert-chest-x-rays (accessed on 10 June 2025).

Acknowledgments: We thank the referee for his comprehensive review and appreciate his comments, corrections, and suggestions, which contributed significantly to improving the quality of the publication.

Conflicts of Interest: Author Erdem Yanar was employed by the ASELSAN. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AHE	Adaptive Histogram Equalization
AI	Artificial Intelligence
BNP	Brain Natriuretic Peptide
CELM	Combined Ensemble Learning Model
CLAHE	Contrast-Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
CQ	Classical-Quantum
CTR	Cardiothoracic Ratio
CVD	Cardiovascular Disease
CVDs	Cardiovascular Diseases
CXR	Chest X-ray
CXRs	Chest X-rays
ECG	Electrocardiogram
EDs	Emergency Departments
GradCAM	Gradient-weighted Class Activation Mapping
ICUs	Intensive Care Units
NLP	Natural Language Processing
NTproBNP	N-Terminal Probnp
PA	Posteroanterior
ResNet50	Residual Network 50
UMLS	Unified Medical Language System

References

1. Lindstrom, M.; DeCleene, N.; Dorsey, H.; Fuster, V.; Johnson, C.O.; LeGrand, K.E.; Mensah, G.A.; Razo, C.; Stark, B.; Turco, J.V.; et al. Global Burden of Cardiovascular Diseases and Risks Collaboration, 1990–2021. *J. Am. Coll. Cardiol.* **2022**, *80*, 2372–2425. [CrossRef] [PubMed]
2. Colizzi, M.; Lasalvia, A.; Ruggeri, M. Prevention and early intervention in youth mental health: Is it time for a multidisciplinary and trans-diagnostic model for care? *Int. J. Ment. Health Syst.* **2020**, *14*, 23. [CrossRef] [PubMed]
3. Mayo Clinic. Enlarged Heart. 2022. Available online: https://www.mayoclinic.org/diseases-conditions/enlarged-heart/symptoms-causes/syc-20355436 (accessed on 1 December 2024).
4. Cao, Z.; Jia, Y.; Zhu, B. BNP and NT-proBNP as diagnostic biomarkers for cardiac dysfunction in both clinical and forensic medicine. *Int. J. Mol. Sci.* **2019**, *20*, 1820. [CrossRef]

5. Martin-Isla, C.M.; Campello, V.; Izquierdo, C.; Raisi-Estabragh, Z.; Baeßler, B.E.; Petersen, S.; Lekadir, K. Image-Based Cardiac Diagnosis with Machine Learning: A Review. *Front. Cardiovasc. Med.* **2020**, *7*, 1. [CrossRef] [PubMed]
6. Amin, H.; Siddiqui, W.J. Cardiomegaly. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2008; pp. 154–196. Available online: https://www.ncbi.nlm.nih.gov/books/NBK542296/ (accessed on 10 June 2025).
7. Truszkiewicz, K.; Poręba, R.; Gać, P. Radiological Cardiothoracic Ratio in Evidence-Based Medicine. *J. Clin. Med.* **2021**, *10*, 2016. [CrossRef]
8. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.S.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying medical diagnoses and treatable diseases by image-based Deep Learning. *Cell* **2018**, *172*, 1122–1131.e9. [CrossRef]
9. Mohare, S.; Dawani, H.; Rathi, V.; Oswal, S. Detection of Cardiomegaly from Chest X-rays using Otsu Algorithm and Convolutional Neural Network. In Proceedings of the 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 29–31 March 2019; pp. 1–5. [CrossRef]
10. Que, Q.; Tang, Z.; Wang, R.; Zeng, Z.; Wang, J.; Chua, M.; Gee, T.S.; Yang, X.; Veeravalli, B. CardioXNet: Automated Detection for Cardiomegaly Based on Deep Learning. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 612–615. [CrossRef]
11. Raghu Kumar, L.; Sravanthi, K.; Kiran, E.S.; Vinith, D.; Siri, D.; Joshi, S.K. Detecting cardiomegaly from CXR images using a 2D and 1d convolutional neural network-based classifier. *E3S Web Conf.* **2023**, *430*, 01156. [CrossRef]
12. Ayalew, A.M.; Enyew, B.; Bezabh, Y.A.; Abuhayi, B.M.; Negashe, G.S. Early stage Cardiomegaly Detection and Classification from X-ray Images Using Convolutional Neural Networks and Transfer Learning. *Intell. Syst. Appl.* **2024**, *24*, 200453. [CrossRef]
13. Sogancioglu, E.; Murphy, K.; Calli, E.; Scholten, E.T.; Schalekamp, S.; Van Ginneken, B. Cardiomegaly detection on chest radiographs: Segmentation versus classification. *IEEE Access* **2020**, *8*, 94631–94642. [CrossRef]
14. Yoo, H.; Han, S.; Chung, K. Diagnosis support model of Cardiomegaly based on CNN using ResNet and explainable feature map. *IEEE Access* **2021**, *9*, 55802–55813. [CrossRef]
15. Gupte, T.; Niljikar, M.; Gawali, M.; Kulkarni, V.; Kharat, A.; Pant, A. Deep learning models for calculation of cardiothoracic ratio from chest radiographs for assisted diagnosis of Cardiomegaly. In Proceedings of the 2021 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), Durban, South Africa, 5–6 August 2021; pp. 1–6. [CrossRef]
16. Candemir, S.; Rajaraman, S.; Thoma, G.; Antani, S. Deep learning for grading cardiomegaly severity in chest X-rays: An investigation. In Proceedings of the 2018 IEEE Life Sciences Conference (LSC), Montreal, QC, Canada, 28–30 October 2018; IEEE: New York, NY, USA; pp. 109–113. [CrossRef]
17. Decoodt, P.; Liang, T.J.; Bopardikar, S.; Santhanam, H.; Eyembe, A.; Garcia-Zapirain, B.; Sierra-Sosa, D. Hybrid classical–quantum transfer learning for cardiomegaly detection in chest X-rays. *J. Imaging* **2023**, *9*, 128. [CrossRef]
18. Chouhan, S.S.; Singhai, J.; Khamparia, A. A novel transfer learning based stacked ensemble for pneumonia detection in chest X-ray images. *Appl. Soft Comput.* **2021**, *98*, 106885. [CrossRef]
19. Hussain, M.; Bird, J.J.; Faria, D.R. A Study on CNN Transfer Learning for Image Classification. *Symmetry* **2021**, *13*, 998. [CrossRef]
20. Majkowska, A.; Mittal, S.; Steiner, D.F.; Reicher, J.J.; McKinney, S.M.; Duggan, G.E.; Eswaran, D.; Peng, L.; Stumpe, M.C. Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiol. Artif. Intell.* **2020**, *2*, e190145. [CrossRef] [PubMed]
21. NIH Chest X-Rays. Available online: https://www.kaggle.com/datasets/nih-chest-xrays/data (accessed on 1 December 2024).
22. ChestX-Ray8. Available online: https://paperswithcode.com/dataset/chestx-ray8 (accessed on 10 June 2025).
23. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. Chestx-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3462–3471. [CrossRef]
24. CheXpert: Chest X-Rays. Available online: https://aimi.stanford.edu/datasets/chexpert-chest-x-rays (accessed on 1 December 2024).
25. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 590–597. [CrossRef]
26. VinBigData Chest X-Ray Abnormalities Detection. Available online: https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection/data (accessed on 1 December 2024).
27. Nguyen, H.Q.; Lam, K.; Le, L.T.; Pham, H.H.; Tran, D.Q.; Nguyen, D.B.; Le, D.D.; Pham, C.M.; Tong, H.T.T.; Dinh, D.H.; et al. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Sci. Data* **2022**, *9*, 429. [CrossRef] [PubMed] [PubMed Central]
28. CXR-PadChest. Available online: https://bimcv.cipf.es/bimcv-projects/padchest/ (accessed on 10 June 2025).
29. Yanar, E.; Hardalaç, F.; Ayturan, K. PELM: A Deep Learning Model for Early Detection of Pneumonia in Chest Radiography. *Appl. Sci.* **2025**, *15*, 6487. [CrossRef]

30. Bustos, A.; Pertusa, A.; Salinas, J.-M.; de la Iglesia-Vayá, M. Padchest: A large chest X-ray image dataset with multi-label annotated reports. *Med. Image Anal.* **2020**, *66*, 101797. [CrossRef]
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and 559 Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [CrossRef]
32. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the 2016 IEEE 561 Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]
33. Huang, W.; Peng, G.; Tang, X. A limit of densely connected convolutional networks V1. *Tech. Rep.* **2019**. [CrossRef]
34. Tasci, E.; Zhuge, Y.; Kaur, H.; Camphausen, K.; Krauze, A.V. Hierarchical Voting-Based Feature Selection and Ensemble Learning Model Scheme for Glioma Grading with Clinical and Molecular Characteristics. *Int. J. Mol. Sci.* **2022**, *23*, 14155. [CrossRef]
35. Mahajan, P.; Uddin, S.; Hajati, F.; Moni, M.A. Ensemble Learning for Disease Prediction: A Review. *Healthcare* **2023**, *11*, 1808. [CrossRef]
36. Yoon, T.; Kang, D. Multi-Modal Stacking Ensemble for the Diagnosis of Cardiovascular Diseases. *J. Pers. Med.* **2023**, *13*, 373. [CrossRef] [PubMed]
37. Innat, M.; Hossain, M.F.; Mader, K.; Kouzani, A.Z. A convolutional attention mapping deep neural network for classification and localization of cardiomegaly on chest X-rays. *Sci. Rep.* **2023**, *13*, 6247. [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.

diagnostics

MDPI

Article

A Hybrid Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM)–Attention Model Architecture for Precise Medical Image Analysis and Disease Diagnosis

Md. Tanvir Hayat [1], Yazan M. Allawi [2,*], Wasan Alamro [3], Salman Md Sultan [1], Ahmad Abadleh [4], Hunseok Kang [5] and Aymen I. Zreikat [5]

1 Innovative Skills Ltd., Dhaka 1207, Bangladesh; tanvir@innovativeskillsbd.com (M.T.H.); salman@innovativeskillsbd.com (S.M.S.)
2 Department of Electrical Engineering, College of Engineering, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
3 Department of Communications and Computer Engineering, Faculty of Engineering, Al-Ahliyya Amman University, Amman 19111, Jordan; w.alamro@ammanu.edu.jo
4 CS Department, Mutah University, Karak 61710, Jordan; ahmad_a@mutah.edu.jo
5 College of Engineering and Technology, American University of the Middle East, Egaila 54200, Kuwait; hunseok.kang@aum.edu.kw (H.K.); aymen.zreikat@aum.edu.kw (A.I.Z.)
* Correspondence: ymallawi@pnu.edu.sa

Abstract: Background: Deep learning (DL)-based medical image classification is becoming increasingly reliable, enabling physicians to make faster and more accurate decisions in diagnosis and treatment. A plethora of algorithms have been developed to classify and analyze various types of medical images. Among them, Convolutional Neural Networks (CNNs) have proven highly effective, particularly in medical image analysis and disease detection. **Methods:** To further enhance these capabilities, this research introduces MediVision, a hybrid DL-based model that integrates a vision backbone based on CNNs for feature extraction, capturing detailed patterns and structures essential for precise classification. These features are then processed through Long Short-Term Memory (LSTM), which identifies sequential dependencies to better recognize disease progression. An attention mechanism is then incorporated that selectively focuses on salient features detected by the LSTM, improving the model's ability to highlight critical abnormalities. Additionally, MediVision utilizes a skip connection, merging attention outputs with LSTM outputs along with Grad-CAM heatmap to visualize the most important regions of the analyzed medical image and further enhance feature representation and classification accuracy. **Results:** Tested on ten diverse medical image datasets (including, Alzheimer's disease, breast ultrasound, blood cell, chest X-ray, chest CT scans, diabetic retinopathy, kidney diseases, bone fracture multi-region, retinal OCT, and brain tumor), MediVision consistently achieved classification accuracies above 95%, with a peak of 98%. **Conclusions:** The proposed MediVision model offers a robust and effective framework for medical image classification, improving interpretability, reliability, and automated disease diagnosis. To support research reproducibility, the codes and datasets used in this study have been publicly made available through an open-access repository.

Keywords: attention mechanism; convolutional neural networks; long short-term memory; medical image analysis; sequential features

1. Introduction

1.1. Overview

The integration of Machine Learning (ML) into medical practice has revolutionized the healthcare field, enabling computers to analyze vast amounts of patients data and derive insights into solving complex clinical problems without explicit programming. As a subset of ML specialized in utilizing artificial neural networks with multiple layers to analyze data, Deep Learning (DL) is playing an increasingly important role in enhancing treatment strategies for chronic conditions, supporting precision medicine, and expediting clinical trials [1,2] and promising significant advances in areas such as diagnostic processes [3–5], prognosis prediction [6–8], and drug discovery [4,5].

Among these applications, medical image classification presents unique challenges due to various factors. The variability in medical images, due to differences in imaging techniques, patient demographics, and the presence of artifacts, complicates the development of reliable and robust models [9,10]. To elaborate, medical images often contain subtle yet critical details that are essential for accurate diagnosis, requiring highly sensitive and precise models. In parallel, tabular medical data, including patient records, lab results, and other clinical information, presents additional challenges [11,12]. Such data are typically high-dimensional, may contain missing values, and are subject to strict privacy regulations, all of which complicate the classification process [13]. Various ML, DL, and hybrid algorithms, such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, Generative Adversarial Networks (GANs), and Support Vector Machine (SVM) [14–18], have advanced automated medical image analysis by improving classification accuracy. However, existing models still face challenges in feature extraction, sequential pattern recognition, generalization across datasets, and computational efficiency [19–21].

1.2. Advances in CNN-Based Medical Image Classification

Medical image analysis has significantly benefited from the advancements in CNNs, as they are highly effective for feature extraction, which is essential for accurate disease detection. The work by P. M. de Sousa et al. in [22] explored Wavelet CNNs, demonstrating improved CT image integrity without requiring data augmentation, resizing, or segmentation techniques. Other CNN-based architectures like C-Net [23] and Convolutional Deep Belief Networks (CDBN) [24,25] have also shown promise in medical image classification. The C-Net model consists of a concatenation of three main components: Outer, Middle, and Inner networks. The Outer and Middle networks function as feature extractors, while the Inner network is responsible for classifying the images as malignant or benign; whereas the CDBN model utilizes SVM as a feature classifier dedicated to classifying CT images to enhance feature transfer and reuse while suppressing less useful ones.

1.3. Emerging Hybrid Architectures for Enhanced Feature Representation

Recognizing the inherent limitations of CNNs in capturing global context and sequential dependencies, several hybrid models have been explored for medical image analysis. Among these are the models that proposed combining CNNs with Transformers, such as MambaConvT [26], which utilizes Transformers for enhanced global context. Other hybrid approaches have incorporated techniques like autoencoders and Particle Swarm Optimization (PSO) [27] or ensemble learning methods (e.g., Random Forest and XGBoost) [28] to optimize feature selection and improve classification performance. Further advancements include the integration of EfficientNet with SVM [29] and the development of Hybrid Quantum CNNs (HQCNNs) with adaptive optimizers [30]. Transfer learning with pretrained CNNs has been explored in [31], enabling deep feature extraction that is more abstract and representative of underlying data patterns compared to traditional methods.

Recent advances continue to emphasize the synergy between CNNs and Vision Transformers (ViTs) for improved medical image analysis. Borji et al. [32] introduced a hybrid CNN-ViT-MLP model tailored for osteosarcoma histopathology image classification, achieving high performance on a four-class Cancer Imaging Archive (TCIA) dataset. Their model effectively captured both local and global features without relying on segmentation, highlighting the strength of combining CNNs for fine-grained pattern recognition and ViTs for long-range dependencies. Similarly, the work in [33] proposed the EFFResNet-ViT framework, which fuses EfficientNet-B0 and ResNet-50 backbones with ViT modules. The model demonstrated strong performance on brain tumor MRI and retinal image datasets, while also emphasizing model interpretability via Grad-CAM visualizations and t-SNE analysis. These works underline a growing trend in hybrid architectures that balance accuracy and explainability, which are considered vital requirements for clinical integration. Nonetheless, despite these promising results, the current models have primarily been evaluated on limited datasets and may lack the generalizability and diversity required for broader clinical applications, highlighting the need for further investigation in more varied and extensive data contexts.

1.4. Sequential and Attention-Based Models for Context-Aware Analysis

Recently, advancements in medical image classification resulted in the development of hybrid models that combine CNNs with RNNs, such as LSTM, which are capable of capturing both spatial and temporal features, such that CNNs excel at extracting local spatial features, while LSTM are designed to handle sequential data, making them suitable for learning temporal patterns. For instance, Salehin et al. [34] introduced MedvLSTM, a hybrid model that integrates CNNs and LSTM to enhance feature representation and classification accuracy in medical image analysis. Complementing these efforts, attention-based models have emerged as powerful tools to improve both accuracy and efficiency in medical imaging tasks. These models utilize attention mechanisms to prioritize relevant features within images, leading to superior classification outcomes across diverse datasets as in [35], where the integration of spatial attention in dual-input CNNs has demonstrated the potential to effectively fuse features from original and modified images. Further innovations in attention mechanisms have expanded their applicability to multi-scale and segmentation tasks. Cheng et al. [36] proposed the Attention-based Multi-scale Nested Network (AMNNet), which incorporates Convolutional Block Attention Modules (CBAM) to capture contextual information at varying scales, highlighting the versatility of attention mechanisms in both segmentation and classification. Similarly, Liu et al. [37] introduced Eff-CTNet, a model that combines CNNs and Transformer architectures with a Group Cascade Attention module, capable of reducing computational costs while maintaining high performance on small-scale datasets, and thus balancing between efficiency and accuracy in medical image analysis.

1.5. Motivation and Research Gaps

The comprehensive literature review presented above underscores the growing importance of hybrid approaches in advancing medical image classification. Techniques such as CNN-LSTM integration, attention mechanisms, and multi-scale feature fusion have demonstrated significant potential in addressing the challenges posed by complex medical datasets. However, while these methods have achieved notable progress, limitations remain, particularly in terms of interpretability, computational efficiency, and robustness across diverse datasets. Existing models often struggle to balance these factors effectively, emphasizing the need for a more unified and adaptive solution. In response to these challenges, we propose the MediVision model, a novel framework that seamlessly integrates the strengths of CNNs, LSTM, attention mechanisms, and interpretability tools like

Grad-CAM heatmaps. By combining these key components, MediVision not only addresses the limitations of current models but also sets a new benchmark for performance and transparency in medical image analysis. In the following section, we detail the architecture and methodology of MediVision, illustrating how its innovative design achieves state-of-the-art results while ensuring interpretability and scalability.

In this paper, we present MediVision, a hybrid CNN-LSTM-attention model specifically designed for medical image classification. Unlike prior models that simply combine CNNs and LSTMs, MediVision integrates three key components: (1) CNNs for extracting spatial features from medical images, (2) an LSTM module that captures sequential dependencies across spatial feature maps, and (3) a tailored attention mechanism applied post-LSTM to selectively emphasize clinically relevant regions. Additionally, a skip connection preserves original CNN features alongside the LSTM–attention pathway, and Grad-CAM visualization enhances interpretability, allowing clinicians to better understand decision-relevant regions. Our proposed model was extensively evaluated over ten diverse real-world medical image datasets, as illustrated in Figure 1, with performance comparison to existing models. The results demonstrate the superior performance of MediVision in achieving classification accuracies for medical image analysis and disease detection.

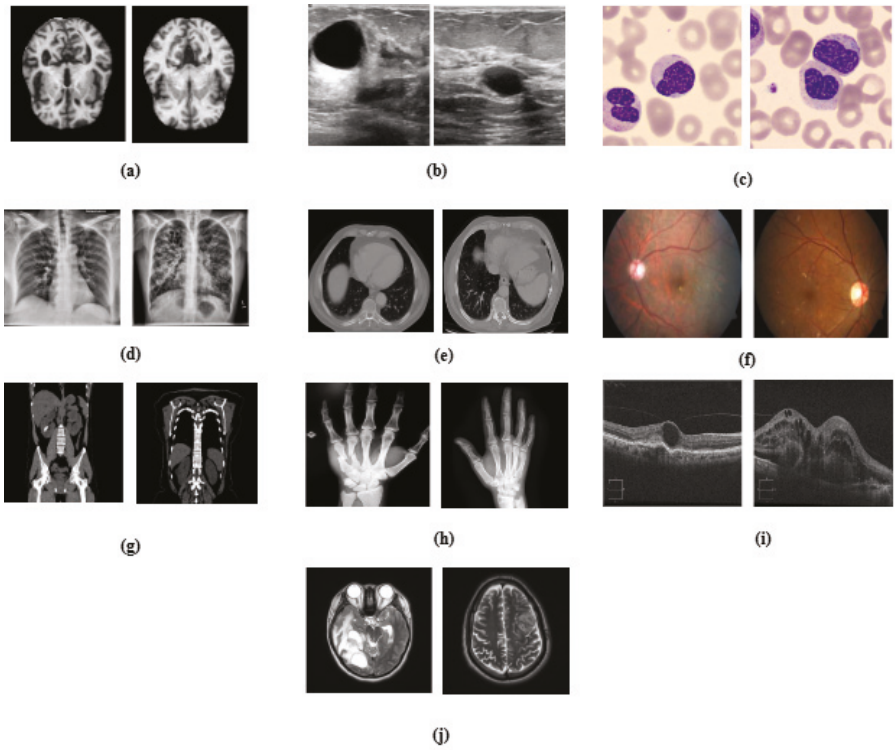


Figure 1. Visual representation of medical datasets used in the study, illustrating sample images for (**a**) Alzheimer's disease, (**b**) breast ultrasound, (**c**) blood cell, (**d**) chest X-ray, (**e**) chest CT scans, (**f**) diabetic retinopathy, (**g**) kidney diseases, (**h**) bone fracture multi-region, (**i**) retinal OCT, and (**j**) brain tumor.

The main contributions of this paper are summarized as follows:

- Present a comprehensive CNN-LSTM-Attention hybrid model for medical image classification;
- Demonstrate the effectiveness and generalizability of MediVision across ten diverse medical imaging datasets;
- Investigate the ability of MediVision to seamlessly integrate and exploit spatial, temporal, and attention-based features;

- Showcase the scalability and adaptability of MediVision for a broad range of medical image analysis tasks;
- Conduct extensive comparative and statistical analyses across diverse medical imaging datasets to validate the consistency, reliability, and generalizability of MediVision.

2. Materials and Methods

The proposed methodology comprises multiple cascaded stages, such that it begins with dataset partitioning, followed by data augmentation, preprocessing, and then model training. Each medical dataset is first split into training (70%), validation (15%), and testing (15%) subsets to prevent data leakage and ensure unbiased evaluation. Subsequently, data augmentation is applied to the training and validation sets to increase the sample diversity and address class imbalance, while the test set remains untouched. After augmentation, the data undergo preprocessing steps including normalization and label encoding. The proposed hybrid model is then trained on the augmented training set, with performance monitored using the validation set. Final evaluation is conducted on the independent test set using standard metrics (i.e., accuracy, precision, recall, and F1-score) to assess both predictive performance and model resilience.

2.1. Dataset Description and Customization

The effectiveness of DL models in medical image analysis heavily relies on the quality and diversity of the datasets used for training, validation and testing. Ten diverse medical imaging datasets were utilized in this study, as summarized in Table 1, encompassing a wide range of medical conditions, including neuro-degenerative disorders, oncological abnormalities, and pulmonary diseases, which provides a comprehensive foundation for robust model use in medical image classification. These datasets were chosen as representatives of various imaging modalities and diagnostic challenges, enabling a thorough evaluation of the proposed framework. While our experiments focus on these conditions, MediVision is a generalizable framework that can be applied to a broader set of diseases with appropriate training data.

Table 1. Summary of the datasets used for training and evaluation.

Dataset Name	Modality	No. of Images	Classes	Subcategories
Alzheimer's Disease [38]	MRI	14,000	4	Non-demented, Very Mild, Mild, Moderate
Breast Ultrasound [39]	Ultrasound	8148	2	Benign, Malignant
Blood Cells [40]	Microscopic	21,841	8	Erythroblast, Neutrophil, Basophil, Eosinophil, Lymphocyte, Monocyte, Immature Granulocyte, Lymphocyte
Chest X-ray [41–45]	X-ray	14,375	4	Normal, COVID-19, Pneumonia, Tuberculosis
Chest CT Scans [46]	CT	4000	4	Normal, Adenocarcinoma, Large Cell, Squamous Cell
Diabetic Retinopathy [47]	Fundus	13,000	2	Diabetic, Non-Diabetic
Kidney Diseases [48,49]	Ultrasound	11,756	4	Normal, Cyst, Stone, Tumor
Bone Fracture [50]	X-ray	9206	2	Fractured, Non-Fractured
Retinal OCT [51]	OCT	6000	5	Normal, AMRD, CSR, DR, MH
Brain Tumor [52]	MRI	10,500	3	Glioma, Meningioma, Pituitary

2.2. Data Augmentation and Normalization

One way to artificially expand the amount and complexity of current data is through data augmentation [53]. To give developers access to more representative training data, data augmentation techniques have been used to increase the size of training sets [54,55]. In this study, we have applied different techniques for data augmentation to enhance

the diversity of the datasets and to achieve a balanced distribution of classes, as in the example illustrated in Figure 2. The augmentation methods that we have employed include rotation, horizontal and vertical flipping, and other transformations to simulate different perspectives and conditions with the parameters summarized in Table 2.

Table 2. Data augmentation parameters.

Augmentation Technique	Description
Rotation	Rotates the image by fixed angles (90°, 180°, 270°).
Horizontal flip	Flips the image along the vertical axis.
Vertical flip	Flips the image along the horizontal axis.
Brightness adjustment	Randomly increases or decreases brightness. Factor range: 0.7–1.3.
Contrast adjustment	Randomly enhances or reduces contrast. Factor range: 0.8–1.5.

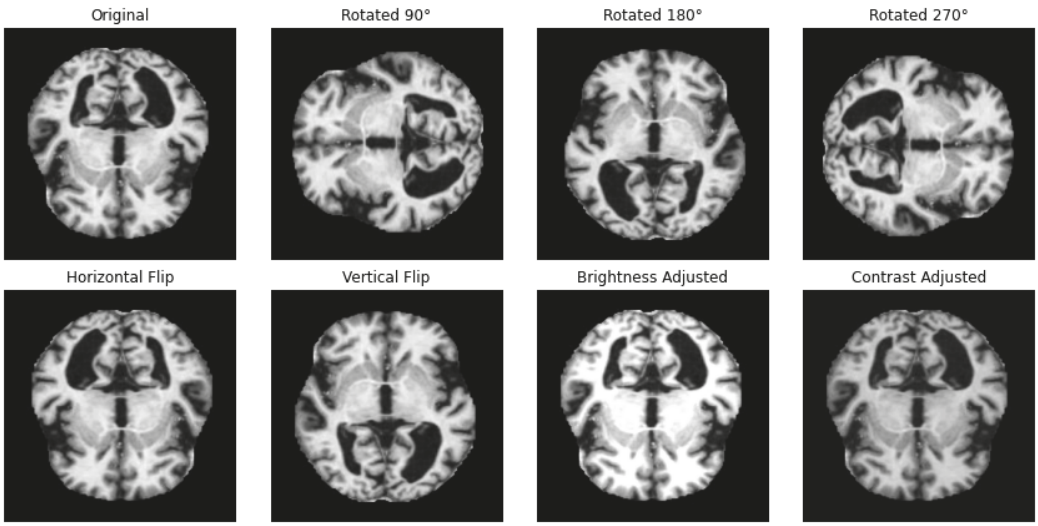


Figure 2. Illustration of applied data augmentation techniques.

In image processing and ML-based approaches, normalizing pixel values is a standard preprocessing step. This process scales the pixel values of an image to a standardized range, typically between 0 and 1 [56]. This normalization is essential, because images are often represented in an 8-bit per channel format, where each pixel can range from 0 (black) to 255 (white) [57]. The range of pixel values is given in the closed interval $[0, 255]$. The normalization equation is as follows:

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

where X represents the pixel value, and X_{min} and X_{max} are the minimum and maximum pixel values, respectively.

2.3. The Proposed MediVision Model

The MediVision model is a hybrid DL-based architecture, designed to combine the strengths of CNNs, LSTM, and attention mechanism, as shown in Figure 3. This combination allows the model to effectively process and analyze visual data, while capturing sequential dependencies and focusing on the most relevant parts of the input sequence. Additionally, a feature selection component is incorporated to refine the learned representations by prioritizing the most informative features and reducing redundancy, thereby enhancing the classification performance and model interpretability. The proposed MediVision model, with details of each block illustrated in Figure 3, is described as follows.

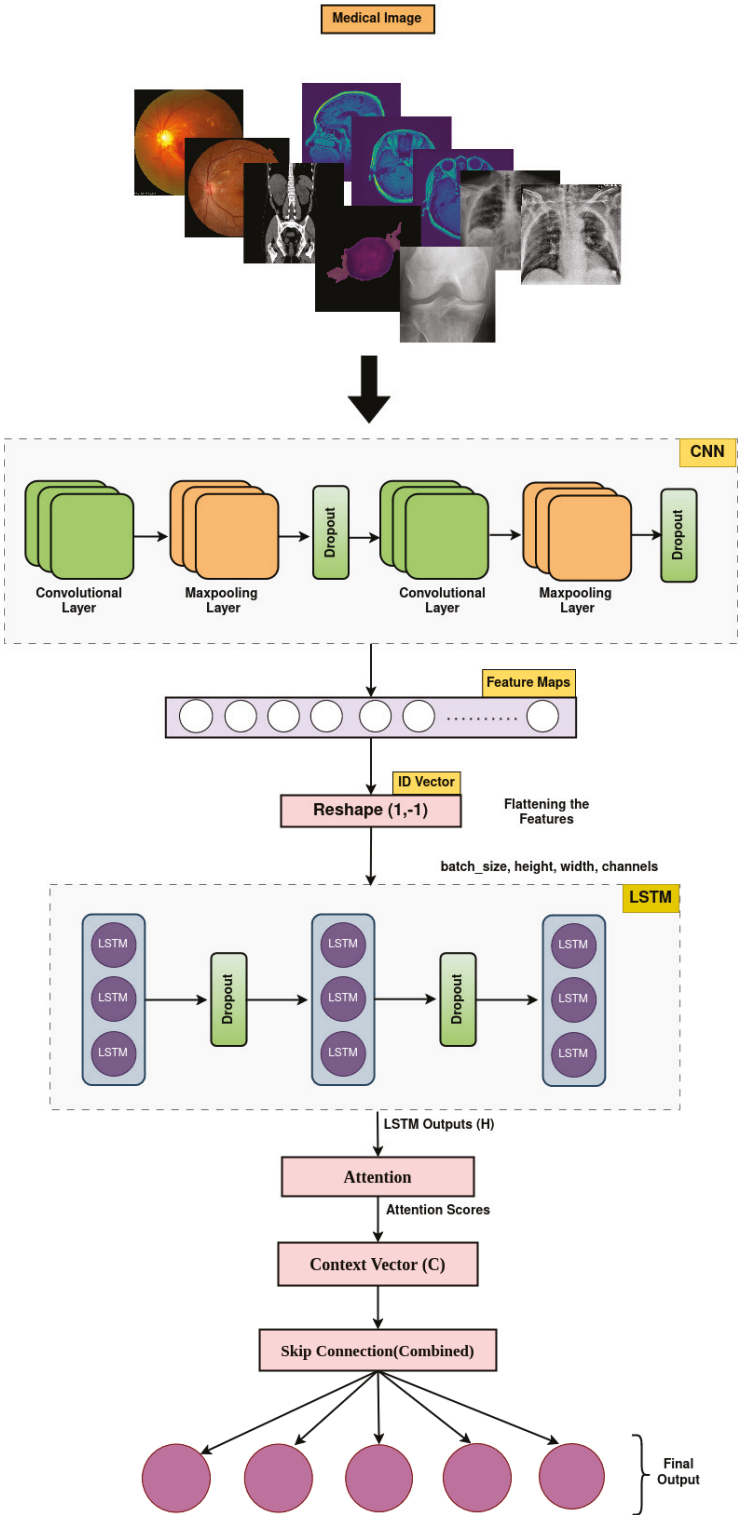


Figure 3. Proposed MediVision architecture.

2.3.1. CNNs Unit

The convolution operation involves sliding a filter (also called a kernel) over an input to produce a feature map. This operation helps detect and preserve spatial hierarchies in images, allowing the learning of features like edges, textures, and more complex patterns [58]. Equation (2) indicates feature maps of a CNN layer.

$$f_{ij}^{k} = \sigma\left(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} w_{mn}^{k} x_{(i+m)(j+n)} + b^{k}\right), \tag{2}$$

where f_{ij}^k is the output feature map at position (i, j) for the k-th filter, w_{mn}^k is the weight at position (m, n) in the k-th filter, $x_{(i+m)(j+n)}$ is the input at position $(i + m, j + n)$, b^k is the bias for the k-th filter, and σ is the ReLU activation function.

2.3.2. Flattening and Reshaping Unit

Flattening and reshaping are essential preprocessing steps in this model architecture, bridging the gap between convolutional feature extraction and sequential analysis using LSTM and attention mechanisms. Flattening converts multidimensional feature maps into a format suitable for further processing, while reshaping tailors the data for sequential LSTM input, optimizing the model's ability to capture sequential dependencies within the task of image classification. A tensor x after convolutional and pooling layers has the following shape:

$$x \in \mathbb{R}^{(batch_size, height, width, channels)}.$$

The flattening operation reshapes x into a 1D vector V:

$$V = \text{flatten}(x), \tag{3}$$

where

$$V \in \mathbb{R}^{(batch_size, height \cdot width \cdot channels)}.$$

The reshaping operation prepares the data for input into the LSTM layer, which expects sequences of vectors. After flattening, V is reshaped into a 3D tensor Y:

$$Y = \text{reshape}(V, batch_size, 1, height, width, channels), \tag{4}$$

where

$$Y \in \mathbb{R}^{(batch_size, 1, height, width, channels)}.$$

Here, 1 indicates that each input sequence to the LSTM consists of a single time step, ensuring compatibility with sequential processing. When replaced by -1, this will indicate that the reshaping operation automatically determines the remaining dimension size based on the total elements of the flattened vector.

2.3.3. LSTM Unit

The LSTM layer in this model processes reshaped input data Y consisting of sequences of feature vectors. It computes hidden states that capture sequential dependencies within each sequence, thus giving it the ability to remember information over long periods.

$$H = \text{LSTM}(Y), \tag{5}$$

where

$$H \in \mathbb{R}^{(batch_size, 1, \text{Units})}.$$

Here, H contains the hidden states for each sequence element (time step). The LSTM computes these hidden states using a set of learned weights and biases, which are updated during training to capture sequential dependencies in the data. The LSTM processes each time step t using the following gate mechanisms:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f). \tag{6}$$

The forget gate f_t decides how much of the previous cell state C_{t-1} to retain for the current time step t based on the previous hidden state h_{t-1} and the current input x_t, which takes the input as described in (7) to (9) below:

$$I_t = \sigma(W_i[h_{t-1}, x_t]), \tag{7}$$

$$\tilde{c}_t = \tanh(W_i[h_{t-1}, x_t] + b_i), \tag{8}$$

$$c_t = f_t \cdot C_{t-1} + I_t \cdot \tilde{c}_t, \tag{9}$$

where the input gate I_t determines how much of the candidate values $\tilde{c}_t$ should be added to the cell state c_t. The candidate cell state $\tilde{c}_t$ represents the new candidate values that could potentially be added to the cell state c_t. The updated cell state c_t combines the decision of the forget gate f_t on what to forget from C_{t-1} and the input gate I_t on what new information to store in c_t.

The output gate O_t in (10) determines how much of the updated cell state c_t should be exposed as h_t, which represents the hidden state computed based on the output gate O_t and the updated cell state c_t as in (11), such that tanh( ) is the hyperbolic tangent activation function.

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{10}$$

$$h_t = O_t \cdot \tanh(c_t) \tag{11}$$

2.3.4. Attention Mechanism

The attention mechanism allows the model to focus on the most relevant parts, which are the selected features of the input sequence by computing attention scores. These scores determine the importance of each element in the sequence, enabling the model to highlight and prioritize important features. The context vector produced by the attention mechanism is a weighted summary of the input sequence [59]. The input to the attention mechanism is the output sequence from the LSTM layer h_t for $t = 1, 2, \ldots, T$, where T is the length of the sequence. The attention mechanism first computes a set of attention scores e_t for each time step t based on (12):

$$e_t = \tanh(W \cdot h_t + b), \tag{12}$$

where W is a weight matrix, and b is a bias vector. The score e_t can be seen as an intermediate representation that helps to determine the importance of each h_t. The scores e_t are then normalized using a softmax function to obtain the attention weights α_t. These weights sum to 1 and represent the relative importance of each time step.

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)}, \tag{13}$$

such that α_t is the attention weight for time step t. The attention weights α_t are used to compute a weighted sum of the LSTM outputs, resulting in a context vector C:

$$C = \sum_{t=1}^{T} \alpha_t \cdot h_t, \tag{14}$$

where C represents the weighted combination of the LSTM hidden states.

2.3.5. Skip Connection

The skip connection combines the output of the attention mechanism with the summed output of the LSTM layer as given in (15). This helps retain important information from the

original sequence while also incorporating the focused information from the attention mechanism. Skip connections can improve the gradient flow and help in training deeper networks.

$$\text{Combined Output} = C + \sum_{t=1}^{T} h_t \tag{15}$$

2.3.6. Prediction Process

This is the final dense layer responsible for producing the classification output. The number of points taken into consideration in this process corresponds to the number of classes in the classification task. The softmax activation function ensures that the output is a probability distribution over the classes, such that

$$Y = \text{softmax}(W_d \cdot \text{Combined Output} + b_d), \tag{16}$$

where W_d is the weight matrix of the dense layer, b_d is the bias of the dense layer, and Y is the final output, which is a probability distribution over the classes. Algorithm A1, provided in Appendix A, summarizes the procedural steps followed by our proposed MediVision model, including input processing, feature extraction using CNNs, LSTM, attention mechanism, skip connection, and predication layers, respectively, to effectively analyze medical images and accurately classify diseases based on learned features.

In this study, we adopted a trial-and-error-based approach to tune hyperparameters, which is a commonly used strategy in DL-based models when automated search methods are found to be of high computational cost [50,56,60]. Specifically, we experimented with various configurations for key parameters such as the number of CNN layers, LSTM units, dropout rates, and learning rates, using validation performance as the guiding metric. Given the heterogeneous nature of the ten medical image datasets investigated in this study, with each having different imaging modality and class distribution, we observed that a uniform configuration often led to suboptimal results on certain datasets. To address this, we performed dataset-specific hyperparameter tuning, and we have documented the codes and final settings for each dataset to enable the reproducibility of our results. While our manual tuning approach yielded competitive results, we acknowledge that a more systematic strategy, such as grid search or Bayesian optimization, could further refine the performance and efficiency. We consider this a promising direction for our future work and plan to explore it in subsequent research efforts. Notably, recent studies, such as the work in [61], have demonstrated the effectiveness of Bayesian optimization in complex model tuning tasks.

3. Results

In this section, the numerical results of the proposed MediVision model are observed and discussed from different perspectives. First, we consider the *Accuracy*, which measures how often the model is correct across all classes. It is the ratio of the number of correct predictions considering True Positives (TP), True Negatives (TN), and the Total number of Instances (TI).

$$Accuracy = \frac{TP + TN}{TI} \tag{17}$$

Precision measures the proportion of TP instances among all instances predicted as positive, which comprises TP and False Positives (FP). It indicates how many of the predicted positives are truly positives.

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

Recall measures the proportion of TP instances among all actual positive instances, which comprises TP and False Negatives (FN). It shows how well the model identifies positive instances.

$$Recall = \frac{TP}{TP + FN} \tag{19}$$

The *F1-Score* is the harmonic mean of *Precision* and *Recall*, providing a balance between them.

$$F1\text{-}Score = 2 \times \frac{Precision \cdot Recall}{Precision + Recall} \tag{20}$$

TPs are the instances that the model correctly identifies as positive, meaning both the prediction and the actual status are positive. TNs represent the instances that the model accurately predicts as negative, where both the prediction and the actual status are negative. Conversely, FPs occur when the model incorrectly predicts a positive outcome for an instance that is negative, while FNs arise when the model mistakenly predicts a negative outcome for an instance that is actually positive.

3.1. Alzheimer's Disease Dataset

The assessment of the proposed MediVision model for Alzheimer's disease classification was demonstrated through accuracy-loss trends, confusion matrix and key metrics such as *Precision*, *Recall*, and *F1-Score*. The accuracy and loss curves, illustrated in Figure 4, reveal a steady improvement in MediVision's performance across training epochs. The training accuracy starts at approximately 60% and gradually increases, reaching around 95% by the final epoch. A similar trend is observed in validation accuracy, which consistently outperforms training accuracy at multiple points, ultimately stabilizing at approximately 96%. This close alignment between the training and validation accuracies suggests strong generalization capabilities, minimizing overfitting risks. Additionally, the loss curves indicate effective learning, with both training and validation losses consistently decreasing and stabilizing around 0.07, which is a large improvement compared to a 40% loss of both the VGG16 and VGG19 models. In addition, the smooth decline, without sharp fluctuations, demonstrates the efficient convergence of our proposed model.

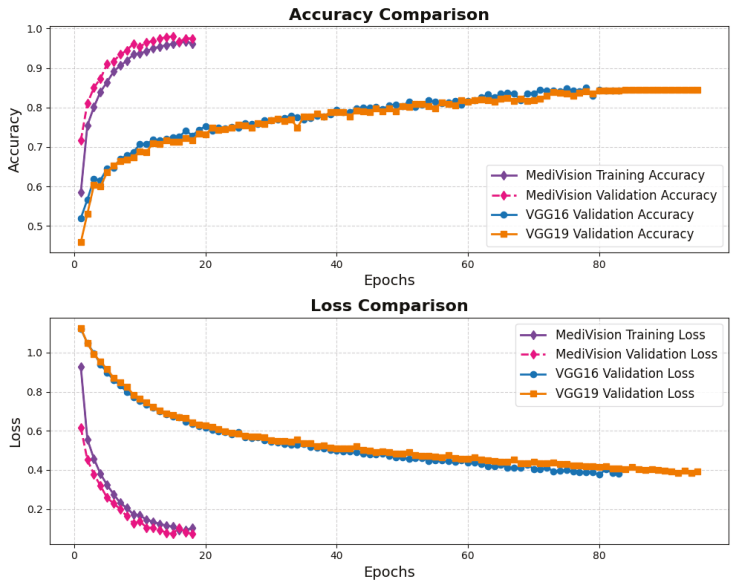


Figure 4. Accuracy and loss graph of Alzheimer's disease dataset.

The confusion matrix, which is illustrated in Figure 5, provides an in-depth analysis of the model's classification performance across four categories: non-demented, very mild demented, mild demented, and moderate demented. The majority of samples are

correctly classified, with 521, 527, 517, and 488 correctly classified instances, respectively. Although most samples are accurately classified, a minor misclassification trend is observed in the very mild demented category, where instances are occasionally misclassified as mild demented. This suggests possible feature similarities between these two conditions. However, the overall classification accuracy remains high, underscoring the model's reliability in distinguishing between different dementia stages. The effectiveness of the model is further validated through the precision, recall, and F1-score metrics. The moderate demented category achieves an exceptional precision and recall of 100%, indicating perfect classification for this group. Similarly, other categories exhibit high performance, with precision values of 96% (mild demented), 98% (non-demented), and 97% (very mild demented). The recall scores remain consistently strong, with mild demented at 99%, moderate demented also at 99%, non-demented at 98%, and very mild demented at 96%. The F1-score, which represents a harmonic mean between precision and recall, follows a similar pattern, with an average score of 98%.

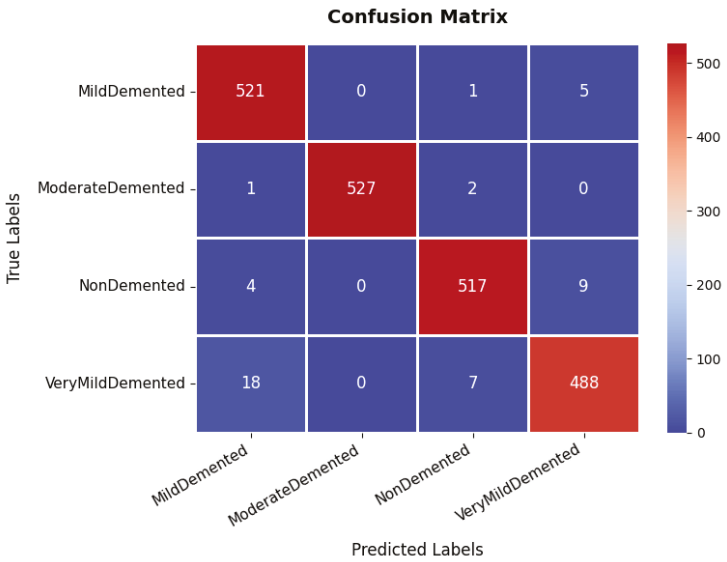


Figure 5. Confusion matrix of Alzheimer's disease dataset.

3.2. Blood Cell Dataset

The performance evaluation for the blood cell image dataset follows a similar approach, with the accuracy and loss results depicted in Figure 6, the training accuracy exhibits a consistent upward trend, beginning at lower levels and converging to over 95%, while the validation accuracy follows a similar trajectory, ultimately stabilizing at around 95% with minimal signs of overfitting. The corresponding loss curves reinforce this stability, showing a smooth and steady decrease over the epochs.

The confusion matrix, presented in Figure 7, reveals that the model accurately classifies the majority of samples across all cell types, including basophils, eosinophils, erythroblasts, Ig (immunoglobulins), lymphocytes, monocytes, neutrophils, and platelets, although occasional misclassifications occur, particularly among closely related classes such as Ig and monocytes. Still, the overall classification accuracy remains high, at approximately 95.45%. Furthermore, the detailed metrics underscore this effectiveness, with most classes attaining precision and recall values above 90%, resulting in F1-scores that consistently indicate strong performance. For instance, the eosinophil and neutrophil classes achieve near-perfect precision and recall, whereas Ig shows slightly lower metrics, suggesting a potential area for further refinement. Collectively, the high accuracy, well-balanced precision and recall, and smoothly converging loss function validate the model's relia-

bility in differentiating various blood cell types, thereby showcasing its utility in clinical diagnostic workflows.

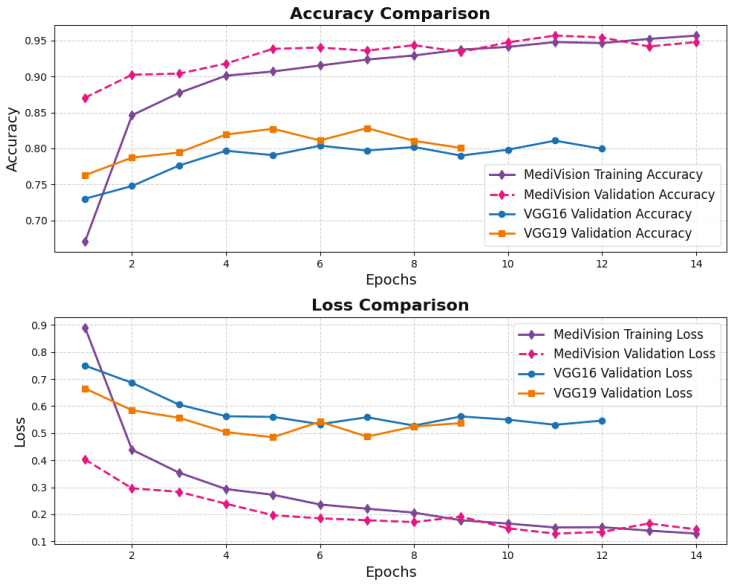


Figure 6. Accuracy and loss graph of blood cell dataset.

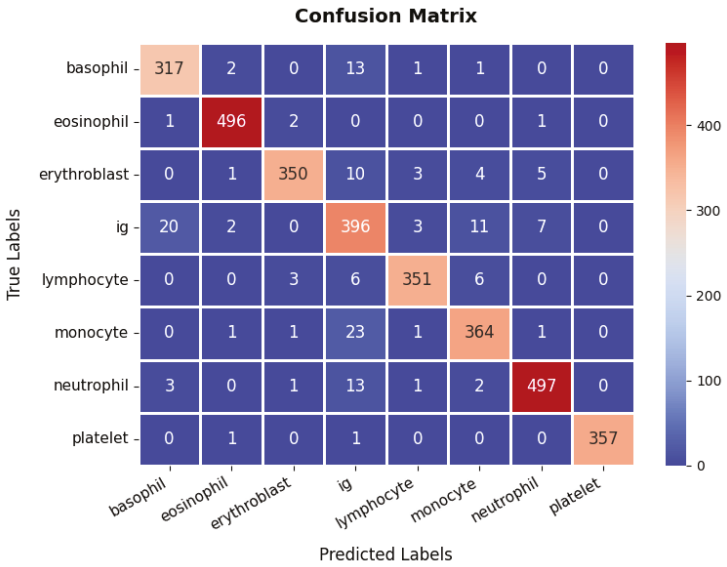


Figure 7. Confusion Matrix of Blood Cell dataset.

3.3. Bone Fracture Multi-Region Dataset

The loss and accuracy curves for the bone fracture multi-region dataset, depicted in Figure 8, indicate high efficiency of our model in learning and generalization, where the training loss drops steeply from above 0.45 to nearly 0.05 within the first three epochs, demonstrating rapid learning, while the validation loss follows a similar declining trend, reaching below 0.05. By the end of training, both training and validation losses stabilize at very low values, signifying minimal errors. The accuracy curves further reinforce the model's effectiveness, with the training accuracy rising sharply from approximately 80% to nearly 95% within the first few epochs, and the validation accuracy surpassing 95% by the second epoch, ultimately reaching nearly 98% at the final epoch with minimal overfitting and strong generalization.

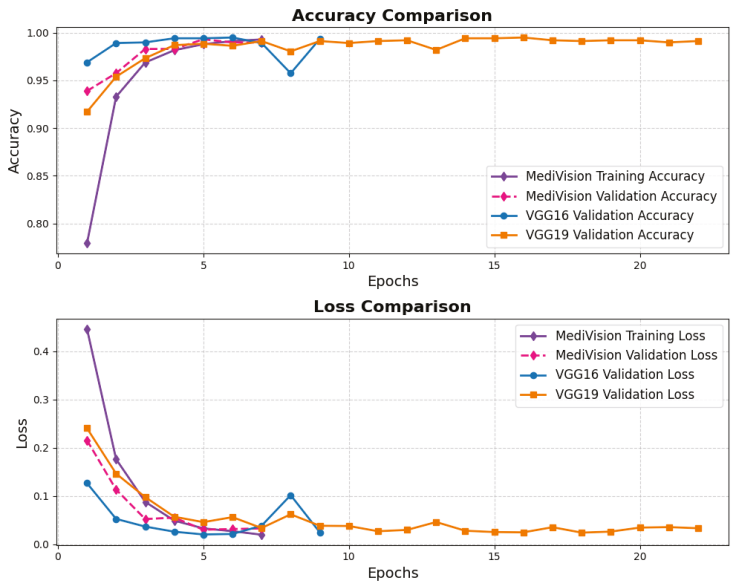


Figure 8. Accuracy and loss graph of bone fracture multi-region dataset.

At the same time, the confusion matrix in Figure 9 provides additional validation, showing that the first class achieves 687 TPs with only eight misclassifications as the second class, while the second class records 680 TPs with just 12 false positives, demonstrating excellent classification performance. The small number of misclassifications highlights the model's high predictive accuracy, making it suitable for real-world bone fracture detection. Furthermore, the precision, recall, and F1-score for each class range between 98% and 99%, with an overall model accuracy of 98.56%, confirming its exceptional reliability.

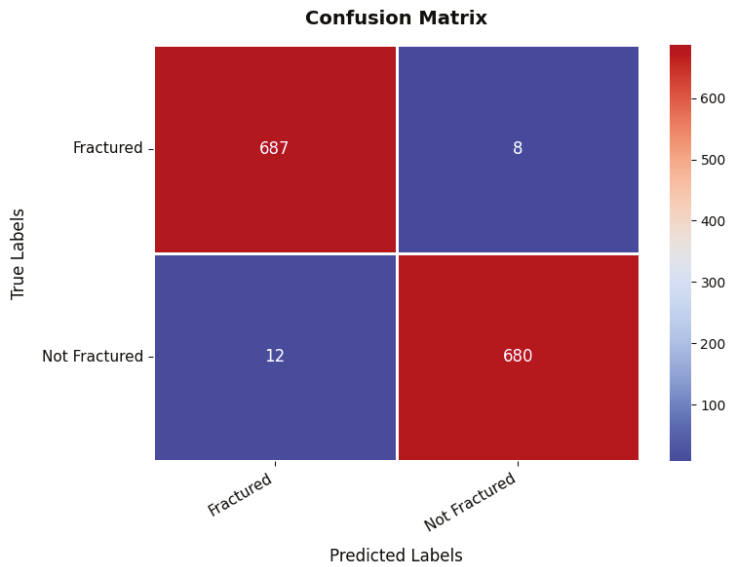


Figure 9. Confusion matrix of bone fracture multi-region dataset.

3.4. Brain Tumor Dataset

The accuracy–loss curves for the brain tumor dataset are given in Figure 10, illustrating a consistent decrease in both training and validation loss over the epochs, with the training loss dropping sharply from approximately 0.7 to around 0.05 by the 20th epoch. This steep decline indicates that the model quickly learns the underlying patterns within the dataset. The validation loss follows a similar decreasing trend but remains slightly higher and fluctuates around 10% in later epochs, suggesting mild overfitting, where the model performs slightly better on training data than on unseen validation data. The accuracy

curves further reinforce this observation, with the training accuracy rising steadily and reaching nearly 98%, while the validation accuracy, though slightly lower and exhibiting minor fluctuations, stabilizes around 95% by the final epochs.

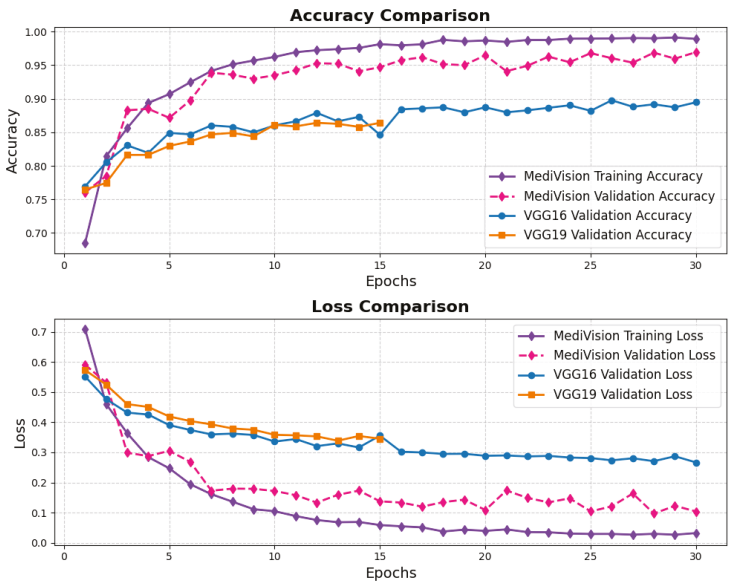


Figure 10. Accuracy and loss graph of brain tumor dataset.

Looking at the confusion matrix in Figure 11, we get further insights into the model's classification performance across different brain tumor classes. The glioma class achieves the highest classification accuracy, with 522 TPs and minimal misclassifications into other categories. Similarly, the meningioma and pituitary classes exhibit strong predictive accuracy, with 508 and 495 TPs, respectively, though some samples are misclassified due to overlapping features among the three tumor types. These classification results are further supported by the performance metrics in Table 3, where all three classes exhibit high precision, recall, and F1-scores ranging between 95% and 98%. The overall accuracy of the model stands at 96.76%, signifying its robustness in distinguishing between different tumor types.

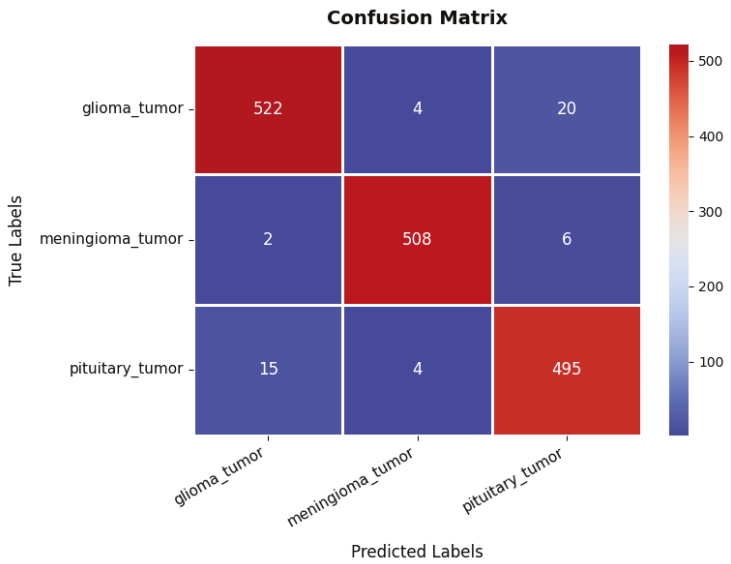


Figure 11. Confusion matrix of brain tumor dataset.

Table 3. Performance of MediVision model across subclasses of all medical image datasets.

Datasets	Subclasses	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Alzheimer's disease	Mild Demented	96.00	99.00	97.00	98.00
	Moderate Demented	100.00	99.00	100.00	
	Non-Demented	98.00	98.00	98.00	
	Very Mild Demented	97.00	95.00	96.00	
Breast ultrasound	Benign	97.00	98.00	97.00	97.45
	Malignant	98.00	97.00	97.00	
Blood cell	Basophil	93.00	95.00	94.00	95.45
	Eosinophil	99.00	99.00	99.00	
	Erythroblast	98.00	94.00	96.00	
	Ig	86.00	90.00	88.00	
	Lymphocyte	97.00	96.00	97.00	
	Monocyte	94.00	93.00	93.00	
	Neutrophil	97.00	96.00	97.00	
	Platelet	100.00	100.00	100.00	
Chest X-ray	COVID19	97.00	97.00	97.00	97.31
	Normal	98.00	97.00	97.00	
	Pneumonia	97.00	98.00	97.00	
	Tuberculosis	97.00	98.00	98.00	
Chest CT scans	Adenocarcinoma	91.00	93.00	92.00	94.83
	Large cell carcinoma	97.00	96.00	97.00	
	Normal	98.00	100.00	99.00	
	Squamous cell carcinoma	93.00	91.00	92.00	
Diabetic retinopathy	DR	96.00	95.00	95.00	95.43
	No DR	94.00	95.00	95.00	
Kidney diseases	Cyst	100.00	100.00	100.00	98.36
	Normal	98.00	96.00	97.00	
	Stone	96.00	98.00	97.00	
	Tumor	100.00	99.00	100.00	
Bone fracture	Fractured	98.00	99.00	99.00	98.56
	Not Fractured	99.00	98.00	99.00	
Retinal OCT	Age-related Macular Degeneration	97.00	100.00	98.00	95.44
	Central Serous Retinopathy	98.00	92.00	95.00	
	Diabetic Retinopathy	89.00	96.00	92.00	
	Macular Hole	95.00	94.00	94.00	
	Normal	98.00	96.00	97.00	
Brain tumor	Glioma	97.00	98.00	96.00	96.76
	Meningioma	98.00	96.00	98.00	
	Pituitary	95.00	98.00	96.00	

3.5. Breast Ultrasound Dataset

The accuracy and loss trends for the case of breast ultrasound image classification are illustrated in Figure 12, in which we notice a steady improvement over the training epochs, indicating effective learning. Initially, the training accuracy starts at approximately 75%, gradually increasing to reach around 98% by the final epoch. The validation accuracy follows a similar trajectory, consistently outperforming the training accuracy at certain points and ultimately stabilizing at approximately 97%, with minimal the risk of overfitting. The loss curves further reinforce this stability, exhibiting a consistent decline without abrupt fluctuations, signifying smooth convergence.

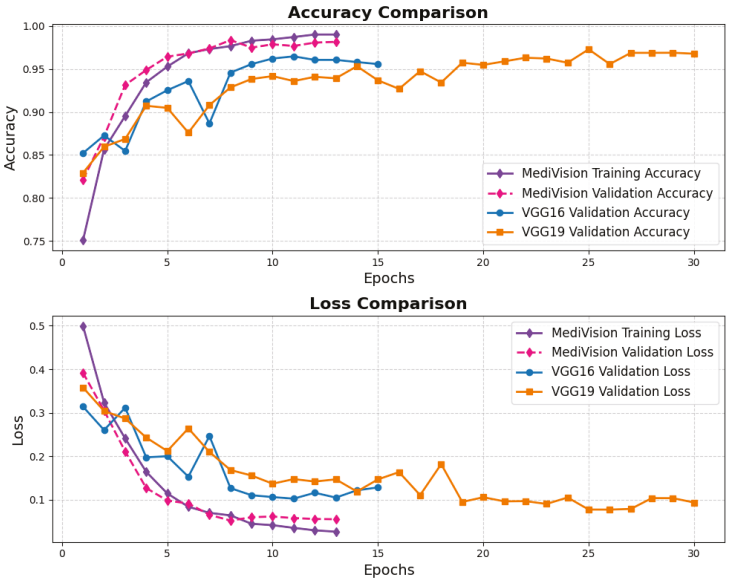


Figure 12. Accuracy and loss graph of breast ultrasound dataset.

The confusion matrix, presented in Figure 13, provides insights into the classification performance for benign and malignant cases. The model correctly classifies 604 benign cases and 583 malignant cases. However, 14 benign samples are misclassified as malignant, while 17 malignant samples are misclassified as benign. Despite these minor misclassifications, the overall classification accuracy remains at an 97%, demonstrating the model's robustness in distinguishing between benign and malignant breast tumors. Furthermore, the precision, recall, and F1-score metrics validate the model's efficiency and reliability, as shown in Table 3. The benign class achieves a precision of 97%, recall of 98%, and F1-score of 97%, indicating strong classification performance. Similarly, the malignant class exhibits a precision of 97.45% and recall of 98%, reflecting a balanced classification ability across both categories. The high recall values indicate that the model is highly effective in identifying malignant cases, which is critical for medical diagnosis.

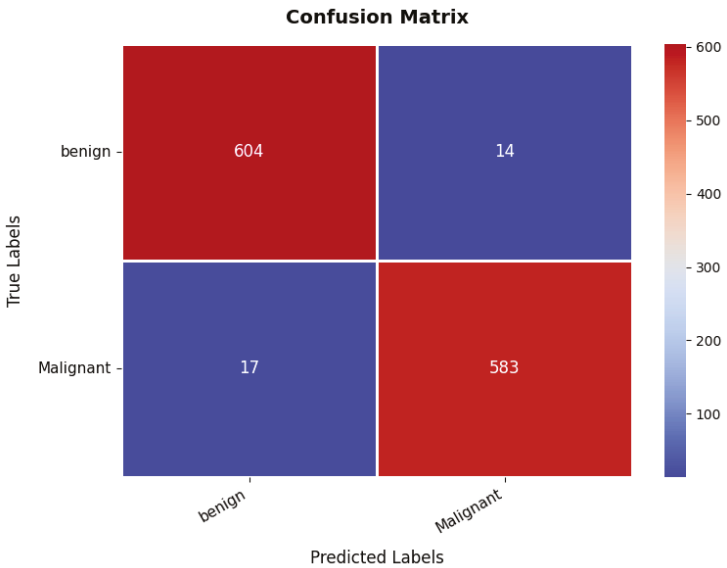


Figure 13. Confusion matrix of breast ultrasound dataset.

3.6. Chest CT Scans Dataset

The loss and accuracy curves, as illustrated in Figure 14 demonstrate strong performance for the chest CT scans dataset. The training loss exhibits a substantial reduction

from approximately 1 to 0.4 within the initial epochs, indicating a rapid learning phase. A similar trend is observed in the validation loss, which decreases from approximately 0.8 to 0.4, suggesting a well-balanced learning process and a reasonable level of generalization. As training progresses, both the training and validation losses continue to decline. Notably, the validation loss experiences fluctuations around epochs 10 and 14. However, in general, the losses remain balanced, affirming the model's stability and robustness. The accuracy curves follow a similar pattern, reinforcing the effectiveness of the model. The training accuracy experiences a sharp rise from approximately 60% to over 90% within the first three epochs, indicating the model's ability to quickly learn the dataset's inherent patterns. The validation accuracy also improves steadily, closely following the training accuracy. After approximately five epochs, both the training and validation accuracy stabilize around 95%, with the training accuracy exhibiting minimal variation and the validation accuracy showing slight oscillations. These minor fluctuations in validation accuracy, reflected in the loss curves, indicate that the model maintains a well-calibrated performance without significant overfitting.

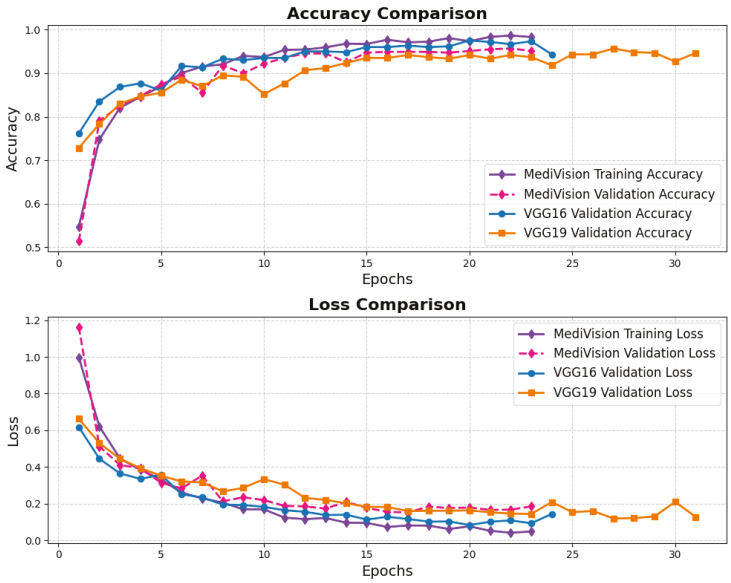


Figure 14. Accuracy and loss graph of chest CT scans dataset.

The confusion matrix, presented in Figure 15, provides further insights into the model's classification performance across the four classes. The first class achieves 149 TPs, though eight instances are misclassified as belonging to the fourth class. The second class is correctly classified in 153 instances, with only five misclassified as the first class, suggesting some overlap in features between these two categories. The third class is handled with high precision, achieving 138 TPs with no misclassifications, underscoring the model's strong ability to distinguish this class. The fourth class records 129 TPs but exhibits minor misclassification errors, with nine instances predicted as the first class and two as the third class. The overall accuracy of the model on this dataset is 94.83%, as shown in Table 3. These results confirm the model's reliability in distinguishing between different categories within the chest CT scan while maintaining a strong balance between training and validation performance.

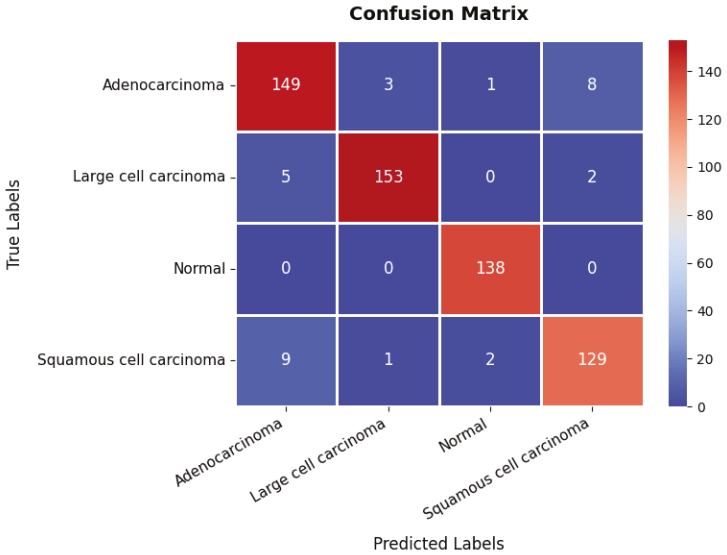


Figure 15. Confusion matrix of chest CT scans dataset.

3.7. Chest X-Ray Dataset

The performance evaluation for the chest X-ray dataset, also follows the same structured approach. The accuracy–loss curves illustrated in Figure 16 indicate that the training accuracy exhibits a consistent upward trajectory, starting at approximately 84% and steadily increasing to nearly 98% over the epochs. Similarly, the validation accuracy follows a comparable trend, stabilizing around 96–97% with minimal signs of overfitting. The corresponding loss curves further reinforce this stability, showing a steady decline in both training and validation loss. This confirms that the model trains efficiently and achieves robust convergence.

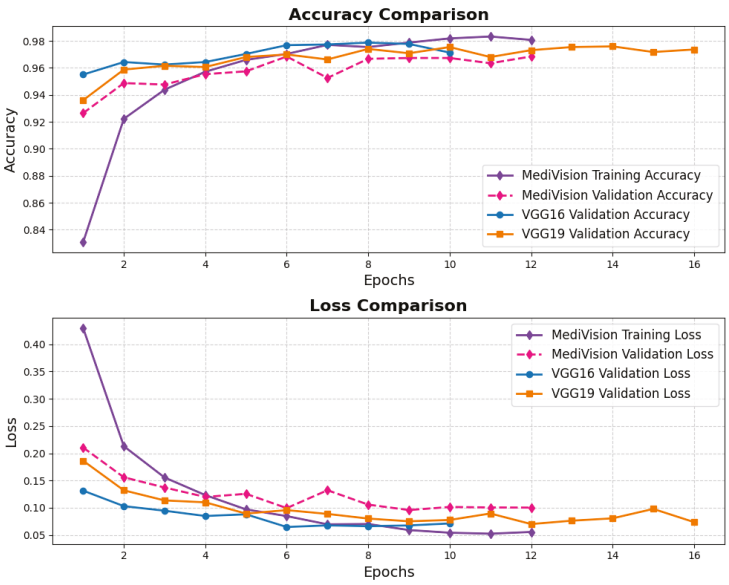


Figure 16. Accuracy and loss graph of chest X-ray dataset.

Looking at the confusion matrix, which is presented in Figure 17, shows that the model classifies the majority of samples correctly across all categories: COVID-19, Normal, Pneumonia, and Tuberculosis. Notably, COVID-19 and Pneumonia classifications have a few misclassifications, with some Pneumonia cases being misclassified as Normal or COVID-19. However, the overall misclassification rate remains low, with a strong emphasis on correctly identifying disease conditions. The detailed evaluation metrics presented in

Table 3 further validate the model's high performance. The precision, recall, and F1-score values consistently exceed 97% across all classes, indicating a well-balanced classification model. The overall classification accuracy is 97.31%, confirming that the model is highly reliable in diagnosing chest-related conditions. While minor misclassifications exist, they do not significantly impact the overall performance, making the model highly suitable for clinical diagnostic applications.

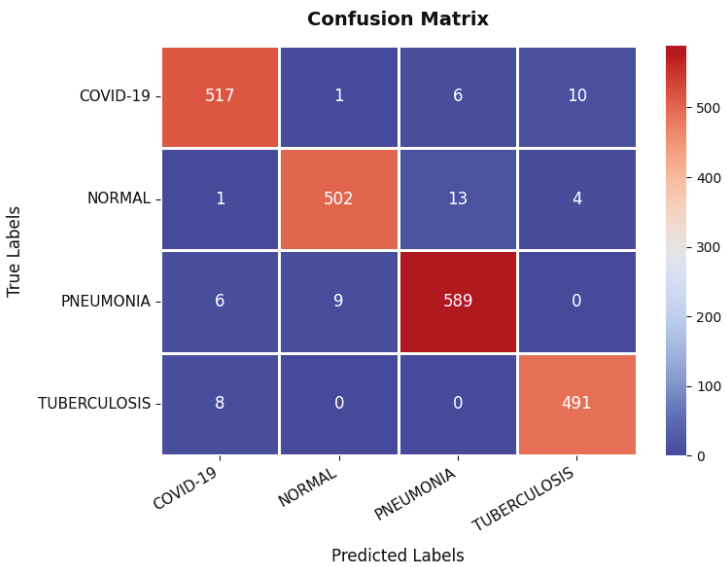


Figure 17. Confusion matrix of chest X-ray dataset.

3.8. Diabetic Retinopathy Dataset

The loss and accuracy curves for the diabetic retinopathy dataset, as illustrated in Figure 18, exhibit similar characteristics to those observed in other datasets. However, the proposed model demonstrates high efficiency despite some oscillations in validation performance. The training loss steadily decreases from approximately 0.28 to 0.12, indicating a consistent learning process, with all log loss values remaining below 0.05, suggesting that the model effectively learns and fits well to the training data. Although the validation loss follows a similar downward trend, it exhibits some fluctuations, particularly between epochs 3 and 6, before stabilizing and converging toward the training loss, demonstrating improved generalization. The training accuracy rapidly increases from approximately 88% to 94% within the first three epochs, signifying the model's ability to quickly adapt to the dataset, while the validation accuracy also improves sharply, reaching around 95% by epoch 3, though it shows some minor oscillations between epochs 4 and 6, mirroring the validation loss. Nonetheless, by the final epochs, both the training and validation accuracy stabilize at approximately 95%, confirming that the model generalizes well without significant overfitting.

Additional insight in to the classification ability of the model is given by the confusion matrix, which is illustrated in Figure 19. The matrix indicates a TP value, which is much higher for both classes with 950 correct predictions for the first class, while 904 for the second class were correctly predicted. However, there are some misclassified, where 53 instances of the first class have been classified wrongly as the second class, and 43 instances of the second class were classified as the first class. These numbers are still low enough that it suggests the overall accuracy is good, but it also hints at some of the times where this model must get confused between the two classes, probably because the images making up the set of diabetic retinopathy are nuanced in their features. The balance between the TPs and the misclassifications suggests that while the model is highly effective, further refinement could

help reduce these errors and improve the overall accuracy of the classification. The range of precision, recall, and F1-score is in between 95% and 96% for both the classes. The overall accuracy is 95.43%.

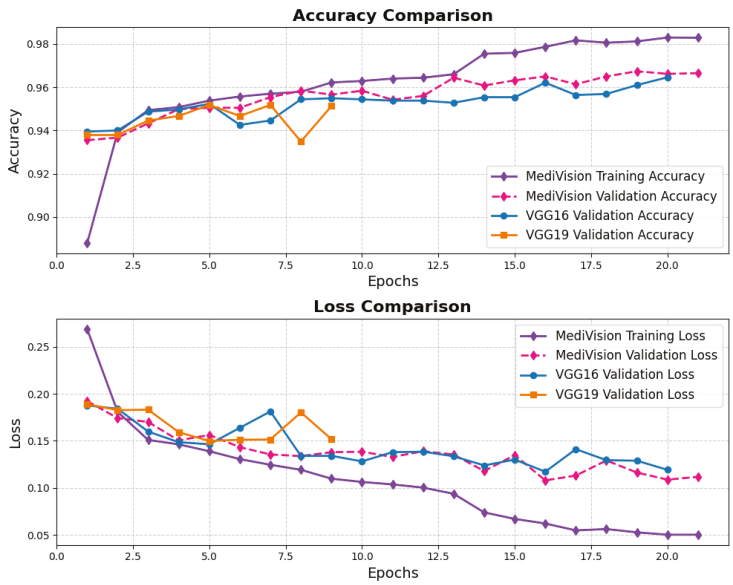


Figure 18. Accuracy and loss graph of diabetic retinopathy dataset.

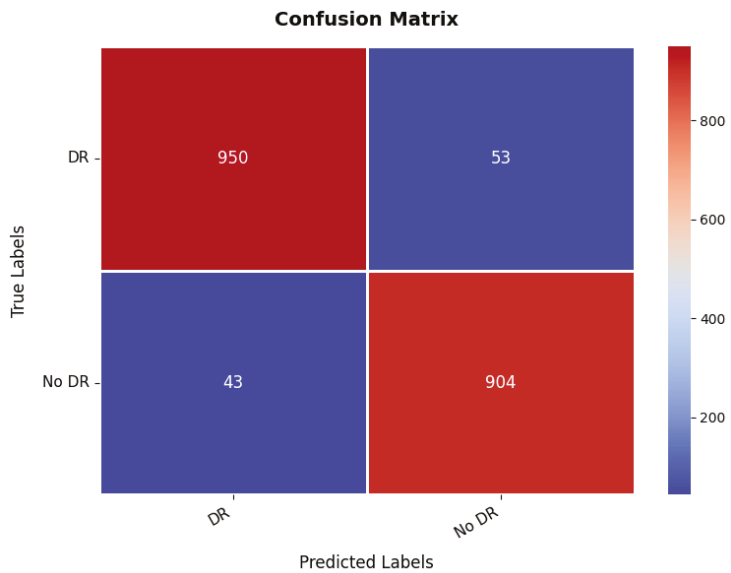


Figure 19. Confusion matrix of diabetic retinopathy dataset.

3.9. Kidney Diseases Dataset

The loss and accuracy curves for the kidney diseases dataset, given in Figure 20, demonstrate a highly optimized model with strong generalization capabilities. The training loss decreases steadily from 0.6189 to 0.0180, while the validation loss shows some fluctuations, starting at 0.2786 and ending at 0.0771. The accuracy trends confirm the model's strong performance. The training accuracy increases sharply from 75% to 95% within the first five epochs, showing the model's quick adaptation, while the validation accuracy closely follows, occasionally surpassing the training accuracy, highlighting strong generalization. After the initial phase, the training accuracy stabilizes around 97% with minor fluctuations, while the validation accuracy oscillates slightly around 94%, maintaining high performance on both sets.

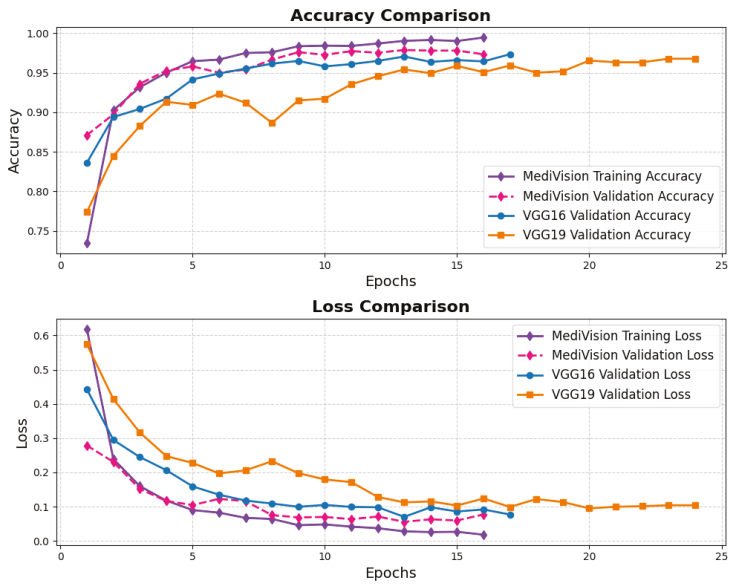


Figure 20. Accuracy and loss graph of kidney diseases dataset.

The confusion matrix in Figure 21 further validates the model's classification effectiveness across different classes. The first class achieves perfect classification with 434 TPs and no FPs. The second class is also well-classified, with 431 TPs, though 18 samples are misclassified as the third class, indicating some difficulty in distinguishing between these two classes. The third class has 424 TPs but shows eight misclassifications as the second class, reinforcing this challenge. The fourth class exhibits excellent classification, with 446 TPs and only three minor misclassifications, with two instances misclassified as the second class. Overall, the model demonstrates strong classification performance, with occasional confusion between the second and third classes. The overall accuracy for the kidney diseases dataset is 98.36%, as summarized in Table 3.

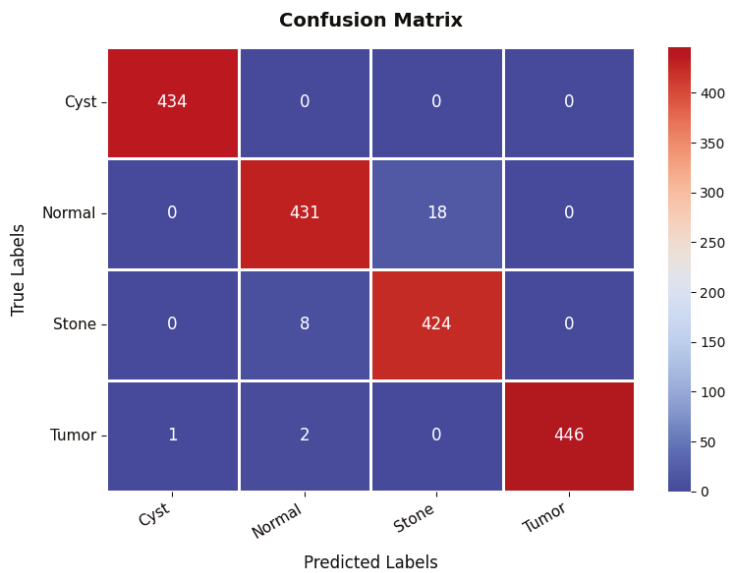


Figure 21. Confusion matrix of kidney diseases dataset.

3.10. Retinal OCT Dataset

The loss curves for the retinal OCT dataset are presented in Figure 22, showing a significant and consistent reduction in training loss, which declines from approximately 1.2 to nearly 0.01 by the 25th epoch. The validation loss follows a similar decreasing trend; however, it exhibits minor fluctuations, particularly between the fifth and tenth epochs.

This variability in validation loss suggests that, while the model effectively learns feature representations, some inconsistencies in the validation set may contribute to occasional performance fluctuations. Despite these variations, the overall decline in validation loss indicates enhanced generalization capability, and both the training and validation losses eventually converge at low values, signifying a stable and well-optimized model. The accuracy curves further support the model's strong performance, demonstrating a rapid increase in training accuracy, surpassing 90% within the first five epochs and continuing to improve steadily. The validation accuracy follows a comparable trajectory, albeit slightly lower than the training accuracy, with minor fluctuations mirroring those observed in validation loss. By the 25th epoch, both the training and validation accuracies stabilize at approximately 95%, which indicates minimal overfitting.

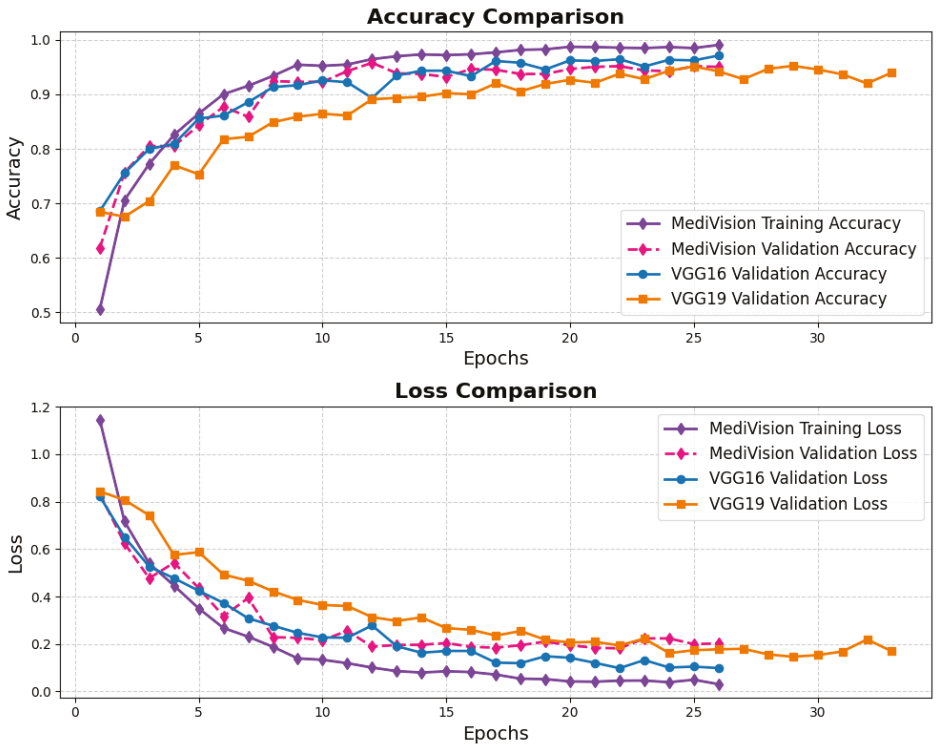


Figure 22. Accuracy and loss graph of retinal OCT dataset.

The confusion matrix in Figure 23 provides a detailed evaluation across different classes in the retinal OCT dataset. The first class exhibits near-perfect classification, achieving 203 TPs with no observed misclassifications, highlighting the model's high precision for this category. The second class also demonstrates strong classification performance with 179 TPs; however, six instances are misclassified as the first class, and a few others are incorrectly assigned to different categories. The third class, with 149 TPs, shows a slightly higher misclassification rate, with certain instances incorrectly categorized as either the second or fourth class. The fourth class attains 174 TPs with minimal classification errors, primarily involving minor confusion with the third class. Overall, the model exhibits robust classification performance, with only minor misclassifications observed in the second and third classes, effectively distinguishing between the majority of classes and providing high precision and recall as summarized in Table 3.

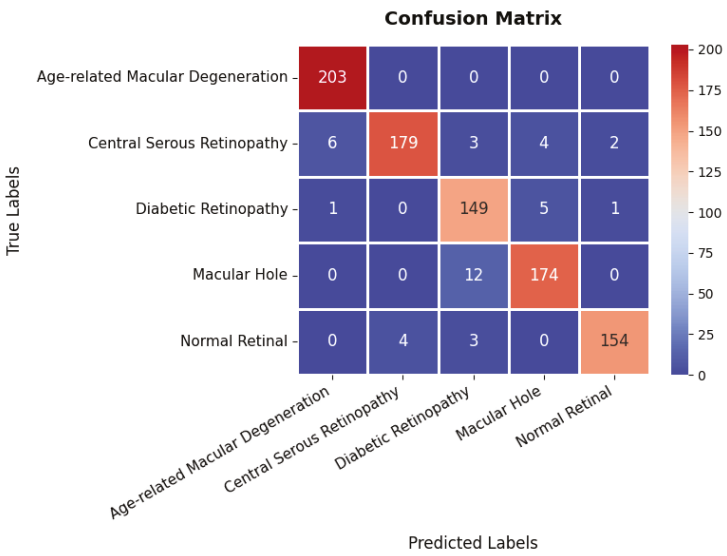


Figure 23. Confusion matrix of retinal OCT dataset.

4. Discussion

4.1. Model Interpretability via Explainable AI

Additionally, we employed Grad-CAM heatmaps to visually explain MediVision's decision-making process across representative medical imaging datasets, namely (a) bone fracture multi-region, (b) chest X-ray, and (c) Alzheimer's disease, as illustrated in Figure 24. These heatmaps highlight the image regions that contribute most significantly to the model's class predictions. Since MediVision performs classification rather than segmentation, the Grad-CAM heatmaps emphasize discriminative features learned by the model, rather than anatomically precise regions of interest. The intermediate convolutional layers act as feature detectors, and the resulting activations, visualized as intensity-based overlays, reveal where the model focuses during classification. This visual interpretation validates that MediVision accurately identifies significant disease characteristics within the input image. The consistency of highlighted regions across these representative datasets supports the model's generalizability and its ability to detect subtle disease patterns across diverse imaging modalities. These findings demonstrate MediVision's potential to enhance diagnostic accuracy and reduce human error in clinical settings.

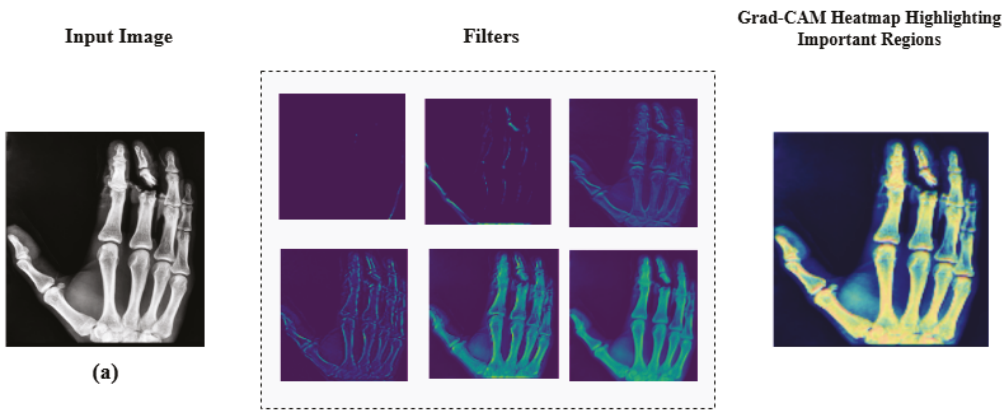


Figure 24. *Cont.*

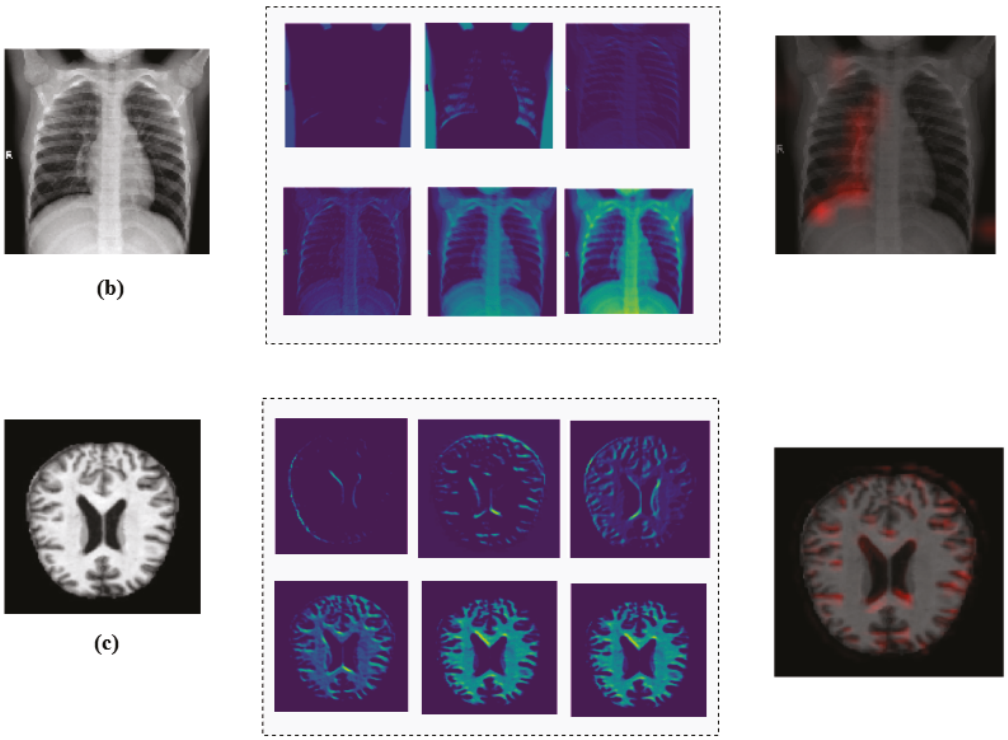


Figure 24. Grad-CAM heatmap visualization showing important regions (in red) for disease classification across representative datasets: (**a**) bone fracture multi-region, (**b**) chest X-ray, and (**c**) Alzheimer's disease.

4.2. Comparative Analysis

The performance comparison between the proposed MediVision model and existing DL-based models, including VGG16, VGG19, and ResNet50, across all ten diverse medical image datasets is summarized in Table 4. Notably, MediVision showed substantial improvements in Alzheimer's disease, achieving 98% compared to the 75–85% range of the other models. The results also demonstrated significant gains in blood cell classification with 95.45% accuracy and brain tumor detection reaching 96.76%, a considerable advancement over ResNet50's by 24.44%. Significant gains were also observed in breast ultrasound, chest X-ray/CT, and kidney disease classification. For retinal OCT classification, while the benchmark models already achieve high accuracy, MediVision still demonstrates a considerable improvement in comparison.

Table 4. Accuracy benchmarking of the proposed MediVision model.

Dataset	Model	Accuracy (%)
Alzheimer's disease	VGG16	85.13
	VGG19	84.03
	ResNet50	75.13
	MediVision	98.00
Blood cell	VGG16	80.20
	VGG19	82.73
	ResNet50	69.09
	MediVision	95.45
Breast ultrasound	VGG16	96.47
	VGG19	97.29
	ResNet50	84.65
	MediVision	97.45

Table 4. *Cont.*

Dataset	Model	Accuracy (%)
Chest X-ray	VGG16	96.00
	VGG19	96.30
	ResNet50	92.12
	MediVision	97.31
Chest CT scans	VGG16	96.00
	VGG19	95.67
	ResNet50	80.83
	MediVision	96.17
Diabetic retinopathy	VGG16	95.21
	VGG19	95.18
	ResNet50	89.13
	MediVision	95.43
Kidney diseases	VGG16	97.05
	VGG19	96.54
	ResNet50	90.08
	MediVision	98.36
Bone fracture multi-region	VGG16	98.42
	VGG19	98.13
	ResNet50	93.80
	MediVision	98.56
Retinal OCT	VGG16	95.44
	VGG19	95.22
	ResNet50	64.56
	MediVision	95.44
Brain tumor	VGG16	89.90
	VGG19	86.79
	ResNet50	72.32
	MediVision	96.76

In addition, Table 5 demonstrates MediVision's competitive edge against state-of-the-art DL-based models, including EfficientNet, DenseNet, and Hybrid CNN-LSTM, across diverse medical image datasets. The table shows that MediVision is capable of achieving superior or comparable accuracy in nearly all categories, as a result of its advanced architecture and learning capabilities, particularly for classifying complex medical images like brain tumor and Alzheimer's disease. These results were achieved through a consistent training strategy based on empirical tuning, where a common set of hyperparameters was identified using a hit-and-trial method. Specifically, we used a batch size of 16, a learning rate of 0.01 with the Adam optimizer, and dropout rates between 0.1 and 0.3. Additionally, early stopping with a patience of three epochs was applied, which helped ensure strong performance across multiple datasets. Overall, the results confirm MediVision's enhanced generalization and accuracy, highlighting its potential for robust and reliable medical image analysis and disease diagnosis.

Table 5. Accuracy comparison of MediVision against existing methods across medical image datasets.

Dataset	Reference	Method	Benchmark (%)	Proposed (%)
Alzheimer's disease	Sethuraman et al. [60]	CNN	96.61	98.00
Breast ultrasound	Raptis et al. [39]	ResNet-50	93.05	97.45
Blood cell	Asghar et al. [62]	EfficientNet	94.70	95.45

Table 5. *Cont.*

Dataset	Reference	Method	Benchmark (%)	Proposed (%)
Chest X-ray	Singh et al. [63]	VGG-16	94.70	97.31
Chest CT scans	Mamun et al. [64]	DenseNet	92.00	94.83
Diabetic retinopathy	Wang et al. [65]	InceptionV3	91.10	95.43
Kidney diseases	Bhandari et al. [48]	MobileNet	99.39	98.36
Bone fracture multi-region	Aldhwani et al. [50]	Xception	97.35	98.56
Retinal OCT	Dai et al. [66]	AlexNet	95.00	95.44
Brain tumor	Khan et al. [67]	Hybrid CNN-LSTM	95.10	96.76

4.3. Statistical Analysis

To validate the effectiveness of the proposed MediVision hybrid CNN–LSTM model, statistical significance testing was performed against three baseline convolutional neural network architectures such as VGG16, VGG19, and ResNet50 across ten diverse medical imaging datasets. As summarized in Table 6, MediVision achieved the highest mean classification accuracy of 96.89%, outperforming VGG16 (92.98%), VGG19 (92.79%), and ResNet50 (81.17%). These consistent improvements demonstrate the superior generalization and stability of MediVision across multiple medical imaging modalities.

Table 6. Mean accuracy across different models.

Model	Mean Accuracy (%)
VGG16	92.98
VGG19	92.79
ResNet50	81.17
MediVision	96.89

Pairwise comparisons between MediVision and each baseline model were conducted using both the paired *t*-test and the non-parametric Wilcoxon signed-rank test, as summarized in Table 7. The Wilcoxon tests yielded statistically significant results ($p < 0.01$) for all comparisons, confirming that the observed improvements were not due to random variation. Although the paired *t*-test for the comparison with VGG16 was marginal ($p = 0.0597$), the non-parametric Wilcoxon result confirmed significance. Furthermore, after applying the Holm–Bonferroni correction to control for multiple comparisons, all adjusted *p*-values remained below 0.01, reinforcing the robustness of the findings.

Table 7. Statistical significance analysis of MediVision compared with baseline models.

Comparison	Mean Difference (%)	Paired *t*-Test (*p*)	Wilcoxon (*p*)	Holm Adj. (*p*)
MediVision vs. VGG16	+3.91	0.0597	0.0039	0.0039
MediVision vs. VGG19	+4.10	0.0489	0.0020	0.0059
MediVision vs. ResNet50	+15.72	0.0006	0.0020	0.0039

Finally, an overall comparison among all four models was performed using the Friedman test, a non-parametric alternative to repeated-measures ANOVA. As shown in Table 8, the test produced a χ^2 value of 27.00 with a *p*-value of 0.000006 ($p < 0.001$), indicating a highly significant difference in performance across models. These results collectively demonstrate that MediVision provides statistically and practically significant improvements over conventional CNN architectures for medical image classification tasks.

Table 8. Overall statistical comparison of all models using the Friedman test.

Test	χ^2	p-Value
Friedman test	27.00	0.000006

5. Conclusions

This study introduces a hybrid framework for medical image classification that seamlessly integrates CNNs, LSTMs, and an attention mechanism with a skip connection to enhance the feature extraction and classification accuracy. Trained and tested on ten diverse representative medical image datasets, MediVision outperforms widely used pretrained models such as VGG16, VGG19, and ResNet50 by 3.91%, 4.10%, and 15.72%, respectively, in terms of the mean classification accuracy. The results of both paired statistical tests and the Friedman test confirm that these improvements are statistically significant (p = 0.000006) across multiple medical imaging modalities. The model's strength lies in its ability to effectively capture both spatial and temporal features, with CNNs extracting spatial information, LSTMs learning sequential dependencies from image sequences, and the attention mechanism refining these features by focusing on the most relevant regions. The inclusion of a skip connection ensures efficient feature propagation, further boosting classification performance. To enhance the model transparency and support the clinical interpretability, Grad-CAM heatmaps were employed to visually highlight the critical image regions influencing classification decisions. Additionally, the datasets were carefully augmented and balanced, ensuring optimal generalization and robustness across different medical imaging domains.

For future work, we plan to conduct a detailed analysis of the computational complexity and runtime performance to evaluate MediVision's scalability in real-world clinical environments. A comprehensive ablation study will also be undertaken to assess the individual and combined contributions of key architectural components (i.e., CNN, LSTM, attention mechanism, and skip connection). Furthermore, we aim to enhance the feature extraction and interpretability by exploring self-supervised learning and transformer-based architectures. To improve the generalizability, we will extend MediVision to multi-modal medical data, such as integrating MRI, CT scans, and histopathological images. In addition, future work will explore the application of federated learning to facilitate collaborative model training across decentralized medical institutions while preserving data privacy and the integration of MediVision into clinical decision support systems to enable real-time AI-driven diagnostics.

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

Funding: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R829), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The codes and datasets used within this study are made available through the public repository Zenodo at https://doi.org/10.5281/zenodo.15498860.

Conflicts of Interest: Authors Md. Tanvir Hayat and Salman Md Sultan are employed by the company Innovative Skills Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A

The following algorithm summarizes the workflow of the proposed MediVision model.

Algorithm A1 Proposed MediVision Procedure

Input: Image I with three channels (H, W, C), number of CNN layers N, LSTM units U, and number of classes K;

Step 1—CNN Feature Extraction:
for $i = 1$ to N **do**
  $x \leftarrow \text{Conv2D}(filters_i, (3,3), \text{ReLU})$ on I;
  $x \leftarrow \text{MaxPooling2D}(2,2)$;
  $x \leftarrow \text{Dropout}(dropout_i)$;
end for
$x \leftarrow \text{Flatten}(x)$;
Step 2—Reshape for LSTM:
$x \leftarrow \text{Reshape}(1, -1)$;
Step 3—LSTM Processing:
$lstm_{out} \leftarrow \text{LSTM}(U, \text{return_sequences} = \text{True})$ on x;
$lstm_{out} \leftarrow \text{Dropout}(dropout_{lstm})$;
Step 4—Attention Mechanism:
$attention_{out} \leftarrow \text{Attention}(lstm_{out})$;
Step 5—Skip Connection:
$combined_{out} \leftarrow \text{Add}(attention_{out}, \sum lstm_{out})$;
Step 6—Final Classification Layer:
$y \leftarrow \text{Dense}(K, \text{softmax or sigmoid})$ on $combined_{out}$;
Step 7—Compilation and Training:
Loss: Sparse Categorical Crossentropy (multiclass) or Binary Crossentropy (binary);
Optimizer: Adam;
Metrics: Accuracy;
Apply Model Checkpoint and Early Stopping Callbacks;

Output: Classified label y;

References

1. Li, M.; Jiang, Y.; Zhang, Y.; Zhu, H. Medical image analysis using deep learning algorithms. *Front. Public Health* **2023**, *11*, 1273253. [CrossRef]
2. Rajula, H.S.R.; Verlato, G.; Manchia, M.; Antonucci, N.; Fanos, V. Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development, and treatment. *Medicina* **2020**, *56*, 455. [CrossRef] [PubMed]
3. Kingslake, J.; Dias, R.; Dawson, G.R.; Simon, J.; Goodwin, G.M.; Harmer, C.J.; Morriss, R.; Brown, S.; Guo, B.; Dourish, C.T.; et al. The effects of using the PReDicT Test to guide the antidepressant treatment of depressed patients: Study protocol for a randomised controlled trial. *Trials* **2017**, *18*, 558. [CrossRef]
4. Chen, C. Ascent of machine learning in medicine. *Nat. Mater.* **2019**, *18*, 407. [CrossRef]
5. Saria, S.; Butte, A.; Sheikh, A. Better medicine through machine learning: What's real, and what's artificial? *PLoS Med.* **2018**, *15*, e1002721. [CrossRef]
6. Churpek, M.M.; Yuen, T.C.; Winslow, C.; Meltzer, D.O.; Kattan, M.W.; Edelson, D.P. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit. Care Med.* **2016**, *44*, 368–374. [CrossRef]
7. Hernesniemi, J.A.; Mahdiani, S.; Tynkkynen, J.A.; Lyytikäinen, L.P.; Mishra, P.P.; Lehtimäki, T.; Eskola, M.; Nikus, K.; Antila, K.; Oksala, N. Extensive phenotype data and machine learning in prediction of mortality in acute coronary syndrome–the MADDEC study. *Ann. Med.* **2019**, *51*, 156–163. [CrossRef]
8. Paik, E.S.; Lee, J.W.; Park, J.Y.; Kim, J.H.; Kim, M.; Kim, T.J.; Choi, C.H.; Kim, B.G.; Bae, D.S.; Seo, S.W. Prediction of survival outcomes in patients with epithelial ovarian cancer using machine learning methods. *J. Gynecol. Oncol.* **2019**, *30*, e65. [CrossRef]

9. Abràmoff, M.D.; Lou, Y.; Erginay, A.; Clarida, W.; Amelon, R.; Folk, J.C.; Niemeijer, M. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investig. Ophthalmol. Vis. Sci.* **2016**, *57*, 5200–5206. [CrossRef]
10. Azizi, S.; Mustafa, B.; Ryan, F.; Beaver, Z.; Freyberg, J.; Deaton, J.; Loh, A.; Karthikesalingam, A.; Kornblith, S.; Chen, T.; et al. Big self-supervised models advance medical image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 3478–3488.
11. Xiao, C.; Choi, E.; Sun, J. Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 1419–1428. [CrossRef]
12. Shickel, B.; Tighe, P.J.; Bihorac, A.; Rashidi, P. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 1589–1604. [CrossRef] [PubMed]
13. Yadav, P.; Steinbach, M.; Kumar, V.; Simon, G. Mining electronic health records (EHRs) A survey. *ACM Comput. Surv. (CSUR)* **2018**, *50*, 1–40. [CrossRef]
14. Kim, H.E.; Cosa-Linan, A.; Santhanam, N.; Jannesari, M.; Maros, M.E.; Ganslandt, T. Transfer learning for medical image classification: A literature review. *BMC Med. Imaging* **2022**, *22*, 69. [CrossRef]
15. Puttagunta, M.; Ravi, S. Medical image analysis based on deep learning approach. *Multimed. Tools Appl.* **2021**, *80*, 24365–24398. [CrossRef]
16. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078. [CrossRef]
17. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Annual Conference on Neural Information Processing Systems 2014, Montreal, QC, Canada, 8–13 December 2014.
18. Chowdhary, C.L.; Mittal, M.; P, K.; Pattanaik, P.A.; Marszalek, Z. An efficient segmentation and classification system in medical images using intuitionist possibilistic fuzzy C-mean clustering and fuzzy SVM algorithm. *Sensors* **2020**, *20*, 3903. [CrossRef]
19. Kumar, R.R.; Shankar, S.V.; Jaiswal, R.; Ray, M.; Budhlakoti, N.; Singh, K. Advances in Deep Learning for Medical Image Analysis: A Comprehensive Investigation. *J. Stat. Theory Pract.* **2025**, *19*, 9. [CrossRef]
20. Kourounis, G.; Elmahmudi, A.A.; Thomson, B.; Hunter, J.; Ugail, H.; Wilson, C. Computer image analysis with artificial intelligence: A practical introduction to convolutional neural networks for medical professionals. *Postgrad. Med. J.* **2023**, *99*, 1287–1294. [CrossRef] [PubMed]
21. Shi, Z.; Hu, Y.; Mo, G.; Wu, J. Attention-based CNN-LSTM and XGBoost hybrid model for stock prediction. *arXiv* **2022**, arXiv:2204.02623.
22. de Sousa, P.M.; Carneiro, P.C.; Pereira, G.M.; Oliveira, M.M.; da Costa Junior, C.A.; de Moura, L.V.; Mattjie, C.; da Silva, A.M.; Macedo, T.A.A.; Patrocinio, A.C. Towards a Classification Model using CNN and Wavelets applied to COVID-19 CT images. *Res. Soc. Dev.* **2022**, *11*, e2411527919. [CrossRef]
23. Barzekar, H.; Yu, Z. C-Net: A reliable convolutional neural network for biomedical image classification. *Expert Syst. Appl.* **2022**, *187*, 116003. [CrossRef]
24. Zebari, N.A.; Mohammed, C.N.; Zebari, D.A.; Mohammed, M.A.; Zeebaree, D.Q.; Marhoon, H.A.; Abdulkareem, K.H.; Kadry, S.; Viriyasitavat, W.; Nedoma, J.; et al. A deep learning fusion model for accurate classification of brain tumours in Magnetic Resonance images. *CAAI Trans. Intell. Technol.* **2024**, *9*, 790–804. [CrossRef]
25. Zhang, Y.; Wang, S.; Zhao, H.; Guo, Z.; Sun, D. CT image classification based on convolutional neural network. *Neural Comput. Appl.* **2021**, *33*, 8191–8200. [CrossRef]
26. Hu, C.; Cao, N.; Zhou, H.; Guo, B. Medical Image Classification with a Hybrid SSM Model Based on CNN and Transformer. *Electronics* **2024**, *13*, 3094. [CrossRef]
27. Kadhim, Y.A.; Guzel, M.S.; Mishra, A. A Novel hybrid machine learning-based system using deep learning techniques and meta-heuristic algorithms for various medical datatypes classification. *Diagnostics* **2024**, *14*, 1469. [CrossRef]
28. Aydın, M.; Kuş, Z.; Akçelik, Z.K. A Hybrid CNN-Tree Based Model for Enhanced Image Classification Performance. In Proceedings of the 2024 32nd Signal Processing and Communications Applications Conference (SIU), Mersin, Türkiye, 15–18 May 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–4.
29. Bouslihim, I.; Cherif, W.; Kissi, M. Application of a hybrid EfficientNet-SVM model to medical image classification. In Proceedings of the 2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA), Casablanca, Morocco, 22–23 November 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6.
30. Ajlouni, N.; Özyavaş, A.; Takaoğlu, M.; Takaoğlu, F.; Ajlouni, F. Medical image diagnosis based on adaptive Hybrid Quantum CNN. *BMC Med. Imaging* **2023**, *23*, 126. [CrossRef]
31. Abbas, Q. An intelligent medical image classification system using few-shot learning. *Concurr. Comput. Pract. Exp.* **2023**, *35*, e7451. [CrossRef]

32. Borji, A.; Kronreif, G.; Angermayr, B.; Hatamikia, S. Advanced hybrid deep learning model for enhanced evaluation of osteosarcoma histopathology images. *Front. Med.* **2025**, *12*, 1555907. [CrossRef]
33. Hussain, T.; Shouno, H.; Hussain, A.; Hussain, D.; Ismail, M.; Mir, T.H.; Hsu, F.R.; Alam, T.; Akhy, S.A. EFFResNet-ViT: A fusion-based convolutional and vision transformer model for explainable medical image classification. *IEEE Access* **2025**, *13*, 54040–54068. [CrossRef]
34. Salehin, I.; Islam, M.S.; Amin, N.; Baten, M.A.; Noman, S.; Saifuzzaman, M.; Yazmyradov, S. Real-Time Medical Image Classification with ML Framework and Dedicated CNN–LSTM Architecture. *J. Sens.* **2023**, *2023*, 3717035. [CrossRef]
35. Gonçalves, T.; Rio-Torto, I.; Teixeira, L.F.; Cardoso, J.S. A survey on attention mechanisms for medical applications: Are we moving toward better algorithms? *IEEE Access* **2022**, *10*, 98909–98935. [CrossRef]
36. Cheng, D.; Deng, J.; Xiao, J.; Yanyan, M.; Kang, J.; Gai, J.; Zhang, B.; Zhao, F. Attention based multi-scale nested network for biomedical image segmentation. *Heliyon* **2024**, *10*, e33892. [CrossRef]
37. Liu, S.; Yue, W.; Guo, Z.; Wang, L. Multi-branch CNN and grouping cascade attention for medical image classification. *Sci. Rep.* **2024**, *14*, 15013. [CrossRef]
38. Altwijri, O.; Alanazi, R.; Aleid, A.; Alhussaini, K.; Aloqalaa, Z.; Almijalli, M.; Saad, A. Novel deep-learning approach for automatic diagnosis of Alzheimer's disease from MRI. *Appl. Sci.* **2023**, *13*, 13051. [CrossRef]
39. 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]
40. Acevedo, A.; Alférez, S.; Merino, A.; Puigví, L.; Rodellar, J. Recognition of peripheral blood cell images using convolutional neural networks. *Comput. Methods Programs Biomed.* **2019**, *180*, 105020. [CrossRef]
41. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **2018**, *172*, 1122–1131. [CrossRef]
42. Rahman, T.; Khandakar, A.; Kadir, M.A.; Islam, K.R.; Islam, K.F.; Mazhar, R.; Hamid, T.; Islam, M.T.; Kashem, S.; Mahbub, Z.B.; et al. Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization. *IEEE Access* **2020**, *8*, 191586–191601. [CrossRef]
43. Bacellar, G.C.; Chandrappa, M.; Kulkarni, R.; Dey, S. COVID-19 chest X-ray image classification using deep learning. *medRxiv* **2021**. [CrossRef]
44. Asif, S.; Wenhui, Y.; Jin, H.; Jinhai, S. Classification of COVID-19 from chest X-ray images using deep convolutional neural network. In Proceedings of the 2020 IEEE 6th International Conference on Computer and Communications (ICCC), Chengdu, China, 11–14 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 426–433.
45. Altayeb, M.; Arabiat, A.; Al-Ghraibah, A. Detection and classification of pneumonia using the Orange3 data mining tool. *Int. J. Electr. Comput. Eng. (2088-8708)* **2024**, *14*, 6894–6903. [CrossRef]
46. Jassim, O.A.; Abed, M.J.; Saied, Z.H.S. Deep learning techniques in the cancer-related medical domain: A transfer deep learning ensemble model for lung cancer prediction. *Baghdad Sci. J.* **2024**, *21*, 1101. [CrossRef]
47. Benítez, V.E.C.; Matto, I.C.; Román, J.C.M.; Noguera, J.L.V.; García-Torres, M.; Ayala, J.; Pinto-Roa, D.P.; Gardel-Sotomayor, P.E.; Facon, J.; Grillo, S.A. Dataset from fundus images for the study of diabetic retinopathy. *Data Brief* **2021**, *36*, 107068. [CrossRef]
48. Bhandari, M.; Yogarajah, P.; Kavitha, M.S.; Condell, J. Exploring the capabilities of a lightweight CNN model in accurately identifying renal abnormalities: Cysts, stones, and tumors, using LIME and SHAP. *Appl. Sci.* **2023**, *13*, 3125. [CrossRef]
49. AlMahadin, G.; Owida, H.A.; Al-Nabulsi, J.I.; Turab, N.; Al Hawamdeh, N. Automated detection of kidney masses lesions using a deep learning approach. *Int. J. Artif. Intell. ISSN* **2024**, *2252*, 2863. [CrossRef]
50. Aldhyani, T.; Ahmed, Z.A.; Alsharbi, B.M.; Ahmad, S.; Al-Adhaileh, M.H.; Kamal, A.H.; Almaiah, M.; Nazeer, J. Diagnosis and detection of bone fracture in radiographic images using deep learning approaches. *Front. Med.* **2025**, *11*, 1506686. [CrossRef]
51. Gholami, P.; Roy, P.; Parthasarathy, M.K.; Lakshminarayanan, V. OCTID: Optical coherence tomography image database. *Comput. Electr. Eng.* **2020**, *81*, 106532. [CrossRef]
52. Hasan, N.; Ahmed, M.F.; Nasif, M.A.; Haq, M.R.; Rahman, M. Hybrid feature extraction approach for robust brain tumor classification: HOG, GLCM, and artificial neural network. In Proceedings of the 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), Dhaka, Bangladesh, 2–4 May 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1292–1297.
53. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
54. Lemley, J.; Bazrafkan, S.; Corcoran, P. Smart augmentation learning an optimal data augmentation strategy. *IEEE Access* **2017**, *5*, 5858–5869. [CrossRef]
55. Moreno-Barea, F.J.; Jerez, J.M.; Franco, L. Improving classification accuracy using data augmentation on small data sets. *Expert Syst. Appl.* **2020**, *161*, 113696. [CrossRef]
56. Singh, D.; Singh, B. Investigating the impact of data normalization on classification performance. *Appl. Soft Comput.* **2020**, *97*, 105524. [CrossRef]

57. Wang, C.; Zhang, L.; Wei, W.; Zhang, Y. Dynamic super-pixel normalization for robust hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5505713. [CrossRef]
58. Chen, B.; Li, H.; Luo, W. Image processing operations identification via convolutional neural network. *arXiv* **2017**, arXiv:1709.02908. [CrossRef]
59. Hao, S.; Lee, D.H.; Zhao, D. Sequence to sequence learning with attention mechanism for short-term passenger flow prediction in large-scale metro system. *Transp. Res. Part C Emerg. Technol.* **2019**, *107*, 287–300. [CrossRef]
60. Sethuraman, S.K.; Malaiyappan, N.; Ramalingam, R.; Basheer, S.; Rashid, M.; Ahmad, N. Predicting Alzheimer's disease using deep neuro-functional networks with resting-state fMRI. *Electronics* **2023**, *12*, 1031. [CrossRef]
61. Iqbal, M.A.; Gil, J.M.; Kim, S.K. Attention-Driven Hybrid Ensemble Approach with Bayesian Optimization for Accurate Energy Forecasting in Jeju Island's Renewable Energy System. *IEEE Access* **2025**, *13*, 7986–8010. [CrossRef]
62. Asghar, R.; Kumar, S.; Hynds, P. Automatic classification of 10 blood cell subtypes using transfer learning via pre-trained convolutional neural networks. *Informatics Med. Unlocked* **2024**, *49*, 101542. [CrossRef]
63. Singh, D.; Kumar, V.; Yadav, V.; Kaur, M. Deep neural network-based screening model for COVID-19-infected patients using chest X-ray images. *Int. J. Pattern Recognit. Artif. Intell.* **2021**, *35*, 2151004. [CrossRef]
64. Mamun, M.; Mahmud, M.I.; Meherin, M.; Abdelgawad, A. Lcdctcnn: Lung cancer diagnosis of ct scan images using cnn based model. In Proceedings of the 2023 10th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 23–24 March 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 205–212.
65. Wang, Z.; Yin, Y.; Shi, J.; Fang, W.; Li, H.; Wang, X. Zoom-in-net: Deep mining lesions for diabetic retinopathy detection. In Proceedings of the Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, 11–13 September 2017; Proceedings, Part III 20; Springer: Berlin/Heidelberg, Germany, 2017; pp. 267–275.
66. Dai, H.; Yang, Y.; Yue, X.; Chen, S. Improving retinal OCT image classification accuracy using medical pre-training and sample replication methods. *Biomed. Signal Process. Control* **2024**, *91*, 106019. [CrossRef]
67. Khan, S.U.R.; Zhao, M.; Asif, S.; Chen, X. Hybrid-NET: A fusion of DenseNet169 and advanced machine learning classifiers for enhanced brain tumor diagnosis. *Int. J. Imaging Syst. Technol.* **2024**, *34*, e22975. [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.

diagnostics

MDPI

Article

Leveraging Large-Scale Public Data for Artificial Intelligence-Driven Chest X-Ray Analysis and Diagnosis

Farzeen Khalid Khan [1,†], Waleed Bin Tahir [1,†], Mu Sook Lee [2], Jin Young Kim [2], Shi Sub Byon [1], Sun-Woo Pi [3] and Byoung-Dai Lee [3,*]

1 AI Laboratory, HealthHub Co., Ltd., Seoul 06524, Republic of Korea
2 Department of Radiology, Keimyung University Dongsan Hospital, Daegu 24601, Republic of Korea
3 Division of AI and Computer Engineering, Kyonggi University, Suwon 16227, Republic of Korea
* Correspondence: blee@kgu.ac.kr; Tel.: +82-31-249-9676
† These authors contributed equally to this work.

Abstract

Background: Chest X-ray (CXR) imaging is crucial for diagnosing thoracic abnormalities; however, the rising demand burdens radiologists, particularly in resource-limited settings. **Method**: We used large-scale, diverse public CXR datasets with noisy labels to train general-purpose deep learning models (ResNet, DenseNet, EfficientNet, and DLAD-10) for multi-label classification of thoracic conditions. Uncertainty quantification was incorporated to assess model reliability. Performance was evaluated on both internal and external validation sets, with analyses of data scale, diversity, and fine-tuning effects. **Result**: EfficientNet achieved the highest overall area under the receiver operating characteristic curve (0.8944) with improved sensitivity and F1-score. Moreover, as training data volume increased—particularly using multi-source datasets—both diagnostic performance and generalizability were enhanced. Although larger datasets reduced predictive uncertainty, conditions such as tuberculosis remained challenging due to limited high-quality samples. **Conclusions**: General-purpose deep learning models can achieve robust CXR diagnostic performance when trained on large-scale, diverse public datasets despite noisy labels. However, further targeted strategies are needed for underrepresented conditions.

Keywords: artificial intelligence; chest X-ray interpretation; deep learning; noisy labels; uncertainty quantification

1. Introduction

Chest X-ray (CXR) is an essential diagnostic tool in healthcare, providing a cost-effective, non-invasive, and widely accessible method for detecting thoracic diseases [1–3]. However, the increasing demand for CXR studies has placed a burden on radiologists, leading to diagnostic delays and highlighting the need for scalable automated diagnostic solutions to enhance efficiency and accuracy [4]. Although deep learning has achieved radiologist-level performance in CXR interpretation [5–8], its clinical integration is hampered by the reliance on private, expert-annotated datasets that limit reproducibility—especially in resource-constrained settings. Public datasets, including National Institute of Health (NIH) Chest X-ray-14 [9], CheXpert [10], and Medical Information Mart for Intensive Care-CXR (MIMIC-CXR) [11], offer diverse training data; however, noisy labels pose challenges for model reliability [12].

In this study, we aimed to demonstrate that the scale and diversity of training data can offset imperfections in label quality and model specificity. We leveraged large-scale, diverse

public CXR datasets to train general-purpose image classification models that robustly classify multiple thoracic conditions without the need for specialized architectures. Moreover, we incorporated uncertainty quantification (UQ) to assess predictive performance and model reliability, offering insights into their real-world clinical applicability. Our findings indicate that, despite noisy labels, general-purpose deep learning models can achieve acceptable diagnostic performance when trained on expansive, varied datasets, thereby supporting the development of accessible, scalable artificial intelligence (AI) diagnostic tools for resource-limited settings.

While our study leverages publicly available datasets and standard deep learning models, its novelty lies in the systematic integration and analysis of these components. Specifically, we:

- Aggregate and harmonize 17 public CXR datasets, addressing label inconsistencies and cross-dataset duplication to enable robust generalization analysis and broaden coverage of rare diseases.
- Perform comprehensive UQ across multiple dataset scales, providing new insights into model stability, calibration, and reliability beyond what prior CXR studies have reported.
- Conduct class-specific uncertainty analysis for clinically important but underexplored categories, including tuberculosis, pneumothorax, masses/nodules, and the heterogeneous "Other" class, offering a granular understanding of class-dependent behavior.
- Characterize scaling-law behavior in both performance and uncertainty, demonstrating how increasing dataset size affects diagnostic accuracy and epistemic uncertainty, and offering actionable guidance for data collection and clinical deployment strategies.

These contributions underscore that our research extends beyond incremental improvements, providing new methodological and practical insights for scalable, uncertainty-aware CXR analysis.

2. Materials and Methods

The Institutional Review Board of Keimyung University Dongsan Medical Center approved this retrospective study and waived the requirement for written informed consent. All procedures followed the approved protocol.

2.1. Target Conditions

To maximize clinical relevance and impact, this study focused on the following six prevalent and clinically significant thoracic conditions: pneumonia, pleural effusion, tuberculosis, mass, consolidation, and pneumothorax. These conditions include both specific diseases (e.g., pneumonia and tuberculosis) and radiological findings (e.g., pleural effusion, mass, and consolidation) that frequently indicate an underlying pathology. The selection was guided by global disease burden, potential for early intervention, and the diagnostic importance of CXR in routine clinical workflows. Pneumonia and tuberculosis remain the leading causes of morbidity and mortality, particularly in low-resource settings with limited diagnostic tools [13]. Pleural effusion, masses, and consolidation are common thoracic abnormalities that, if detected early, can alter patient management and improve outcomes. Pneumothorax, although less prevalent, is a life-threatening condition requiring immediate identification and intervention [14]. For cases involving abnormal findings not otherwise classified among these six target diseases, an "Other" disease category was included to ensure broad diagnostic coverage without compromising specificity. This strategy enhances the clinical utility while maintaining focus on conditions with the greatest potential impact on patient care.

https://doi.org/10.3390/diagnostics16010146

2.2. Data Collection and Preparation

2.2.1. Development Dataset

To construct a comprehensive training dataset, we aggregated 17 public CXR datasets (Table 1) into 894,373 unique frontal-view images, excluding non-frontal projections to ensure clinical consistency. This strategy enhanced model generalizability by incorporating diverse patient demographics, imaging protocols, and disease prevalence. Five major datasets (NIH Chest X-Ray-14 [9], CheXpert [10], MIMIC-CXR [11], PadChest [15], and BRAX [16]) provided a substantial portion of the development dataset. These datasets employed natural language processing (NLP) tools or deep learning algorithms to automatically extract labels from radiological reports. These methods facilitate large-scale label generation and introduce variability in label quality, particularly in disease categories requiring subjective interpretation. To address this, we harmonized labels across datasets, treated low-confidence labels consistently, and ensured mutual exclusivity in training and test sets (see Appendix B for full details).

Table 1 highlights class imbalance—a pervasive challenge in public CXR datasets—across target diseases. To address this, we strategically adjusted per-disease sample counts and augmented underrepresented classes (tuberculosis, mass, consolidation, pneumothorax) via image duplication and online augmentation techniques. Augmentation was restricted to training and validation sets to preserve the integrity of the uniquely composed test set, thereby preventing data leakage. Table 2 provides the final class distribution across all dataset splits.

2.2.2. External Validation Dataset

To assess generalization and fine-tuning efficacy, we constructed an independent validation dataset of 3031 unique-patient, posterior–anterior CXR images (July 2017–December 2019) from our hospital's picture archiving and communication system (Table 2), ensuring no overlap with training data. This dataset enabled (1) an unbiased assessment of baseline model performance and (2) evaluation of domain-specific fine-tuning. To evaluate domain-specific fine-tuning, we used 10-fold cross-validation (training on $k-1$ folds and validating on the remaining fold), whereas non-fine-tuned models were directly tested on the external set to isolate domain adaptation effects.

https://doi.org/10.3390/diagnostics16010146

Table 1. Summary of publicly available CXR datasets used in the study (NLP: Natural Language Processing, NN: Neural Network).

Datasets	No. of Labels	Annotation Method	No. of CXRs								
			Pneumonia	Pleural Effusion	Tuberculosis	Mass	Consolidation	Pneumothorax	Other Diseases	No Finding	Total
NIH Chest X-ray-14 [9]	14	An NLP Tool	1431	13,317	0	11,207	4667	5302	36,251	60,361	112,120
CheXpert [10]	14	An NLP Tool	3738	79,713	0	0	12,170	16,257	144,789	15,892	224,316
MIMIC-CXR [11]	14	An NLP Tool	26,221	76,954	0	0	14,675	14,257	180,251	143,351	377,110
PadChest [15]	19	27% of reports were manually annotated and the rest using a supervised NN.	4138	5441	647	3034	1426	345	52,515	28,999	160,000
VinDr-CXR [17]	28	Radiologists	1229	1430	1006	1080	444	133	6809	10,606	100,000
RSNA Pneumonia Detection Challenge [18]	2	Radiologists	6012	0	0	0	0	0	0	20,672	26,684
SIIM-ACR Pneumothorax Segmentation [19]	2	Radiologists	0	0	0	0	0	2379	0	8296	10,675
JSRT [20]	2	Radiologists	0	0	0	154	0	0	93	0	247
Shenzhen Hospital CXR Set [21]	4	Radiologists	0	34	336	132	18	0	272	326	662
Montgomery County chest X-ray set (MC) [21]	2	Radiologists	0	0	58	0	0	0	0	80	138
BRAX [16]	14	An NLP Tool	264	532	0	0	1120	38	3612	14,782	40,967
COVID-19, Pneumonia and Normal Chest X-ray PA Dataset [22]	3	Radiologists	1525	0	0	0	0	0	0	1525	4575
Chest X-Ray Images (Pneumonia) [23]	2	Radiologists	3875	0	0	0	0	0	0	1341	5856
Tuberculosis (TB) Chest X-ray Database [24]	2	Radiologists	0	0	700	0	0	0	0	3500	4200
TBX11K [25]	5	Radiologists	0	0	800	0	0	0	3800	3800	11,200
Belarus Dataset [26]	1	Radiologists	0	0	304	0	0	0	0	0	304
Chest X-rays tuberculosis from India [27]	2	Radiologists	0	0	78	0	0	0	0	77	156
Total	–	–	48,433	177,421	3929	15,607	34,520	38,711	428,392	313,608	1,079,210

https://doi.org/10.3390/diagnostics16010146

Table 2. Data distribution per target disease for the training, validation, and testing sets of the deep learning models.

Datasets	Development Dataset				External Validation Dataset
	No. of Total CXR Images	Training	Validation	Testing	No. of Total CXR Images
Pneumonia	48,433	40,000	9237	2421	10
Pleural Effusion	177,421	40,000	9721	2990	422
Tuberculosis	3929	40,000	7820	1237	318
Mass	15,607	40,000	9763	2945	69
Consolidation	34,520	40,000	9884	2988	0
Pneumothorax	38,711	40,000	9982	2998	1312
Other diseases	428,299	68,525	12,277	3024	843
No finding	313,701	39,999	9807	2936	240

2.3. Deep Learning Models

The three deep learning models used in this study—ResNet-50 [28], DenseNet-121 [29], and EfficientNet-B5 [30]—were selected for their proven effectiveness in natural image classification tasks and served as robust baselines for comparison. Furthermore, Deep Learning-based Algorithm for Detecting 10 Abnormalities (DLAD-10) [5], a high-performance model designed for CXR classification and recognized as a base model in commercial services due to its superior performance, was also included in the experiment.

All classification models were initialized with ImageNet-pretrained weights, and their inputs were normalized using the standard ImageNet mean and standard deviation. This preprocessing preserves the activation scaling expected by pretrained backbones and contributes to stable optimization during fine-tuning. In contrast, the U-Net segmentation model used to extract lung masks was trained entirely from scratch, without ImageNet initialization; therefore, segmentation inputs were processed using only min–max intensity normalization, ensuring consistency between preprocessing and initialization strategies.

For classification experiments, all models were trained and evaluated on the same pooled dataset to ensure consistent comparisons. Training was performed using the Adam optimizer with an initial learning rate of 1×10^{-3} and a batch size of 32. A comprehensive list of all hyperparameters, ensuring full reproducibility, is provided in Appendix A.

Although more recent architectures—including Vision Transformers [31], ConvNeXt [32], and large foundation models [33]—have demonstrated strong performance in medical image analysis, they require substantially greater computational resources and specialized pretraining pipelines. Consistent with our objective of evaluating scalable and broadly deployable diagnostic models, we focused on representative convolutional architectures that are widely used, computationally accessible, and well supported in open-source frameworks. This design choice is particularly relevant for resource-limited clinical environments, where training or deploying heavy transformer-based models or large foundation models may not be feasible. Accordingly, our evaluation centers on general-purpose convolutional neural network-based classifiers, including DLAD-10, an early commercial model, to assess whether large-scale public datasets alone can enable clinically acceptable performance without reliance on computationally intensive architectures.

2.4. UQ Using Monte Carlo (MC) Dropout

We employed MC dropout [34] to estimate model uncertainty by enabling stochastic dropout during inference, thereby generating multiple predictions for each input. This approach captures both aleatoric uncertainty (inherent data noise) and epistemic uncertainty (model or data limitations). Multiple forward passes were performed, and their predictions

were aggregated to compute predictive entropy (PE) [35]—reflecting total uncertainty—and variance of predictions (VP) [36]—a direct measure of epistemic uncertainty. For robust estimation, we used a dropout rate of 0.4 and 20 stochastic passes, balancing computational efficiency with reliable metrics. The final class prediction was selected by averaging the probabilities across all passes. The dropout rate follows the standard configuration in previous studies [30,37], and the number of stochastic passes was chosen based on sensitivity analysis on the test set (see Appendix C for details).

2.5. Statistical Analysis

Diagnostic performance was evaluated using precision, sensitivity, specificity, F1-score, and area under the receiver operating characteristic curve (AUROC). These metrics provided a comprehensive assessment of the ability of the model to accurately classify disease states in chest radiographs, offering insights into discriminative capability and clinical utility. A 10-fold cross-validation was conducted, with performance reported as the average values across all folds, and 95% confidence intervals were calculated to assess variability and reliability. Stratified splitting was applied to ensure that each fold was representative of all target diseases, and multiple random seeds were evaluated to prevent folds with zero samples in low-prevalence classes; a seed was ultimately selected that preserved at least two samples per disease in every fold.

Class-specific decision thresholds (provided in Tables 3 and 4) were optimized on the validation set by selecting operating points that favor sensitivity while avoiding excessive loss of specificity. Consistent with the clinical imperative to minimize false-negative findings—particularly for conditions such as tuberculosis, pneumonia, or pneumothorax, where missed diagnoses carry substantial risk—sensitivity was modestly prioritized. This approach acknowledges that while false positives necessitate additional confirmatory evaluation, they present significantly lower clinical risk than delayed diagnosis and treatment.

Table 3. Performance of deep learning models and selected thresholds for each disease using the development dataset.

Target Diseases	Threshold				ResNet	DenseNet	EfficientNet	DLAD-10
	ResNet	DenseNet	EfficientNet	DLAD-10				
Pneumonia	0.28	0.24	0.255	0.20	(0.31, 0.74, 0.74, 0.43)	(0.30, 0.74, 0.74, 0.43)	(0.37, 0.80, 0.79, 0.51)	(0.33, 0.74, 0.76, 0.46)
Pleural Effusion	0.25	0.145	0.215	0.16	(0.45, 0.80, 0.80, 0.57)	(0.45, 0.78, 0.81, 0.57)	(0.47, 0.83, 0.81, 0.60)	(0.46, 0.81, 0.81, 0.59)
Tuberculosis	0.50	0.50	0.50	0.50	(0.98, 0.73, 1.0, 0.84)	(0.98, 0.81, 1.00, 0.89)	(0.98, 0.89, 1.00, 0.93)	(0.99, 0.86, 1.00, 0.92)
Mass	0.35	0.29	0.10	0.29	(0.51, 0.82, 0.84, 0.63)	(0.51, 0.84, 0.84, 0.64)	(0.57, 0.87, 0.87, 0.69)	(0.56, 0.86, 0.87, 0.68)
Consolidation	0.18	0.19	0.20	0.20	(0.36, 0.75, 0.74, 0.49)	(0.36, 0.74, 0.74, 0.48)	(0.40, 0.76, 0.77, 0.52)	(0.38, 0.76, 0.76, 0.51)
Pneumothorax	0.18	0.15	0.18	0.18	(0.45, 0.81, 0.81, 0.58)	(0.47, 0.81, 0.82, 0.60)	(0.58, 0.86, 0.88, 0.70)	(0.59, 0.86, 0.88, 0.70)
Other diseases	0.35	0.30	0.38	0.38	(0.29, 0.66, 0.67, 0.40)	(0.27, 0.65, 0.65, 0.38)	(0.28, 0.65, 0.67, 0.39)	(0.27, 0.64, 0.66, 0.38)
No finding	0.23	0.24	0.30	0.24	(0.40, 0.77, 0.78, 0.53)	(0.40, 0.77, 0.77, 0.53)	(0.45, 0.80, 0.81, 0.58)	(0.44, 0.80, 0.80, 0.57)

Data in parentheses indicate precision, sensitivity, specificity, and F1-score.

https://doi.org/10.3390/diagnostics16010146

Table 4. Performance of deep learning models and selected thresholds for each disease using the external validation dataset.

Target Diseases	With Fine-Tuning				Without Fine-Tuning	
	Thresholds		EfficientNet	DLAD-10	EfficientNet	DLAD-10
	EfficientNet	DLAD-10				
Pneumonia	0.255	0.20	(0.00, 0.00, 1.00, 0.00)	(0.00, 0.00, 1.00, 0.00)	(0.01, 0.40, 0.76, 0.01)	(0.01, 0.70, 0.68, 0.01)
Pleural Effusion	0.215	0.16	(0.59, 0.78, 0.91, 0.67)	(0.55, 0.83, 0.89, 0.66)	(0.30, 0.70, 0.74, 0.42)	(0.31, 0.74, 0.73, 0.44)
Tuberculosis	0.5	0.50	(0.82, 0.88, 0.98, 0.84)	(0.84, 0.80, 0.98, 0.80)	(0.15, 0.08, 0.95, 0.10)	(0.23, 0.08, 0.97, 0.12)
Mass	0.1	0.29	(0.23, 0.48, 0.96, 0.31)	(0.40, 0.19, 0.99, 0.21)	(0.04, 0.41, 0.74, 0.07)	(0.05, 0.39, 0.84, 0.09)
Consolidation	–	–	–	–	–	–
Pneumothorax	0.18	0.18	(0.94, 0.97, 0.96, 0.96)	(0.90, 0.98, 0.92, 0.94)	(0.88, 0.84, 0.91, 0.86)	(0.88, 0.86, 0.91, 0.87)
Other diseases	0.38	0.38	(0.73, 0.87, 0.87, 0.79)	(0.73, 0.84, 0.88, 0.78)	(0.31, 0.68, 0.42, 0.43)	(0.32, 0.63, 0.48, 0.42)
No finding	0.3	0.24	(0.61, 0.77, 0.96, 0.68)	(0.60, 0.73, 0.96, 0.65)	(0.24, 0.87, 0.76, 0.37)	(0.23, 0.92, 0.74, 0.37)

Data in parentheses indicate precision, sensitivity, specificity, and F1-Score.

3. Results

3.1. Model Performance on Internal Validation Sets

Figure 1 and Table 3 summarize the performance of four deep learning models—ResNet, DenseNet, EfficientNet, and DLAD-10—on the internal validation sets. EfficientNet achieved the highest overall performance, with an average AUROC (0.8944) among all models. This model consistently outperformed the other models in sensitivity (0.81) and F1-score (0.61), reflecting its ability to accurately detect diseases while maintaining a balance between precision and recall (Table 3). DLAD-10 closely followed with an average AUROC of 0.8810, and although EfficientNet had a slight edge overall, disease-specific analysis revealed instances where DLAD-10 outperformed EfficientNet in certain categories, indicating that the relative advantage can vary by disease. In contrast, ResNet and DenseNet had slightly lower average AUROCs (0.8654 and 0.8626, respectively), with DenseNet exhibiting a marginally better F1-score than ResNet.

3.2. Model Performance on External Validation Set

For the external validation set, analysis focused on EfficientNet, which demonstrated the best performance using the development dataset. As shown in Figure 2 and Table 4, EfficientNet demonstrated comparable or superior performance to DLAD-10 across most disease categories, consistent with its performance on the development dataset. Fine-tuning with a portion of the external validation dataset consistently improved the results of both models across most target diseases. However, performance variations were observed between internal and external datasets. Pneumonia, tuberculosis, and mass showed higher AUROC scores internally, whereas the external dataset yielded better performance for the remaining diseases. These differences likely stemmed from the smaller number of samples for certain diseases in the external dataset, affecting generalizability. Overall, EfficientNet maintained strong diagnostic accuracy on the external validation set, demonstrating adaptability to diverse clinical scenarios.

3.3. Impact of Training Data Scale and Diversity on Model Effectiveness

3.3.1. Diagnostic Performance Across Data Scales

To evaluate the impact of training data scale, we incrementally increased the dataset size (20%, 40%, 60%, and 100% of 40,000 CXR images per disease) while maintaining proportional representation from source datasets. Experiments were conducted under controlled conditions using a development dataset, with MC dropout applied during inference to estimate uncertainty.

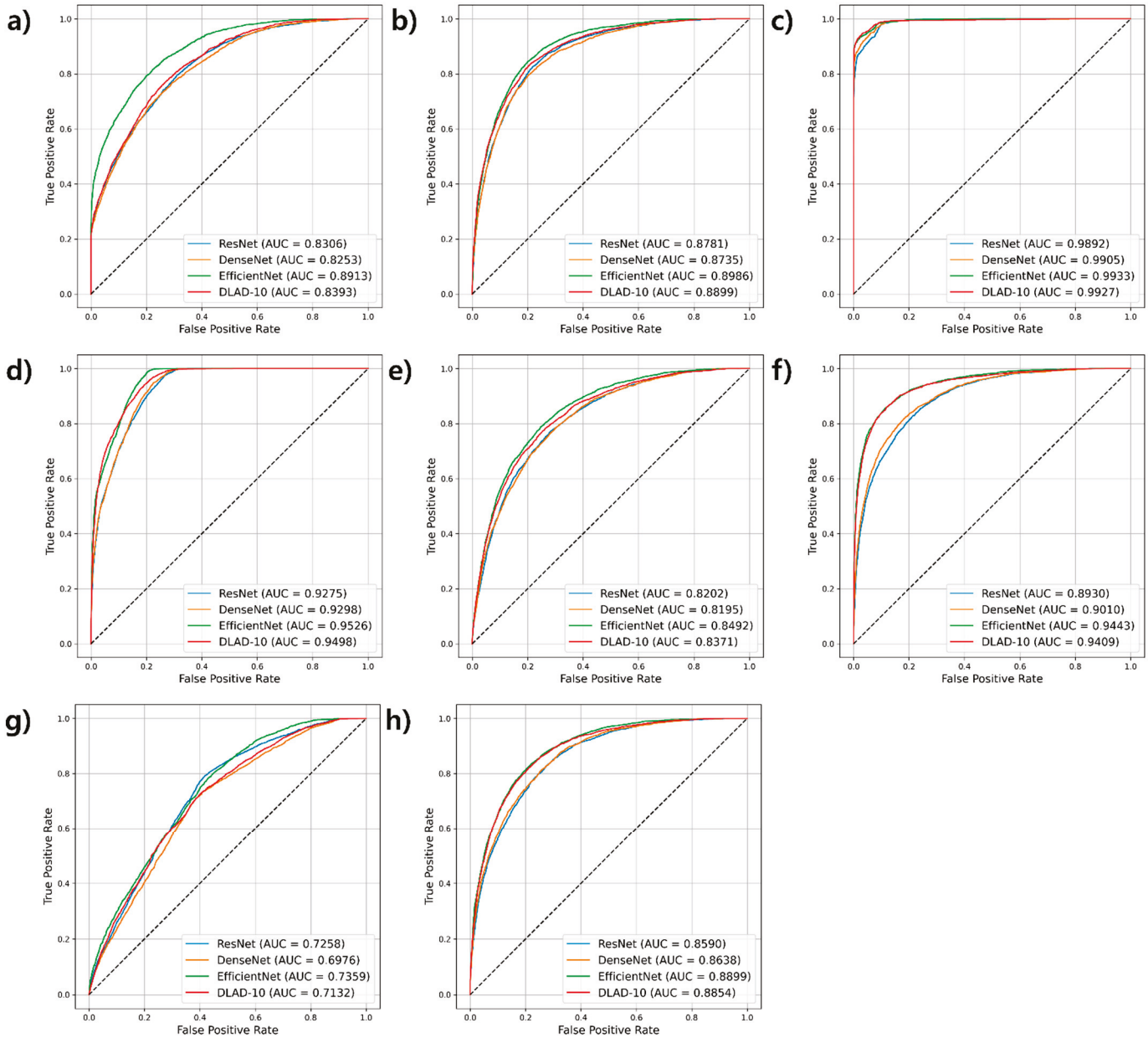


Figure 1. Comparison of AUROCs for deep learning models across different diseases using the development dataset. (**a**) Pneumonia, (**b**) Pleural effusion, (**c**) Tuberculosis, (**d**) Mass, (**e**) Consolidation, (**f**) Pneumothorax, (**g**) Other diseases, and (**h**) No finding (AUROC: area under the receiver operating characteristic curve).

As presented in Table 5, model performance consistently improved with larger datasets: the mean AUROC rose from 0.8680 (20%) to 0.8937 (100%), and precision (0.47–0.51), sensitivity (0.80–0.82), specificity (0.78–0.81), and F1-score (0.57–0.61) all increased. Disease-specific results revealed significant gains for pneumothorax (0.9115–0.9437), pleural effusion (0.8657–0.8978), mass (0.9416–0.9520), and pneumonia (0.8414–0.8905), underscoring the benefit of expanded training data. In contrast, tuberculosis showed a slight AUROC decline (0.9937–0.9926). Given the relatively small number of unique tuberculosis cases (3929 studies) and the substantial reliance on duplicated samples to reach the target dataset size, this trend may reflect the limited diversity of underlying image features. However, other factors—including label variability across datasets or differences in acquisition protocols—may also contribute, and therefore this behavior should be interpreted with caution. Consolidation and the "Other" disease category also

improved, although less dramatically. Overall, these findings confirm that larger, more diverse training datasets enhance model performance for well-represented diseases but underscore the need for targeted strategies when addressing rare or underrepresented conditions, including tuberculosis.

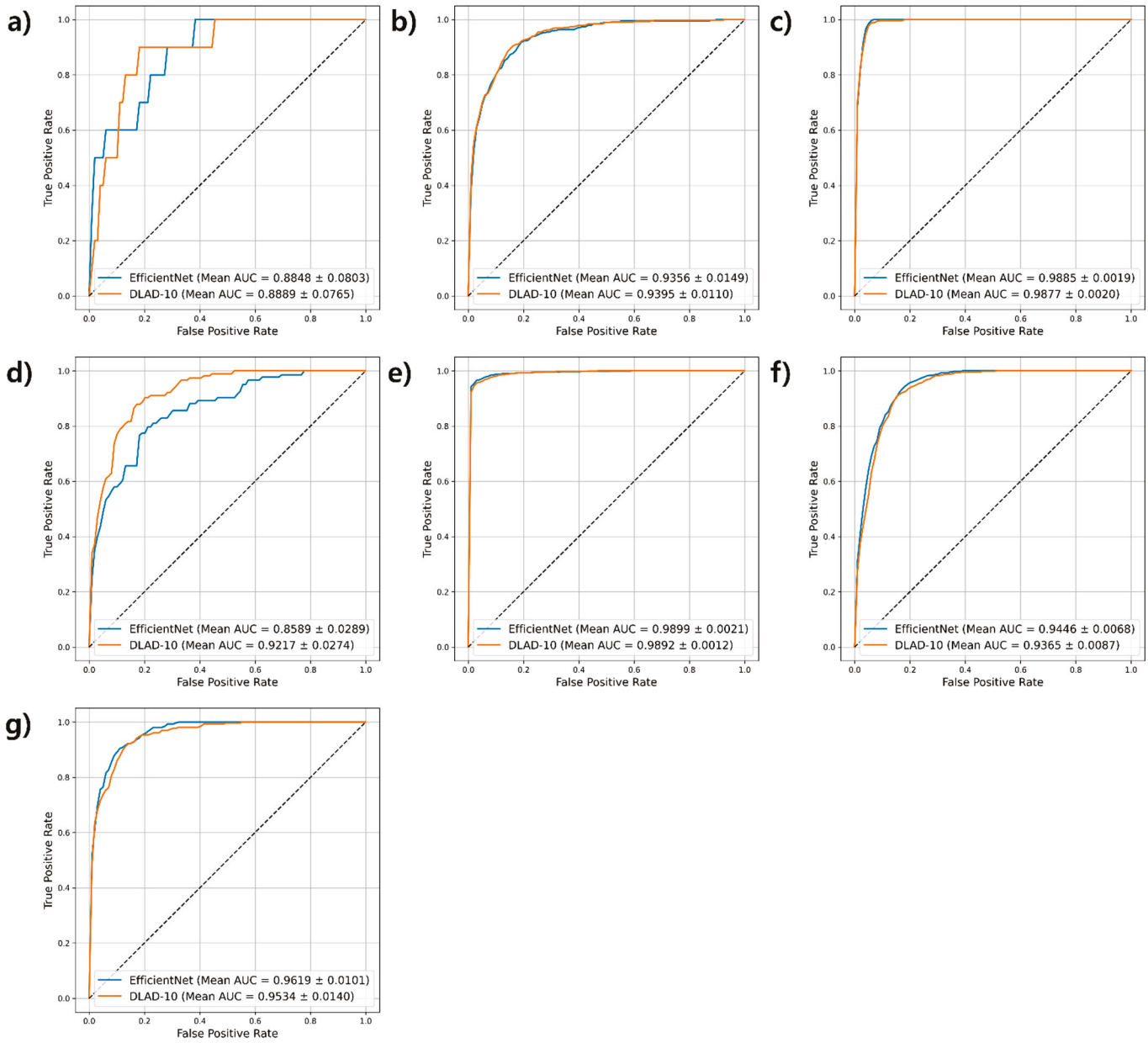


Figure 2. Comparison of AUROCs for deep learning models across different diseases using the external validation dataset. (**a**) Pneumonia, (**b**) Pleural effusion, (**c**) Tuberculosis, (**d**) Mass, (**e**) Pneumothorax, (**f**) Other diseases, and (**g**) No finding (AUROC: area under the receiver operating characteristic curve).

Table 5. Diagnostic performance for each target condition across different training data scales using the external validation dataset (*N*: number of samples used for training, AUROC: area under the receiver operating characteristic curve).

Target Diseases	Case−1			Case−2			Case−3			Case−4		
	N	Performance	AUROC	*N*	Performance	AUROC	*N*	Performance	AUROC	*N*	Performance	AUROC
Pneumonia	7913	(0.32, 0.75, 0.75, 0.45)	0.8414	16,077	(0.34, 0.76, 0.79, 0.48)	0.8689	24,017	(0.37, 0.79, 0.79, 0.50)	0.8819	40,000	(0.37, 0.79, 0.80, 0.51)	0.8905
Pleural Effusion	8005	(0.42, 0.78, 0.78, 0.55)	0.8657	15,908	(0.45, 0.81, 0.81, 0.58)	0.8881	24,031	(0.46, 0.81, 0.82, 0.59)	0.8900	40,000	(0.47, 0.81, 0.82, 0.57)	0.8978
Tuberculosis	7977	(0.97, 0.99, 0.87, 0.92)	0.9937	16,071	(0.96, 0.99, 0.90, 0.93)	0.9948	24,029	(0.96, 0.99, 0.90, 0.93)	0.9930	40,000	(0.98, 0.99, 0.88, 0.93)	0.9926
Mass	7964	(0.54, 0.86, 0.86, 0.66)	0.9416	15,993	(0.55, 0.87, 0.86, 0.67)	0.9494	24,039	(0.58, 0.88, 0.88, 0.70)	0.9580	40,000	(0.57, 0.87, 0.87, 0.69)	0.9520
Consolidation	8090	(0.36, 0.74, 0.73, 0.48)	0.8139	16,109	(0.38, 0.76, 0.76, 0.51)	0.8365	24,029	(0.39, 0.77, 0.77, 0.52)	0.8488	40,000	(0.40, 0.77, 0.76, 0.52)	0.8485
Pneumothorax	8097	(0.49, 0.83, 0.82, 0.61)	0.9115	15,958	(0.53, 0.85, 0.84, 0.65)	0.9267	23,928	(0.54, 0.85, 0.86, 0.67)	0.9375	40,000	(0.58, 0.88, 0.86, 0.70)	0.9437
Other diseases	13,777	(0.27, 0.63, 0.67, 0.38)	0.7141	27,447	(0.26, 0.64, 0.64, 0.37)	0.7098	41,072	(0.27, 0.64, 0.65, 0.38)	0.7186	68,525	(0.28, 0.67, 0.65, 0.39)	0.7354
No finding	7933	(0.41, 0.78, 0.79, 0.54)	0.8623	15,972	(0.43, 0.80, 0.79, 0.56)	0.8745	23,995	(0.43, 0.79, 0.80, 0.56)	0.8817	39,999	(0.44, 0.80, 0.80, 0.57)	0.8891

Data in parentheses indicate precision, sensitivity, specificity, and F1-score.

https://doi.org/10.3390/diagnostics16010146

3.3.2. Diagnostic Performance Across Data Diversity

We trained models using two configurations to assess the impact of data diversity. In the single-source setup, each disease category comprised 5000 samples drawn from one public dataset (Chest X-Ray-14, CheXpert, MIMIC-CXR, or PadChest), with duplicates added as needed. In the multiple-source setup, 1250 samples per disease were selected from each dataset to form a combined set of 5000 images, thereby increasing diversity. An independent test dataset (Table 2) ensured consistent evaluation under uniform conditions. All models were trained under uniform conditions.

Table 6 indicates that diagnostic performance generally improved with the multiple-source dataset: mean AUROC increased from 0.7061 (Chest X-Ray-14-only) and 0.7523 (MIMIC-CXR-only) to 0.7708, with sensitivity and specificity rising to 0.63 and 0.75, respectively. However, the degree of improvement varied across diseases and metrics. For example, the AUROC for mass increased substantially from 0.6765 to 0.8850, and similar trends were observed for pneumonia and pleural effusion, whereas the "Other" disease category exhibited only modest gains (AUROC from 0.5324 to 0.5669). In some cases, single-source datasets outperformed the multiple-source configuration for specific metrics (e.g., MIMIC-CXR-only achieved a higher specificity for pneumonia: 0.72 vs. 0.65).

Tuberculosis exhibited unique behavior because its samples were exclusively derived from the PadChest dataset, rendering its data effectively single-source even in the multiple-source configuration. The AUROC improved from 0.3891 (PadChest-only) to 0.6440 with the multiple-source dataset, while precision and specificity showed inconsistent trends. This discrepancy likely reflects the influence of diverse non-tuberculosis samples on overall model calibration, while limited variability in tuberculosis-specific data constrained further improvements. Overall, these results demonstrate that data diversity enhances performance for common thoracic conditions, while highlighting the need for targeted data collection strategies to improve generalization for rare or heterogeneous diseases.

3.3.3. Uncertainty and Data Scale Relationship

As presented in Table 7, a descriptive analysis of PE and VP across increasing training data scales (20%, 40%, 60%, and 100%) reveals that VP generally decreases or stabilizes across all disease categories. Pleural effusion, consolidation, other diseases, and no finding exhibited clear decreases followed by stable low VP values (e.g., consolidation: 0.0007 at 20% to 0.0004 from 40% onward). Mass showed the largest reduction (0.0009 to 0.0005 and then to 0.0004), while pneumothorax remained unchanged through 60% (0.0004) and then decreased at 100% (0.0003). Pneumonia displayed a mild non-monotonic pattern, and tuberculosis remained constant at 0.0001 across all data scales.

Conversely, PE varied substantially by class and did not follow a monotonic trend. Most diseases showed a noticeable reduction between 20% and 40% data—such as pneumonia (1.9592 to 1.616) and consolidation (2.2179 to 1.893)—but increased again at 60% or 100%. Additionally, pneumothorax decreased sharply at 40% and partially rebounded at higher scales, while tuberculosis demonstrated the largest fluctuation relative to its scale ($0.1858 \rightarrow 0.0812 \rightarrow 0.3547 \rightarrow 0.2172$) despite stable VP. The "Other" disease category exhibited a distinct pattern, with PE increasing from 20% to 60% (1.0883 to 1.3439) but decreasing sharply at 100% (0.6445). This substantial drop is attributable to the broad heterogeneity of the class: with limited data, the model is highly uncertain because it has insufficient exposure to the diverse subpatterns within this category. As data volume increases, the model encounters a wider and more representative range of these patterns, reducing epistemic uncertainty and yielding more stable calibration across this heterogeneous class.

https://doi.org/10.3390/diagnostics16010146

Table 6. Diagnostic performance for each target condition across datasets with varying diversity (AUROC: area under the receiver operating characteristic curve).

Target Diseases	NIH Chest X-Ray-14		MIMIC-CXR		PadChest		CheXpert		Multiple-Source	
	Performance	AUROC	Performance	AUROC	Performance	AUROC	Performance	AUROC	Performance	AUROC
Pneumonia	(0.21, 0.62, 0.63, 0.31)	0.6695	(0.21, 0.48, 0.72, 0.29)	0.6593	(0.13, 0.48, 0.51, 0.21)	0.4829	(0.21, 0.62, 0.65, 0.32)	0.6723	(0.22, 0.66, 0.65, 0.33)	0.7184
Pleural Effusion	(0.32, 0.72, 0.69, 0.44)	0.7915	(0.40, 0.78, 0.77, 0.53)	0.8479	(0.19, 0.40, 0.66, 0.26)	0.5511	(0.37, 0.75, 0.75, 0.50)	0.8209	(0.38, 0.83, 0.73, 0.52)	0.8574
Tuberculosis	–	–	–	–	(0.05, 0.59, 0.26, 0.10)	0.3891	–	–	(0.05, 0.01, 0.99, 0.01)	0.6440
Mass	(0.23, 0.62, 0.60, 0.34)	0.6765	–	–	(0.08, 0.14, 0.68, 0.10)	0.3969	–	–	(0.41, 0.79, 0.78, 0.54)	0.8850
Consolidation	(0.24, 0.63, 0.60, 0.34)	0.6656	(0.27, 0.66, 0.64, 0.38)	0.7059	(0.15, 0.40, 0.56, 0.22)	0.4804	(0.27, 0.67, 0.64, 0.38)	0.7131	(0.31, 0.70, 0.70, 0.43)	0.7636
Pneumothorax	(0.39, 0.75, 0.77, 0.52)	0.8393	(0.44, 0.77, 0.80, 0.56)	0.8625	(0.15, 0.28, 0.69, 0.20)	0.4997	(0.44, 0.76, 0.80, 0.56)	0.8514	(0.46, 0.83, 0.81, 0.60)	0.9002
Other diseases	(0.17, 0.52, 0.51, 0.26)	0.5324	(0.22, 0.60, 0.57, 0.32)	0.6214	(0.16, 0.31, 0.66, 0.21)	0.4560	(0.18, 0.52, 0.51, 0.26)	0.5538	(0.20, 0.51, 0.59, 0.29)	0.5669
No finding	(0.30, 0.72, 0.67, 0.42)	0.7680	(0.36, 0.74, 0.74, 0.49)	0.8167	(0.16, 0.42, 0.56, 0.23)	0.4943	(0.31, 0.70, 0.70, 0.43)	0.7687	(0.37, 0.76, 0.75, 0.50)	0.8312

The models trained on Chest X-ray-14, MIMIC-CXR, and CheXpert lacked tuberculosis classification capability, whereas those trained on MIMIC-CXR and CheXpert lacked mass classification because of the absence of corresponding samples in these datasets. Data in parentheses indicate precision, sensitivity, specificity, and F1-score.

Table 7. Predictive entropy (PE) and variance of predictions (VP) for each target condition across different training data scales using the external validation dataset (*N*: number of samples used for training).

Target Diseases	Case−1			Case−2			Case−3			Case−4		
	N	PE	VP	*N*	PE	VP	*N*	PE	VP	*N*	PE	VP
Pneumonia	7913	1.9592	0.0007	16,077	1.616	0.0004	24,017	1.9135	0.0005	40,000	1.772	0.0004
Pleural Effusion	8005	1.8092	0.0006	15,908	1.506	0.0005	24,031	1.5154	0.0004	40,000	1.7032	0.0004
Tuberculosis	7977	0.1858	0.0001	16,071	0.0812	0.0001	24,029	0.3547	0.0001	40,000	0.2172	0.0001
Mass	7964	2.1449	0.0009	15,993	2.0338	0.0005	24,039	1.9871	0.0004	40,000	2.228	0.0004
Consolidation	8090	2.2179	0.0007	16,109	1.893	0.0004	24,029	1.6253	0.0004	40,000	1.9582	0.0004
Pneumothorax	8097	2.2859	0.0004	15,958	1.8744	0.0004	23,928	2.0114	0.0004	40,000	2.0566	0.0003
Other diseases	13,777	1.0883	0.0009	27,447	1.1631	0.0006	41,072	1.3439	0.0005	68,525	0.6445	0.0005
No finding	7933	1.4209	0.0007	15,972	0.6132	0.0005	23,995	1.446	0.0005	39,999	0.7514	0.0005

https://doi.org/10.3390/diagnostics16010146

Overall, these results demonstrate that scaling the training data reliably reduces VP across diseases, whereas PE is driven predominantly by class-specific factors including heterogeneity, rarity, and label consistency. Therefore, total predictive uncertainty does not necessarily decrease with more data, even when epistemic uncertainty becomes stable.

4. Discussion

Our study demonstrates that large-scale, diverse public CXR datasets with noisy labels can effectively train deep learning models for disease classification, overcoming limitations of imperfect data through scalability and diversity. However, rare and under-represented conditions remain challenging, underscoring the need for targeted strategies, including higher-quality data collection and advanced model architectures tailored to their characteristics.

A key contribution of our study is the integration of UQ to analyze how predictive certainty evolves with increasing data sizes and across disease categories. As shown in our results, VP consistently decreased or stabilized as the training data expanded from 20% to 40%, 60%, and 100%, indicating improved parameter certainty under more balanced data regimes. In contrast, PE displayed heterogeneous and frequently non-monotonic trajectories, demonstrating that total predictive uncertainty is shaped not only by dataset size but also by inherent class characteristics, including heterogeneity, annotation quality, and imaging variability. For example, pneumonia, pleural effusion, and consolidation showed an initial reduction in PE but exhibited secondary increases at larger data scales, suggesting the presence of persistent aleatoric uncertainty that cannot be resolved solely by increasing training samples. Pneumothorax exhibited a similar pattern, with PE decreasing substantially at 40% data and subsequently increasing at higher scales despite stable VP.

The tuberculosis category exhibited a distinctive uncertainty pattern: VP remained constant at 0.0001 across all data scales, whereas PE fluctuated substantially. This behavior likely reflects the limited diversity of tuberculosis images in the original datasets and the extensive oversampling required to match the training volume of other diseases. Because the model repeatedly encounters a narrow set of visual patterns during training, epistemic variability remains minimal, resulting in stable VP values. However, PE remains sensitive to residual data-level ambiguities—such as heterogeneous imaging conditions, variations in disease presentation, and inconsistencies in text-derived labels—leading to the observed fluctuations. This contrast highlights how VP and PE capture different aspects of model uncertainty: while VP reflects stability in parameter estimates under repeated sampling, PE is influenced more strongly by intrinsic variability within the underlying data distribution.

These class-specific uncertainty profiles can be further interpreted by examining the dataset composition and adjusted training counts presented in Tables 1 and 2, respectively. Several rare diseases—most notably tuberculosis—were severely underrepresented in the original datasets (~3.9 k of ~1.08 M) and required extensive oversampling and augmentation to balance the training distribution. Under these conditions, VP naturally stabilized due to repeated sampling of similar examples, whereas PE remained sensitive to unresolved label inconsistencies and limited intra-class diversity. Mass and pneumothorax, both of which rely heavily on NLP-derived labels, also showed PE fluctuations consistent with known variability in automated labeling pipelines and cross-dataset domain heterogeneity. Conversely, the substantial PE reduction observed for the "Other" diseases category at 100% data reflects the stabilizing effect of its exceptionally large and diverse original sample pool; only when exposed to the full dataset could the model approximate the underlying distribution sufficiently to reduce total uncertainty.

Direct comparisons with previous studies are challenging because of differences in dataset composition, labeling quality, and model architectures. Many previous studies

https://doi.org/10.3390/diagnostics16010146

relied on private or mixed private-public datasets, whereas our approach used only public datasets, enhancing accessibility but introducing variability and noise. Moreover, our study incorporated UQ to assess predictive certainty, a dimension frequently overlooked in similar research. Tang et al. [38] demonstrated robust binary classification using well-established models, while our study extends this approach to multiclass classification using an even larger dataset and highlights the benefits of fine-tuning for domain adaptation. Wu et al. [6] developed a custom model to classify 72 thoracic diseases from 353,818 CXR images from the NIH and MIMIC datasets using NLP-based automated labeling, achieving a mean AUROC of 0.807—comparable to a third-year radiology resident. Similarly, Cid et al. [7] utilized 1,896,034 images from three UK hospitals to classify 37 diseases, attaining a mean AUROC of 0.864, underscoring the impact of dataset scale on diagnostic accuracy. More recently, Seah et al. [8] trained EfficientNet on 821,681 images from five public datasets (including NIH Chest X-ray-14 and MIMIC) with radiologist-labeled high-quality data, achieving an outstanding mean AUROC of 0.957. A key distinction of our study is the methodological focus imposed by our open-source approach. While these large-scale studies [7,8] utilized private or highly curated institutional data and concentrated on maximizing accuracy within custom, complex taxonomies, our use of entirely public datasets necessitates and introduces multi-dataset harmonization and a focus on cross-dataset generalization. This methodological foundation, combined with our uncertainty-aware analysis, demonstrates high diagnostic performance using fully open-source resources across clinically significant multiclass conditions, broadening the applicability of automated diagnostic solutions, particularly in resource-constrained settings.

This study had some limitations. First, the external validation dataset was derived from a single institution, and the uneven distribution of cases for different diseases may not reflect real-world prevalence, potentially limiting generalizability. Second, although we aggregated 17 public datasets, substantial upsampling was required for several rare disease classes to maintain controlled comparisons across categories. While all unique and duplicated sample counts are transparently reported, reliance on duplicated images may constrain feature diversity for these underrepresented classes. Third, the "Other" diseases category encompasses a broad and heterogeneous set of abnormalities, and although we provide detailed label mappings, such heterogeneity inevitably reduces class-specific interpretability compared with well-defined disease categories. Fourth, although labels were harmonized and cross-dataset duplicates were removed, residual domain shifts arising from differences in imaging protocols, patient demographics, and disease prevalence across the 17 public datasets may still affect model generalizability, particularly for conditions with inconsistent labeling strategies. Fifth, we focused on six thoracic conditions, restricting their applicability to various diseases. Sixth, we employed MC dropout as the sole UQ method, leaving alternative techniques unexplored. Finally, we utilized a subset of state-of-the-art deep learning models, which may not fully leverage more recent architectures, including Vision Transformers or large foundation models; while our findings on scalability and uncertainty are expected to generalize, evaluating these newer models remains an avenue for future investigation.

5. Conclusions

This study demonstrated the feasibility of leveraging public datasets and open-source deep learning models to achieve robust diagnostic performance in CXR analysis. Using data scale and diversity, our approach achieved results comparable to those obtained with proprietary datasets while providing a fully reproducible and accessible framework. Our uncertainty analysis further highlighted that, although epistemic uncertainty decreases with increasing data, total predictive uncertainty remains strongly influenced by class-

specific characteristics, label quality, and inherent data heterogeneity. These findings underscore the need for more diverse and high-quality datasets—particularly for rare or underrepresented conditions—to further improve model reliability and stability. Despite these challenges, our study advances the democratization of AI tools for medical imaging by providing a scalable, uncertainty-aware framework that facilitates broader adoption in diverse and resource-constrained clinical settings.

Author Contributions: Conceptualization, B.-D.L.; Data curation, F.K.K., W.B.T., M.S.L., J.Y.K. and B.-D.L.; Formal analysis, F.K.K., W.B.T., M.S.L., and B.-D.L.; Investigation, all authors; Methodology, F.K.K., W.B.T., M.S.L. and B.-D.L.; Project administration, S.S.B. and B.-D.L.; Resources, S.S.B. and B.-D.L.; Software, F.K.K., W.B.T. and S.-W.P.; Validation, all authors; writing—original draft, all authors; writing—review and editing, all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This retrospective study was approved by the Institutional Review Board of Keimyung University Dongsan Medical Center (DSMC-2021-02-021, date of approval 21 February 2021). All procedures were performed in accordance with the ethical standards of the Declaration of Helsinki.

Informed Consent Statement: Informed consent was waived due to the retrospective nature of this study. The data used in this study were fully de-identified to protect patient confidentiality.

Data Availability Statement: Public datasets used in this study are described in the text and tables and are available from the corresponding author upon request. The in-house dataset supporting these findings is maintained at Keimyung University Dongsan Hospital. Due to licensing restrictions, these data are not publicly available; however, access may be granted upon reasonable request and subject to approval by Keimyung University Dongsan Hospital.

Conflicts of Interest: Authors Farzeen Khalid Khan, Waleed Bin Tahir and Shi Sub Byon were employed by the company HealthHub Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AUROC	Area Under the Receiver Operating Characteristics Curve
CXR	Chest X-ray
DLAD-10	Deep Learning-based Algorithm for Detecting 10 Abnormalities
MC	Monte Carlo
MIMIC-CXR	Medical Information Mart for Intensive Care—Chest X-ray
NIH	National Institute of Health
NLP	Natural Language Processing
PA	Posterior–Anterior
PE	Predictive Entropy
TB	Tuberculosis
UQ	Uncertainty quantification
VP	Variance of Predictions

Appendix A

We implemented four deep learning architectures for CXR classification—ResNet-50, DenseNet-121, EfficientNet-B5 and DLAD-10—chosen for their demonstrated effectiveness in natural image tasks and established performance in chest radiograph analysis. The origi-

nal classification layer of each model was replaced with a linear layer corresponding to the dataset labels. The models were trained as multi-label classifiers, producing independent sigmoid outputs for each class, using binary cross-entropy with logit loss. This loss is suitable for multi-label classification, where each class is treated independently. The model parameters were optimized using the Adam optimizer. Class-specific decision thresholds were determined by evaluating performance metrics across a range of thresholds (0.0–1.0) on the validation set. Thresholds were selected at points where sensitivity and specificity were approximately balanced, with slight prioritization of sensitivity to minimize false negatives. The resulting disease-specific thresholds were subsequently applied unchanged to the test set to ensure unbiased evaluation.

Before input into the models, each input CXR image was subjected to lung segmentation using U-Net [39] architecture trained specifically for this study. The model was initialized with random weights and trained on a combined dataset of 7951 images from the Mendeley Chest X-ray Lung Segmentation dataset [40] and the MIMIC-CXR segmentation subset [41], split into 70–20–10 training, validation, and test sets. During segmentation model training, data augmentation included horizontal flipping, color jittering (brightness, contrast, and saturation $\pm$ 0.5; hue $\pm$ 0.3), conversion to grayscale, and resizing to 512 $\times$ 512 pixels. The model achieved a Dice coefficient of 0.9251 on an independent test set (n = 796), indicating accurate delineation of lung fields (see Table A1). The resulting segmentation masks were used to crop each image to the lung region, removing non-informative background areas. The cropped images were subsequently resized to 448 $\times$ 448 pixels.

For classification model training, online data augmentation was applied using PyTorch 2.3.0, deliberately employing milder transformations, random rotations up to $\pm 20^\circ$ and color jittering (brightness, contrast, and saturation $\pm$ 0.2, hue $\pm$ 0.1) than those used for segmentation, ensuring preservation of critical anatomical detail while still promoting generalization. To reduce overfitting, each image was uniquely augmented at every iteration. Subsequently, input images were resized and normalized using the standard ImageNet mean and standard deviation to ensure compatibility with the ImageNet-pretrained backbone used for initialization, preventing a distribution mismatch that could degrade performance. While ImageNet statistics originate from natural RGB images, their use in chest X-rays is widely adopted and empirically validated when using pretrained models. Chest X-rays are inherently single-channel; they are duplicated across three channels to match the input requirements of ImageNet-pretrained models. Therefore, normalization primarily standardizes the input range and stabilizes gradients without distorting diagnostic structures. Furthermore, widely used medical imaging benchmarks—including CheXpert [10], MIMIC-CXR [11], RSNA [18], and SIIM-ACR pneumothorax [19] datasets—commonly adopt ImageNet normalization with pretrained backbones, demonstrating its practicality and suitability for radiography tasks. The models' training was configured for up to 30 epochs with early stopping using a batch size of 32. The initial learning rate started at 0.001 and was reduced upon validation loss plateau. Fine-tuning employed the same parameters with a reduced learning rate of 0.0001.

Table A1. Performance metrics of the U-Net architecture used for lung segmentation (preprocessing) on the independent test set (n = 796).

Metrics	Value
Dice Coefficient	0.9251
IoU	0.8573
Precision	0.9396
Recall	0.9073

https://doi.org/10.3390/diagnostics16010146

Appendix B

To ensure consistent and high-quality annotations across the 17 aggregated CXR datasets, several harmonization steps were applied. Only frontal-view CXRs (AP/PA) were included to maintain clinical consistency. Labels with low confidence or uncertainty (e.g., CheXpert's "Uncertain") were excluded from the training set to maintain label reliability. Annotation protocols varied across datasets in both naming conventions and label granularity. To ensure consistency, labels referring to related abnormalities (e.g., "mass" and "nodule") were consolidated into a single Mass category. Findings outside the primary target disease set, including cardiomegaly, atelectasis, and fibrosis, were grouped into an "Other" disease category to maintain a unified label structure. PadChest [15] provides 174 radiographic findings extracted via NLP from full radiology reports; for this dataset, a target disease was marked positive whenever its corresponding condition appeared within PadChest's annotated label list. Duplicate studies across datasets were identified and removed to prevent overlap between the training and evaluation sets, particularly for large public datasets, including CheXpert [10] and MIMIC-CXR [11].

To mitigate class imbalance, images from underrepresented classes were duplicated in the training set. Each image was copied without alteration to preserve its original characteristics. The distribution of original images, duplicates added, and total images after duplication is summarized in Table A2.

All harmonization rules and label mappings are mentioned in Table A3, enabling full reproducibility and transparency in dataset construction.

Table A2. Class distribution before and after image duplication: (**a**) Training Set, (**b**) Validation Set, and (**c**) Test Set.

(a)			
Class	**Unique Images**	**Duplicates Added**	**Total Images**
Pneumonia	35,987	4013	40,000
Pleural Effusion	33,360	6640	40,000
Tuberculosis	2301	37,699	40,000
Mass	12,156	27,844	40,000
Consolidation	31,002	8998	40,000
Pneumothorax	35,160	4840	40,000
Other diseases	67,576	949	68,525
No finding	39,999	0	39,999
(b)			
Class	**Unique Images**	**Duplicates added**	**Total Images**
Pneumonia	1479	7758	9237
Pleural Effusion	2996	6725	9721
Tuberculosis	391	7429	7820
Mass	506	9257	9763
Consolidation	530	9354	9884
Pneumothorax	553	9429	9982
Other diseases	1420	10,857	12,277
No finding	9807	0	9807

https://doi.org/10.3390/diagnostics16010146

Table A2. *Cont.*

(c)			
Class	**Unique Images**	**Duplicates added**	**Total Images**
Pneumonia	2421	0	2421
Pleural Effusion	2990	0	2990
Tuberculosis	1237	0	1237
Mass	2945	0	2945
Consolidation	2988	0	2988
Pneumothorax	2998	0	2998
Other diseases	3024	0	3024
No finding	2936	0	2936

Table A3. Label Mapping Across Datasets.

Dataset	Original Dataset Label	Harmonized Label	Notes
NIH Chest X-ray-14 [9]	Atelectasis	Other	Not part of target diseases
	Cardiomegaly	Other	Not part of target diseases
	Effusion	Pleural Effusion	Retained as-is
	Infiltration	Other	Not part of target diseases
	Mass	Mass	Combined for consistency
	Nodule	Mass	Combined for consistency
	Pneumonia	Pneumonia	Retained as-is
CheXpert [10], MIMIC-CXR [11], BRAX [16]	Enlarged Cardiomediastinum	Other	Not part of target diseases
	Cardiomegaly	Other	Not part of target diseases
	Lung Lesion	Other	Not part of target diseases
	Lung Opacity	Other	Not part of target diseases
	Edema	Other	Not part of target diseases
	Consolidation	Consolidation	Retained as-is
	Pneumonia	Pneumonia	Retained as-is
	Atelectasis	Other	Not part of target diseases
	Pneumothorax	Pneumothorax	Retained as-is
	Pleural Effusion	Pleural Effusion	Retained as-is
	Pleural Other	Other	Not part of target diseases
	Fracture	Other	Not part of target diseases
	Support Devices	Support Devices	Retained as-is
PadChest [15] *	effusion	Pleural Effusion	Target disease assigned positive if present in PadChest label list
	nodule	Mass	
	mass	Nodule	
	consolidation	Consolidation	
	pneumonia	Pneumonia	
	tuberculosis	Tuberculosis	
VinDr-CXR [17]	Aortic enlargement	Other	Not part of target diseases
	Atelectasis	Other	Not part of target diseases
	Cardiomegaly	Other	Not part of target diseases
	Calcification	Other	Not part of target diseases
	Clavicle fracture	Other	Not part of target diseases
	Consolidation	Consolidation	Retained as-is
	Edema	Other	Not part of target diseases
	Emphysema	Other	Not part of target diseases
	Enlarged PA	Other	Not part of target diseases
	Interstitial lung disease (ILD)	Other	Not part of target diseases

https://doi.org/10.3390/diagnostics16010146

Table A3. *Cont.*

Dataset	Original Dataset Label	Harmonized Label	Notes
VinDr-CXR [17]	Infiltration	Other	Not part of target diseases
	Lung cavity	Other	Not part of target diseases
	Lung cyst	Other	Not part of target diseases
	Lung opacity	Other	Not part of target diseases
	Mediastinal shift	Other	Not part of target diseases
	Nodule/Mass	Mass	Renamed for consistency
	Pulmonary fibrosis	Other	Not part of target diseases
	Pneumothorax	Pneumothorax	Retained as-is
	Pleural thickening	Other	Not part of target diseases
	Pleural effusion	Pleural Effusion	Retained as-is
	Rib fracture	Other	Not part of target diseases
	Other lesion	Other	Not part of target diseases
	Lung tumor	Mass	Combined for consistency
	Pneumonia	Pneuonia	Retained as-is
	Tuberculosis	Tuberculosis	Retained as-is
	Other diseases	Other	Not part of target diseases
	Chronic obstructive pulmonary disease (COPD)	Other	Not part of target diseases
RSNA Pneumonia Detection Challenge [18]	Pneumonia	Pneumonia	Retained as-is
SIIM-ACR Pneumothorax Segmentation [19]	Pneummothorax	Pneumothorax	Retained as-is
JSRT [20]	Lung Nodule	Mass	Retained as-is
Shenzhen Hospital CXR Set [21] **	Tuberculosis	Tuberculosis	Retained as-is
	Effusion	Pleural Effusion	Extracted from 'findings'
	Mass	Mass	Extracted from 'findings'
	Consolidation	Consolidation	Extracted from 'findings'
	Pleural thickening	Other	Not part of target diseases
	Fibrous lesions	Other	Not part of target diseases
Montgomery County chest X-ray set (MC) [21]	Tuberculosis	Tuberculosis	Retained as -is
COVID-19, Pneumonia and Normal Chest X-ray PA Dataset [22]	Pneumonia	Pneumonia	Retained as -is
	COVID-19	Other	Not part of target diseases
Chest X-Ray Images (Pneumonia) [23]	Pneumonia	Pneumonia	Retained as-is
Tuberculosis (TB) Chest X-ray Database [24], TBX11K [25], Belarus Dataset [26], Chest X-rays tuberculosis from India [27]	Tuberculosis	Tuberculosis	Retained as-is

* PadChest annotations comprise 174 NLP-derived radiographic findings; target disease labels in this study were assigned as positive if the matching condition was present in the PadChest findings. ** Shenzhen annotations are recorded in the findings column, including descriptors for conditions other than tuberculosis.

Appendix C

To ensure the reliability of the uncertainty estimation reported in this study, we performed a justification and sensitivity analysis for the two key hyperparameters of the Monte Carlo Dropout (MCD) method: the dropout rate and the number of stochastic forward passes.

The dropout rate was set to 0.4. This selection strictly adheres to the original architectural design of EfficientNet-B5 proposed by Tan and Le [30] to balance model capacity and regularization. This value is also the established default in standard deep learning framework implementations of this architecture [37], ensuring that feature distributions remain consistent with the pre-trained weights used for transfer learning.

We determined the optimal number of forward stochastic passes (N) for inference by analyzing the convergence of the uncertainty metrics (Predictive Entropy and Variance). We evaluated the model predictions with 5, 10, 15, and 20 passes to identify the point where the estimation of epistemic uncertainty stabilized. As presented in Table A4, the Mean Predictive Variance converged significantly by 10 passes (3.75×10^{-4}) and remained constant through 15 and 20 passes. Similarly, Predictive Entropy showed asymptotic stability between 15 and 20 passes. Based on this "diminishing returns" analysis, 20 stochastic passes were selected as the operating point. This ensures that the uncertainty trends reported in our results are attributable to data properties rather than stochastic estimation noise.

Table A4. Sensitivity analysis of uncertainty metrics with respect to the number of stochastic passes (N).

Stochastic Passes (N)	Mean Predictive Entropy	Mean Variance of Predictions	Relative Change (Variance)
5	1.408	3.13×10^{-4}	-
10	1.142	3.75×10^{-4}	+19.8%
15	1.146	3.75×10^{-4}	0.0% (Converged)
20 (Selected)	1.146	3.75×10^{-4}	0.0% (Stable)

References

1. Raoof, S.; Feigin, D.; Sung, A.; Raoof, S.; Irugulpati, L.; Rosenow, E.C. Interpretation of plain chest roentgenogram. *Chest* **2012**, *141*, 545–558. [CrossRef] [PubMed]
2. WHO (World Health Organization). *Chest Radiography in Tuberculosis Detection: Summary of Current WHO Recommendations and Guidance on Programmatic Approaches*; World Health Organization: Geneva, Switzerland, 2021. Available online: https://www.who.int/publications/i/item/9789241511506 (accessed on 30 October 2025).
3. Ellis, S.; Aziz, Z. Radiology as an aid to diagnosis in lung disease. *Postgrad. Med. J.* **2016**, *92*, 620–623. [CrossRef]
4. Rimmer, A. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ* **2017**, *359*, j4683. [CrossRef]
5. Nam, J.G.; Kim, M.; Park, J.; Hwang, E.J.; Lee, J.H.; Hong, J.H.; Goo, J.M.; Park, C.M. Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs. *Eur. Respir. J.* **2021**, *57*, 2003061. [CrossRef] [PubMed]
6. Wu, J.T.; Wong, K.C.L.; Gur, Y.; Ansari, N.; Karargyris, A.; Sharma, A.; Morris, M.; Saboury, B.; Ahmad, H.; Boyko, O.; et al. Comparison of chest radiograph interpretations by artificial intelligence algorithm vs radiology residents. *JAMA Netw. Open* **2020**, *3*, e2022779. [CrossRef] [PubMed]
7. Cid, Y.D.; Macpherson, M.; Gervais-Andre, L.; Zhu, Y.; Franco, G.; Santeramo, R.; Lim, C.; Selby, I.; Muthuswamy, K.; Amlani, A.; et al. Development and validation of open-source deep neural networks for comprehensive chest x-ray reading: A retrospective, multicentre study. *Lancet Digit. Health* **2024**, *6*, e44–e57. [CrossRef]
8. Seah, J.C.Y.; Tang, C.H.M.; Buchlak, Q.D.; Holt, X.G.; Wardman, J.B.; Aimoldin, A.; Esmaili, N.; Ahmad, H.; Pham, H.; Lambert, J.F.; et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: A retrospective, multireader multicase study. *Lancet Digit. Health* **2021**, *3*, e496–e506. [CrossRef]
9. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 3462–3471. [CrossRef]
10. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 590–597. [CrossRef]
11. Johnson, A.E.W.; Pollard, T.J.; Berkowitz, S.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.Y.; Mark, R.G.; Horng, S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **2019**, *6*, 317. [CrossRef]

12. Khan, R.F.; Lee, B.D.; Lee, M.S. Transformers in medical image segmentation: A narrative review. *Quant. Imaging Med. Surg.* **2023**, *13*, 8747–8767. [CrossRef]
13. World Health Organisation. *Global Tuberculosis Report 2020*; World Health Organization: Geneva, Switzerland, 2020. Available online: https://www.who.int/publications/i/item/9789240013131 (accessed on 30 October 2025).
14. Rueckel, J.; Huemmer, C.; Fieselmann, A.; Ghesu, F.C.; Mansoor, A.; Schachtner, B.; Wesp, P.; Trappmann, L.; Munawwar, B.; Ricke, J.; et al. Pneumothorax detection in chest radiographs: Optimizing artificial intelligence system for accuracy and confounding bias reduction using in-image annotations in algorithm training. *Eur. Radiol.* **2021**, *31*, 7888–7900. [CrossRef]
15. Bustos, A.; Pertusa, A.; Salinas, J.M.; de la Iglesia-Vayá, M. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Med. Image Anal.* **2020**, *66*, 101797. [CrossRef]
16. Reis, E.P.; Paiva, J.; Bueno da Silva, M.C.; Sousa Ribeiro, G.A.; Fornasiero Paiva, V.; Bulgarelli, L.; Lee, H.; Dos Santos, P.V.; brito v Amaral, L.; Beraldo, G.; et al. BRAX, a Brazilian Labeled Chest X-Ray Dataset (version 1.1.0). *PhysioNet*, 17 June 2022. [CrossRef]
17. Nguyen, H.Q.; Pham, H.H.; Linh, T.; Dao, M.; Khanh, L. VinDr-CXR: An Open Dataset of Chest X-Rays with Radiologist Annotations (version 1.0.0). *PhysioNet*, 22 June 2021. [CrossRef]
18. Stein, A.; Wu, C.; Carr, C.; Shih, G.; Dulkowski, J.; Kalpathy, L.C.; Prevedello, L.; Kohli, M.; McDonald, M.; Peter, P.C.; et al. RSNA Pneumonia Detection Challenge. Kaggle. 2018. Available online: https://kaggle.com/competitions/rsna-pneumonia-detection-challenge (accessed on 30 October 2025).
19. Zawacki, A.; Wu, C.; Shih, G.; Elliott, J.; Fomitchev, M.; Hussain, M.; ParasLakhani, P.; Culliton, S.B.; Siim, A.C.R. Pneumothorax Segmentation. Kaggle. 2019. Available online: https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation (accessed on 30 October 2025).
20. Shiraishi, J.; Katsuragawa, S.; Ikezoe, J.; Matsumoto, T.; Kobayashi, T.; Komatsu, K.; Matsui, M.; Fujita, H.; Kodera, Y.; Doi, K. Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *AJR Am. J. Roentgenol.* **2000**, *174*, 71–74. [CrossRef] [PubMed]
21. Jaeger, S.; Candemir, S.; Antani, S.; Wáng, Y.X.J.; Lu, P.X.; Thoma, G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant. Imaging Med. Surg.* **2014**, *4*, 475–477. [CrossRef] [PubMed]
22. Asraf, A.; Islam, Z. COVID19, Pneumonia and Normal Chest X-Ray PA Dataset. 2021. Version 1. Available online: https://data.mendeley.com/datasets/jctsfj2sfn/1 (accessed on 30 October 2025).
23. Mooney, P.T. Chest X-Ray Images (Pneumonia). Kaggle. 2018. Available online: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia (accessed on 30 October 2025).
24. Rahman, T.; Khandakar, A.; Chowdhury, M.E.H. Tuberculosis (TB) chest X-ray database. *IEEE Dataport*, 2020. [CrossRef]
25. Liu, Y.; Wu, Y.-H.; Ban, Y.; Wang, H.; Cheng, M.-M. Rethinking computer-aided tuberculosis diagnosis. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2643–2652. [CrossRef]
26. 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] [PubMed]
27. Raddar. Chest X-Rays Tuberculosis from India. Kaggle. 2021. Available online: https://www.kaggle.com/datasets/raddar/chest-xrays-tuberculosis-from-india/data (accessed on 30 October 2025).
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
29. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]
30. Tan, M.; Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019. Available online: https://proceedings.mlr.press/v97/tan19a.html (accessed on 30 October 2025).
31. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
32. Todi, A.; Narula, N.; Sharma, M.; Gupta, U. ConvNext: A contemporary architecture for convolutional neural networks for image classification. In Proceedings of the 2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT), Dehradun, India, 8–9 September 2023; pp. 1–6. [CrossRef]
33. Zhang, S.; Xu, Y.; Usuyama, N.; Xu, H.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; et al. BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image–text pairs. *arXiv* **2023**, arXiv:2303.00915.
34. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the ICML'16: Proceedings of the 33rd International Conference on International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016.

35. Tyralis, H.; Papacharalampous, G. A review of predictive uncertainty estimation with machine learning. *Artif. Intell. Rev.* **2024**, *57*, 94. [CrossRef]
36. Choubineh, A.; Chen, J.; Coenen, F.; Ma, F. Applying Monte Carlo dropout to quantify the uncertainty of skip connection-based convolutional neural networks optimized by big data. *Electronics* **2023**, *12*, 1453. [CrossRef]
37. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Vancouver, BC, Canada, 2019; Volume 32, pp. 8024–8035.
38. Tang, Y.X.; Tang, Y.B.; Peng, Y.; Yan, K.; Bagheri, M.; Redd, B.A.; Brandon, C.J.; Lu, Z.; Han, M.; Xiao, J.; et al. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *npj Digit. Med.* **2020**, *3*, 70. [CrossRef] [PubMed]
39. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W., Frangi, A., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2015; Volume 9351. [CrossRef]
40. Danilov, V.; Proutski, A.; Kirpich, A.; Litmanovich, D.; Gankin, Y. Chest X-Ray Dataset for Lung Segmentation (version 2). *Mendeley Data*, 3 October 2022. [CrossRef]
41. Chen, L.; Kuo, P.; Wang, R.; Gichoya, J.; Celi, L.A. Version 1.0.0. Chest X-Ray Segmentation Images Based on MIMIC-CXR (version 1.0.0). *PhysioNet*, 18 August 2022; RRID:SCR_007345. [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.

diagnostics

MDPI

Article

Precision Through Detail: Radiomics and Windowing Techniques as Key for Detecting Dens Axis Fractures in CT Scans

Karl Ludger Radke [1,*], Anja Müller-Lutz [1], Daniel B. Abrar [1], Marius Vach [1], Christian Rubbert [1], David Latz [2], Gerald Antoch [1], Hans-Jörg Wittsack [1], Sven Nebelung [1,3] and Lena Marie Wilms [1]

[1] Department of Diagnostic and Interventional Radiology, Medical Faculty and University Hospital Düsseldorf, D-40225 Düsseldorf, Germany; anja.lutz@med.uni-duesseldorf.de (A.M.-L.); danielbenjamin.abrar@med.uni-duesseldorf.de (D.B.A.); marius.vach@med.uni-duesseldorf.de (M.V.); christian.rubbert@med.uni-duesseldorf.de (C.R.); antoch@med.uni-duesseldorf.de (G.A.); hans-joerg.wittsack@med.uni-duesseldorf.de (H.-J.W.); snebelung@ukaachen.de (S.N.); lena.wilms@med.uni-duesseldorf.de (L.M.W.)

[2] Department of Orthopedic and Trauma Surgery, University Hospital, D-40225 Düsseldorf, Germany; david.latz@med.uni-duesseldorf.de

[3] Department of Diagnostic and Interventional Radiology, University Hospital Aachen, D-52074 Aachen, Germany

* Correspondence: ludger.radke@med.uni-duesseldorf.de

Abstract: Background/Objectives: The present study investigates the influence of advanced windowing techniques and the combination of different classification methods on the accuracy of dens axis fracture detection in computed tomography (CT) images. The aim was to evaluate and compare the diagnostic performance of two different computational models—a pure deep learning (DL) approach and a combined approach of DL segmentation, windowing, and radiomics. **Methods**: In this retrospective study, CT datasets of the upper cervical spine of 366 patients were included. All datasets were further divided into training, validation, and test sets. Model 1 (M1) relied on a pure DL method using a Convolutional Neural Network (CNN) and a Feedforward Neural Network (FNN), without prior manual segmentation. Model 2 (M2) incorporated a fully automatic U-Net-based segmentation followed by radiomics feature extraction and classification using a Machine Learning (ML) Classifier. The performance of both models was measured by classification accuracy, with a particular focus on the impact of CT windowing parameters and the chosen ML classification strategies. **Results**: M1 achieved a maximum classification accuracy of 93.7%, while M2 accomplished a classification accuracy of up to 95.7% by using ROI-based windowing and advanced feature extraction. **Conclusions**: Integrating advanced windowing techniques, U-Net segmentation, and radiomics improves the detection of dens axis fractures in CT imaging. This approach could enhance diagnostic accuracy and warrants further exploration and clinical integration.

Keywords: dens axis fracture; CT imaging; deep learning; radiomics; machine learning

1. Introduction

Computed tomography (CT) imaging remains a cornerstone of diagnostic precision in medical practice, particularly in the detection of bony injuries such as dens axis fractures. The dens axis, a critical structure within the cervical spine, presents a diagnostic challenge in cases where fractures are subtle. If missed, these can lead to serious clinical implications including instability or worse neurological deficits, due to its proximity to

the upper cervical spinal cord and lower brainstem [1]. Therefore, enhancing image processing and visualization techniques is essential to facilitate accurate diagnosis and timely clinical intervention.

In recent years, advancements in radiological imaging and computational techniques have significantly improved diagnostic accuracy in medical imaging [2,3]. AI-based methodologies, specifically deep learning (DL), have proven highly effective for automatic segmentation and classification tasks in medical images, achieving robust results across various imaging modalities and clinical applications [4–6].

However, despite impressive capabilities, DL-only approaches cannot fully exploit the complex and nuanced information inherent in medical images, particularly when it comes to detecting subtle pathologies.

Radiomics, a technique involving the extraction of high-dimensional quantitative features from medical images, has emerged as a complementary strategy to DL, enhancing interpretability and diagnostic accuracy [7–12]. Radiomics has shown substantial promise across multiple diagnostic domains, including oncology, neurology, and musculoskeletal disorders, by capturing intricate texture-, morphology-, and intensity-based patterns not readily discernible by visual inspection alone [8,9,12].

Recent work in bladder cancer has further demonstrated that radiomics can non-invasively differentiate tumor grade and stage with clinically meaningful accuracy, even in small patient cohorts [13].

Recent studies emphasize the advantages of combining segmentation-driven approaches with radiomic analysis to improve diagnostic outcomes [7–12]. For example, Liang et al. demonstrated enhanced accuracy in breast cancer classification through integrating radiomics with segmentation [14], while Gutsche et al. illustrated superior predictive performance for brain metastases using radiomics compared to semantic features alone [15]. Such findings underscore the potential benefit of combining precise anatomical delineation through segmentation with detailed radiomic analysis.

Thus, in this study, we propose an integrated analytical pipeline that combines sophisticated CT windowing techniques, advanced convolutional neural network-based segmentation, and radiomics-driven classification to enhance the detection accuracy of dens axis fractures in a diverse CT dataset. This approach capitalizes on the strengths of both DL and radiomics, leveraging advanced image processing strategies—including Contrast-Limited Adaptive Histogram Equalization (CLAHE), histogram-based windowing, and region-of-interest (ROI)-based windowing—to optimize image visualization and fracture delineation [16–19]. Our objective is to systematically evaluate the effectiveness of this integrated pipeline compared to traditional DL-only methods, assessing its robustness, accuracy, and clinical applicability with different image processing methods and classification algorithms. Our hypothesis is that this integrated approach will significantly outperform traditional DL-only methods in terms of diagnostic accuracy and robustness.

2. Materials and Methods

2.1. Study Design and Patient Selection

This retrospective comparative imaging study was conducted in accordance with stringent local ethical guidelines and approved by the Ethics Committee of the Faculty of Medicine at Heinrich Heine University, Düsseldorf, Germany (study reference number: 2020-850).

A total of 400 CT scans of the upper cervical spine, acquired at the University Hospital Düsseldorf between October 2016 and March 2020, were reviewed. These scans were retrieved using various search terms (i.e., “upper cervical spine”, “CT skull + cervical spine”, and “cervical spine”) to ensure a deliberate diversity of imaging perspectives. As

a result, the dataset included a wide range of CT scans—from those solely depicting the upper cervical spine to those encompassing both part of the skull and the whole cervical spine—thereby increasing the complexity and realism of the data and highlighting challenges in accurate segmentation and analysis. Each scan was evaluated by an experienced clinical radiologist (L.M.W., with seven years of experience in musculoskeletal imaging) to confirm complete visualization of the dens axis. Exclusion criteria included postoperative CT scans, fractures other than dens axis fractures, other pathological findings, incomplete or truncated imaging, and pediatric CT examinations.

Following the application of these criteria, 366 CT scans were selected for further analysis, comprising 287 without dens fractures and 79 with dens fractures. The dataset was then randomly divided into training (70%, n = 256), validation (10%, n = 37), and test (20%, n = 73) sets, ensuring a systematic and balanced approach to model development and evaluation procedure.

2.2. Image Data Acquisition and Data Processing

All CT scans were acquired as part of routine clinical practice at the University Hospital Düsseldorf, Germany. To ensure a representative variety of imaging conditions, three different Siemens Healthineers CT scanners were utilized: SOMATOM Definition Flash (n = 245); SOMATOM Definition AS (n = 113); SOMATOM Definition Edge (n = 8).

The scans exhibited an average slice thickness of 1.94 ± 0.23 mm (1–2 mm), with a mean in-plane pixel resolution of 0.55 ± 0.16 mm (0.17–0.98 mm). The computed tomography dose index volume (CTDI_vol) averaged 13.19 ± 1.91 mGy (3.10–31.12 mGy).

2.2.1. Manual Segmentations

Manual segmentation of the dens and, when applicable, the associated fractures, was conducted on axial CT slices using ITK-SNAP (v3.8, Cognitica, Philadelphia, PA, USA) [20]. Segmentations were performed by L.M.W., who delineated each axial slice containing the dens to create reference ground truth segmentations. Following initial segmentation, all masks underwent visual quality control to ensure that only pixels unequivocally belonging to the dens were retained, thereby reducing partial volume effects (Figure 1).

2.2.2. Data Pre-Processing and Contrast Adjustment

To standardize the dataset, all CT images and corresponding segmentation masks were resampled to an isotropic resolution of 1 × 1 × 1 mm. Linear interpolation was used for CT images, while nearest-neighbor interpolation was applied to the segmentation masks to preserve label integrity. Following this spatial normalization, a range of contrast adjustment and windowing optimization techniques was applied to enhance image quality and create a comprehensive dataset for further analysis (Figure 2).

- **Bone Windowing:** A conventional windowing technique that enhances bone visualization by using a standard clinical window setting of 400 Hounsfield units (HU) for the window level and 2000 HU for the window width [21]. This setting highlights bone density and structure, facilitating better identification and assessment of bone pathologies.
- **Bone Gamma Correction:** This post-processing technique modifies the intensity distribution of bone-windowed images by applying gamma correction. Two gamma values were tested: γ = 0.5 to enhance brightness and γ = 2.0 to reduce it [16]. Gamma values less than 1 emphasize darker areas, while values greater than 1 reduce overall brightness, allowing finer details in both low- and high-intensity regions to emerge, particularly aiding in the detection of bone fractures.

- **Histogram-Based Windowing:** In this method, the window boundaries are determined by analyzing the histogram of the entire image. The 5th and 95th percentile values of the image's pixel intensity data are used to set the window boundaries, effectively filtering out extreme pixel values [22]. This technique optimizes the contrast distribution across the entire image, reducing overexposure and underexposure and thereby improving image quality.
- **Contrast-Limited Adaptive Histogram Equalization (CLAHE):** CLAHE increases local contrast by dividing the image into smaller segments and applying histogram equalization to each segment individually [17,18]. This method improves the visibility of subtle structures, especially in areas of low contrast, without excessively increasing global contrast. While commonly applied in lung imaging [18,19], recent evidence supports its utility in bone imaging—for example, Park et al. demonstrated that CLAHE improved classification performance in scintigraphy [19].
- **ROI-Based Windowing:** This targeted method utilizes the segmentation mask of the dens region to calculate the 5th and 95th percentile intensity values within the ROI and its immediate surroundings. These values are then applied uniformly to the entire image to maintain consistent contrast and prevent localized discrepancies. This approach ensures clinical readability and, in a two-stage pipeline such as Model M2, can be fully automated using upstream segmentation results.

2.2.3. Data Augmentation

The performance of neural networks is strongly influenced by the size and variability of the training dataset [23]. To enhance model robustness and mitigate overfitting, several data augmentation strategies were applied exclusively to the training set.

The augmentation pipeline began with Z-score normalization, where the mean intensity was subtracted and divided by the standard deviation for each image. This process standardized the dataset to have a mean of zero and a standard deviation of one, thereby stabilizing the training process through a uniform intensity distribution.

Subsequently, all images were resampled and standardized to a uniform voxel resolution of $256 \times 256 \times 256$, with zero-padding applied as needed to maintain anatomical completeness. To introduce variability in anatomical presentation and imaging conditions, random scaling transformations were then performed, resizing the images by a factor ranging from 0.8 to 1.2. Additionally, mirror flipping along the sagittal, coronal, and axial planes was performed to improve the model's orientation invariance.

Finally, to emphasize the ROI while still preserving contextual variability, the augmented training images were randomly cropped from their $256 \times 256 \times 256$ format to a size of $120 \times 120 \times 120$ voxels. This cropping ensured that at least 90% of the dens axis remained visible in each volume, thereby preserving diagnostic relevance.

Augmentations were applied on-the-fly during training, rather than by generating a permanently augmented dataset. Each patient case was thus presented multiple times per epoch in randomized, slightly altered forms. During training, fracture cases were sampled more frequently (10 times per epoch) compared to non-fracture cases (3 times per epoch), which contributed to balancing the classes and increasing robustness without inflating the dataset size.

In contrast, the validation and test datasets were not subjected to random cropping. Instead, a single, centered $120 \times 120 \times 120$ voxel region was extracted from each standardized $256 \times 256 \times 256$ image volume, centered around the dens. This consistent and standardized cropping approach ensured that all evaluation samples were processed under identical conditions, thereby providing a stable and fair benchmark for model comparison and performance assessment.

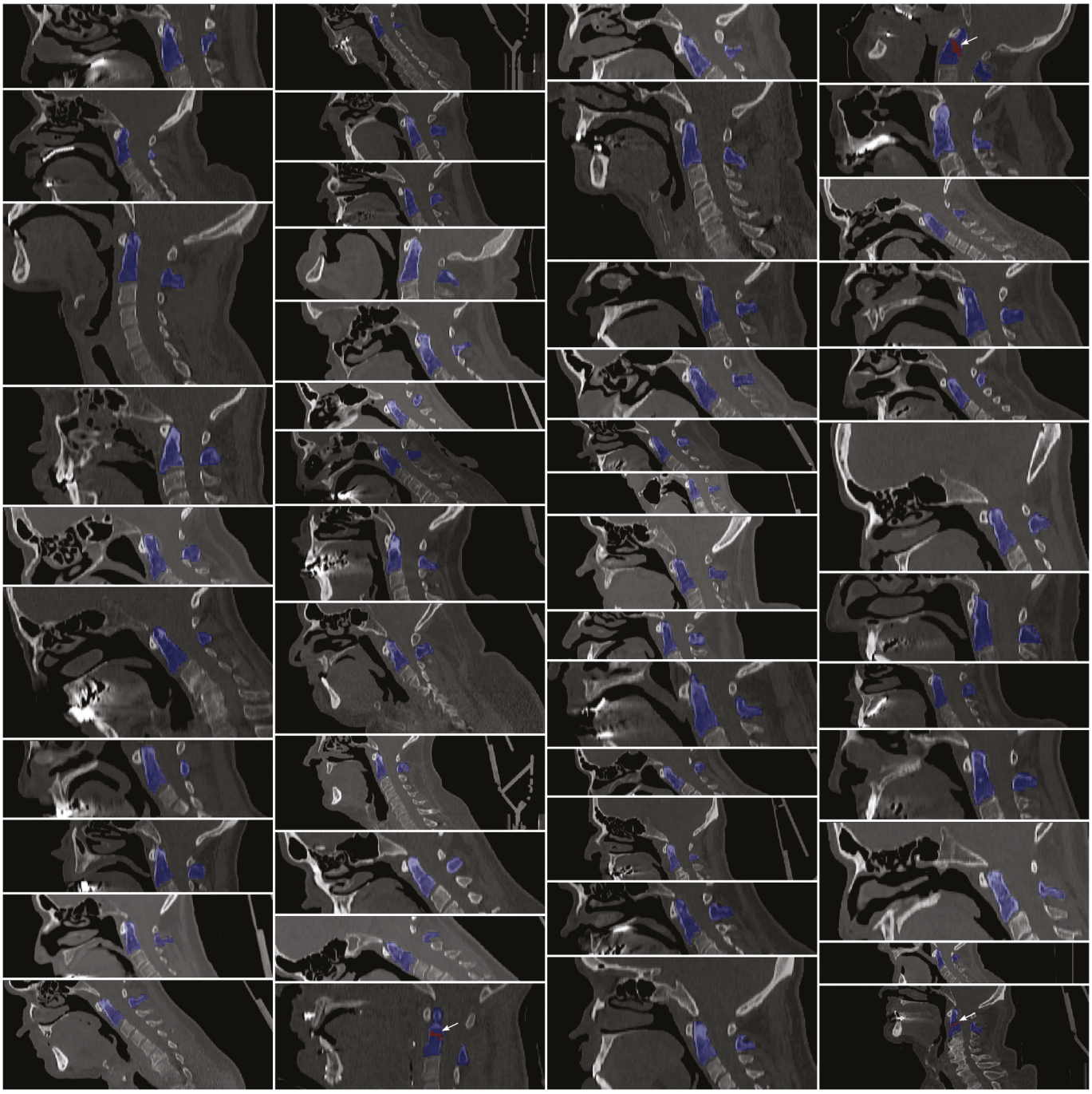


Figure 1. Representative visualization of 48 selected computed tomography (CT) images of the upper cervical spine. The images are sagittally reformatted and display overlaid manual segmentations of the dens axis (blue), and, if present, the corresponding fracture (red). Following manual segmentation, all images were processed according to the data post-processing procedures described in Section 2.2.2 and resampled to an isotropic resolution of 1 × 1 × 1 mm. The selected CT images illustrate the substantial variability within the dataset, including differences in resolutions and anatomical coverage—from scans focused solely on the upper cervical spine to those encompassing the entire skull and cervical spine. This heterogeneity underscores the complexity of the data and highlights the challenges encountered during segmentation, as well as the level of precision required to accurately delineate and analyze these structures.

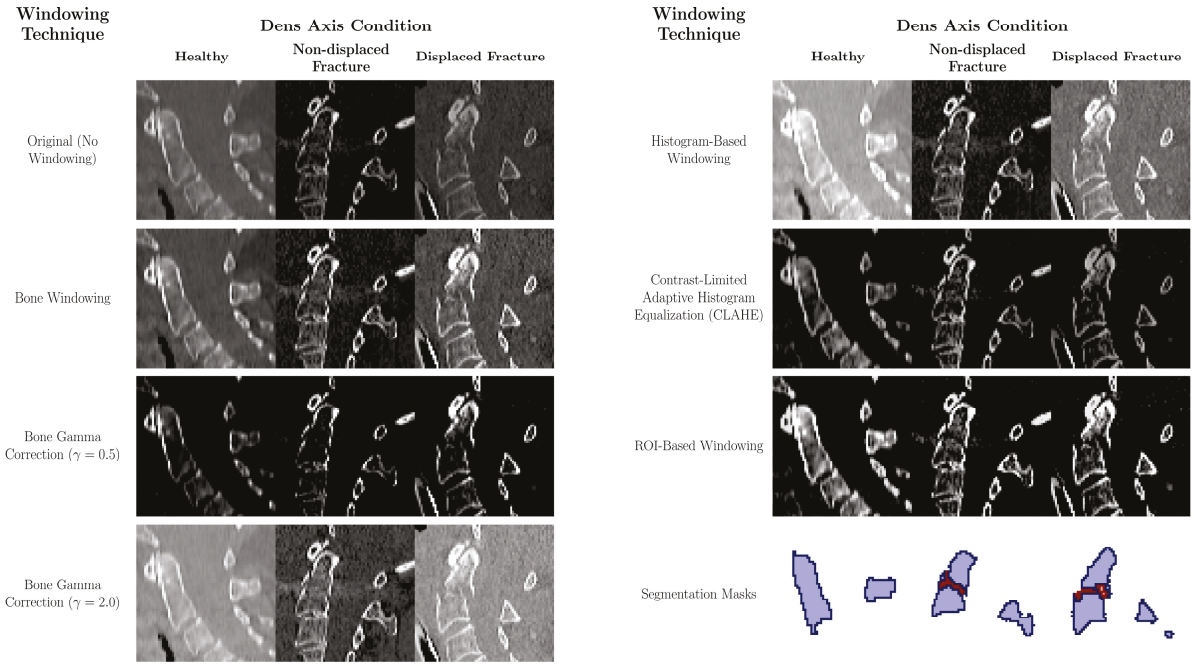


Figure 2. Representative examples illustrating the effects of different contrast adjustment methods on CT images of the dens axis. Shown are three sagittal CT views: one of a healthy dens, one with a non-displaced fracture, and one with a displaced fracture. Each case is presented using both the original contrast and various modified windowing techniques. Of all methods evaluated, ROI-based windowing most effectively enhances fracture visibility. CLAHE and bone windowing also yield noticeable improvements over unadjusted images, though to a lesser extent.

2.3. CNN-Based Segmentation and Classification

Convolutional neural networks (CNNs) have already demonstrated outstanding performance in medical image analysis, particularly for CT-based segmentation tasks [24,25]. In this study, we compare two distinct DL architectures to fully automate the detection of dens axis fractures from CT images, each leveraging different computational strategies to enhance diagnostic accuracy.

Model 1 (M1) is a single-stage architecture that combines a CNN with a feedforward neural network (FNN). It processes entire CT volumes without predefined ROIs, relying solely on learned features for direct classification into "fracture" or "no fracture" categories.

Model 2 (M2) adopts a two-stage pipeline. First, it employs U-Net-based segmentation to first identify the dens axis from the CT image. In the second stage, radiomic features are extracted from the segmented ROI and used to train a variety of ML classifiers. During training, M2 utilizes manually annotated ROIs to guide the segmentation network. Once trained, it autonomously segments unseen CT scans, enabling a fully automated end-to-end workflow suitable for clinical deployment.

2.3.1. M1—CNN- and FNN-Based Classification

Model M1 is a single-stage architecture that integrates a CNN with a FNNto classify dens axis CT images into binary categories: "with fracture" or "without fracture" (Figure 3, top). The CNN component incorporates DoubleConv modules, each consisting of two consecutive 3×3 convolutional layers, followed by batch normalization and ReLU activation functions. Max-pooling layers are inserted between modules to progressively downsample the feature maps while extracting spatial dimensions and hierarchical features. Following feature extraction, the resulting feature maps are flattened and passed through a five-layer FNN. Dropout layers (rate = 0.1) are applied between these layers to reduce overfitting.

The final output layer uses a softmax activation function to generate class probabilities, enabling robust binary classification.

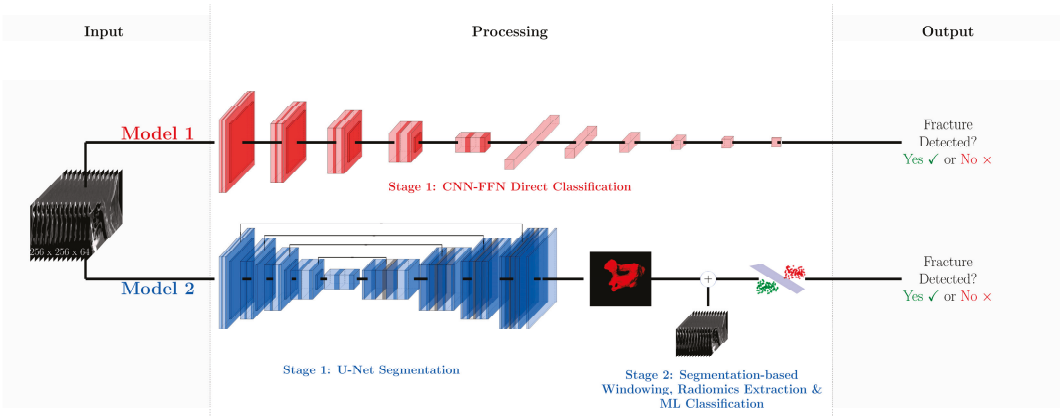


Figure 3. Comparative workflow of Model M1 and Model M2 for dens axis fracture classification. Both models process the same 3D input image. Model M1 (**top**) employs a convolutional neural network and a feedforward neural network (FNN) for direct classification, using DoubleConv modules and max-pooling for feature extraction, followed by multiple FNN layers with dropout to prevent overfitting. In contrast, Model M2 (**bottom**) follows a two-stage approach; a U-Net architecture first segments the dens axis (Stage 1), after which radiomic features are extracted for machine learning-based classification (Stage 2), enabling a more targeted and potentially more accurate detection of fractures.

2.3.2. M2—CNN-Based Segmentation and Radiomics Analysis

Model M2 incorporates a two-step methodology to improve upon M1 by incorporating a dedicated segmentation stage prior to classification (Figure 3, bottom). In the first step, a 3D U-Net architecture segments the dens region from the CT image. This encoder–decoder structure with skip connections captures both global context and fine-grained details, ensuring accurate ROI isolation for subsequent analysis.

In the second step, radiomics features are extracted from the segmented dens region using the PyRadiomics framework [26]. These features include quantitative descriptors such as texture patterns (e.g., gray-level co-occurrence matrix features), shape metrics (e.g., surface area, volume), and intensity statistics (e.g., first-order histogram features). This comprehensive feature set encapsulates the complex characteristics of the dens that may indicate the presence of fractures. The extracted radiomics features are then input into multiple machine learning classifiers—decision trees, extra trees, gradient boosting machines, k-nearest neighbors (KNN), Gaussian Naive Bayes, and random forests. By evaluating these diverse algorithms, we identify the most effective classifier, leveraging the strengths of ensemble and distance-based methods to manage the high-dimensional radiomic feature space.

2.3.3. Computational Implementation and Evaluation

All models were trained on a high-performance workstation equipped with two Intel Xeon Gold 6242R CPUs (Intel Corporation, Santa Clara, CA, USA), 376 GB RAM, and four NVIDIA RTX 3090 GPUs (NVIDIA, Santa Clara, CA, USA). Model development was performed in PyTorch (v2.3.1) [27], with PyTorch Lightning (v2.3.2) used to facilitate streamlined and efficient training procedures [28].

The Adam optimizer was employed with an initial learning rate of 0.001 for M1 and 0.01 for M2 to accommodate the differing complexities of the two models. A weight decay of 10^{-6} was applied to mitigate overfitting [29]. Learning rates were dynamically adjusted using a "ReduceLROnPlateau" scheduler, configured with a patience of 10 epochs,

a cooldown period of 10 epochs, and a reduction factor of 0.1. Training was conducted for a maximum of 50 epochs with a batch size of four, using input volumes standardized to $120 \times 120 \times 120$ voxels.

For the segmentation tasks in M2, a combined Dice and cross-entropy loss was applied [30] to address class imbalance and improve segmentation accuracy. For classification tasks in both M1 and M2, standard cross-entropy loss was utilized [31]. To compensate for the unequal distribution of fracture and non-fracture cases, class weights were introduced into the loss function based on the observed prevalence in the dataset. This weighting scheme ensured that misclassified fracture cases were penalized more strongly, thereby maintaining sensitivity despite the lower number of positive cases. In combination with the on-the-fly augmentation strategy, this approach provided a balanced optimization during training. The model achieving the best performance on the validation dataset was selected for final evaluation on the independent test set.

2.4. Evaluation Metrics and Statistical Analysis

Model performance was evaluated using distinct metrics tailored to the two key tasks of this study: segmentation and classification. For segmentation in M2, the Dice Similarity Coefficient (DSC) was used to quantify the spatial overlap between automatically generated segmentations and the manually annotated ground truth. Ranging from 0 (no overlap) to 1 (perfect overlap), the DSC quantifies directly reflect the model's ability to accurately localize the dens—a prerequisite for reliable radiomic analysis.

For classification performance, accuracy was used as the principal metric. Defined as the proportion of correctly classified cases (fractured vs. no fracture) out of the total number of cases, accuracy provides an intuitive and interpretable measure of diagnostic performance.

To further investigate the dependence of radiomics-based classification on segmentation quality, a stability analysis was performed. Radiomics feature extraction was conducted after segmentation, and the DSC was systematically varied within a controlled range. This procedure enabled assessment of how deviations from perfect segmentation impacted classification performance across different machine learning classifiers, as segmentation methodology has previously been shown to influence radiomics feature stability and reproducibility in imaging studies [32].

Additionally, descriptive statistical methods—including the use of heat maps—were used to visualize model performance across different windowing techniques, ML classifiers, and segmentation accuracies. These visual tools facilitated intuitive interpretation of interdependencies among model components and reinforced confidence in the system's reliability and clinical applicability.

3. Results

3.1. Segmentation Performance

M2's U-Net-based segmentation demonstrated consistently high precision, with DSC values ranging from 0.82 to 0.94 across the test dataset. As summarized in Table 1, the histogram-based windowing technique yielded the highest mean DSC (0.94 ± 0.01), closely followed by CLAHE (0.92 ± 0.01). Both techniques significantly enhanced local contrast around the dens axis, facilitating more accurate delineation of this critical anatomical structure.

Table 1. Comparison of classification accuracy and balanced accuracy for Model 1 (M1) and segmentation performance plus classification accuracy (using a K-nearest neighbor classifier) for Model 2 (M2) across different CT windowing techniques. For extended M2 classification results using additional classifiers, see Figure 4.

Windowing	Model Performance		
	Model 1	Model 2	
	Accuracy/Balanced Accuracy	Segmentation DSC (Mean ± SD)	Accuracy/Balanced Accuracy
without windowing	0.937/0.811	0.91 ± 0.01	0.853/0.865
CLAHE	0.875/0.734	0.92 ± 0.01	0.895/0.817
Bone (γ = 1.0)	0.918/0.834	0.85 ± 0.01	0.849/0.817
Bone (γ = 0.5)	0.932/0.862	0.82 ± 0.03	0.853/0.817
Bone (γ = 2.0)	0.891/0.817	0.90 ± 0.01	0.849/0.817
Histogram-based	0.902/0.825	0.94 ± 0.01	0.862/0.800
ROI-based [1]	(0.942/0.881)	n/a	0.957/0.937

Abbreviations: M1—Model 1; M2—Model 2; CLAHE—Contrast Limited Adaptive Histogram Equalization; Bone—windowed bone image; γ—gamma adjustment parameter; DSC—Dice Similarity Coefficient; SD—standard deviation. [1] ROI-based windowing applies optimal contrast adjustment by calculating the 5th and 95th percentiles of Hounsfield values within a given segmentation. Since this approach requires prior segmentation, no segmentation accuracy was evaluated. The values for M1 were only included to enable comparison with M2. However, ROI-based windowing inherently requires a two-stage approach and is therefore not compatible with the single-stage nature of M1. n/a—Not available because segmentation is necessary for contrast adjustment, which was only generated by the model during segmentation.

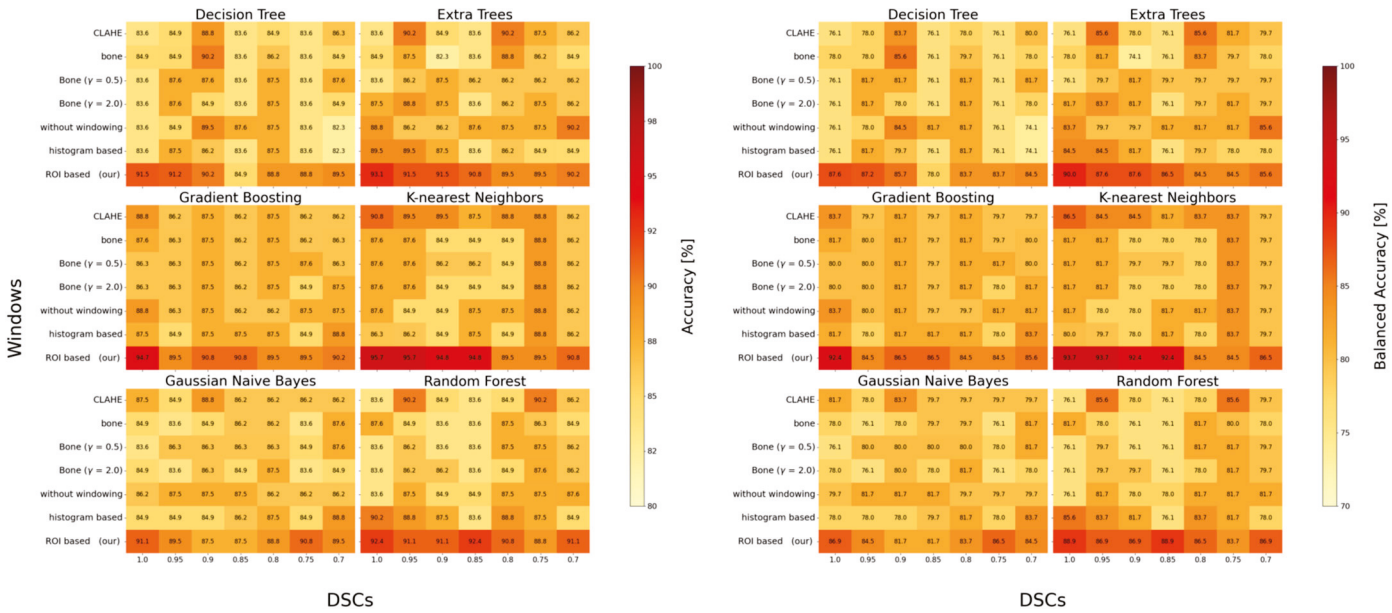


Figure 4. Heatmap illustrating classification accuracy (**left**) and balanced accuracy (**right**) of Model 2 (M2) across varying Dice Similarity Coefficient (DSC) levels, windowing techniques, and machine learning classifiers. Each cell represents the classification performance under a specific combination of segmentation quality and preprocessing method. Darker shades indicate higher accuracy. The visualization highlights M2's robustness and adaptability, demonstrating consistent performance across multiple classifiers and preprocessing strategies—even when segmentation quality varies. This stability analysis emphasizes the importance of verifying segmentation masks when applying such a two-stage pipeline in clinical routine. Abbreviations: CLAHE—Contrast Limited Adaptive Histogram Equalization; Bone—windowed bone image; γ—gamma adjustment parameter; DSC—Dice Similarity Coefficient.

3.2. Classification Performance

M1 Performance: M1, which relies solely on a CNN-FNNpipeline without predefined segmentation, achieved a peak classification accuracy of 93.7% on unprocessed images (Table 1). Applying ROI-based windowing marginally improved this result to 94.2%, while other preprocessing techniques showed limited added benefit. The corresponding balanced accuracy values were consistently lower (73.4–88.1%), indicating that M1 was more affected by the underlying class imbalance and tended to favor the majority class.

M2 Performance: In contrast, M2 consistently outperformed M1, achieving classification accuracies above 95% across all windowing methods tested. The highest accuracy—95.7%—was obtained using ROI-based windowing combined with a KNN classifier. As illustrated in Figure 4 (left), M2 maintained high classification performance across multiple ML classifiers—including Decision Trees, Extra Trees, Gradient Boosting, KNN, Gaussian Naive Bayes, and Random Forest. ROI-based windowing consistently emerged as the most effective preprocessing method, enabling detection of even minimally displaced fractures. A dedicated stability analysis revealed that classification accuracy declined by less than 1% as the DSC values dropped from 0.94 to 0.82. Furthermore, Figure 4 (right) highlights the stability of M2 K-nearest Neighbors classification across segmentation conditions. Importantly here is that the difference between accuracy and balanced accuracy remained marginal (<2%), which further supporting that M2 maintains a high true-positive rate despite the unequal class distribution. It is important to emphasize that in settings with lower overall accuracy, the corresponding balanced accuracy declined even more markedly. This indicates that, for these specific windowing–classifier combinations, the primary source of performance loss was a reduced sensitivity, i.e., a considerable number of fractures would have been missed. This finding underlines the clinical relevance of classifier choice, as inappropriate configurations may compromise the detection of subtle but critical fractures despite seemingly acceptable overall accuracy values.

4. Discussion

This study demonstrates that a two-stage approach—combining U-Net-based segmentation with radiomics-driven classification and targeted windowing adaptations—significantly improves the detection of dens axis fractures in CT imaging within the context of our dataset. Compared to a conventional single-stage DL-based model (M1), the integrated pipeline (M2) consistently achieved higher classification accuracies, reaching up to 95.7%, while maintaining robust performance across multiple preprocessing techniques and ML classifiers. Importantly, this performance was accomplished even with segmentation DSC values ranging from 0.82 to 0.94, indicating the model's strong resilience to imperfect ROI delineation. Such results highlight the potential of blending advanced segmentation, feature extraction, and ML classification for addressing complex radiological tasks involving subtle anatomical changes.

Our findings support emerging evidence that pure one-stage DL strategies, though powerful, may not fully leverage the complexity of medical imaging data—particularly in identifying subtle pathologies [33,34]. Dens axis fractures exemplify such diagnostic challenges, as they might be morphologically inconspicuous and can easily be overlooked. By isolating the ROI and extracting interpretable, high-dimensional radiomic features, M2 effectively capitalizes on subtle textural, shape, and intensity patterns. This synergy between anatomical localization and radiomic analysis improves the detection of even minimally displaced fractures—an area where end-to-end DL-only models like M1 may underperform due to limited spatial specificity. This advantage resonates with observations from other domains. Our results corroborate the added value of segmentation-driven radiomics in other contexts [14,15] and extend these findings to dens-axis fracture detection.

While our study focuses on fracture detection, this approach is broadly applicable to other diagnostic and prognostic imaging tasks.

To place our findings into context, Table 2 provides a comparative summary of previously published studies that investigated AI-based methods for fracture detection in the cervical and vertebral column using CT imaging. Most prior work has concentrated on the entire cervical spine or on thoracolumbar levels, where fractures tend to be larger and therefore more conspicuous. Reported accuracies range between 70% and 98%, depending on dataset size, class balance, and methodological design. Importantly, none of these studies has specifically focused on dens axis fractures. Our work therefore represents the first dens-specific study in this field, highlighting the importance of developing dedicated models for small but clinically critical anatomical structures. Given the potentially life-threatening consequences of missed dens fractures, such tailored approaches may provide substantial added value in clinical decision-making.

Table 2. Comparative overview of AI-based studies on cervical/vertebral fracture detection in CT.

Study	Region	Preprocessing	Pipeline/Classifier	Performance
Salehinejad et al., 2021 [35]	Cervical fracture (C1–C7)	CT	ResNet50-Cnn + BLSTM	Acc: 70.9–79.2%
Small et al., 2021 [36]	Cervical fracture (C1–C7)	CT	CNN (Aidoc)	Acc: 92%
Chłąd et al., 2023 [37]	Cervical fracture (slice based)	CT, bone window	Yolov5 + Vision Transformer	Acc: 98%
Li et al., 2023 [34]	Occult vertebral fractures (cervical + thoracolumbar)	CT, bone window	Radiomics + ML	Acc: 84.6%
Zhang et al., 2024 [38]	Osteoporotic vertebral fractures (cervical + thoracolumbar)	CT, RadImageNet feature extraction	CNN (RadImageNet vs. ImageNet)	C-Index: 0.795
Singh et al., 2025 [39]	Cervical fracture (C1–C7)	HU normalization	Inception-ResNet-v2 + U-Net decoder	Acc: 98.4%
Liu et al., 2025 [40]	Vertebral fractures + osteoporosis (incl. cervical, thoracolumbar)	CT, RadImageNet pretrained features	CNN (RadImageNet) vs. ImageNet	Acc: 76–80% (per class)
our	Dens axis fracture	ROI-based	U-Net segmentation + Radiomics + KNN	Acc: 95.7%

Beyond the methodological comparisons, our findings also carry important clinical implications. Missed dens fractures can lead to severe cervical instability and potentially life-threatening neurological complications. An automated pipeline with high diagnostic accuracy could therefore provide valuable support in emergency and trauma settings, where rapid and reliable fracture detection is critical. When used as a second reader or triage system, such tools may help reduce the rate of missed fractures, accelerate decision-making, and enhance efficiency in high-throughput radiology departments. While the present study represents a proof of concept, our results underscore the value of developing dedicated artificial intelligence (AI) solutions for small but clinically decisive anatomical structures such as the dens. Prospective multicenter studies will be required to confirm these benefits in real-world workflows and to quantify their impact on patient outcomes.

In addition to model architecture, our findings underscore the role of targeted contrast enhancement techniques, such as ROI-based and histogram-based windowing, which enhanced segmentation performance by improving local contrast around the dens axis

itself, thereby improving diagnostic accuracy as reflected by higher DSC scores and lower standard deviations. Although global contrast enhancement methods like CLAHE are well-established to improve visibility of subtle features [18], their incremental benefit was limited in this context—possibly due to the already optimized internal representation of CNN-based feature extractors. Nonetheless, contrast techniques that enhance local contrast were associated with improvements in both segmentation precision and classification accuracy. This targeted strategy may also be adaptable to other clinical scenarios requiring focused anatomical analysis.

Moreover, M2 demonstrated high performance stability across a range of segmentation qualities, underscoring its clinical applicability. A decline in DSC from 0.94 to 0.82 resulted in less than a 1% decrease in classification accuracy, suggesting that M2 does not rely on perfect segmentations. Rather, its diagnostic strength lies in the integration of reasonably accurate anatomical localization with high-dimensional robust radiomic feature analysis. This resilience is particularly relevant in real-world clinical settings, where image quality may vary due to factors such as heterogeneous imaging protocols, patient motion, and scanner settings—a variability also reflected in our study's deliberately heterogeneous population. These findings indicate M2's strong potential for clinical translation, pending prospective validation.

The flexibility of M2's radiomics-based classification stage is another important strength. While the KNN classifier yielded the highest accuracy in our dataset, M2 also performed well across various ML algorithms, including decision trees, random forests, and gradient boosting machines. This adaptability suggests that M2 is well-suited for further optimization and customization for various clinical use cases. Nonetheless, no single classifier is guaranteed to universally outperform others and future investigations may explore ensemble strategies, advanced boosting methods, or hybrid DL-ML architectures [41–43], to further optimize classification performance for specific clinical needs, patient populations, or imaging conditions.

Despite these promising results, several limitations must be acknowledged. First, this study should primarily be regarded as a proof-of-concept. Accordingly, no evaluation of runtime feasibility—including inference speed, PACS integration, or workflow impact on radiologists—was performed. Given the retrospective nature and limited size of the dataset, such an assessment was beyond the scope of this work. Nevertheless, clinical translation will ultimately depend not only on accuracy but also on seamless integration into daily radiological workflows and compliance with regulatory frameworks. Second, all data were derived from a single institution and acquired on scanners from the same vendor. Although we deliberately included examinations from different CT systems of one manufacturer to increase heterogeneity, this still constitutes a single-center, single-vendor setting. Such homogeneity limits generalizability to other populations, imaging protocols, and equipment. Variability in scanner hardware, reconstruction algorithms and acquisition parameters is well known to substantially influence radiomics features and thus model performance and remains one of the major challenges for translating AI into clinical routine. To address this, future work should incorporate multi-center, multi-vendor datasets with harmonized acquisition protocols to rigorously evaluate robustness and external validity [44]. Third, while a range of ML classifiers was tested, our selection deliberately focused on comparatively simple but well-established algorithms, reflecting the study's proof-of-concept character and the limited dataset size. The rationale was to illustrate that ML-based classifiers can still outperform complex DL-only approaches in settings with small sample sizes and subtle imaging patterns. Nevertheless, future work on larger, multi-center datasets should systematically include additional ML methods such as XGBoost or LightGBM to further optimize performance and evaluate robustness across institutions.

Fourth, while radiomics features were extracted, no dedicated feature importance analysis was performed. We deliberately refrained from reporting single feature rankings, as their interpretability outside of a high-dimensional multivariate context is limited and strongly model-dependent. Future multi-center studies should systematically assess feature stability and reproducibility across different scanners and protocols. In such a context, feature importance analyses may not only enhance scientific transparency but also support clinical interpretability, by providing radiologists with concrete radiomics-based markers to contextualize model decisions. Fifth, the dataset exhibited class imbalance, with relatively few positive cases of dens axis fracture. Although this reflects clinical reality, it may reduce sensitivity for rare fracture patterns. Strategies such as data augmentation, oversampling, or cost-sensitive training may help mitigate this limitation and refine the model's capacity to detect subtle or infrequent fracture patterns. Sixth, the range of classifiers evaluated, while representative, was not exhaustive. Additional testing with advanced ensemble methods, dimensionality reduction techniques, or advanced DL architectures could uncover further performance gains and may yield further improvements in accuracy or interpretability [43]. Seventh, only the validation-optimal model was stored and evaluated, which means that complete epoch-wise accuracy and loss curves are not available. This choice reflects a clinically oriented workflow but limits the possibility to illustrate training dynamics in detail. Finally, although the pipeline demonstrated strong retrospective performance, its clinical utility has not yet been validated prospectively. Future studies should assess how M2 integrates into existing PACS workflows, impacts radiologists' efficiency, and influences patient outcomes. Establishing practical benefits such as reduced diagnostic delays or improved decision-making will be essential to demonstrating clinical value and encouraging broader adoption.

5. Conclusions

In conclusion, our integrated pipeline—combining ROI-based windowing, radiomics-driven feature extraction, and advanced machine learning—significantly improves dens axis fracture detection in CT imaging. The two-stage model (M2) outperformed a conventional CNN-based approach, demonstrating strong accuracy and robustness across various preprocessing techniques and classifiers. While these findings highlight potential for improved diagnostic accuracy, prospective studies on larger datasets are needed to validate the approach and assess its clinical integration.

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

Funding: S.N. and L.M.W. were supported by grants from the local Research Committee of the Medical Faculty of Heinrich-Heine-University Düsseldorf, Germany. A.M-L. and S.N. were additionally supported by "Deutsche Forschungsgemeinschaft" (DFG) (530863408 and NE 2136/3-1).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Ethics Committee of Heinrich-Heine-University of Düsseldorf (protocol code 2020-850, approval date: 8 April 2020).

Informed Consent Statement: Patient consent was waived due to the retrospective study design and the complete anonymization of all imaging and clinical data.

Data Availability Statement: Data and evaluation scripts can be provided by the authors upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

CT	Computed Tomography
DL	Deep Learning
CNN	Convolutional Neural Network
FNN	Feedforward Neural Network
ROI	Region of Interest
ML	Machine Learning
U-Net	U-shaped Convolutional Neural Network Architecture
3D	Three-Dimensional
CLAHE	Contrast Limited Adaptive Histogram Equalization
DSC	Dice Similarity Coefficient
SD	Standard Deviation
HU	Hounsfield Unit
CTDI_vol	Computed Tomography Dose Index Volume
PACS	Picture Archiving and Communication System
Adam	Adaptive Moment Estimation (optimizer)
CPU	Central Processing Unit
GPU	Graphics Processing Unit
MRI	Magnetic Resonance Imaging
AI	Artificial intelligence

References

1. Good, A.E.; Ramponi, D.R. Odontoid/Dens Fractures. *Adv. Emerg. Nurs. J.* **2024**, *46*, 38–43. [CrossRef]
2. Krupinski, E.A.; Williams, M.B.; Andriole, K.; Strauss, K.J.; Applegate, K.; Wyatt, M.; Bjork, S.; Seibert, J.A. Digital radiography image quality: Image processing and display. *J. Am. Coll. Radiol.* **2007**, *4*, 389–400. [CrossRef]
3. Kabasawa, H. MR Imaging in the 21st Century: Technical Innovation over the First Two Decades. *Magn. Reson. Med. Sci.* **2022**, *21*, 71–82. [CrossRef]
4. Pruessmann, K.P.; Weiger, M.; Scheidegger, M.B.; Boesiger, P. SENSE: Sensitivity encoding for fast MRI. *Magn. Reson. Med.* **1999**, *42*, 952–962. [CrossRef]
5. van Vaals, J.J.; Brummer, M.E.; Dixon, W.T.; Tuithof, H.H.; Engels, H.; Nelson, R.C.; Gerety, B.M.; Chezmar, J.L.; den Boer, J.A. "Keyhole" method for accelerating imaging of contrast agent uptake. *J. Magn. Reson. Imaging* **1993**, *3*, 671–675. [CrossRef] [PubMed]
6. Chen, G.-H.; Tang, J.; Leng, S. Prior image constrained compressed sensing (PICCS): A method to accurately reconstruct dynamic CT images from highly undersampled projection data sets. *Med. Phys.* **2008**, *35*, 660–663. [CrossRef] [PubMed]
7. Radke, K.L.; Hußmann, J.; Röwer, L.; Voit, D.; Frahm, J.; Antoch, G.; Klee, D.; Pillekamp, F.; Wittsack, H.-J. "shortCardiac"—An open-source framework for fast and standardized assessment of cardiac function. *SoftwareX* **2023**, *23*, 101453. [CrossRef]
8. Zhang, R.; Shi, J.; Liu, S.; Chen, B.; Li, W. Performance of radiomics models derived from different CT reconstruction parameters for lung cancer risk prediction. *BMC Pulm. Med.* **2023**, *23*, 132. [CrossRef]
9. Wang, W.; Fan, Z.; Zhen, J. MRI radiomics-based evaluation of tuberculous and brucella spondylitis. *J. Int. Med. Res.* **2023**, *51*, 3000605231195156. [CrossRef]
10. Salahuddin, Z.; Woodruff, H.C.; Chatterjee, A.; Lambin, P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Comput. Biol. Med.* **2022**, *140*, 105111. [CrossRef]
11. Bhatt, R.B.; Gopal, M. Neuro-fuzzy decision trees. *Int. J. Neural Syst.* **2006**, *16*, 63–78. [CrossRef] [PubMed]
12. Zheng, Y.; Zhou, D.; Liu, H.; Wen, M. CT-based radiomics analysis of different machine learning models for differentiating benign and malignant parotid tumors. *Eur. Radiol.* **2022**, *32*, 6953–6964. [CrossRef] [PubMed]

13. Benfante, V.; Salvaggio, G.; Ali, M.; Cutaia, G.; Salvaggio, L.; Salerno, S.; Busè, G.; Tulone, G.; Pavan, N.; Di Raimondo, D.; et al. Grading and Staging of Bladder Tumors Using Radiomics Analysis in Magnetic Resonance Imaging. In *Image Analysis and Processing-ICIAP 2023 Workshops*; Foresti, G.L., Fusiello, A., Hancock, E., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2024; pp. 93–103. ISBN 978-3-031-51025-0.
14. Liang, C.; Cheng, Z.; Huang, Y.; He, L.; Chen, X.; Ma, Z.; Huang, X.; Liang, C.; Liu, Z. An MRI-based Radiomics Classifier for Preoperative Prediction of Ki-67 Status in Breast Cancer. *Acad. Radiol.* **2018**, *25*, 1111–1117. [CrossRef]
15. Gutsche, R.; Lohmann, P.; Hoevels, M.; Ruess, D.; Galldiks, N.; Visser-Vandewalle, V.; Treuer, H.; Ruge, M.; Kocher, M. Radiomics outperforms semantic features for prediction of response to stereotactic radiosurgery in brain metastases. *Radiother. Oncol.* **2022**, *166*, 37–43. [CrossRef]
16. Zhou, Y.; Shi, C.; Lai, B.; Jimenez, G. Contrast enhancement of medical images using a new version of the World Cup Optimization algorithm. *Quant. Imaging Med. Surg.* **2019**, *9*, 1528–1547. [CrossRef]
17. Yoshimi, Y.; Mine, Y.; Ito, S.; Takeda, S.; Okazaki, S.; Nakamoto, T.; Nagasaki, T.; Kakimoto, N.; Murayama, T.; Tanimoto, K. Image preprocessing with contrast-limited adaptive histogram equalization improves the segmentation performance of deep learning for the articular disk of the temporomandibular joint on magnetic resonance images. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **2023**, *138*, 128–141. [CrossRef]
18. Alwakid, G.; Gouda, W.; Humayun, M. Deep Learning-Based Prediction of Diabetic Retinopathy Using CLAHE and ESRGAN for Enhancement. *Healthcare* **2023**, *11*, 863. [CrossRef]
19. Park, K.S.; Cho, S.-G.; Kim, J.; Song, H.-C. Effect of contrast-limited adaptive histogram equalization on deep learning models for classifying bone scans. *J. Nucl. Med.* **2022**, *63*, 3240.
20. Yushkevich, P.A.; Piven, J.; Hazlett, H.C.; Smith, R.G.; Ho, S.; Gee, J.C.; Gerig, G. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage* **2006**, *31*, 1116–1128. [CrossRef]
21. O'Broin, E.S.; Morrin, M.; Breathnach, E.; Allcutt, D.; Earley, M.J. Titanium mesh and bone dust calvarial patch during cranioplasty. *Cleft Palate Craniofac. J.* **1997**, *34*, 354–356. [CrossRef]
22. Hemminger, B.M.; Zong, S.; Muller, K.E.; Coffey, C.S.; DeLuca, M.C.; Johnston, R.E.; Pisano, E.D. Improving the detection of simulated masses in mammograms through two different image-processing techniques. *Acad. Radiol.* **2001**, *8*, 845–855. [CrossRef] [PubMed]
23. Khan, Z.; Yahya, N.; Alsaih, K.; Ali, S.S.A.; Meriaudeau, F. Evaluation of Deep Neural Networks for Semantic Segmentation of Prostate in T2W MRI. *Sensors* **2020**, *20*, 3183. [CrossRef] [PubMed]
24. Radke, K.L.; Wollschläger, L.M.; Nebelung, S.; Abrar, D.B.; Schleich, C.; Boschheidgen, M.; Frenken, M.; Schock, J.; Klee, D.; Frahm, J.; et al. Deep Learning-Based Post-Processing of Real-Time MRI to Assess and Quantify Dynamic Wrist Movement in Health and Disease. *Diagnostics* **2021**, *11*, 1077. [CrossRef] [PubMed]
25. Turečková, A.; Tureček, T.; Komínková Oplatková, Z.; Rodríguez-Sánchez, A. Improving CT Image Tumor Segmentation Through Deep Supervision and Attentional Gates. *Front. Robot. AI* **2020**, *7*, 106. [CrossRef]
26. van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.-C.; Pieper, S.; Aerts, H.J.W.L. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* **2017**, *77*, e104–e107. [CrossRef]
27. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 8024–8035.
28. Falcon, W.; Borovec, J.; Wälchli, A.; Eggert, N.; Schock, J.; Jordan, J.; Skafte, N.; Ir1dXD; Bereznyuk, V.; Harris, E.; et al. PyTorchLightning/Pytorch-Lightning: 0.7.6 Release. 2020. Available online: https://zenodo.org/records/3828935 (accessed on 5 September 2025).
29. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. 2014. Available online: https://arxiv.org/pdf/1412.6980.pdf (accessed on 5 September 2025).
30. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Jorge Cardoso, M. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. 2017. Available online: https://arxiv.org/pdf/1707.03237.pdf (accessed on 13 May 2025).
31. Ma, Y.; Liu, Q.; Quan, Z. Automated image segmentation using improved PCNN model based on cross-entropy. In Proceedings of the 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, ISIMP 2004, Hong Kong, China, 20–22 October 2004; IEEE Operations Center: Piscataway, NJ, USA, 2004; pp. 743–746, ISBN 0-7803-8687-6.
32. Giaccone, P.; Benfante, V.; Stefano, A.; Cammarata, F.P.; Russo, G.; Comelli, A. PET Images Atlas-Based Segmentation Performed in Native and in Template Space: A Radiomics Repeatability Study in Mouse Models. In *Image Analysis and Processing. ICIAP 2022 Workshops*; Mazzeo, P.L., Frontoni, E., Sclaroff, S., Distante, C., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 351–361. ISBN 978-3-031-13320-6.
33. Ren, M.; Yi, P.H. Deep learning detection of subtle fractures using staged algorithms to mimic radiologist search pattern. *Skeletal Radiol.* **2022**, *51*, 345–353. [CrossRef]

34. Li, W.-G.; Zeng, R.; Lu, Y.; Li, W.-X.; Wang, T.-T.; Lin, H.; Peng, Y.; Gong, L.-G. The value of radiomics-based CT combined with machine learning in the diagnosis of occult vertebral fractures. *BMC Musculoskelet. Disord.* **2023**, *24*, 819. [CrossRef]
35. Salehinejad, H.; Ho, E.; Lin, H.-M.; Crivellaro, P.; Samorodova, O.; Arciniegas, M.T.; Merali, Z.; Suthiphosuwan, S.; Bharatha, A.; Yeom, K.; et al. Deep Sequential Learning for Cervical Spine Fracture Detection on Computed Tomography Imaging. 2020. Available online: http://arxiv.org/pdf/2010.13336v4 (accessed on 3 October 2025).
36. Small, J.E.; Osler, P.; Paul, A.B.; Kunst, M. CT Cervical Spine Fracture Detection Using a Convolutional Neural Network. *AJNR Am. J. Neuroradiol.* **2021**, *42*, 1341–1347. [CrossRef]
37. Chłąd, P.; Ogiela, M.R. Deep Learning and Cloud-Based Computation for Cervical Spine Fracture Detection System. *Electronics* **2023**, *12*, 2056. [CrossRef]
38. Zhang, J.; Xia, L.; Zhang, X.; Liu, J.; Tang, J.; Xia, J.; Liu, Y.; Zhang, W.; Liang, Z.; Tang, G.; et al. Development and validation of a predictive model for vertebral fracture risk in osteoporosis patients. *Eur. Spine J.* **2024**, *33*, 3242–3260. [CrossRef]
39. Singh, M.; Tripathi, U.; Patel, K.K.; Mohit, K.; Pathak, S. An efficient deep learning based approach for automated identification of cervical vertebrae fracture as a clinical support aid. *Sci. Rep.* **2025**, *15*, 25651. [CrossRef]
40. Liu, J.; Zhang, L.; Yuan, Y.; Tang, J.; Liu, Y.; Xia, L.; Zhang, J. Deep Learning Radiomics Model Based on Computed Tomography Image for Predicting the Classification of Osteoporotic Vertebral Fractures: Algorithm Development and Validation. *JMIR Med. Inform.* **2025**, *13*, e75665. [CrossRef]
41. Bruschetta, R.; Tartarisco, G.; Lucca, L.F.; Leto, E.; Ursino, M.; Tonin, P.; Pioggia, G.; Cerasa, A. Predicting Outcome of Traumatic Brain Injury: Is Machine Learning the Best Way? *Biomedicines* **2022**, *10*, 686. [CrossRef]
42. Fricke, C.; Alizadeh, J.; Zakhary, N.; Woost, T.B.; Bogdan, M.; Classen, J. Evaluation of Three Machine Learning Algorithms for the Automatic Classification of EMG Patterns in Gait Disorders. *Front. Neurol.* **2021**, *12*, 666458. [CrossRef]
43. Choy, G.; Khalilzadeh, O.; Michalski, M.; Do, S.; Samir, A.E.; Pianykh, O.S.; Geis, J.R.; Pandharipande, P.V.; Brink, J.A.; Dreyer, K.J. Current Applications and Future Impact of Machine Learning in Radiology. *Radiology* **2018**, *288*, 318–328. [CrossRef]
44. Koçak, B.; Durmaz, E.Ş.; Ateş, E.; Kılıçkesmez, Ö. Radiomics with artificial intelligence: A practical guide for beginners. *Diagn. Interv. Radiol.* **2019**, *25*, 485–495. [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.

diagnostics

MDPI

Article

Computed Tomography-Based Radiomics Diagnostic Model for Fat-Poor Small Renal Tumor Subtypes

Seokhwan Bang [1,†], Heehwan Wang [2,†], Hoyoung Bae [3], Sung-Hoo Hong [1], Jiook Cha [2,4,*] and Moon Hyung Choi [5,*]

[1] Department of Urology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea; drvion@gmail.com (S.B.); toomey@catholic.ac.kr (S.-H.H.)
[2] Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul 08826, Republic of Korea; dhkdgmlghks@snu.ac.kr
[3] Department of Urology, Seoul Metropolitan Government, Seoul National University Boramae Medical Center, Seoul 07061, Republic of Korea; hoyoung.b@gmail.com
[4] Department of Phychology, Seoul National University, Seoul 08826, Republic of Korea
[5] Department of Radiology, Eunpyung St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea
* Correspondence: connectome@snu.ac.kr (J.C.); choimh1205@gmail.com (M.H.C.)
† These authors contributed equally to this work.

Abstract: Background: Differentiating histologic subtypes of fat-poor small renal masses using conventional imaging remains difficult due to their overlapping radiologic characteristics. We aimed to develop a machine learning-based diagnostic model using CT-derived radiomic features to classify the five most common renal tumor subtypes: clear cell RCC (ccRCC), papillary RCC (pRCC), chromophobe RCC (chRCC), angiomyolipoma (AML), and oncocytoma. **Methods:** A total of 499 patients with pathologically confirmed renal tumors who underwent preoperative contrast-enhanced CT and nephrectomy were retrospectively analyzed. **Results:** We extracted and analyzed radiomic features from 1548 multi-phase CT scans from 499 patients, focusing on fat-poor tumors. Five machine learning classifiers including Linear SVM, Rbf SVM, Random Forest, and XGBoost were involved. Among the models, XGBoost showed the best classification performance, with an average AU-PRC: mean = 0.757, standard error = 0.033 and a renal angiomyolipoma-specific AU-ROC: mean = 0.824, standard error = 0.023. These results outperformed other single-phase CT radiomic feature-based machine learning models trained with 20% of principal components. **Conclusions:** This study demonstrates the effectiveness of radiomics-based machine learning in classifying renal tumor subtypes and highlights the potential of AI in medical imaging. The findings, particularly the utility of single-phase CT and feature optimization, offer valuable insights for future precision medicine approaches. Such methods may support more personalized diagnosis and treatment planning in renal oncology.

Keywords: RCC; small renal mass; radiomics; kidney cancer; diagnostics

1. Introduction

Kidney cancer represents a significant global health burden, with the American Cancer Society reporting an estimated 81,610 new cases diagnosed in the United States alone in 2024. This highlights the urgent need for advanced diagnostic and treatment strategies to improve patient outcomes. Kidney cancer, predominantly renal cell carcinoma (RCC), is complex due to its multiple subtypes, each with unique prognostic implications [1,2].

However, considered anatomically, the kidney is surrounded by a membrane called Gerota's fascia, and this membrane plays a role in preventing the spread of kidney cancer,

so inserting a needle to perform a biopsy is taboo. For this reason, imaging diagnosis is very important in kidney cancer. Kidney cancer is usually discovered by ultrasound sonography or CT scan accidentally, and when a renal mass is diagnosed as cancer, the accuracy is currently known to be about 90% [3]. However, this high diagnosis rate includes all cases that can be easily diagnosed as cancer, such as very large kidney tumors. In cases such as small kidney tumors or fat-poor tumors, the diagnosis rate drops sharply [4]. When pathological results cannot be predicted, partial nephrectomy or radical nephrectomy, which are chosen for proactive surgical management, are highly invasive procedures that require advanced surgical expertise. If the lesion turns out to be benign, these surgeries can impose significant health and economic burdens on the patient. Therefore, accurate screening through imaging studies is of paramount importance.

Accurate diagnosis of kidney cancer subtypes remains challenging due to the variability in their appearance on imaging modalities like computed tomography (CT). This variability can lead to differential diagnostic outcomes, affecting treatment decisions [5,6]. Studies such as those by Bauman et al. [7] and Tanaka et al. [8] have documented the incidence and challenges of accurately classifying these tumors, emphasizing the risk of unnecessary interventions due to diagnostic inaccuracies.

The integration of machine learning with radiomic data represents a cutting-edge approach in the diagnosis of kidney cancer. Deep learning models, particularly those employing convolutional neural networks (CNNs), have demonstrated superior accuracy in identifying and classifying renal tumors compared to traditional methods [9,10]. Such models not only streamline the diagnostic process, but also reduce the variability associated with human interpretation [11,12].

Our research team has also made various attempts to utilize AI in interpreting CT scans to classify the presence of renal tumors and their subtypes. Our team first attempted prediction using Convolutional Neural Networks (CNN), one of the AI techniques known for achieving the highest predictive accuracy. Although we achieved an accuracy of nearly 62%, we were unable to obtain actionable feedback from this approach. However, this study confirmed that predictions are feasible, and that the predictive accuracy of the AI model exceeded that of current radiologists [13]. This model addresses critical gaps in current diagnostic approaches by providing a robust, automated system that offers high accuracy and reproducibility [14]. However, this study concluded that it could not provide feedback substantial enough to advance current medical practice. While the CNN demonstrated good predictive performance, it lacked the capability to offer actionable feedback or highlight specific components for radiologists to focus on.

Radiomics has emerged as a revolutionary technique in medical imaging, offering the potential to extract quantifiable data from images that are beyond human visual detection. This approach has significantly advanced the classification of renal masses, aiding in distinguishing between benign and malignant forms [15]. Sun et al. [16] and Leon et al. [17] provide evidence of the effectiveness of radiomic features in enhancing the diagnostic precision of RCC.

Radiomics, an AI-driven technology, offers significant potential as a supportive tool for radiologists. Its diagnostic decisions are based on features that can be validated by human experts, making it a reliable feedback mechanism. Leveraging this technology, we explored its application in diagnosing subtypes of renal cancer. We hypothesized that integrating radiomics into clinical practice could enhance the diagnostic accuracy of radiologists. As a precursor to this study, we first conducted research focused solely on determining the presence or absence of tumors [18]. Through this approach, we aimed to identify, using radiomics, features that should be prioritized for classification. This study is also part of the process of verifying those features.

By improving the accuracy of subtype classification, our model contributes to more tailored treatment strategies, potentially leading to better patient outcomes. The adaptability of our model across different imaging modalities and its application to other cancer types also suggests a broad potential impact, paving the way for future innovations in oncological diagnostics [19,20].

2. Materials and Methods

2.1. Patients

We utilized data from patients who underwent total or partial renal resection at a single institution between 2003 and 2021. Excluded from the study were cases of kidney tumor with a fat component exceeding 30%. Only patients with fat-poor tumors were included, resulting in a total selection of 499 patients. The average age of these patients was 56.02 ± 12.18 years, and the mean tumor size on CT was 3.515 ± 2.42 cm. The study encompassed patients with benign tumors like oncocytoma and AML, as well as those with malignant tumors such as clear cell, chromophobe, and papillary-type RCC. Each patient underwent CT imaging in one to four phases, resulting in a total of 1548 sets of CT images. Details of the patient demographics are provided in Table 1.

Table 1. Demographic and pathological characteristics of the patients.

	Training Dataset	**Test Dataset**	***p* Value**
Sex	289	210	0.639
Male (%)	162 (49.2%)	123 (58.6%)	
Female (%)	127 (50.8%)	87 (41.4%)	
Age (years) (range)	57.0 (22, 83)	57.0 (27, 81)	0.816
Cancer size (cm) (range)	3.629 (0.8, 17.5)	3.357 (0.7, 12.0)	0.216
Renal cell tumor subtype			0.881
oncocytoma (%)	25 (8.6%)	23 (11%)	
angiomyolipoma (%)	47 (16.3%)	31 (14.7%)	
chromophobe renal cell carcinoma (%)	48 (16.6%)	36 (17.1%)	
papillary renal cell carcinoma (%)	50 (17.3%)	39 (18.6%)	
clear cell renal cell carcinoma (%)	119 (41.2%)	81 (38.6%)	

We used paired the sample t-test or chi-squared test to hold-out test the differences in demographic and pathological characteristics for continuous variables (i.e., age, cancer size) and categorical variables (i.e., sex, kidney cancer subtype).

CT scans were obtained in various configurations: non-contrast, arterial phase (20–30 s after contrast injection), portal phase (60–70 s), and delayed phase (>180 s). For each CT scan, voxel-level segmentation labels were collected, with trained annotators manually outlining kidneys and tumors in the images. Annotations were further reviewed and refined by a radiologist with 11 years of experience. A second diagnosis was provided only if the radiologist had significant doubts about the initial diagnosis. Radiologist performance was evaluated based on the first diagnosis (top-one performance) and both the first and second diagnoses (top-two performance).

2.2. Radiomics Feature Analysis Workflow

Radiomics quantitatively extract textural information from medical images, which can then be utilized further by machine learning algorithms to support clinical decision-making. To develop multi-phase CT scan radiomic feature-based machine learning models for classifying subtypes of renal cell tumors, we initially obtained various types of radiomic features from each phase of the CT scans. Considering that some participants only underwent one or two phases of CT scans, such as the non-contrast and portal phases, we split participants with multi-phase and those with limited phase CT scans differently into training

and hold-out test sets to align with our objective of developing machine learning models based on multi-phase CT scan radiomic features. We first randomly split participants with multi-phase CT scans into 75% of the training set and 25% of the hold-out test set in a stratified manner, and then we included those with single-phase or two-phase phase CT scans only in the hold-out test set. This resulted in both the training and the hold-out test set having a similar number of participants. The training and hold-out test sets consisted of similar proportions of renal cell tumor subtypes. Only the training set was used for feature selection and model training with 10-fold cross-validation. We compared the following four most popular machine learning algorithms: Linear Support Vector Machine (Linear SVM), Radial basis function Support Vector Machine (Rbf SVM), Random Forest, and XGBoost [21]. We conducted hyperparameter tuning through 50 trials of a random hyperparameter search for renal cell tumor subtype classification. We searched hyperparameters with Optuna (version 3.1.0) [22]. Codes used for our analyses are available for reproducibility testing (https://github.com/Transconnectome/Kidney_Radiomics; accessed on 9 February 2025). Due to the class imbalance among renal cell tumor subtypes, with a great number of participants having specific renal cell tumor subtypes such as clear cell renal cell carcinoma compared to other subtypes, we employed oversampling during the training phase of model-based feature selection and model development. We implemented oversampling with the Synthetic Minority Over-sampling Technique (SMOTE) using imbalanced-learn (version 0.10.1) [23]. We evaluated the classification performance of our machine learning models based on their ability to accurately differentiate each renal cell tumor subtype from the others. To assess this, we calculated the evaluation metrics, including accuracy, f1 score, Area Under Receiving Operating Characteristic Curves, and Area Under the Precision and Recall Curve, using a One-vs-Rest approach and the "micro" options in scikit-learn. The hold-out test set was exclusively utilized to evaluate the final performance of machine learning models. By using scikit-learn (version 1.2.1) [24] and Python (version 3.10.8), we conducted feature selection, machine learning model development, and model evaluation processes. We followed the Checklist for Artificial Intelligence in Medical Imaging (CLAIMS) to ensure the reproducibility of our study [25]. A schematic overview of the radiomics feature analysis workflow is illustrated in Figure 1.

2.3. Radiomics Feature Extraction

We resampled 1548 multi-phase CT scans with a resolution of 1 mm $\times$ 1 mm $\times$ 1 mm, and then we extracted 1288 radiomic features from the segmented region of interests (ROIs) of original CT scans, wavelet-filtered CT scans, and Laplacian of Gaussian-filtered scans. We extracted radiomic features utilizing the Python package PyRadiomics (version 3.1.0) [26] and Python (version 3.7). The extracted radiomic features included first-order features, three-dimensional shape features, Gray Level Cooccurence Matrix (GLCM), Gray Level Run Length Matrix (GLRM), Gray Level Size Zone Matrix (GLSZM), Neighboring Gray Tone Difference Matrix (NGTDM), and Gray Level Dependence Matrix (GLDM). Representative multi-phase CT scans with corresponding kidney ROI masks used for radiomic feature extraction are shown in Figure 2.

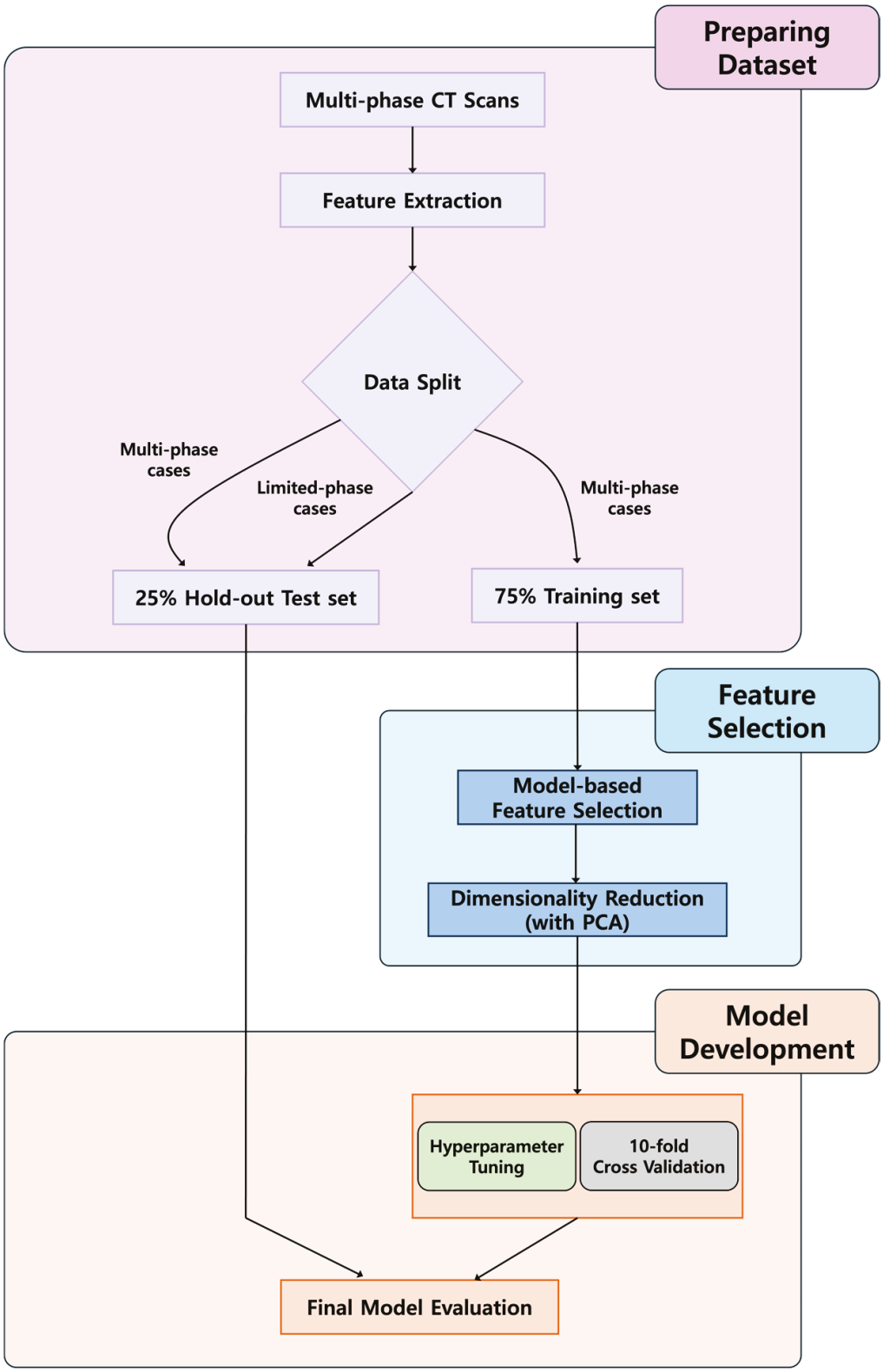
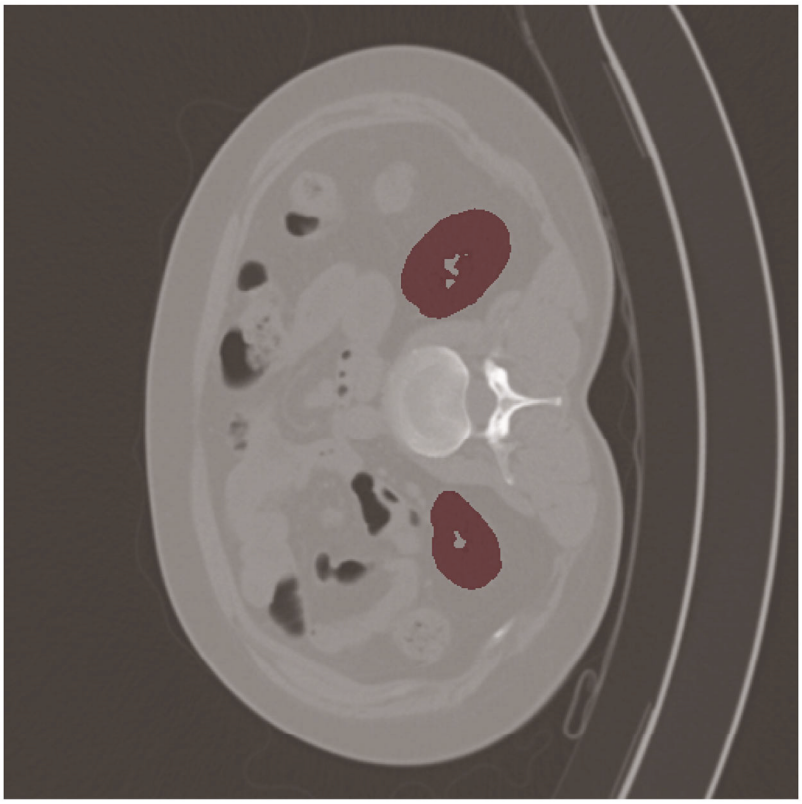
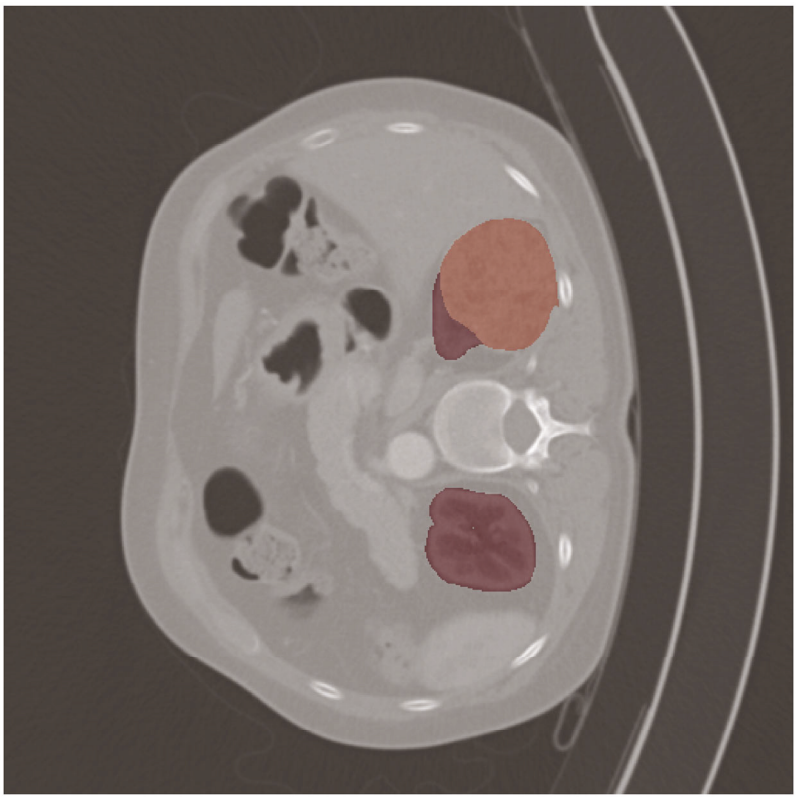


Figure 1. Flow chart of radiomic feature analysis. Radiomics features extracted from multi-phase CT scans were used to develop machine learning models for classifying renal cell tumor subtypes. Participants with multi-phase scans were split into training (75%) and hold-out test (25%) sets, while those with single- or two-phase scans were included only in the test set. To address class imbalance, synthetic oversampling was applied during model training and feature selection.

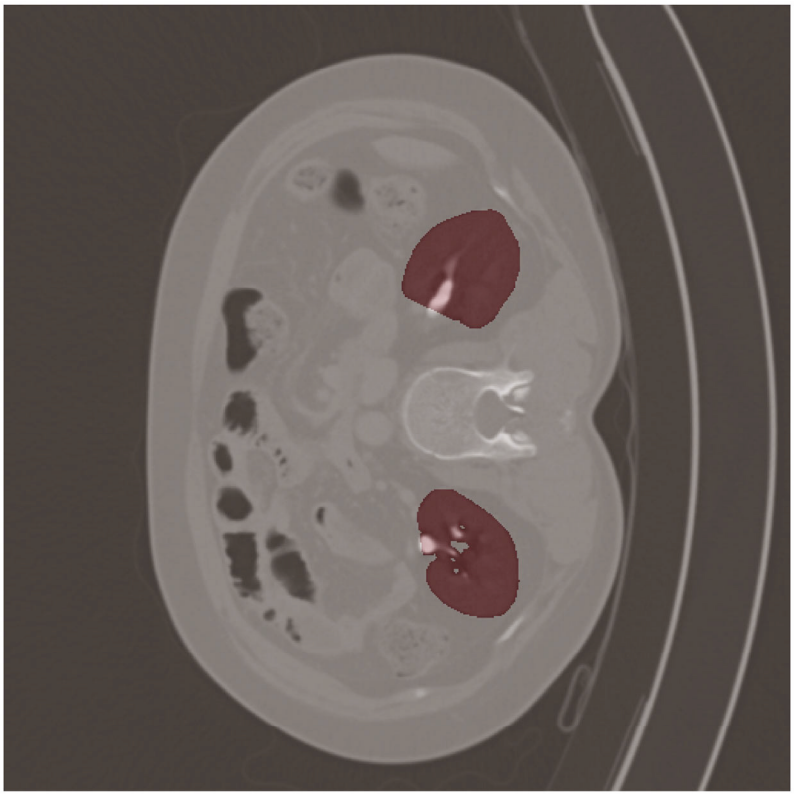


(**A**)

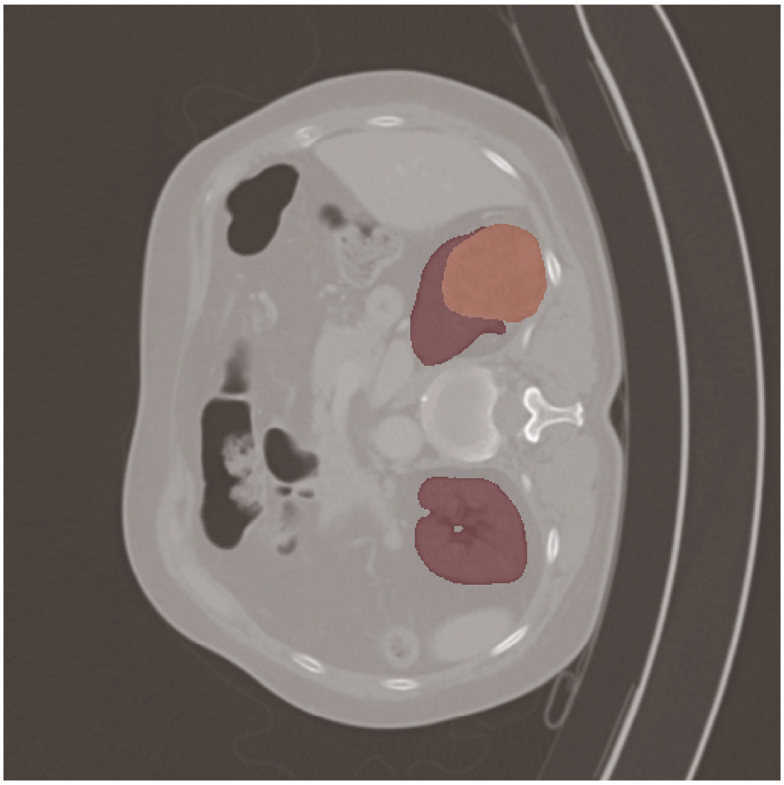


(**B**)

Figure 2. *Cont.*



(**C**)



(**D**)

Figure 2. Region of interest (ROI) masks for radiomic feature extraction. (**A**) Non-contrast phase CT scans overlaid with kidney ROI mask. (**B**) Arterial phase CT scans overlaid with kidney ROI mask. (**C**) Delayed phase CT scans overlaid with kidney ROI mask. (**D**) Portal phase CT scans overlaid with kidney ROI mask.

2.4. Feature Selection and Dimensionality Reduction

To ensure that the hold-out test set remained independent and unbiased for evaluating the final model's performance, we utilized several feature selection methods and a dimensionality reduction method solely on the training set. In the feature selection process, ANOVA F-test with FDR correction using the Benjamin–Hochberg procedure ($P_{fdr} = 0.05$) was performed on 1288 z-score normalized radiomic features derived from each CT phase. This selected 846, 773, 660, and 972 features from the arterial, delayed, non-contrast, and portal phase scans, respectively. Principal Component Analysis (PCA) was then performed on those selected radiomic features to reduce dimensionality. In this study, we aimed to compare the predictive performance of machine learning models with respect to the number of principal components used, a hyperparameter typically determined by researchers using rule-of-thumb approaches [27,28]. We set the number of principal components to values corresponding to 10% and 20% of the count of features selected via the F-test. When the number of principal components was set to 10% of the selected feature count, the final feature counts were 84, 77, 66, and 97 for arterial, delayed, non-contrast, and portal phase scans, respectively. When set to 20%, the final feature counts were 169, 154, 132, and 194 for the corresponding phases. The PCA results from the training set indicated that setting the number of principal components to 10% and 20% of the count of selected features was appropriate for representing radiomic features selected via the F-test while effectively reducing the dimensionality of the feature space (with the accumulated variance of components were almost 100%) (Figures S1 and S2). During development of the machine learning models, we assessed the evaluation performance of machine learning models in relation to the number of principal components used. We standardized these principal components (i.e., PCA whitening) for further analyses. Additionally, we employed the Sequential Feature Selector (SFS) on these standardized principal components to identify those that could optimize model evaluation performance on the validation set. The SFS method systematically chose features through a sequential removal process, based on their impact on model performance in the validation set.

2.5. Reproducibility and Ethical Considerations

All processes from feature extraction to model evaluation adhered to Checklist for Artificial Intelligence in Medical Imaging (CLAIM) standards to ensure reproducibility and ethical compliance [25]. Our codes are publicly available for validation and further research at GitHub Repository.

3. Results

3.1. Demographic Characteristics of Patients

In total, 499 patients were included in the development of our multi-phase CT radiomics machine learning model for renal cell tumor subtype classification. Participants with either single-phase or two-phase CT scans were included solely in the hold-out test set. Meanwhile, those with multi-phase CT scans were divided using a stratified approach, with 75% allocated to the training dataset and the remaining 25% reserved for the hold-out test dataset ($N_{training} = 289$, $N_{test} = 210$). There were no significant statistical differences in demographic or pathological characteristics between the training and hold-out test datasets ($p > 0.05$) (Table 1).

3.2. Multi-Phase CT Radiomic Features for Renal Cell Tumor Subtype Classification

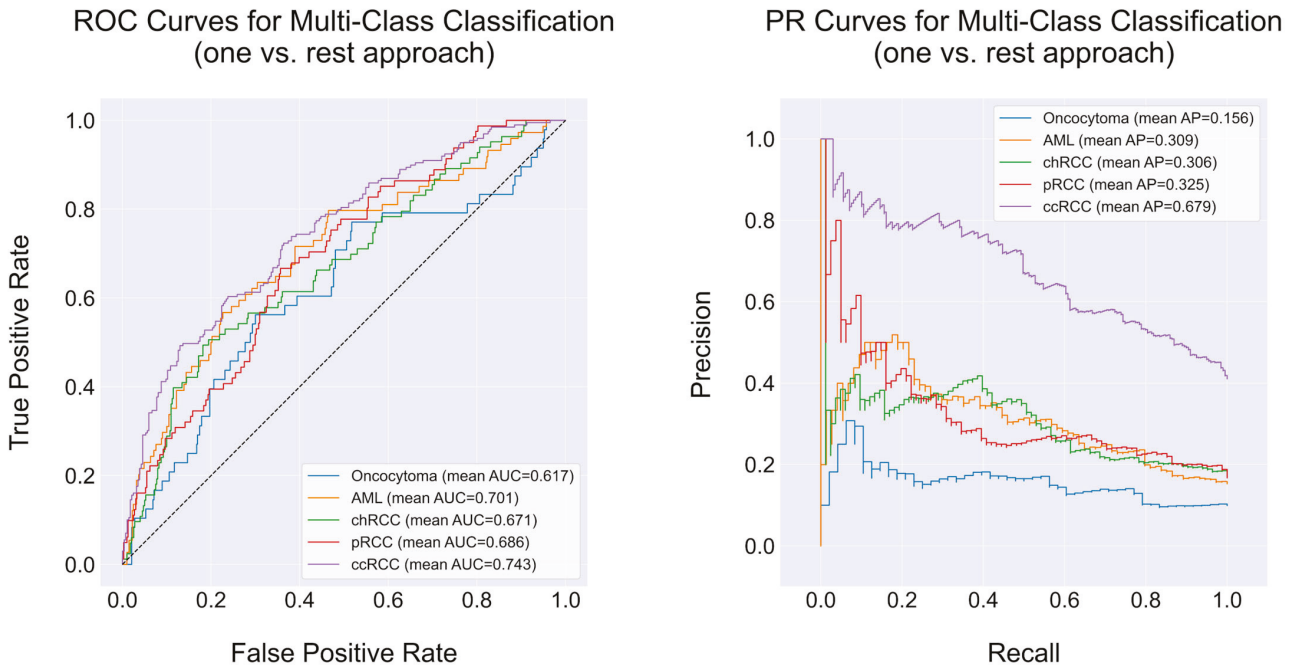
We compared the classification performance on a hold-out test set across various machine learning models with multiple evaluation metrics. The XGBoost model, trained on 20% of principal components from selected radiomic features ($XGBoost_{20\%}$), outper-

formed other algorithms in classifying renal cell tumor subtypes, achieving the highest average accuracy (ACC: mean = 0.475, standard error = 0.011), Area Under Receiving Operating Characteristic Curves (AU-ROC: mean = 0.744, standard error = 0.004), Area Under the Precision and Recall Curve (AU-PRC: mean = 0.474, standard error = 0.006), and F1 score (mean = 0.453, standard error = 0.007). Similarly, XGBoost trained on 10% of principal components from selected radiomic features ($XGBoost_{10\%}$) also performed well, achieving the highest performance evaluation metrics across ACC (mean = 0.453, standard error = 0.007), AU-ROC (mean = 0.744, standard error = 0.004), AU-PRC (mean = 0.474, standard error = 0.006), and F1 (mean = 0.453, standard error = 0.007). However, $XGBoost_{20\%}$ showed higher overall metrics than $XGBoost_{10\%}$. These results are detailed in Table 2.

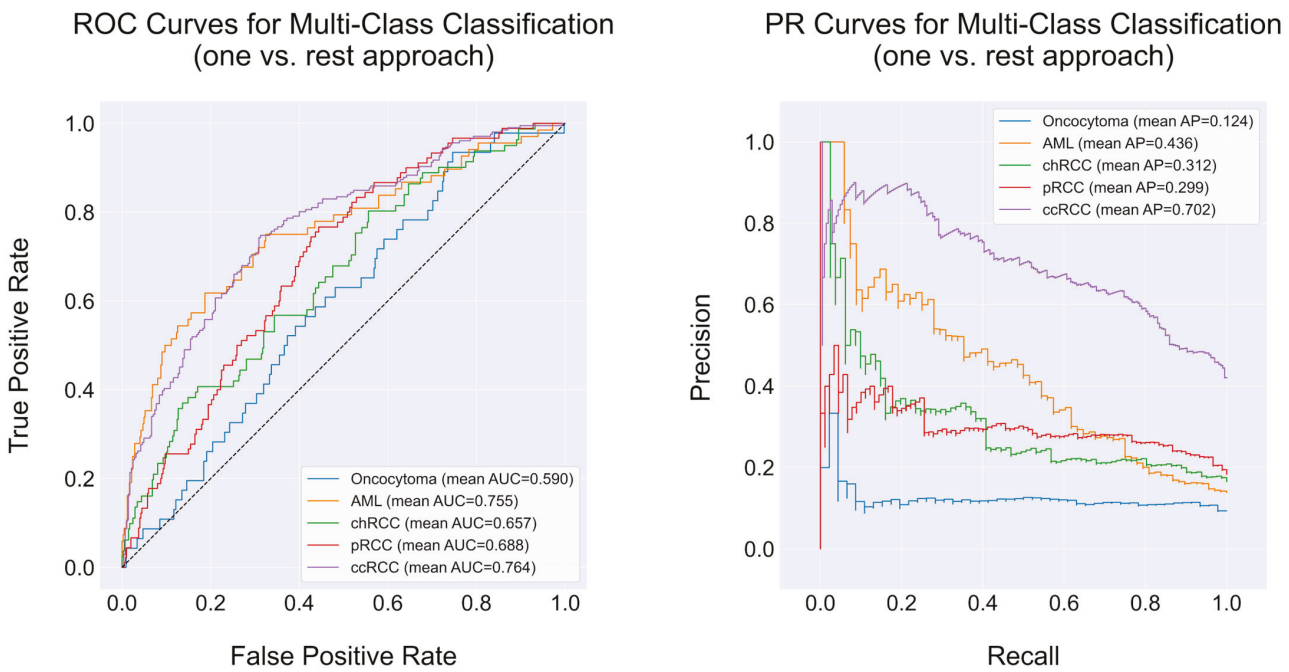
Of note, although the evaluation performance in classifying a specific subtype with other subtypes varied according to the machine learning model and the number of principal components used, XGBoost showed the best performance in classifying clear cell renal cell carcinoma against other subtypes when using either 10% (AU-PRC of $XGBoost_{10\%}$: mean = 0.673, standard error = 0.008; AU-ROC of $XGBoost_{10\%}$: mean = 0.738, standard error = 0.007) or 20% (AU-PRC of $XGBoost_{20\%}$: mean = 0.686, standard error = 0.012; AU-ROC of $XGBoost_{20\%}$: mean = 0.750, standard error = 0.011) of principal components for training. For angiomyolipoma (AML), Random Forest (10% principal components) trained with 10% of principal components achieved the highest AU-ROC in classifying AML against other subtypes (mean = 0.769, standard error = 0.009). However, it showed lower AU-PRC in classifying renal angiomyolipoma compared to clear cell renal cell carcinoma when distinguishing these subtypes from others (mean = 0.366, standard error = 0.01). These results indicated that machine learning models were biased towards predicting clear cell renal cell carcinoma, which constitutes the majority of the dataset. The results are detailed in Table 2 and Figure 3.

3.3. Single-Phase CT Radiomic Features for Renal Cell Tumor Subtype Classification

In contrast with the results from the multi-phase models, the machine learning models trained on 10% of principal components from selected radiomic features generally outperformed machine learning models trained on 20% of principal components for multiple evaluation metrics. XGBoost, trained on radiomic features derived from arterial and portal phase CT scans, achieved the highest F1 scores (arterial phase: mean = 0.594, standard error = 0.037; portal phase: mean = 0.456, standard error = 0.015) and ACC (arterial phase: mean = 0.594, standard error = 0.037; portal phase: mean = 0.456, standard error = 0.015) compared to other machine learning models trained on arterial and portal phase CT scans radiomic features. Random Forest trained on portal phase CT scan radiomic features showed the highest AU-PRC (portal phase: mean = 0.473, standard error = 0.015) and AU-ROC (portal phase: mean = 0.756, standard error = 0.008). Random Forest trained on radiomic features derived from delayed and non-contrastive phase CT scans also achieved the highest F1 (delayed phase: mean = 0.493, standard error = 0.015; non-contrastive phase: mean = 0.461, standard error = 0.011), AU-PRC (delayed phase: mean = 0.492, standard error = 0.031; non-contrastive phase: mean = 0.471, standard error = 0.013), ACC (delayed phase: mean = 0.493, standard error = 0.015; non-contrastive phase: mean = 0.461, standard error = 0.011), and AU-ROC (delayed phase: mean = 0.766, standard error = 0.016; non-contrastive phase: mean = 0.744, standard error = 0.008).



(**A**)



(**B**)

Figure 3. Renal cell tumor subtype classification performance of XGBoost. (**A**) Evaluation performance of machine learning models trained with 10% of the principal components derived from selected radiomic features in stratified 10-fold cross-validation. (**B**) Evaluation performance of machine learning models trained with 20% of the principal components derived from selected radiomic features in stratified 10-fold cross-validation. Abbreviations: AML = renal angiomyolipoma, chRCC = chromophobe renal cell carcinoma, pRCC = papillary renal cell carcinoma, ccRCC = clear cell renal cell carcinoma.

Similar to models trained on 10% of principal components from selected radiomic features, XGBoost and Random Forest, when trained on 20% of principal components, showed the highest evaluation performance among other machine learning models using the same proportion of principal components. However, their performance was slightly lower compared to models trained with only 10% of principal components. When trained

with 20% of principal components derived from arterial phase CT scan radiomic features, XGBoost achieved the best performance across multiple metrics: F1 (arterial phase: mean = 0.531, standard error = 0.002), AU-PRC (arterial phase: mean = 0.565, standard error = 0.014), and ACC (arterial phase: mean = 0.531, standard error = 0.02). XGBoost, when trained with 20% of principal components derived from delayed phase CT scans, achieved the highest AU-ROC (delayed phase: mean = 0.766, standard error = 0.001) compared to other machine learning models trained with the same components. Similarly, XGBoost trained with 20% of principal components derived from portal phase CT scans showed the highest F1 (portal phase: mean = 0.447, standard error = 0.019) and ACC (portal phase: mean = 0.447, standard error = 0.019), while Random Forest achieved the highest AU-PRC (portal phase: mean = 0.449, standard error = 0.014) and AU-ROC (portal phase: mean = 0.735, standard error = 0.007) using the same components. Additionally, Random Forest achieved the highest performance in all metrics when trained with 20% of principal components derived from delayed and non-contrastive phase CT scans: F1 (delayed phase: mean = 0.497, standard error = 0.016; non-contrastive phase: mean = 0.45, standard error = 0.006), AU-PRC (delayed phase: mean = 0.529, standard error = 0.02; non-contrastive phase: mean = 0.433, standard error = 0.009), and ACC (delayed phase: mean = 0.497, standard error = 0.016; non-contrastive phase: mean = 0.45, standard error = 0.006). In the case of AU-ROC, Random Forest, when trained with 20% of principal components derived from either the arterial or the non-contrastive phase, showed the best performance (arterial phase: mean = 0.797, standard error = 0.018; non-contrastive phase: mean = 0.716, standard error = 0.007). The results from single-phase CT radiomic feature-based machine learning models are detailed in Table 3.

In subtype-specific performance, single-phase CT radiomic feature-based machine learning models also performed well on classifying clear cell renal cell carcinoma, similar to the results from multi-phase CT radiomics feature-based machine learning models. Notably, XGBoost, trained with 10% of principal components derived from selected radiomics features extracted from arterial phase CT scans, achieved the highest performance in classifying clear cell renal cell carcinoma (AU-PRC: mean = 0.757, standard error = 0.033) and renal angiomyolipoma (AU-ROC: mean = 0.824, standard error = 0.023). These results outperformed other single-phase CT radiomics feature-based machine learning models trained with 20% of principal components (Table 4A–D; Supplementary Figures S3–S6).

3.4. Comparison of Multi-Phase and Single-Phase Models

In summary, performance of machine learning models varied depending on which phase of the CT scan was used for training and which evaluation metric was applied. However, single-phase CT radiomics feature-based models generally outperformed multi-phase models. XGBoost and Random Forest consistently showed better performance than linear SVM and rbf SVM. For single-phase CT scans, machine learning models using 10% principal components derived from selected radiomic features yielded better results. In contrast, for multi-phase scans, models using 20% principal components derived from selected radiomic features performed better. Of note, regardless of the proportion of principal components used, models trained with components derived from single-phase CT scans consistently outperformed those trained with components from multi-phase CT scans. The superior performance of single-phase models, particularly those based on arterial phase scans, suggests that these may be sufficient for accurate renal cell tumor subtype classification, potentially reducing the need for multi-phase CT imaging in clinical practice (Table 4A–D; Supplementary Figures S3–S6).

Table 2. Model performance of multi-phase CT radiomic features-based machine learning models in renal cell tumor subtype classification. We compared the model performance of four different types of machine learning algorithms in classifying the subtype of malignant kidney cancer. The average and standard error over test performance of each model in 10-fold cross-validation on a hold-out test set were calculated. (**A**,**C**) show overall evaluation performance of machine learning algorithms trained with 10% and 20% of principal components, respectively. (**B**,**D**) show the performance of machine learning algorithms in classifying renal cell tumor subtype via a one vs. rest approach, trained with 10% and 20% of principal components, respectively. Abbreviations: AML = renal angiomyolipoma, chRCC = chromophobe renal cell carcinoma, pRCC = papillary renal cell carcinoma, ccRCC = clear cell renal cell carcinoma. Values following signs indicate standard error.

(A)

	F1	**AU-PRC**	**ACC**	**AU-ROC**
Linear SVM	0.423 ± 0.008	0.393 ± 0.016	0.423 ± 0.008	0.703 ± 0.009
Rbf SVM	0.433 ± 0.017	0.408 ± 0.011	0.433 ± 0.017	0.722 ± 0.006
XGBoost	0.453 ± 0.007	0.474 ± 0.006	0.453 ± 0.007	0.744 ± 0.004
Random Forest	0.444 ± 0.009	0.443 ± 0.006	0.444 ± 0.009	0.742 ± 0.003

(B)

	AU-PRC					**AU-ROC**				
	Oncocytoma vs. Rest	**AML vs. Rest**	**chRCC vs. Rest**	**pRCC vs. Rest**	**ccRCC vs. Rest**	**Oncocytoma vs. Rest**	**AML vs. Rest**	**chRCC vs. Rest**	**pRCC vs. Rest**	**ccRCC vs. Rest**
Linear SVM	0.096 ± 0.01	0.285 ± 0.039	0.257 ± 0.044	0.246 ± 0.022	0.613 ± 0.019	0.511 ± 0.042	0.688 ± 0.03	0.605 ± 0.037	0.607 ± 0.023	0.697 ± 0.011
Rbf SVM	0.101 ± 0.014	0.292 ± 0.017	0.242 ± 0.013	0.276 ± 0.02	0.570 ± 0.023	0.488 ± 0.042	0.707 ± 0.017	0.618 ± 0.01	0.648 ± 0.022	0.663 ± 0.015
XGBoost	0.151 ± 0.018	0.31 ± 0.009	0.3 ± 0.015	0.318 ± 0.015	0.673 ± 0.08	0.606 ± 0.021	0.698 ± 0.009	0.666 ± 0.007	0.681 ± 0.014	0.738 ± 0.007
Random Forest	0.142 ± 0021	0.366 ± 0.01	0.259 ± 0.022	0.298 ± 0.017	0.617 ± 0.011	0.616 ± 0.017	0.769 ± 0.009	0.607 ± 0.014	0.686 ± 0.016	0.695 ± 0.01

(C)

	F1	**AU-PRC**	**ACC**	**AU-ROC**
Linear SVM	0.423 ± 0.011	0.396 ± 0.019	0.423 ± 0.011	0.716 ± 0.001
Rbf SVM	0.424 ± 0.0	0.322 ± 0.001	0.424 ± 0.0	0.66 ± 0.004
XGBoost	0.475 ± 0.011	0.495 ± 0.009	0.475 ± 0.011	0.751 ± 0.006
Random Forest	0.433 ± 0.005	0.47 ± 0.011	0.433 ± 0.005	0.75 ± 0.003

Table 2. *Cont.*

(D)										
	AU-PRC					AU-ROC				
	Oncocytoma vs. Rest	**AML vs. Rest**	**chRCC vs. Rest**	**pRCC vs. Rest**	**ccRCC vs. Rest**	**Oncocytoma vs. Rest**	**AML vs. Rest**	**chRCC vs. Rest**	**pRCC vs. Rest**	**ccRCC vs. Rest**
Linear SVM	0.141 ± 0.015	0.278 ± 0.022	0.230 ± 0.032	0.288 ± 0.020	0.598 ± 0.04	0.631 ± 0.027	0.734 ± 0.013	0.574 ± 0.042	0.658 ± 0.017	0.673 ± 0.026
Rbf SVM	0.096 ± 0.0	0.148 ± 0.0	0.164 ± 0.0	0.168 ± 0.0	0.424 ± 0.0	0.5 ± 0.0	0.5 ± 0.0	0.5 ± 0.0	0.5 ± 0.0	0.5 ± 0.0
XGBoost	0.124 ± 0.012	0.410 ± 0.029	0.289 ± 0.017	0.294 ± 0.017	0.686 ± 0.012	0.578 ± 0.029	0.748 ± 0.006	0.646 ± 0.013	0.674 ± 0.02	0.750 ± 0.011
Random Forest	0.183 ± 0.033	0.33 ± 0.015	0.258 ± 0.012	0.316 ± 0.017	0.677 ± 0.020	0.645 ± 0.026	0.743 ± 0.014	0.672 ± 0.016	0.672 ± 0.018	0.742 ± 0.013

Table 3. The model performance of single-phase CT radiomic feature-based machine learning models in renal cell tumor subtype classification. (**A**) shows the number of CT scans used for training and evaluating machine learning algorithms. (**B**,**C**) show the evaluation performance of machine learning algorithms trained with 10% and 20% of principal components, respectively. We compared the model performance of four difference types of machine learning algorithms in classifying renal cell tumor subtypes with regards to the number of principal components used for training models. We evaluated the model performance of each model in 10-fold cross-validation on a hold-out test set. Sample size indicates the number of labeled images used for training and evaluating machine learning algorithms.

(A)								
	Arterial Phase		**Delayed Phase**		**Non-Contrast Phase**		**Portal Phase**	
	Training Dataset	**Test Dataset**	**Training Dataset**	**Test Dataset**	**Training Dataset**	**Test Dataset**	**Training Dataset**	**Test Dataset**
Renal cell tumor subtype	243	75	257	83	280	139	253	188
oncocytoma (%)	22 (9%)	6 (8%)	23 (9%)	8 (10%)	25 (9%)	13 (10%)	20 (8%)	21 (11%)
angiomyolipoma (%)	40 (16%)	10 (13%)	42 (16%)	10 (12%)	44 (16%)	20 (14%)	42 (17%)	27 (14%)
chromophobe renal cell carcinoma (%)	43 (18%)	13 (17%)	42 (16%)	15 (18%)	48 (17%)	25 (18%)	40 (16%)	28 (15%)
papillary renal cell carcinoma (%)	36 (15%)	14 (19%)	43 (17%)	13 (16%)	48 (17%)	24 (17%)	46 (18%)	37 (20%)
clear cell renal cell carcinoma (%)	102 (42%)	32 (43%)	107 (42%)	37 (44%)	115 (41%)	57 (41%)	105 (41%)	75 (40%)

Table 3. *Cont.*

(B)																
	F1 score				**AU-PRC**				**ACC**				**AU-ROC**			
CT Phase	**Arterial**	**Delayed**	**Non-contrast**	**Portal**	**Arterial**	**Delayed**	**Non-contrast**	**Portal**	**Arterial**	**Delayed**	**Non-contrast**	**Portal**	**Arterial**	**Delayed**	**Non-contrast**	**Portal**
Linear SVM	0.421 ± 0.036	0.378 ± 0.032	0.397 ± 0.02	0.371 ± 0.011	0.42 ± 0.026	0.368 ± 0.032	0.345 ± 0.018	0.341 ± 0.019	0.421 ± 0.036	0.378 ± 0.032	0.397 ± 0.02	0.371 ± 0.011	0.719 ± 0.022	0.676 ± 0.024	0.671 ± 0.019	0.659 ± 0.012
Rbf SVM	0.461 ± 0.028	0.477 ± 0.009	0.425 ± 0.006	0.386 ± 0.004	0.48 ± 0.051	0.416 ± 0.031	0.303 ± 0.024	0.525 ± 0.05	0.461 ± 0.028	0.477 ± 0.09	0.425 ± 0.006	0.386 ± 0.04	0.756 ± 0.027	0.725 ± 0.021	0.67 ± 0.023	0.629 ± 0.018
XGBoost	0.594 ± 0.037	0.479 ± 0.009	0.439 ± 0.015	0.456 ± 0.015	0.608 ± 0.022	0.442 ± 0.18	0.419 ± 0.014	0.463 ± 0.012	0.594 ± 0.037	0.479 ± 0.009	0.439 ± 0.015	0.456 ± 0.015	0.83 ± 0.012	0.728 ± 0.013	0.712 ± 0.006	0.739 ± 0.007
Random Forest	0.478 ± 0.013	0.493 ± 0.015	0.461 ± 0.011	0.441 ± 0.014	0.511 ± 0.016	0.492 ± 0.31	0.471 ± 0.013	0.473 ± 0.015	0.478 ± 0.013	0.493 ± 0.015	0.461 ± 0.011	0.441 ± 0.014	0.778 ± 0.013	0.766 ± 0.016	0.744 ± 0.008	0.756 ± 0.008

(C)																
	F1 score				**AU-PRC**				**ACC**				**AU-ROC**			
CT Phase	**Arterial**	**Delayed**	**Non-contrast**	**Portal**	**Arterial**	**Delayed**	**Non-contrast**	**Portal**	**Arterial**	**Delayed**	**Non-contrast**	**Portal**	**Arterial**	**Delayed**	**Non-contrast**	**Portal**
Linear SVM	0.396 ± 0.028	0.385 ± 0.027	0.414 ± 0.022	0.297 ± 0.045	0.353 ± 0.048	0.321 ± 0.036	0.354 ± 0.031	0.267 ± 0.031	0.396 ± 0.028	0.385 ± 0.027	0.414 ± 0.022	0.297 ± 0.045	0.669 ± 0.036	0.627 ± 0.032	0.688 ± 0.022	0.581 ± 0.026
Rbf SVM	0.437 ± 0.0	0.462 ± 0.0	0.413 ± 0.0	0.382 ± 0.0	0.337 ± 0.01	0.337 ± 0.016	0.317 ± 0.002	0.296 ± 0.007	0.437 ± 0.0	0.462 ± 0.0	0.413 ± 0.0	0.382 ± 0.0	0.677 ± 0.006	0.67 ± 0.008	0.665 ± 0.005	0.633 ± 0.007
XGBoost	0.531 ± 0.02	0.466 ± 0.029	0.354 ± 0.024	0.447 ± 0.019	0.565 ± 0.014	0.507 ± 0.021	0.382 ± 0.022	0.429 ± 0.013	0.531 ± 0.002	0.466 ± 0.029	0.354 ± 0.024	0.447 ± 0.019	0793 ± 0.005	0.766 ± 0.001	0.686 ± 0.023	0.732 ± 0.008
Random Forest	0.504 ± 0.022	0.47 ± 0.016	0.45 ± 0.006	0.427 ± 0.013	0.52 ± 0.026	0.529 ± 0.02	0.433 ± 0.009	0.449 ± 0.014	0.504 ± 0.022	0.497 ± 0.016	0.45 ± 0.006	0.427 ± 0.013	0.797 ± 0.018	0.73 ± 0.014	0.16 ± 0.007	0.735 ± 0.007

Table 4. The model performance of single-phase CT radiomics feature-based machine learning models in renal cell tumor subtype classification (**A**,**B**) show the AU-PRC and AU-ROC, respectively, of singe-phase CT radiomics feature-based machine learning models trained with 10% of principal components in classifying a specific subtype against other subtypes. (**C**,**D**) show the AU-PRC and AU-ROC, respectively, of models trained with 20% of principal components in the same task.

(A)

	Arterial					Delayed					Non-Contrast					Portal				
CT Phase	**Onco-cytoma**	**AML**	**chRCC**	**pRCC**	**ccRCC**	**Onco-cytoma**	**AML**	**chRCC**	**pRCC**	**ccRCC**	**Onco-cytoma**	**AML**	**chRCC**	**pRCC**	**ccRCC**	**Onco-cytoma**	**AML**	**chRCC**	**pRCC**	**ccRCC**
Linear SVM	0.147 ± 0.068	0.290 ± 0.077	0.280 ± 0.028	0.298 ± 0.078	0.656 ± 0.065	0.151 ± 0.056	0.205 ± 0.041	0.310 ± 0.079	0.260 ± 0.059	0.564 ± 0.037	0.109 ± 0.016	0.214 ± 0.025	0.202 ± 0.051	0.206 ± 0.036	0.503 ± 0.030	0.141 ± 0.016	0.194 ± 0.027	0.217 ± 0.019	0.275 ± 0.028	0.550 ± 0.054
Rbf SVM	0.133 ± 0.038	0.382 ± 0.135	0.285 ± 0.041	0.598 ± 0.119	0.638 ± 0.085	0.128 ± 0.059	0.241 ± 0.044	0.268 ± 0.1	0.421 ± 0.095	0.547 ± 0.07	0.084 ± 0.008	0.340 ± 0.105	0.170 ± 0.013	0.168 ± 0.038	0.463 ± 0.034	0.117 ± 0.014	0.190 ± 0.060	0.148 ± 0.008	0.200 ± 0.025	0.384 ± 0.048
XGBoost	0.089 ± 0.017	0.605 ± 0.076	0.590 ± 0.063	0.393 ± 0.1	0.757 ± 0.033	0.105 ± 0.01	0.420 ± 0.028	0.351 ± 0.038	0.346 ± 0.049	0.586 ± 0.023	0.111 ± 0.015	0.401 ± 0.022	0.320 ± 0.033	0.196 ± 0.016	0.556 ± 0.024	0.189 ± 0.033	0.411 ± 0.016	0.273 ± 0.035	0.333 ± 0.31	0.682 ± 0.02
Random Forest	0.163 ± 0.026	0.384 ± 0.035	0.265 ± 0.049	0.413 ± 0.052	0.711 ± 0.026	0.158 ± 0.023	0.356 ± 0.021	0.277 ± 0.038	0.344 ± 0.093	0.712 ± 0.046	0.149 ± 0.029	0.499 ± 0.023	0.239 ± 0.027	0.238 ± 0.019	0.637 ± 0.039	0.214 ± 0.032	0.479 ± 0.052	0.271 ± 0.027	0.295 ± 0.016	0.643 ± 0.02

(B)

	Arterial					Delayed					Non-Contrast					Portal				
CT Phase	**Onco-cytoma**	**AML**	**chRCC**	**pRCC**	**ccRCC**	**Onco-cytoma**	**AML**	**chRCC**	**pRCC**	**ccRCC**	**Onco-cytoma**	**AML**	**chRCC**	**pRCC**	**ccRCC**	**Onco-cytoma**	**AML**	**chRCC**	**pRCC**	**ccRCC**
Linear SVM	0.625 ± 0.149	0.681 ± 0.073	0.572 ± 0.053	0.687 ± 0.056	0.663 ± 0.049	0.584 ± 0.094	0.631 ± 0.048	0.561 ± 0.084	0.645 ± 0.067	0.595 ± 0.042	0.529 ± 0.056	0.685 ± 0.049	0.465 ± 0.075	0.492 ± 0.049	0.595 ± 0.022	0.561 ± 0.040	0.639 ± 0.047	0.537 ± 0.053	0.584 ± 0.046	0.655 ± 0.044
Rbf SVM	0.133 ± 0.038	0.382 ± 0.135	0.285 ± 0.041	0.598 ± 0.119	0.638 ± 0.084	0.444 ± 0.100	0.570 ± 0.079	0.547 ± 0.117	0.699 ± 0.113	0.635 ± 0.082	0.458 ± 0.064	0.580 ± 0.139	0.415 ± 0.065	0.452 ± 0.090	0.524 ± 0.047	0.504 ± 0.055	0.424 ± 0.133	0.459 ± 0.018	0.442 ± 0.018	0.500 ± 0.074
XGBoost	0.507 ± 0.05	0.843 ± 0.028	0.775 ± 0.032	0.781 ± 0.04	0.824 ± 0.023	0.500 ± 0.042	0.715 ± 0.019	0.707 ± 0.029	0.616 ± 0.041	0.675 ± 0.011	0.513 ± 0.049	0.693 ± 0.02	0.668 ± 0.026	0.522 ± 0.014	0.645 ± 0.018	0.580 ± 0.033	0.785 ± 0.016	0.659 ± 0.024	0.667 ± 0.015	0.761 ± 0.013
Random Forest	0.664 ± 0.052	0.755 ± 0.045	0.636 ± 0.048	0.769 ± 0.033	0.775 ± 0.026	0.644 ± 0.059	0.749 ± 0.025	0.659 ± 0.037	0.637 ± 0.113	0.785 ± 0.037	0.603 ± 0.056	0.780 ± 0.026	0.585 ± 0.019	0.577 ± 0.038	0.706 ± 0.025	0.669 ± 0.033	0.806 ± 0.017	0.651 ± 0.022	0.656 ± 0.023	0.731 ± 0.017

Table 4. *Cont.*

(C)

	Arterial					Delayed					Non-Contrast					Portal				
CT Phase	**Onco-cytoma**	**AML**	**chRCC**	**pRCC**	**ccRCC**	**Onco-cytoma**	**AML**	**chRCC**	**pRCC**	**ccRCC**	**Onco-cytoma**	**AML**	**chRCC**	**pRCC**	**ccRCC**	**Onco-cytoma**	**AML**	**chRCC**	**pRCC**	**ccRCC**
Linear SVM	0.106 ± 0.06	0.217 ± 0.085	0.188 ± 0.075	0.207 ± 0.063	0.583 ± 0.059	0.111 ± 0.025	0.189 ± 0.082	0.196 ± 0.049	0.218 ± 0.054	0.469 ± 0.061	0.129 ± 0.030	0.331 ± 0.055	0.207 ± 0.04	0.199 ± 0.035	0.512 ± 0.042	0.111 ± 0.017	0.163 ± 0.029	0.17 ± 0.018	0.217 ± 0.025	0.408 ± 0.04
Rbf SVM	0.185 ± 0.072	0.215 ± 0.046	0.165 ± 0.007	0.168 ± 0.015	0.420 ± 0.019	0.18 ± 0.049	0.212 ± 0.097	0.162 ± 0.024	0.181 ± 0.043	0.412 ± 0.031	0.091 ± 0.0	0.147 ± 0.0	0.175 ± 0.0	0.175 ± 0.0	0.413 ± 0.0	0.111 ± 0.009	0.248 ± 0.034	0.154 ± 0.009	0.205 ± 0.024	0.361 ± 0.02
XGBoost	0.177 ± 0.056	0.584 ± 0.037	0.261 ± 0.023	0.393 ± 0.067	0.726 ± 0.026	0.184 ± 0.077	0.399 ± 0.048	0.239 ± 0.026	0.34 ± 0.036	0.714 ± 0.032	0.119 ± 0.025	0.443 ± 0.048	0.22 ± 0.023	0.222 ± 0.025	0.549 ± 0.029	0.147 ± 0.016	0.452 ± 0.038	0.318 ± 0.045	0.394 ± 0.027	0.552 ± 0.025
Random Forest	0.249 ± 0.072	0.451 ± 0.028	0.326 ± 0.079	0.368 ± 0.076	0.671 ± 0.041	0.136 ± 0.02	0.533 ± 0.072	0.252 ± 0.052	0.388 ± 0.051	0.741 ± 0.032	0.141 ± 0.025	0.395 ± 0.021	0.212 ± 0.04	0.221 ± 0.032	0.627 ± 0.022	0.123 ± 0.021	0.428 ± 0.044	0.261 ± 0.027	0.318 ± 0.031	0.622 ± 0.034

(D)

	Arterial					Delayed					Non-Contrast					Portal				
CT Phase	**Onco-cytoma**	**AML**	**chRCC**	**pRCC**	**ccRCC**	**Onco-cytoma**	**AML**	**chRCC**	**pRCC**	**ccRCC**	**Onco-cytoma**	**AML**	**chRCC**	**pRCC**	**ccRCC**	**Onco-cytoma**	**AML**	**chRCC**	**pRCC**	**ccRCC**
Linear SVM	0.509 ± 0.135	0.566 ± 0.135	0.478 ± 0.132	0.537 ± 0.139	0.580 ± 0.061	0.466 ± 0.094	0.514 ± 0.161	0.481 ± 0.079	0.475 ± 0.094	0.5 ± 0.066	0.598 ± 0.068	0.744 ± 0.036	0.53 ± 0.044	0.488 ± 0.062	0.605 ± 0.034	0.46 ± 0.058	0.495 ± 0.079	0.484 ± 0.043	0.493 ± 0.041	0.491 ± 0.055
Rbf SVM	0.497 ± 0.085	0.56 ± 0.105	0.457 ± 0.045	0.441 ± 0.039	0.43 ± 0.049	0.492 ± 0.061	0.529 ± 0.115	0.429 ± 0.066	0.433 ± 0.082	0.387 ± 0.072	0.5 ± 0.0	0.5 ± 0.	0.5 ± 0.	0.5 ± 0.	0.5 ± 0.	0.483 ± 0.041	0.587 ± 0.042	0.412 ± 0.05	0.518 ± 0.031	0.409 ± 0.045
XGBoost	0.489 ± 0.039	0.805 ± 0.028	0.671 ± 0.027	0.707 ± 0.029	0.751 ± 0.015	0.594 ± 0.054	0.762 ± 0.019	0.630 ± 0.016	0.766 ± 0.029	0.731 ± 0.023	0.503 ± 0.045	0.747 ± 0.021	0.545 ± 0.022	0.607 ± 0.041	0.632 ± 0.025	0.560 ± 0.028	0.796 ± 0.019	0.678 ± 0.031	0.726 ± 0.021	0.648 ± 0.031
Random Forest	0.648 ± 0.074	0.78 ± 0.028	0.717 ± 0.077	0.717 ± 0.077	0.716 ± 0.024	0.541 ± 0.062	0.799 ± 0.033	0.609 ± 0.045	0.712 ± 0.061	0.777 ± 0.03	0.581 ± 0.037	0.757 ± 0.016	0.513 ± 0031	0.554 ± 0.048	0.723 ± 0.028	0.567 ± 0.049	0.804 ± 0.015	0.658 ± 0.036	0.669 ± 0.025	0.706 ± 0.019

4. Discussion

This study represents a significant step forward in the application of radiomics and machine learning for the classification of renal cell tumor subtypes. By leveraging multi-phase CT scans and advanced machine learning techniques, we have demonstrated a novel approach that has the potential to transform the diagnostic landscape in renal oncology [29–31]. Our findings address several critical challenges in the current management of renal tumors. The XGBoost model, trained on 20% of principal components from multi-phase CT radiomic features (XGBoost20%), demonstrated the best overall performance in classifying renal cell tumor subtypes.

Accurate preoperative classification of renal tumor subtypes is a significant clinical challenge, impacting treatment planning and patient outcomes [32,33]. Our radiomics-based approach offers a non-invasive and cross-sectional method to enhance diagnostic accuracy, potentially reducing the need for invasive biopsies and their associated risks [34,35].

The superior performance in identifying clear cell renal cell carcinoma (ccRCC) is particularly noteworthy, given that this subtype is associated with the worst prognosis among RCC subtypes [32]. Early and accurate identification could lead to more aggressive treatment strategies and improved patient outcomes. Our multi-phase CT scans radiomic feature-based machine learning models achieved an AU-PRC of 0.69 and AU-ROC of 0.75 in the independent testing cohort, which represents significant prediction performance.

A key finding of our study is that single-phase CT scans, particularly those from the arterial phase, can achieve comparable or even superior performance to multi-phase models for certain subtypes. The XGBoost model trained on arterial phase CT scans achieved the highest performance in classifying clear cell renal cell carcinoma (AU-PRC = 0.757) and renal angiomyolipoma (AU-ROC = 0.824). This suggests the potential for reduced radiation exposure and shorter imaging protocols, aligning with the broader goals of minimizing patient risk and optimizing resource utilization in healthcare [31,34].

Interestingly, models trained on 10% of principal components generally outperformed those trained on 20% for single-phase CT scans. This suggests that a more parsimonious feature set may suffice for accurate classification, potentially reducing computational complexity and improving model interpretability [33]. This finding opens new avenues for optimizing feature selection in radiomics research.

Our study also highlights the varying levels of model performance across different renal tumor subtypes. While clear cell renal cell carcinoma was consistently well-classified, other subtypes proved more challenging. This pattern aligns with the known heterogeneity of renal tumors and underscores the need for subtype-specific approaches in radiomics research [32,34]. Furthermore, our study adds to the growing body of evidence supporting the utility of radiomics in oncology [30]. By demonstrating that quantitative imaging features can capture clinically relevant information not visible to the naked eye, we reinforce the potential of radiomics to serve as a powerful tool in precision medicine [33].

Despite the promising results, our study has several limitations that warrant discussion. The predominance of ccRCC cases (64.3% in the training cohort) reflects a common challenge in medical imaging studies but may have introduced bias in our models. Our use of single-institution data may limit the generalizability of our findings to different patient populations or imaging protocols. The retrospective nature of our study could have introduced selection bias, and prospective validation studies are necessary to confirm the clinical utility of our approach. We encountered challenges in differentiating among specific malignant subtypes, particularly in identifying oncocytomas (AUC = 0.57–0.69), highlighting the need for further refinement of our models and potentially the incorporation of additional data types. Lastly, while we used an independent hold-out test set,

our study lacks external validation on datasets from different institutions or populations, which is crucial for assessing the robustness and generalizability of our findings [33,34].

The challenges we encountered in differentiating among specific malignant subtypes, particularly in identifying oncocytomas (AUC = 0.57–0.69), mirror the difficulties faced in clinical practice. This alignment between computational and human challenges in tumor classification suggests that future improvements in AI models could have direct and significant impacts on clinical decision-making.

Looking ahead, our findings open up several exciting avenues for future research. The exploration of deep learning models, which can learn directly from image data without the need for hand-crafted features, represents a promising next step. Additionally, the integration of radiomics with other data types, such as genomics or clinical information, could lead to more comprehensive and accurate predictive models.

Our study has several significant limitations. First, the predictive accuracy of our research did not reach the commonly accepted threshold of 80%. However, we believe this limitation can be overcome by expanding the dataset, and we are actively collecting additional data to address this issue. Additionally, our study was conducted retrospectively, which means its efficacy has not yet been fully validated. We anticipate that conducting a prospective study based on these data will demonstrate its practical value.

Furthermore, the goal of this study was to identify recommended features using radiomics techniques and utilize feedback to enhance radiologists' predictive accuracy. However, due to the scope of this research, we were only able to confirm the model's ability to predict subtypes. We plan to include these broader aspects in future studies to fully realize the potential of this approach.

Ultimately, if we are able to validate our hypothesis step by step, and this predictive model achieves successful outcomes, it will allow us to determine the pathology of renal cancer through imaging studies alone, without the need for biopsy. This advancement would enable us to predict the optimal timing for surgery, prepare appropriate chemotherapy protocols, avoid unnecessary surgeries and their associated risks, and achieve significant economic savings.

5. Conclusions

In conclusion, our study not only demonstrates the potential of radiomics-based machine learning in renal tumor classification, but also contributes to the broader dialogue about the role of AI in medical imaging. The patterns we observed, particularly the efficacy of single-phase CT and the potential for feature set optimization, provide valuable insights for future research directions. As we continue to refine these techniques and address current limitations, the promise of more personalized, accurate, and efficient diagnosis and treatment planning in renal oncology comes increasingly within reach. This work represents a crucial step towards realizing the potential of precision medicine in the management of renal tumors, with far-reaching implications for patient care and outcomes.

Code Availability

The prediction of malignancy in renal tumors by using CT radiomics is found to be feasible. Based on this technology, it is expected that there will be future advances in the diagnosis of renal tumors. The relevant codes are freely available for reproducibility.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/diagnostics15111365/s1, Figure S1: Scree plots of principal components derived from PCA using 10% of selected radiomics features via F-test for each CT phase, Figure S2: Scree plots of principal components derived from PCA using 10% of selected radiomics features via F-test for each CT phase, Figure S3: Renal cell tumor subtype classification performance

of XGBoost trained with 10% of principal components derived from arterial phase CT scan radiomic features, Figure S4: Renal cell tumor subtype classification performance of XGBoost trained with 10% of principal components derived from delayed phase CT scan radiomic features, Figure S5: Renal cell tumor subtype classification performance of XGBoost trained with 10% of principal components derived from non-contrast phase CT scan radiomic features, Figure S6: Renal cell tumor subtype classification performance of XGBoost trained with 10% of principal components derived from portal phase CT scan radiomic features

Author Contributions: Conceptualization, S.B.; Methodology, H.W.; Software, H.W., S.-H.H. and M.H.C.; Formal analysis, S.B., H.B. and J.C.; Investigation, S.-H.H.; Resources, H.B. and J.C.; Data curation, H.B. and M.H.C.; Writing—original draft, S.B. and H.W.; Writing—review & editing, M.H.C.; Project administration, S.-H.H.; Funding acquisition, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Korea Medical Device Development Fund grant funded by the Korean government (Ministry of Science and ICT, Ministry of Trade, Industry and Energy, Ministry of Health & Welfare, Republic of Korea, Ministry of Food and Drug Safety) (Project Number: KMDF_PR_20200901_0096) and the Technology development Program (RS-2024-00460263) funded by the Ministry of SMEs and Startups(MSS, Korea). This work was also supported by Creative-Pioneering Researchers Program through Seoul National University (No. 200-20240057, 200-20240135) and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [NO. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)].

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of the Catholic University of Korea, Seoul St Mary's Hospital (Protocol code KC22RISI0753) on 29 October 2024.

Informed Consent Statement: Patient consent was waived due to retrospective study.

Data Availability Statement: The relevant code is freely available for reproducibility (https://github.com/Transconnectome/Kidney_Radiomics, accessed on 16 October 2023).

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

References

1. Moch, H.; Cubilla, A.L.; Humphrey, P.A.; Reuter, V.E.; Ulbright, T.M. The 2016 WHO classification of tumours of the urinary system and male genital organs—Part A: Renal, penile, and testicular tumours. *Eur. Urol.* **2016**, *70*, 93–105. [CrossRef] [PubMed]
2. American Cancer Society. Key Statistics About Kidney Cancer. Available online: https://www.cancer.org/cancer/types/kidney-cancer/about/key-statistics.html (accessed on 9 February 2025).
3. Capitanio, U.; Bensalah, K.; Bex, A.; Boorjian, S.A.; Bray, F.; Coleman, J.; Gore, J.L.; Sun, M.; Wood, C.; Russo, P. Epidemiology of Renal Cell Carcinoma. *Eur. Urol.* **2019**, *75*, 74–84. (In English) [CrossRef] [PubMed]
4. Kang, S.K.; Huang, W.C.; Pandharipande, P.V.; Chandarana, H. Solid renal masses: What the numbers tell us. *AJR Am. J. Roentgenol.* **2014**, *202*, 1196–1206. (In English) [CrossRef]
5. Prasad, S.R.; Dalrymple, N.C.; Surabhi, V.R. Cross-sectional imaging evaluation of renal masses. *Radiol. Clin. N. Am.* **2008**, *46*, 95–111. [CrossRef]
6. Young, J.R.; Margolis, D.; Sauk, S.; Pantuck, A.J.; Sayre, J.; Raman, S.S. Clear cell renal cell carcinoma: Discrimination from other renal cell carcinoma subtypes and oncocytoma at multiphasic multidetector CT. *Radiology* **2013**, *267*, 444–453. [CrossRef]
7. Bauman, T.M.; Potretzke, A.M.; Wright, A.J.; Knight, B.A.; Vetter, J.M.; Figenshau, R.S. Partial nephrectomy for presumed renal-cell carcinoma: Incidence, predictors, and perioperative outcomes of benign lesions. *J. Endourol.* **2017**, *31*, 412–417. [CrossRef]
8. Tanaka, T.; Huang, Y.; Marukawa, Y.; Tsuboi, Y.; Masaoka, Y.; Kojima, K.; Iguchi, T.; Hiraki, T.; Gobara, H.; Yanai, H. Differentiation of small ($\leq$4 cm) renal masses on multiphase contrast-enhanced CT by deep learning. *Am. J. Roentgenol.* **2020**, *214*, 605–612. [CrossRef]
9. De Fauw, J.; Ledsam, J.R.; Romera-Paredes, B.; Nikolov, S.; Tomasev, N.; Blackwell, S.; Askham, H.; Glorot, X.; O'Donoghue, B.; Visentin, D. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **2018**, *24*, 1342–1350. [CrossRef]

10. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef]
11. Mei, X.; Lee, H.-C.; Diao, K.-y.; Huang, M.; Lin, B.; Liu, C.; Xie, Z.; Ma, Y.; Robson, P.M.; Chung, M. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. *Nat. Med.* **2020**, *26*, 1224–1228. [CrossRef]
12. Ardila, D.; Kiraly, A.P.; Bharadwaj, S.; Choi, B.; Reicher, J.J.; Peng, L.; Tse, D.; Etemadi, M.; Ye, W.; Corrado, G. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **2019**, *25*, 954–961. [CrossRef] [PubMed]
13. Uhm, K.H.; Jung, S.W.; Choi, M.H.; Shin, H.K.; Yoo, J.I.; Oh, S.W.; Kim, J.Y.; Kim, H.G.; Lee, Y.J.; Youn, S.Y.; et al. Deep learning for end-to-end kidney cancer diagnosis on multi-phase abdominal computed tomography. *NPJ Precis. Oncol.* **2021**, *5*, 54. (In English) [CrossRef] [PubMed]
14. Zhao, W.; Jiang, D.; Queralta, J.P.; Westerlund, T. MSS U-Net: 3D segmentation of kidneys and tumors from CT images with a multi-scale supervised U-Net. *Inform. Med. Unlocked* **2020**, *19*, 100357. [CrossRef]
15. Kim, S.H.; Kim, C.S.; Kim, M.J.; Cho, J.Y.; Cho, S.H. Differentiation of clear cell renal cell carcinoma from other subtypes and fat-poor angiomyolipoma by use of quantitative enhancement measurement during three-phase MDCT. *Am. J. Roentgenol.* **2016**, *206*, W21–W28. [CrossRef]
16. Sun, X.-Y.; Feng, Q.-X.; Xu, X.; Zhang, J.; Zhu, F.-P.; Yang, Y.-H.; Zhang, Y.-D. Radiologic-radiomic machine learning models for differentiation of benign and malignant solid renal masses: Comparison with expert-level radiologists. *Am. J. Roentgenol.* **2020**, *214*, W44–W54. [CrossRef]
17. de Leon, A.D.; Pedrosa, I. Imaging and screening of kidney cancer. *Radiol. Clin.* **2017**, *55*, 1235–1250. [CrossRef]
18. Bang, S.; Wang, H.-H.; Kim, H.; Choi, M.H.; Cha, J.; Choi, Y.; Hong, S.-H. Development and Validation of a Prediction Model for Differentiation of Benign and Malignant Fat-Poor Renal Tumors Using CT Radiomics. *Appl. Sci.* **2023**, *13*, 11345. [CrossRef]
19. Oberai, A.; Varghese, B.; Cen, S.; Angelini, T.; Hwang, D.; Gill, I.; Aron, M.; Lau, C.; Duddalwar, V. Deep learning based classification of solid lipid-poor contrast enhancing renal masses using contrast enhanced CT. *Br. J. Radiol.* **2020**, *93*, 20200002. [CrossRef]
20. Kaur, R.; Juneja, M.; Mandal, A.K. Computer-aided diagnosis of renal lesions in CT images: A comprehensive survey and future prospects. *Comput. Electr. Eng.* **2019**, *77*, 423–434. [CrossRef]
21. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
22. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631.
23. LemaÃŽtre, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **2017**, *18*, 1–5.
24. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
25. Klontzas, M.E.; Gatti, A.A.; Tejani, A.S.; Kahn, C.E., Jr. AI reporting guidelines: How to select the best one for your research. *Radiol. Artif. Intell.* **2023**, *5*, e230055. [CrossRef] [PubMed]
26. Van Griethuysen, J.J.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.; Fillion-Robin, J.-C.; Pieper, S.; Aerts, H.J. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **2017**, *77*, e104–e107. [CrossRef]
27. van Timmeren, J.E.; Cester, D.; Tanadini-Lang, S.; Alkadhi, H.; Baessler, B. Radiomics in medical imaging-"how-to" guide and critical reflection. *Insights Imaging* **2020**, *11*, 91. (In English) [CrossRef]
28. Li, S.; Wang, K.; Hou, Z.; Yang, J.; Ren, W.; Gao, S.; Meng, F.; Wu, P.; Liu, B.; Liu, J.; et al. Use of Radiomics Combined With Machine Learning Method in the Recurrence Patterns After Intensity-Modulated Radiotherapy for Nasopharyngeal Carcinoma: A Preliminary Study. *Front. Oncol.* **2018**, *8*, 648. (In English) [CrossRef]
29. Shi, Y.; Zou, Y.; Liu, J.; Wang, Y.; Chen, Y.; Sun, F.; Yang, Z.; Cui, G.; Zhu, X.; Cui, X. Ultrasound-based radiomics XGBoost model to assess the risk of central cervical lymph node metastasis in patients with papillary thyroid carcinoma: Individual application of SHAP. *Front. Oncol.* **2022**, *12*, 897596. [CrossRef]
30. Mühlbauer, J.; Egen, L.; Kowalewski, K.-F.; Grilli, M.; Walach, M.T.; Westhoff, N.; Nuhn, P.; Laqua, F.C.; Baessler, B.; Kriegmair, M.C. Radiomics in renal cell carcinoma—A systematic review and meta-analysis. *Cancers* **2021**, *13*, 1348. [CrossRef]
31. Bertsimas, D.; Wiberg, H. Machine learning in oncology: Methods, applications, and challenges. *JCO Clin. Cancer Inform.* **2020**, *4*, CCI-20. [CrossRef]
32. Alhussaini, A.J.; Steele, J.D.; Jawli, A.; Nabi, G. Radiomics Machine Learning Analysis of Clear Cell Renal Cell Carcinoma for Tumour Grade Prediction Based on Intra-Tumoural Sub-Region Heterogeneity. *Cancers* **2024**, *16*, 1454. [CrossRef]

33. Lambin, P.; Leijenaar, R.T.; Deist, T.M.; Peerlings, J.; De Jong, E.E.; Van Timmeren, J.; Sanduleanu, S.; Larue, R.T.; Even, A.J.; Jochems, A. Radiomics: The bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* **2017**, *14*, 749–762. [CrossRef]
34. Uhlig, A.; Uhlig, J.; Leha, A.; Biggemann, L.; Bachanek, S.; Stöckle, M.; Reichert, M.; Lotz, J.; Zeuschner, P.; Maßmann, A. Radiomics and machine learning for renal tumor subtype assessment using multiphase computed tomography in a multicenter setting. *Eur. Radiol.* **2024**, *34*, 6254–6263. [CrossRef]
35. Jeon, H.G.; Seo, S.I.; Jeong, B.C.; Jeon, S.S.; Lee, H.M.; Choi, H.-Y.; Song, C.; Hong, J.H.; Kim, C.-S.; Ahn, H. Percutaneous kidney biopsy for a small renal mass: A critical appraisal of results. *J. Urol.* **2016**, *195*, 568–573. [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.

diagnostics

MDPI

Article

Temporal Lobe Epilepsy Focus Detection Based on the Correlation Between Brain MR Images and EEG Recordings with a Decision Tree

Cansel Ficici [1,*], Ziya Telatar [2], Osman Erogul [3] and Onur Kocak [2]

[1] Department of Electrical and Electronics Engineering, Ankara University, 06830 Ankara, Turkey
[2] Department of Biomedical Engineering, Başkent University, 06790 Ankara, Turkey; ztelatar@baskent.edu.tr (Z.T.); okocak@baskent.edu.tr (O.K.)
[3] Department of Biomedical Engineering, TOBB University of Economics and Technology, 06560 Ankara, Turkey; erogul@etu.edu.tr
* Correspondence: ogretmenoglu@ankara.edu.tr

Abstract: Background/Objectives: In this study, a medical decision support system is presented to assist physicians in epileptic focus detection by correlating MRI and EEG data of temporal lobe epilepsy patients. Methods: By exploiting the asymmetry in the hippocampus in MRI images and using voxel-based morphometry analysis, gray matter reduction in the temporal and limbic lobes is detected, and epileptic focus prediction is realized. In addition, an epileptic focus is also determined by calculating the asymmetry score from EEG channels. Finally, epileptic focus detection was performed by associating MRI and EEG data with a decision tree. Results: The results obtained from the proposed algorithm provide 100% overlap with the physician's finding on the EEG data. Conclusions: MRI and EEG correlation in epileptic focus detection was improved compared with physicians. The proposed algorithm can be used as a medical decision support system for epilepsy diagnosis, treatment, and surgery planning.

Keywords: EEG; MRI; decision tree; temporal lobe epilepsy; voxel-based morphometry; epileptic focus

1. Introduction

In traditional epilepsy studies, epilepsy detection is usually based on EEG data. In cases where the onset and focus of seizures are difficult to distinguish visually from electroencephalography (EEG) recordings, the diagnosis of epilepsy may vary from physician to physician. In this case, the physician examines the patient's magnetic resonance imaging (MRI) data to detect morphologic abnormalities and volume loss in the brain regions. However, this may be too subtle to be recognized on MRI. In addition, analyzing MRI and EEG data takes a long time, causing delays in the diagnosis and treatment process. In the literature, there are studies that detect epilepsy from EEG recordings and MR images separately, but there are few studies that use these two neuroimaging methods together and detect the epilepsy focus by correlating them.

Recently, some studies have been presented in the literature to detect epileptic seizures from EEG recordings by different feature extraction and classification approaches based on machine learning [1–8] and deep learning methods [9–16]. Ficici et al. [1] presented a study to discriminate epileptic and non-epileptic seizures from EEG signals using DWT-based feature extraction and the boosted tree method. Their highest accuracy was given as 97.2%. Altunay et al. [2] proposed a study to detect epileptic seizures from EEG signals by the linear prediction error energy (LPEE) approach. They obtained 93.6% success in detecting the presence of seizures. Ficici et al. [3] classified interictal, ictal, and normal state periods of EEG signals using support vector machines (SVMs) with nonlinear and linear features. Zarei and Asl [4] presented a method based on DWT and orthogonal matching pursuit

(OMP) techniques to detect epilepsy by SVM. Their classifier yielded an average accuracy of 97%. In Slimen et al.'s [5] study, an automated epilepsy detection method based on dual-tree complex wavelet transform was given. Khan et al. [6] introduced the classification of EEG recordings as epileptic and non-epileptic with discriminant analysis. They reported their classification accuracy as a mean accuracy of 99.6% on a clear dataset. Raghu et al. [7] proposed an approach to detect epilepsy with SVM. Their accuracy rate was given as 96.34%. Richhariya et al. [8] presented a method to discriminate seizure and non-seizure epochs by universum support vector machine (USVM).

Natu et al. [9] presented a deep learning-based model combining CNN and RNN to detect epilepsy seizures by extracting spatial and temporal features in EEG signals. They attained a classification accuracy of 98.5%. Ficici et al. [10] detected ictal epochs using a long short-term memory (LSTM) network fed with EEG sub-band features based on discrete wavelet transform (DWT). They achieved classification accuracies of 86.84% and 96.67% on two different EEG datasets. In addition, they identified epileptic focus with an accuracy of 96.10%. Qiu et al. [11] introduced a study for epileptic seizure detection by using a ResNet-LSTM network on the Bonn University dataset, and they achieved an accuracy of 99.78%. Poorani et al. [12] conducted epilepsy detection by using CNN and LSTM methods on the CHB-MIT database. They obtained a classification accuracy of 94.83%. McCallan et al. [17] presented a review of multi-seizure type classification performance based on the Temple University Hospital Seizure Corpus (TUSZ) dataset for focal and generalized classification and multi-seizure type classification.

Also, MRI-based epilepsy detection studies exist in the literature. Ficici et al. [18] presented a study to localize epileptic focus from MR images by using voxel-based morphometry (VBM). Their TLE focus detection results and seizure lateralization from EEG recordings realized by the expert overlapped at a rate of 91.7%. In addition, they obtained a sensitivity of 100% and 80% for right TLE and left TLE detection, respectively. Fearns et al. [19] presented a study to quantitatively assess the diagnostic yield of Morphometric Analysis Program (MAP) feature maps in patients with un-equivocal localization of the epileptic zone and MRI-negative patients. They reported that MAP was useful in detecting previously overlooked subtle structural lesions. However, in truly MRI-negative patients, the additional diagnostic yield was very limited. Yin et al. [20] conducted voxel-based morphology (VBM) and fractional amplitude of low-frequency fluctuation (fALFF) analyses to compare cortical morphology and spontaneous brain activity between drug-naïve benign childhood epilepsy and healthy controls. Aslam et al. [21] described statistical thresholds for asymmetry analysis of hippocampal volumetry and PET to improve the detection of TLE. They stated that threshold parameters warranted further validation in prospective studies. They reported that clinical EEG findings, combined with seizure semiology, could overcome scalp EEG's limitations and leaned towards MRI lateralization in specific cases. Jber et al. [22] conducted a VBM analysis-based study to find morphometric GM changes in TLE patients. They detected epileptogenic focus by using hippocampal volume. They stated that more subcortical volume and cortical thickness alterations were observed in the left TLE patients than in the right TLE patients by comparison with the healthy control cases. Li et al. [23] proposed a study to evaluate GM abnormalities in MTLE patients. They applied corrected VBM and Surfaced-Based Morphometry (SBM) analysis methods to MR images. They reported that ipsilateral hippocampal atrophy was a characteristic morphological biomarker of MTLE. Shigemoto et al. [24] used a graph theory-based GM network analysis to identify TLE patients. They reported that in TLE groups, GM volume reduction was observed mainly in the ipsilateral hippocampus. Chen et al. [25] detected pediatric MTLE by using the SVM method. They conducted VBM analysis to extract features from MR images. Riederer et al. [26] used the VBM analysis method to detect hippocampal atrophy and seizure focus of MTLE patients. They also applied high-dimensional warping and nonparametric statistics methods. Uher et al. [27] presented a review study that described the recent state-of-the-art computational approaches for image analysis and

their potential for epileptogenic zone detection. Also, they reported the current limitations of the methods and possible future directions to augment epileptogenic zone detection. Gupta et al. [28] proposed a study to test the epilepsy prediction accuracy of individual MRI-based tissue types and a novel composite ratio parameter [(GM + WM)/CSF]. They estimated the prediction accuracy of individual MRI tissue types and the composite ratio parameter using EEG results alone or EEG results plus seizure semiology. They reported that the composite ratio provides the highest accuracy of 85% when confirmed by EEG results, while accuracy improved further to 88% when EEG findings and seizure semiology were combined.

On the other hand, there are few studies in the literature in which MRI and EEG findings were compared or evaluated together for the detection of epilepsy. Fallahi et al. [29] presented a decision-making approach to evaluate and compare the performance of neuroimaging markers with clinical findings for the lateralization of mesial temporal lobe epilepsy (mTLE). They used descriptions of seizure semiology, ictal EEG, the ictal epileptogenic zone, the interictal–irritative zone, and MRI findings as clinical features in the Self-Organizing Map (SOM) clustering method. They also used quantitative measurements based on MRI images of T1, FLAIR, and DTI in classification by the SOM method. They used clinical findings and neuroimaging marker features separately. Then, they compared the classification accuracies of neuroimaging markers and clinical findings. Their results showed that the SOM results based on the clinical features could identify all the cases correctly, while the neuroimaging markers made a correct identification at a rate of 94%. Jing et al. [30] examined the relationship between nonlinear analysis of EEG signals, MRI, and single-photon emission computed tomography (SPECT) analyses in epilepsy lateralization of TLE patients. According to the study, they concluded that lateralization by EEG analysis and lateralization by MRI analysis were compatible at 73.9%, EEG analysis and SPECT analysis were compatible at 78.2%, and MRI and SPECT analysis were compatible at 65.2%.

In this study, an automated decision support system is presented to guide the physician in determining the epileptic focus in TLE patients. In the proposed algorithm, a decision tree method is used based on the correlation of the findings of EEG recordings and MRI images. The analysis of EEG and MRI data was based on the symmetrical nature of the brain. Gray matter reduction and asymmetry of hippocampus volumes were calculated in MRI images. Also, the asymmetry in the energies of symmetric channels was calculated from EEG recordings. The found parameters provide information about the epileptic focus. The experimental results show that the proposed decision tree algorithm achieves 100% success in finding the epileptic focus of TLE patients compared with the physician's findings.

This article is organized as follows: Section 2 introduces the EEG and MRI datasets used in this study. Furthermore, the algorithm is explained in detail using flowcharts and theoretical backgrounds. In Section 3, MRI and EEG analysis results are reported. The novelty of this paper and contributions and conclusions are presented in Section 4.

2. Materials and Methods

2.1. Description of the Dataset and Framework of This Study

In this study, multi-channel EEG recordings and brain MR images of TLE patients obtained from the Ankara University hospital were used. EEG signals with a sampling frequency of 500 Hz were received from 18 channels according to a 10–20 electrode placement system. EEG signals were recorded from Fp1-F7, F7-T3, T3-T5, T5-O1, Fp2-F8, F8-T4, T4-T6, T6-O2, Fp1-F3, F3-C3, C3-P3, P3-O1, Fp2-F4, F4-C4, C4-P4, P4-O2, FZ-CZ, and CZ-PZ channels. The EEG database used in this study includes 76 seizure recordings (48 right TLE, 28 left TLE) of 26 TLE patients. Epileptic seizure and seizure intervals were labeled on EEG signals by the expert retrospectively. Rhythmic theta and beta activities, repetitive spikes, sharp waves, and low-voltage rapid activity were the epileptic seizure indicators for the expert. Also, left TLE and right TLE patients were labeled by the expert based on the findings of EEG recordings and seizure semiology. The MRI dataset used

in this study was obtained from 15 TLE patients diagnosed with MTS, HS, hippocampal atrophy, or MRI-negative, and 14 control subjects. Hippocampus areas were labeled on MR images by the expert. Atrophy observed on the hippocampus area on coronal T1-weighted MRI, increased T2 signal reflecting gliosis, and abnormal hippocampus morphology on T1-weighted images were the indicators of hippocampal sclerosis diagnosis for the expert. Expert labeling on EEG recordings and MR images was considered the gold standard for the performance evaluation of this study.

This proposed study was implemented with Matlab 2021a [31] via a computer, Intel Core i7, 2.60 GHz processor, with 16.0 GB RAM. A flowchart of the proposed approach for TLE detection based on the correlation between EEG and MRI is given in Figure 1 [32]. Firstly, in the hippocampal volume estimation and pixel-based analysis part, the mean and standard deviation of the gray-level pixel intensities of the left and right hippocampus areas on T2-weighted MR images were calculated. In addition, gray-level MR images were masked with the hippocampus regions labeled by the expert and then converted into binary images. After that, the left and right hippocampus volumes were automatically calculated by the algorithm. Left and right hippocampus volume, mean value of pixel intensities, and standard deviation features were used to label MR images as left TLE and right TLE by using machine learning. Secondly, in the VBM analysis part, T1-weighted MR images were used to determine whether TLE patients had GM reduction in the temporal lobe or limbic lobe regions. If GM reduction was detected, the patient was labeled as either left TLE or right TLE based on the cerebral hemisphere where GM reduction was found. Thirdly, in the symmetry analysis of EEG channels part, energies of ictal EEG signals were calculated for each channel. Then, the asymmetry scores were calculated from the left and right symmetric EEG channels. According to the asymmetry score, seizure recordings were labeled as either left TLE or right TLE. In the decision tree, L1, L2, and L3 represent the lateralization labels of Analysis-1 (hippocampal volume estimation and pixel-based analysis), Analysis-2 (VBM analysis), and Analysis-3 (symmetry analysis of EEG channels). The number "−1" was assigned to the left TLE class, the number "1" to the right TLE class, and the number "0" to MR-negative. Finally, the epileptic focus was determined by correlating the decisions obtained from MR images and EEG recordings by using a decision tree. In the decision tree, label numbers assigned in Analysis-1, Analysis-2, and Analysis-3 were summed. If the total number was greater than zero, the epileptic focus for that patient was ultimately determined to be on the left brain hemisphere; otherwise, the epileptic focus was on the right hemisphere.

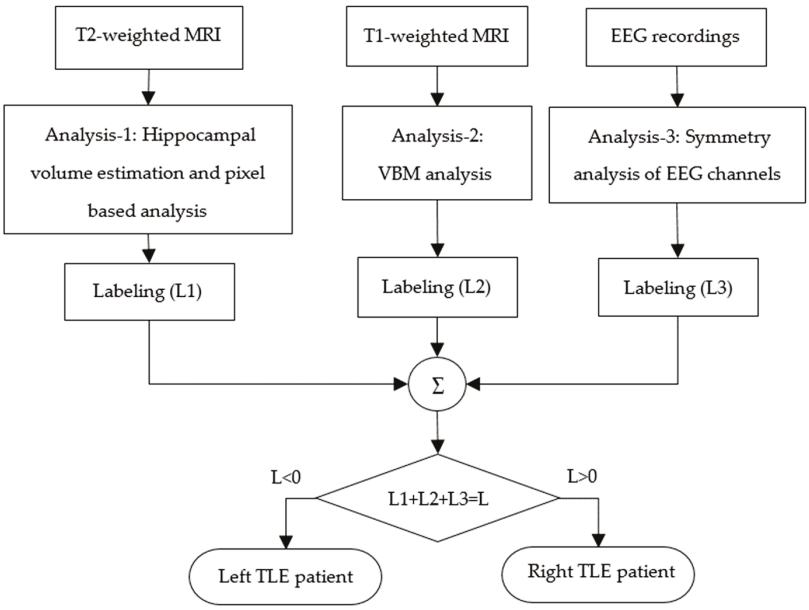


Figure 1. Flowchart of TLE detection based on correlation of EEG and MRI.

2.2. Hippocampal Volume Estimation and Pixel-Based Analysis of MR Images

Abnormal conditions such as atrophy and gliosis in the hippocampus of TLE patients cause an increase in signal intensity and a decrease in hippocampus volumes on T2-weighted MR images. Therefore, in this part of this study, epileptic focus detection from subtle signal increases in T2-weighted images and volume reductions in hippocampus regions were realized. A flowchart of the TLE localization from hippocampus on MRI is given in Figure 2 [32].

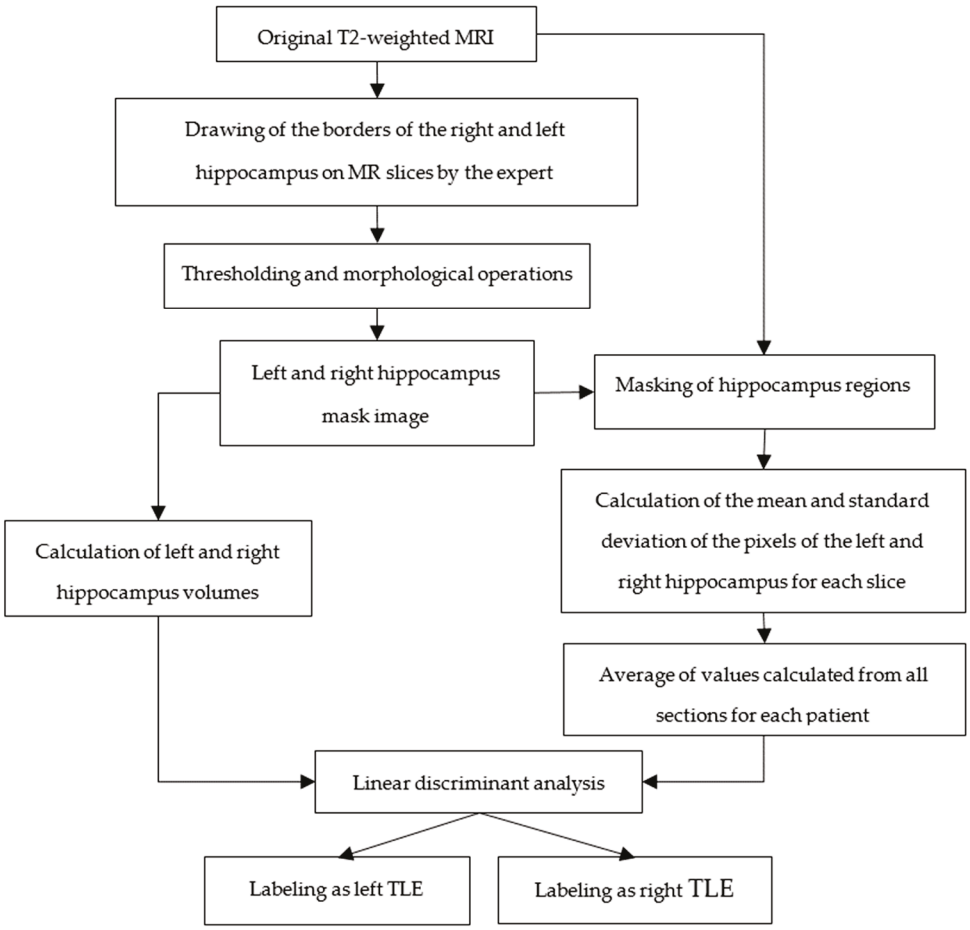


Figure 2. Flowchart of TLE localization from the hippocampus on MRI.

First of all, hippocampus boundaries were drawn by the expert in T2-weighted coronal MR images of TLE patients. A sample of a T2-weighted coronal MR slice of a TLE patient with hippocampus boundaries drawn by the expert is shown in Figure 3. Then, binary hippocampus mask images were obtained by thresholding the images labeled by the expert. Left and right hippocampus volumes were calculated from these binary hippocampus images. In addition, the original T2-weighted MR images were masked with the binary hippocampus images. The mean and standard deviation of the pixel intensities of the left and right hippocampus regions were calculated for each gray level MR slice. Volume estimation from hippocampus mask images was realized with (1) and (2). The mean and standard deviation of the pixel intensities of the left and right hippocampus were calculated for each section by using (3) and (4), respectively. The values calculated from all slices of each patient were averaged. These features were used in the linear discriminant analysis (LDA) method; thus, left TLE and right TLE classification was conducted.

$$Area = (number\ of\ pixels)\ x\ (pixel\ area), \tag{1}$$

$$Volume = \sum_i Area_i\ \ x\ slice\ thickness, \tag{2}$$

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x(n), \tag{3}$$

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(x(n)-\mu)^2}. \tag{4}$$

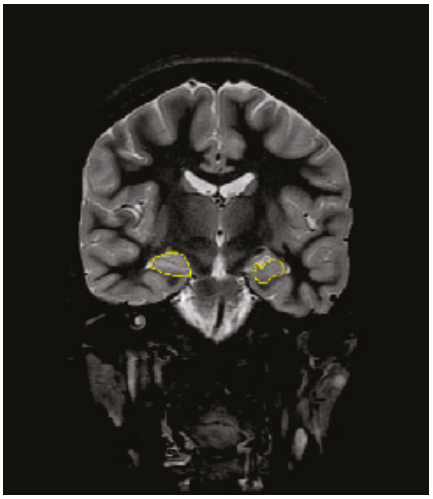


Figure 3. MRI slice of a TLE patient with hippocampus boundaries drawn in yellow by an expert.

2.3. VBM Analysis from MR Images

In TLE patients, GM reduction generally occurs in the hippocampus and structures of the limbic system and mesial temporal lobe [33]. So, the region of interest (ROI) of VBM analysis in this study is the limbic lobe and temporal lobe. By applying the VBM method, GM reduction areas were detected from the T1-weighted MR images of TLE patients by comparing them with the T1-weighted MR images of the control group [18]. A flowchart of the TLE localization using VBM analysis is given in Figure 4. The region with GM reduction is labeled as the epileptic focus. If no GM reduction was found in the temporal lobe or limbic lobe, that patient was labeled as MR-negative.

VBM analysis is based on the statistical *t*-test. The decrease in GM in MRI images of epilepsy patients compared to the control group was calculated using the *t*-test.

The first step of VBM analysis is image registration enabling the overlapping of boundaries of MR images taken from different databases [34]. The second step is normalization and then smoothing GM images with a Gaussian filter. The final step of the VBM method is the segmentation of white matter, gray matter, and cerebrospinal fluid parts in MR images by applying the unified segmentation based on the Gaussian Mixture Model (GMM) [35]. Threshold values at voxel and cluster levels were determined experimentally. In the *t*-test analysis in this study, threshold values of $p < 0.001$ or $p < 0.003$ were considered significantly different to detect GM changes in each patient with TLE. Furthermore, a cluster size greater than 50 voxels (N value given in Figure 3) was considered an appropriate threshold value to determine the epilepsy focus [32].

An area with significant gray matter reduction in the limbic lobe and the temporal lobe area detected by VBM analysis for two patients from the AU dataset are shown in Figures 5a and 5b, respectively. Figure 5 shows the glass views of the MR slices (left column) and the corresponding gray-level images (right column). In the glass view images, voxels in GM reduction areas are indicated by black-shaded areas, while the gray-shaded areas are specific regions selected by the user. In this study, temporal lobe and limbic lobe areas were selected for TLE analysis. In the gray-level images, GM atrophy is shown with white shaded areas, while temporal lobe and limbic lobe areas are shown in yellow. Once the highest *t*-test value was determined, the cluster containing this voxel was found and marked with a red arrow (Figure 5).

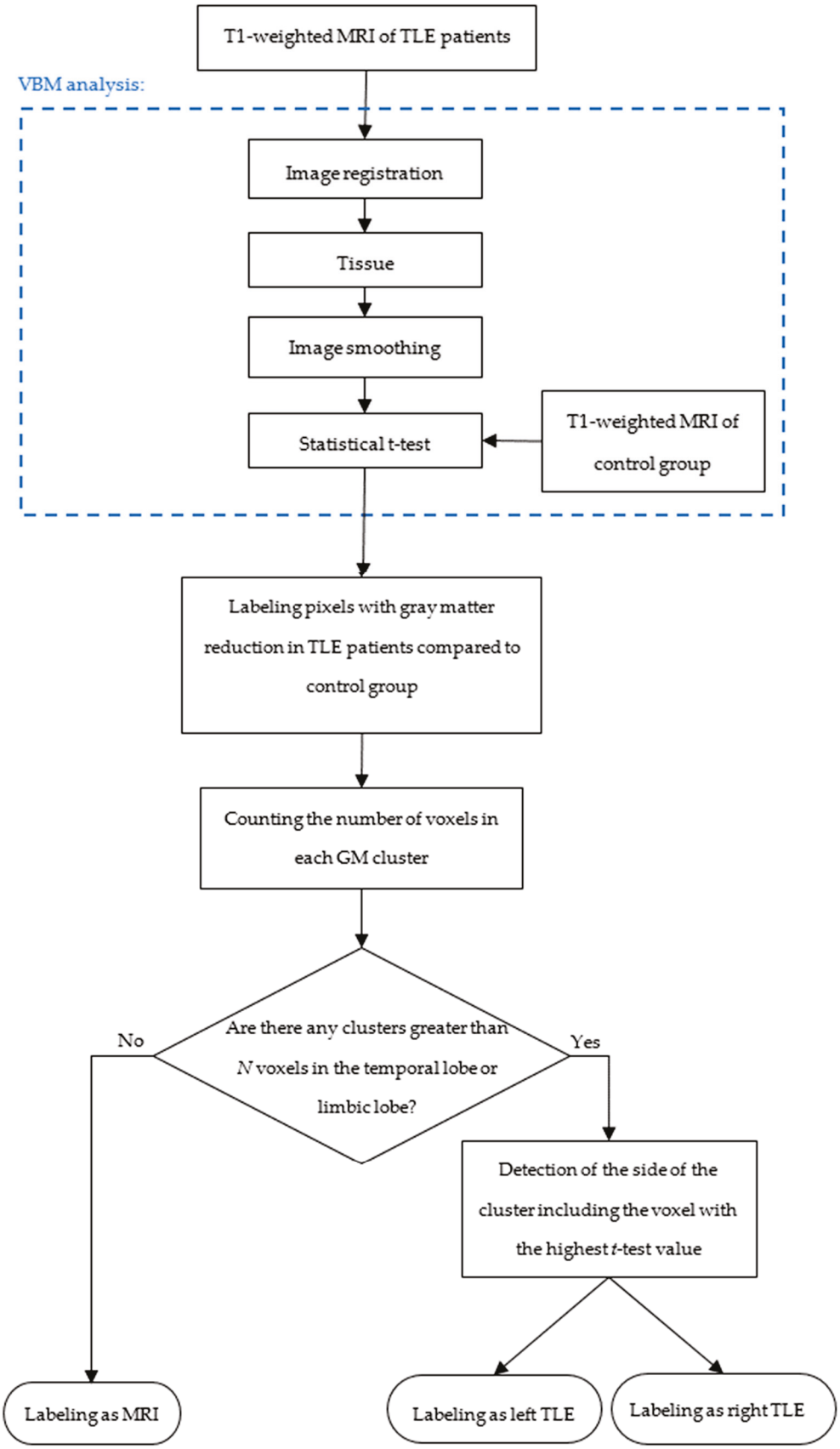


Figure 4. Flowchart of TLE localization using VBM analysis applied on MRI.

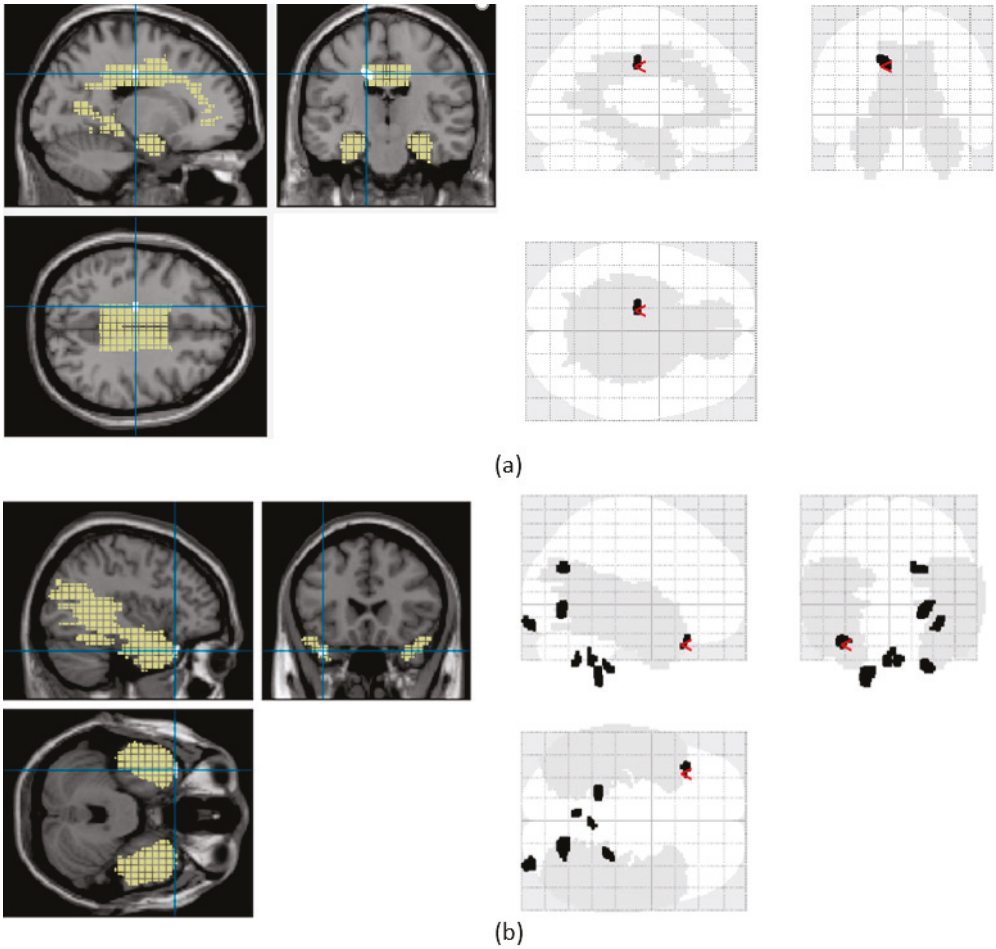


Figure 5. Area with significant gray matter reduction in the (**a**) limbic lobe area and the (**b**) temporal lobe area detected by VBM analysis for two patients from the AU dataset.

2.4. Symmetry Analysis of EEG Signals

In this part, ictal EEG signal energies were calculated for each left and right symmetric EEG channel; then, the asymmetry score was calculated. A flowchart of TLE localization using symmetry analysis of EEG signal channels is given in Figure 6 [32]. The asymmetry score was calculated from the energy ratios of these symmetrical left and right bipolar EEG channels. Symmetrical EEG channel pairs are given in Table 1. The asymmetry score represented by A was calculated by using (5) for each channel [10]. x_{left} and x_{right} represent symmetric left and right EEG channel signals for epoch length, respectively.

$$A = \frac{\sum x_{left}(n)^2}{\sum x_{right}(n)^2} \tag{5}$$

Table 1. Symmetrical EEG channels.

	Left Side	Right side
1	Fp1-F7	Fp2-F8
2	F7-T3	F8-T4
3	T3-T5	T4-T6
4	T5-O1	T6-O2
5	Fp1-F3	Fp2-F4
6	F3-C3	F4-C4
7	C3-P3	C4-P4
8	P3-O1	P4-O2

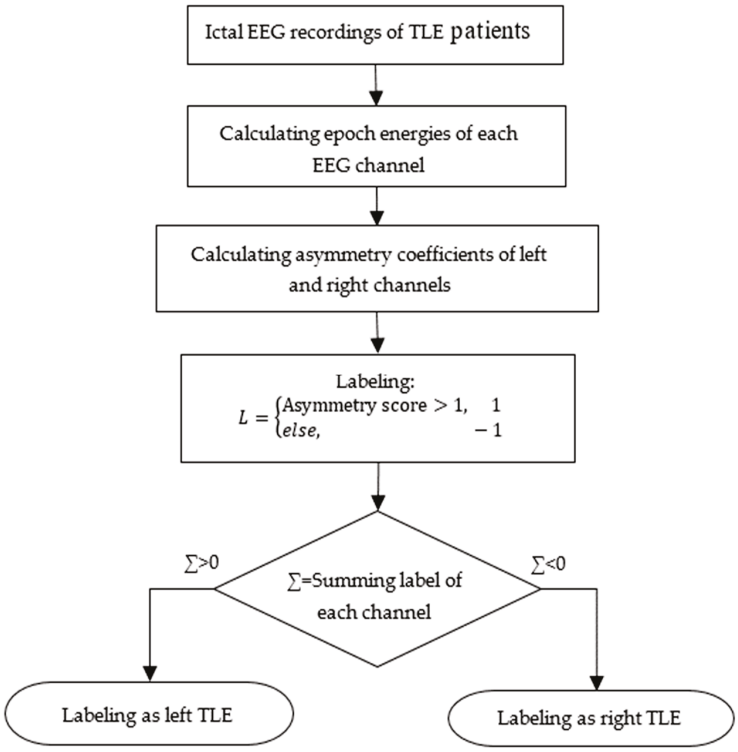


Figure 6. Flowchart of TLE localization by using symmetry analysis of EEG signal channels.

Examples of EEG recordings in seizure-free and epileptic seizure intervals are given in Figures 7 and 8, respectively. These figures were obtained by using Matlab [36]. By comparing Figures 7 and 8, it can be seen that excessive neural activity during the seizure causes an increase in EEG signal energy on the epileptic focus side. So, if the asymmetry score was greater than 1, EEG channel pair of the patient was labeled as "1"; otherwise, it was labeled as "−1". Then, the label numbers were summed. If the total number was greater than zero, that recording was labeled as left sided focus; otherwise; it was labeled as right sided focus. This labeling was realized for each seizure recording of each patient. If the seizure recordings labeled as the left focus exceed the number of seizure recordings labeled as the right focus, that patient was determined as a left TLE patient; otherwise, the patient was labeled as a right TLE patient.

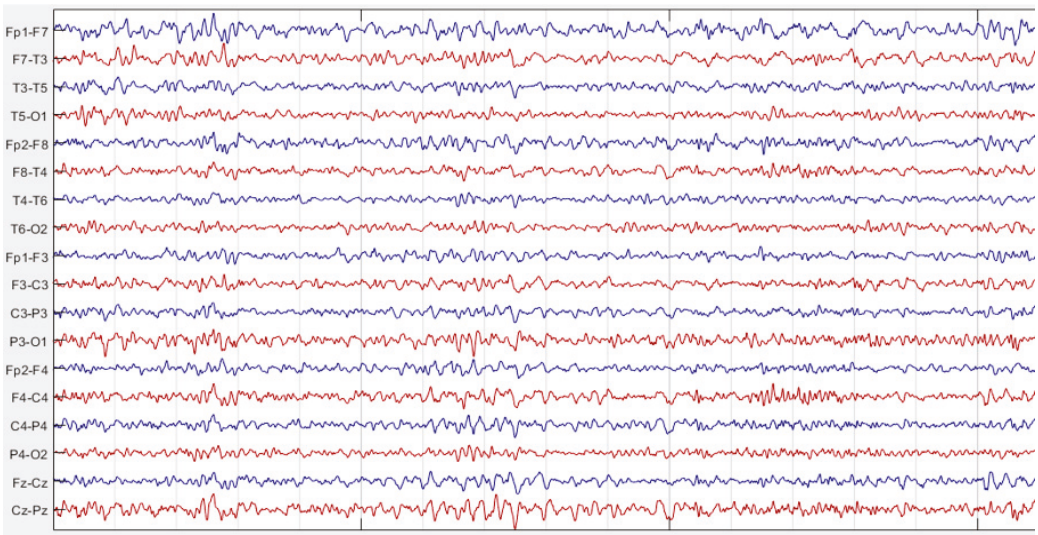


Figure 7. EEG recording of a patient in a seizure-free interval.

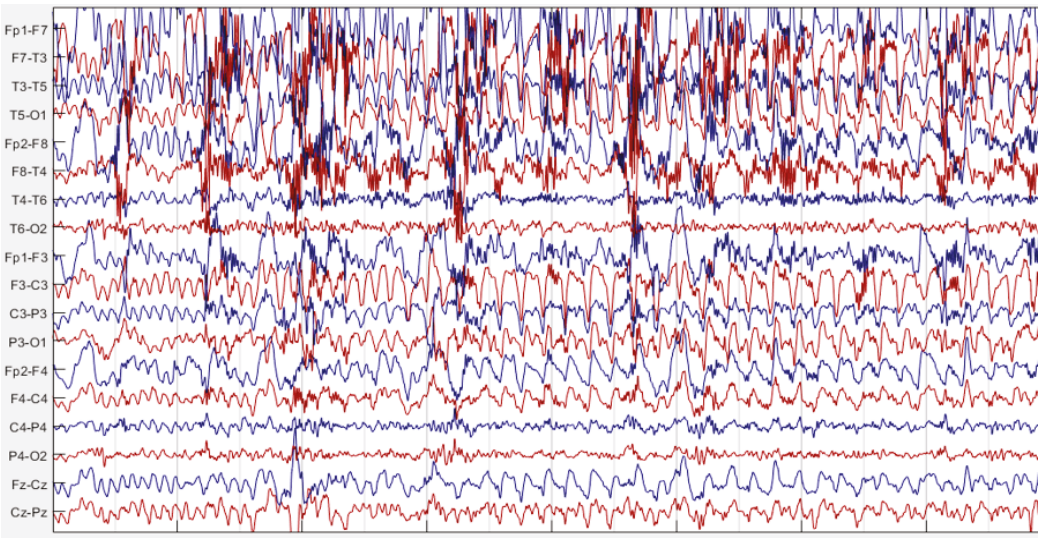


Figure 8. EEG recording of a patient during an epileptic seizure.

3. Results

In this study, MRI and EEG data were analyzed, and results were obtained using the decision tree algorithm. For EEG data, the energies of symmetric EEG channels were calculated, and asymmetry coefficients were calculated and labeled as left or right epileptic focus. In addition, separate epileptic focus detection analyses were performed with T2- and T1-weighted MR images, and as a result, left and right foci were labeled. Afterward, a final decision was determined for the epileptic focus with the decision tree algorithm.

The results for Analysis-1 given in Figures 1 and 2 are presented in Tables 2–4. Left and right hippocampus volumes, statistical features of hippocampal regions, which are the mean value and standard deviation of pixels in the left and right hippocampus regions, were used as feature vectors in different machine learning methods (decision tree, discriminant analysis, Naïve Bayes, SVM, k-nearest neighbor, boosting, bagging, random subsampling, ANN) with different combinations for TLE classification. The results are given in Tables 2–4. The results of Analysis-1 using hippocampus volumes and statistical features of hippocampal regions for two-class classification (TLE and control group), three-class classification (LTLE, RTLE, and control group), and another two-class classification (LTLE and RTLE) are given in Table 2, Table 3, and Table 4, respectively. As can be seen from these tables, the common feature with the highest classification accuracy across all classification groups is the left and right hippocampus volumes.

In addition, the LDA method produced the highest accuracy, sensitivity, and PPV (Positive Prediction Value) by handling left and right hippocampus volume features.

Table 2. Results of Analysis-1 using hippocampus volumes and statistical features of hippocampal regions in TLE and control group classification.

Classes	Feature Combinations	Method	Accuracy (%)	Sensitivity (%)	PPV (%)
14 TLE, 14 control group	mean	Naïve Bayes	78.6	85.7	75.0
	standard deviation	KNN	82.1	92.9	76.5
	mean and standard deviation	Decision tree	82.1	85.7	80.0
	mean and standard deviation and hippocampus volumes	KNN	85.7	92.9	83.1
	hippocampus volumes	Naïve Bayes	89.3	92.9	86.7
	mean and hippocampus volumes	Naïve Bayes	89.3	92.9	86.7
	standard deviation and hippocampus volumes	KNN	82.1	92.9	76.5

Table 3. Results of Analysis-1 using hippocampus volumes and statistical features of hippocampal regions in LTLE, RTLE, and control group classification.

Classes	Feature Combinations	Method	Accuracy (%)	Sensitivity (%)	PPV (%)
six LTLE, eight RTLE, 14 control group	mean	LDA	67.9	58.9	74.1
	standard deviation	Naïve Bayes	57.1	47.2	43.2
	mean and standard deviation	ANN	57.1	45.8	47.2
	mean and standard deviation and hippocampus volumes	KNN	78.6	73.9	76.0
	hippocampus volumes	LDA	78.6	74.0	77.1
	mean and hippocampus volumes	Naïve Bayes	71.4	60.6	60.3
	standard deviation and hippocampus volumes	LDA	67.9	60.8	62.1

Table 4. Results of Analysis-1 using hippocampus volumes and statistical features of hippocampal regions in LTLE and RTLE classification.

Classes	Feature Combinations	Method	Accuracy (%)	Sensitivity (%)	PPV (%)
six LTLE, eight RTLE	mean	SVM	64.3	83.3	55.6
	standard deviation	Naïve Bayes	57.1	50.0	50.0
	mean and standard deviation	SVM	57.1	0	-
	mean and standard deviation and hippocampus volumes	LDA	71.4	66.7	75.0
	hippocampus volumes	LDA	78.6	66.7	80.0
	mean and hippocampus volumes	KNN	71.4	50.0	75.0
	standard deviation and hippocampus volumes	ANN	78.6	66.7	80.0

The results for Analysis-2 given in Figures 1 and 4 are presented in Table 5 [18]. In Table 5, the location of the highest *t*-test value in the limbic lobe and temporal lobe areas detected by using VBM analysis are given. Also, cluster size, *t*-test values, and MNI coordinates are given. According to the location of the peak voxel, the TLE focus side is determined as the left or right hemisphere. Since the absence of an epileptogenic lesion or atrophy on visual inspection of the MRI conducted by the expert was defined as MRI-negative TLE, the absence of the GM reduction in the temporal lobe or the limbic lobe was defined as MRI-negative in the VBM analysis.

Table 5. Detection of the highest *t*-test value and its region by using VBM analysis.

Patient No.	Location of Peak Voxels	Cluster Size (Number of Voxels)	Voxel *t* Statistic (Peak Voxel)	MNI Coordinates (x, y, z)	TLE Focus Side Decision
1	//Right Cerebrum//Limbic Lobe//Right Hippocampus	19837	13.07	30, −20, −15	Right
2	//Left Cerebrum//Temporal Lobe//Temporal Gyrus	51	4.56	−38, 24, −28	Left
3	//Left Cerebrum//Limbic Lobe//Cingulate Gyrus	74	4.3026	−16.5, −18, 37.5	Left
4	NaN*	NaN	NaN	NaN	MRI-negative
5	//Right Cerebrum//Limbic Lobe//Right Hippocampus	172	6.53	32, −12, −20	Right
6	//Right Cerebrum//Temporal Lobe//Middle Temporal Gyrus	83	5.0754	48, −36, 0	Right
7	//Left Cerebrum//Limbic Lobe//Left Hippocampus	337	4.63	−33, −14, −20	Left
8	//Left Cerebrum//Temporal Lobe//Temporal Gyrus	4096	11.71	−40, 8, −26	Left
9	//Right Cerebrum//Temporal Lobe//Right Hippocampus	108	4.7375	−27, −39, 0	Right
10	//Right Cerebrum//Limbic Lobe//Right Hippocampus	54	4.6149	31.5, −12, −19.5	Right
11	//Left Cerebrum//Temporal Lobe//Left Middle Temporal Gyrus	14862	16.10	−38, 20, −36	Left
12	//Left Cerebrum//Limbic Lobe// Parahippocampal Gyrus	50	4.5347	−28.5, −58.5, −7.5	Left
13	NaN	NaN	NaN	NaN	MRI-negative
14	NaN	NaN	NaN	NaN	MRI-negative
15	//Left Cerebrum//Limbic Lobe//Left Anterior Cingulate and Paracingulate Gyri	61	4.88	−4, 48, 2	Left

NaN* indicates that GM reduction was not detected in the temporal lobe or in the limbic lobe.

The physician's results of epileptic focus detection separately from EEG and MRI are compared with the results of Analysis-2. It was seen that the VBM method gives more consistent results with the expert decision from EEG (accuracy of 91.7%) than the physician's evaluation for separate EEG and MRI data (accuracy of 50%).

The results for Analysis-3 shown in Figures 1 and 6 are presented in Table 6 [10]. As seen in Table 6, the accuracy, sensitivity, and specificity values for left and right epileptic focus detection from EEG signals are 96.1%, 100%, and 93.8%, respectively.

Table 6. Results of symmetry analysis of the EEG signal algorithm.

Number of EEG Data	Result
76 (48 RTLE, 28 LTLE)	Accuracy = 96.1% Sensitivity = 100% Specificity = 93.8%

The MRI and EEG lateralization of the physician and the proposed method are given in Table 7. Since the MRI of the first patient in this table consists of images taken after temporal lobectomy, the hippocampus volume calculation could not be performed, so Analysis-1 could not be applied to this patient.

Table 7. Expert findings and decisions of the proposed method on TLE diagnosis.

TLE Patient No	Expert Findings from EEG	Expert Findings from MRI	Decision of Analysis-1	Decision of Analysis-2	Decision of Analysis-3	Decision Tree Final Result
1	Right TLE	Right Mesial Temporal Sclerosis	NaN*	Right TLE	Right TLE	Right TLE
2	Left TLE	Right Hippocampal Sclerosis	Right TLE	Left TLE	Left TLE	Left TLE
3	Left TLE	MRI-negative	Left TLE	Left TLE	Left TLE	Left TLE
4	Right TLE	MRI-negative	Right TLE	MRI negative	Left TLE	Right TLE
5	Right TLE	MRI-negative	Right TLE	Right TLE	Right TLE	Right TLE
6	Right TLE	MRI-negative	Right TLE	Right TLE	Right TLE	Right TLE
7	Left TLE	MRI-negative	Left TLE	Left TLE	Left TLE	Left TLE
8	Right TLE	Bilateral Hippocampal Atrophy	Right TLE	Left TLE	Right TLE	Right TLE
9	Right TLE	Bilateral Hippocampal Atrophy	Right TLE	Right TLE	Right TLE	Right TLE
10	Right TLE	MRI-negative	Left TLE	Right TLE	Right TLE	Right TLE
11	Left TLE	Left Hippocampal Atrophy	Left TLE	Left TLE	Left TLE	Left TLE
12	Left TLE	MRI-negative	Right TLE	Left TLE	Left TLE	Left TLE
13	Right TLE	Right Hippocampal Atrophy	Right TLE	MRI negative	Left TLE	Right TLE
14	Right TLE	MRI-negative	Right TLE	MRI negative	Right TLE	Right TLE
15	Left TLE	MRI-negative	Left TLE	Left TLE	Left TLE	Left TLE

NaN* indicates that Patient-1 was excluded from hippocampal volume estimation (Analysis-1) due to a history of temporal lobectomy.

The results of the method of correlating MRI and EEG data are shown in Table 8. This table indicates the following:

- The physician detected epilepsy findings (HS, hippocampal atrophy, mesial temporal sclerosis) in the MR images of 6 out of the 15 TLE patients. The MRI lateralization of patients with epilepsy detected by the physician on MRI images and the physician's EEG lateralization in these patients overlap by 50% (3/6).
- The physician's lateralization of epilepsy from EEG recordings and the MRI lateralization of Analysis-1 overlap by 78.6% (11/14).

- Analysis-2 detected epilepsy on MRI in 80.0% (12/15) of the TLE patients. The MRI lateralization of patients with epilepsy detected by the VBM algorithm (Analysis-2) and the lateralization made by the physician from EEG data overlap by 91.7% (11/12).
- The epilepsy lateralization made by the physician from the EEG recordings and the EEG lateralization of Analysis-3 overlap by 86.7% (13/15).
- The lateralization of epilepsy made by the physician from EEG recordings and the lateralization of the proposed MRI and EEG data correlation algorithm (decision tree) overlap by 100% (15/15).
- The MRI lateralization of patients with epilepsy detected by the physician on MRI images and the MRI lateralization of Analysis-1 overlap by 60% (3/5) (when calculating this overlap rate, patient 1 with temporal lobectomy was not included).
- The MRI lateralization of patients with gray matter reduction detected by the VBM analysis (Analysis-2) and the EEG lateralization of Analysis-3 in these patients overlap by 91.7% (11/12).
- The proposed VBM analysis algorithm (Analysis-2) detected epilepsy in 80.0% (12/15) of the TLE patients, while the physician detected epilepsy in the MR images of 40% (6/15) of the TLE patients.

Table 8. The rates of overlap between expert findings and decisions of the proposed method on TLE.

Gold Standard		Proposed Method				Results
Expert Lateralization from EEG	**Expert Lateralization from MRI**	**Lateralization of Analysis-1**	**Lateralization of Analysis-2**	**Lateralization of Analysis-3**	**Lateralization of the Decision Tree**	**Lateralization Overlap Ratio**
√	√					50%
√		√				78.6%
√			√			91.7%
√				√		86.7%
√					√	100%
	√	√				60%
	√		√			50%
			√	√		91.7%

The overlap ratio of the proposed MRI and EEG correlation algorithm (decision tree) with the proposed EEG lateralization algorithm (Analysis-3) is 100% in patients with epilepsy. The overlap rate of EEG and MRI lateralization by the physician in patients where the MRI findings are detected is 50%. Therefore, both MRI finding detection and MRI and EEG lateralization overlap rates were improved compared with the physician.

Table 9 shows a comparison of the epileptic focus detection studies available in the literature with the proposed method. To the best of our knowledge, there are few automatic epileptic focus detection studies in the literature that use MRI and EEG data together. In the studies given in Table 9, epileptic focus detections were performed on separate MRI or EEG data and on MRI and EEG data together. There is a 100% match between the proposed algorithm and the physician evaluation results. When compared with the EEG and MRI studies in the literature, the highest accuracy in the detection of epileptic focus is achieved by the proposed algorithm.

Table 9. Table showing a comparison of the proposed method with existing epileptic focus detection studies.

Authors	Dataset	Algorithm/Method	Results
Chen et al. [25]	Shenzhen Children's Hospital: 22 MTLE-HS patients (16 left, six right) and 15 healthy controls	Right-HS and control group classification by using volumes of regions with GM reduction from MRI/VBM and SVM	Accuracy: 93.3% Sensitivity: 100% Specificity: 90.7%
		Left-HS and control group classification by using volumes of regions with GM reduction from MRI/VBM and SVM	Accuracy: 80.0% Sensitivity: 80.0% Specificity: 80.0%
Behesti et al. [37]	National Center of Neurology and Psychiatry Hospital: 63 participants, including 24 HCs, 19 MRI-negative, PET-positive left TLE patients and 20 MRI-negative, PET-positive right TLE	Discrimination of left TLE, right TLE by using MR images obtained by SPM12 toolbox/SVM	Accuracy: 66.7% Sensitivity: 63.2% Specificity: 70%
Nazem-Zadeh et al. [38]	Dataset source not available in the original manuscript. The authors implemented the study in Henry Ford Hospital: 10 LTLE, 10 RTLE, 45 control group	Classification of left and right TLE patients from MRI by using hippocampus volumes/Hemispheric Variation Uncertainty (HVU) analysis	Accuracy: 84.9%
Jafari-Khouzani et al. [39]	Dataset source not available in the original manuscript. The authors implemented the study in Henry Ford Hospital.: 20 LTLE, 16 RTLE, 25 control group	Classification of left and right TLE patients from MRI by using hippocampus volumes/LDA	Accuracy: 75.3%
Türk et al. [40]	Bonn University: 500 EEG segments with 23.6 s duration for each	Epileptic focus identification from EEG/convolutional neural network (CNN)	Accuracy: 98.5%
Daoud et al. [41]	Bern–Barcelona and Bonn University: 7500 EEG segments with 20 sec duration for each	Epileptic focus identification from EEG/Deep convolutional autoencoder	Accuracy: 96.0%
Fallahi et al. [28]	Tehran University: 35 mesial TLE patients	Presurgical lateralization of mesial temporal lobe epilepsy from EEG and MRI/self-organizing maps	Accuracy: 94.0%
Jing et al. [30]	Dataset source not available in the paper: 23 TLE patient	Investigated the correlation between lateralization MRI, SPECT, and EEG/nonlinear analysis	Concordance between EEG and MRI: 73.9% Concordance between EEG and SPECT: 78.2% Concordance between MRI and SPECT: 65.2%
Proposed study	University Hospital: 15 TLE (six LTLE, nine RTLE)	Epileptic focus detection/decision tree	Accuracy: 100%

4. Discussion and Conclusions

In this study, by correlating MRI and EEG data of TLE patients, an automated decision support system is obtained to assist physicians in epileptic focus detection. In this respect, EEG signals and MRI images of TLE patients were analyzed, and epilepsy disease was detected with the decision tree algorithm.

T1- and T2-weighted images were used in the analysis of MR images. By performing VBM analysis using T1-weighted images, areas with gray matter reduction in the brains of epilepsy patients were detected. Using T2-weighted images, the hippocampus volumes of epilepsy patients and the asymmetry parameters of the right and left hippocampus were

calculated. Thus, it was decided whether there was an asymmetry between the left and right hippocampus by using the LDA method. In addition, the asymmetry coefficient was obtained from ictal EEG signals recorded through symmetrical EEG channels in the left and right brain lobes. In the last stage, labels were assigned to the results obtained from the analysis of MRI images and EEG signals. These labels formed the inputs of the decision tree algorithm. The output of the algorithm made the final decision of epileptic focus.

There are many studies on epilepsy detection from EEG signals alone in the literature. In addition, there are also studies that applied VBM analysis to MR images alone. The novelty of this study is that it utilizes the correlation in the information provided by both neuroimaging techniques as a decision tree for epilepsy detection.

Our study outperforms other related studies because of its comprehensive approach to detecting TLE by combining multiple advanced neuroimaging and electrophysiological features that capture unique aspects of the disorder. Specifically, our study integrates precise gray matter reductions identified in both the limbic and temporal lobes, advanced statistical analysis of hippocampal pixel properties in MR images, and asymmetry measures of both hippocampal volumes and EEG signal energies across channels. This multifaceted approach enhances the accuracy of TLE detection by leveraging complementary insights from structural, functional, and statistical data—something most related studies may not achieve when focusing on a limited set of features. Consequently, this comprehensive feature set enables our model to more effectively capture TLE-related abnormalities, leading to better diagnostic performance and potentially higher clinical relevance.

The main contributions of this study are as follows: (1) It was shown that the detection of epileptic focus by correlation of EEG and MR images provides higher success compared to the evaluation of EEG and MRI analysis separately in the detection of epileptic focus. (2) VBM analysis enabled the detection of GM reductions in MR images of epilepsy patients that cannot be visually detected by the physician. (3) Unique distinguishing features were obtained in the detection of TLE patients. These features are GM reductions detected in the limbic lobe and temporal lobe, statistical properties of pixels belonging to hippocampus regions in MR images, asymmetry in hippocampus volumes, and asymmetry in EEG signal channel energies.

Author Contributions: Writing—original draft, C.F. and Z.T.; methodology, C.F., O.E. and Z.T.; conceptualization, C.F., O.E., Z.T. and O.K.; formal analysis, C.F. and Z.T.; investigation, O.E. and Z.T.; software, C.F.; supervision, O.E. and Z.T.; validation, O.E., Z.T. and O.K.; visualization, C.F. and O.K.; writing—review and editing, O.E. and Z.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: This study was conducted in accordance with the Declaration of Helsinki and approved by the Clinical Research Ethics Committee of the Faculty of Medicine of Ankara University (i5-225-19, 14 November 2019).

Data Availability Statement: Data are unavailable because of privacy and ethical restrictions.

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

References

1. Ficici, C.; Telatar, Z.; Eroğul, O. Automated temporal lobe epilepsy and psychogenic nonepileptic seizure patient discrimination from multichannel EEG recordings using DWT based analysis. *Biomed. Signal Process. Control* **2022**, *77*, 103755. [CrossRef]
2. Altunay, S.; Telatar, Z.; Erogul, O. Epileptic EEG detection using the linear prediction error energy. *Expert Syst. Appl.* **2010**, *37*, 5661–5665. [CrossRef]
3. Ficici, C.; Erogul, O.; Telatar, Z. Epileptic Activity Detection in EEG Signals using Linear and Non-linear Feature Extraction Methods. In Proceedings of the 2019 11th International Conference on Electrical and Electronics Engineering (ELECO), Bursa, Turkey, 28–30 November 2019.
4. Zarei, A.; Asl, B.M. Automatic seizure detection using orthogonal matching pursuit, discrete wavelet transform, and entropy based features of EEG signals. *Comput. Biol. Med.* **2021**, *131*, 104250. [CrossRef] [PubMed]

5. Slimen, I.B.; Boubchir, L.; Mbarki, Z.; Seddik, H. Algorithms, EEG epileptic seizure detection and classification based on dual-tree complex wavelet transform and machine learning. *J. Biomed. Res.* **2020**, *34*, 151–161. [CrossRef]
6. Khan, K.A.; Shanir, P.P.; Khan, Y.U.; Farooq, O. A hybrid Local Binary Pattern and wavelets based approach for EEG classification for diagnosing epilepsy. *Expert Syst. Appl.* **2020**, *140*, 112895. [CrossRef]
7. Raghu, S.; Sriraam, N.; Temel, Y.; Rao, S.V.; Hegde, A.S.; Kubben, P.L. Performance evaluation of DWT based sigmoid entropy in time and frequency domains for automated detection of epileptic seizures using SVM classifier. *Comput. Biol. Med.* **2019**, *110*, 127–143. [CrossRef]
8. Richhariya, B.; Tanveer, M. EEG signal classification using universum support vector machine. *Expert Syst. Appl.* **2018**, *106*, 169–182. [CrossRef]
9. Natu, M.; Bachute, M.; Kotecha, K. HCLA_CBiGRU: Hybrid Convolutional Bidirectional GRU Based Model for Epileptic Seizure Detection. *Neurosci. J.* **2023**, *3*, 100135. [CrossRef]
10. Ficici, C.; Telatar, Z.; Kocak, O.; Erogul, O. Identification of TLE Focus from EEG Signals by Using Deep Learning Approach. *Diagnostics* **2023**, *13*, 2261. [CrossRef]
11. Qiu, X.; Yan, F.; Liu, H. A difference attention ResNet-LSTM network for epileptic seizure detection using EEG signal. *Biomed. Signal Process. Control* **2023**, *83*, 104652. [CrossRef]
12. Poorani, S.; Balasubramanie, P. Deep learning based epileptic seizure detection with EEG data. *Int. J. Syst. Assur. Eng. Manag.* **2023**. [CrossRef]
13. Varlı, M.; Yılmaz, H. Multiple classification of EEG signals and epileptic seizure diagnosis with combined deep learning. *J. Comput. Sci.* **2023**, *67*, 101943. [CrossRef]
14. Mir, W.A.; Anjum, M.; Shahab, S. Deep-EEG: An optimized and robust framework and method for EEG-based di-agnosis of epileptic seizure. *Diagnostics* **2023**, *13*, 773. [CrossRef]
15. Lebal, A.; Moussaoui, A.; Rezgui, A. Epilepsy-Net: Attention-based 1D-inception network model for epilepsy detection using one-channel and multi-channel EEG signals. *Multimed. Tools Appl.* **2023**, *82*, 17391–17413. [CrossRef]
16. Ilias, L.; Askounis, D.; Psarras, J. Multimodal detection of epilepsy with deep neural networks. *Expert Syst. Appl.* **2023**, *213*, 119010. [CrossRef]
17. McCallan, N.; Davidson, S.; Ng, K.Y.; Biglarbeigi, P.; Finlay, D.; Lan, B.L.; McLaughlin, J. Epileptic multi-seizure type classification using electroencephalogram signals from the Temple University Hospital Seizure Corpus: A review. *Expert Syst. Appl.* **2023**, *234*, 121040. [CrossRef]
18. Ficici, C.; Telatar, Z.; Erogul, O. Localization of epileptic focus by gray matter reduction analysis from brain MR images for temporal lobe epilepsy patients. *Biomed. Signal Process. Control* **2023**, *83*, 104716. [CrossRef]
19. Fearns, N.; Birk, D.; Bartkiewicz, J.; Rémi, J.; Noachtar, S.; Vollmar, C. Quantitative analysis of the morphometric analysis program MAP in patients with truly MRI-negative focal epilepsy. *Epilepsy Res.* **2023**, *192*, 107133. [CrossRef] [PubMed]
20. Yin, Y.; Wang, F.; Ma, Y.; Yang, J.; Li, R.; Li, Y.; Liu, H. Structural and functional changes in drug-naïve benign childhood epilepsy with centro-temporal spikes and their associated gene expression profiles. *Cereb. Cortex* **2023**, *33*, 5774–5782. [CrossRef]
21. Aslam, S.; Rajeshkannan, R.; Sandya, C.J.; Sarma, M.; Gopinath, S.; Pillai, A. Statistical asymmetry analysis of volumetric MRI and FDG PET in temporal lobe epilepsy. *Epilepsy Behav.* **2022**, *134*, 108810. [CrossRef]
22. Jber, M.; Habibabadi, J.M.; Sharifpour, R.; Marzbani, H.; Hassanpour, M.; Seyfi, M.; Mobarakeh, N.M.; Keihani, A.; Hashemi-Fesharaki, S.S.; Ay, M.; et al. Temporal and extratemporal atrophic manifestation of temporal lobe epilepsy using voxel-based morphometry and corticometry: Clinical application in lateralization of epileptogenic zone. *Neurol. Sci.* **2021**, *42*, 3305–3325. [CrossRef] [PubMed]
23. Li, Z.; Gao, Q.; Peng, K.; Lin, J.; Wang, W.; Wang, W.; Deng, B. Quantitative evaluation of gray matter alterations in patients with mesial temporal lobe epilepsy (MTLE). *Neurosci. Inform.* **2021**, *1*, 100018. [CrossRef]
24. Shigemoto, Y.; Sato, N.; Sone, D.; Maikusa, N.; Yamao, T.; Kimura, Y.; Chiba, E.; Suzuki, F.; Fujii, H.; Takayama, Y.; et al. Single-subject gray matter networks in temporal lobe epilepsy patients with hippocampal sclerosis. *Epilepsy Res.* **2021**, *177*, 106766. [CrossRef]
25. Chen, S.; Zhang, J.; Ruan, X.; Deng, K.; Zhang, J.; Zou, D.; He, X.; Feng, L.; Bin, G.; Zeng, H.; et al. Voxel-based morphometry analysis and machine learning based classification in pediatric mesial temporal lobe epilepsy with hippocampal sclerosis. *Brain Imaging Behav.* **2020**, *14*, 1945–1954. [CrossRef]
26. Riederer, F.; Seiger, R.; Lanzenberger, R.; Pataraia, E.; Kasprian, G.; Michels, L.; Beiersdorf, J.; Kollias, S.; Thomas, C.; Johannes, H.; et al. Voxel-Based Morphometry—From Hype to Hope. A Study on Hippocampal Atrophy in Mesial Temporal Lobe Epilepsy. *AJNR Am. J. Neuroradiol.* **2020**, *41*, 987–993. [CrossRef]
27. Uher, D.; Drenthen, G.S.; Schijns, O.E.; Colon, A.J.; Hofman, P.A.; van Lanen, R.H.; Backes, W.H. Advances in Image Processing for Epileptogenic Zone Detection with MRI. *Radiology* **2023**, *307*, 220927. [CrossRef]
28. Gupta, S.; Razdan, R.; Hanumanthu, R.; Tomycz, L.; Ghesani, N.; Pak, J.; Kannurpatti, S.S. MRI based composite parameter of multiple tissue types for improved patient-level hemispheric and regional level lateralization in pediatric epilepsy. *Magn. Reson. Imaging* **2022**, *94*, 174–180. [CrossRef]
29. Fallahi, A.; Pooyan, M.; Habibabadi, J.M.; Nazem-Zadeh, M.R. Comparison of multimodal findings on epileptogenic side in temporal lobe epilepsy using self-organizing maps. *Magn. Reson. Mater. Phys. Biol. Med.* **2021**, *35*, 249–266. [CrossRef]
30. Jing, H.; Takigawa, M.; Benasich, A.A. Relationship of nonlinear analysis, MRI and SPECT in the lateralization of temporal lobe epilepsy. *Eur. Neurol.* **2022**, *48*, 11–19. [CrossRef]

31. The MathWorks Inc. Deep Learning Toolbox (R2021a). Available online: https://www.mathworks.com (accessed on 12 April 2023).
32. Ficici, C. MRI and EEG Data Correlation for Epilepsy Detection. Ph.D. Thesis, Electrical and Electronics Engineering Department, Ankara University, Ankara, Türkiye, 2022.
33. Chan, S.; Erickson, J.K.; Yoon, S.S. Limbic system abnormalities associated with mesial temporal sclerosis: A model of chronic cerebral changes due to seizures. *Radiographics* **1997**, *17*, 1095–1110. [CrossRef]
34. Ashburner, J. A fast diffeomorphic image registration algorithm. *NeuroImage* **2007**, *38*, 95–113. [CrossRef]
35. Ashburner, J.; Friston, K.J. Unified segmentation. *NeuroImage* **2005**, *26*, 839–851. [CrossRef]
36. GitHub. Available online: https://github.com/otoolej/eeg_viewer (accessed on 1 February 2023).
37. Beheshti, I.; Sone, D.; Maikusa, N.; Kimura, Y.; Shigemoto, Y.; Sato, N.; Matsuda, H. Accurate lateralization and classification of MRI-negative 18F-FDG-PET-positive. *Comput. Biol. Med.* **2021**, *137*, 104805. [CrossRef]
38. Nazem-Zadeh, M.R.; Schwalb, J.M.; Elisevich, K.V.; Bagher-Ebadian, H.; Hamidian, H.; Akhondi-Asl, A.R.; Jafari-Khouzani, K.; Soltanian-Zadeh, H. Lateralization of temporal lobe epilepsy using a novel uncertainty analysis of MR diffusion in hippocampus, cingulum, and fornix, and hippocampal volume and FLAIR intensity. *J. Neurol. Sci.* **2014**, *342*, 152–161. [CrossRef]
39. Jafari-Khouzani, K.; Elisevich, K.; Patel, S.; Smith, B.; Soltanian-Zadeh, H. FLAIR signal and texture analysis for lateralizing mesial temporal lobe epilepsy. *NeuroImage* **2010**, *49*, 1559–1571. [CrossRef]
40. Türk, Ö.; Özerdem, M.S. Epilepsy detection by using scalogram based convolutional neural network from EEG signals. *Brain Sci.* **2019**, *9*, 115. [CrossRef]
41. Daoud, H.; Bayoumi, M. Deep learning approach for epileptic focus localization. *IEEE Trans. Biomed. Circuits Syst.* **2019**, *S14*, 209–220. [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.

diagnostics

MDPI

Article

Diagnostic Accuracy of a Multi-Target Artificial Intelligence Service for the Simultaneous Assessment of 16 Pathological Features on Chest and Abdominal CT

Valentin A. Nechaev [1], Nataliya Y. Kashtanova [1], Evgenii V. Kopeykin [1], Umamat M. Magomedova [1], Maria S. Gribkova [2], Anton V. Hardin [1], Marina I. Sekacheva [3], Varvara D. Sanikovich [3], Valeria Y. Chernina [4,5] and Victor A. Gombolevskiy [3,5,6,7,*]

1 S.S. Yudin City Clinical Hospital, Moscow Department of Health, Moscow 129090, Russia; dfkz2005@gmail.com (V.A.N.)
2 Multidisciplinary Medical Center of the Bank of Russia, Moscow 117647, Russia
3 World-Class Reserch Centre "Digital Biodesign and Personalized Healthcare", I.M. Sechenov First Moscow State Medical University (Sechenov University), Moscow 119991, Russia
4 Intelligent Radiology Assistance Laboratories LLC (IRA LABS), Moscow 143026, Russia
5 Northern State Medical University, Arkhangelsk 163000, Russia
6 AIRI, Moscow 105064, Russia
7 Russian Medical Academy of Continuous Professional Education, Moscow 125993, Russia
* Correspondence: gombolevskiy@airi.net

Abstract: Background/Objectives: Chest, abdominal, and pelvic computed tomography (CT) with intravenous contrast is widely used for tumor staging, treatment planning, and therapy monitoring. The integration of artificial intelligence (AI) services is expected to improve diagnostic accuracy across multiple anatomical regions simultaneously. We aimed to evaluate the diagnostic accuracy of a multi-target AI service in detecting 16 pathological features on chest and abdominal CT images. **Methods**: We conducted a retrospective study using anonymized CT data from an open dataset. A total of 229 CT scans were independently interpreted by four radiologists with more than 5 years of experience and analyzed by the AI service. Sixteen pathological features were assessed. AI errors were classified as minor, intermediate, or clinically significant. Diagnostic accuracy was evaluated using the area under the receiver operating characteristic curve (AUC). **Results**: Across 229 CT scans, the AI service made 423 errors (11.5% of all evaluated features, n = 3664). False positives accounted for 262 cases (61.9%) and false negatives for 161 (38.1%). Most errors were minor (62.9%) or intermediate (31.7%), while clinically significant errors comprised only 5.4%. The overall AUC of the AI service was 0.88 (95% CI: 0.87–0.89), compared with 0.78–0.81 for radiologists. For clinically significant findings, the AI AUC was 0.90 (95% CI: 0.71–1.00). Diagnostic accuracy was unsatisfactory only for urolithiasis. **Conclusions**: The multi-target AI service demonstrated high diagnostic accuracy for chest and abdominal CT interpretation, with most errors being clinically negligible; performance was limited for urolithiasis.

Keywords: artificial intelligence; computer vision; computed tomography; chest; abdomen

1. Introduction

Artificial intelligence (AI) technologies have introduced novel opportunities in radiology, including enhanced diagnostic accuracy, workflow optimization, and the advancement of medical research [1,2]. The integration of AI algorithms into clinical practice is necessitated by the steadily increasing volume of imaging studies, the persistent shortage of

radiologists, and the ongoing demand for greater diagnostic precision in imaging modalities [1,2]. Among these modalities, contrast-enhanced computed tomography (CT) of the chest, abdomen, and pelvis remains one of the most accessible and accurate techniques. CT is recommended for the assessment of tumor extent, treatment planning, and evaluation of therapeutic efficacy and is therefore considered an indispensable component of comprehensive patient management [3]. Nonetheless, radiology reports based on chest CT [4] and abdominal CT [5] are frequently subject to diagnostic errors, most commonly false-negative findings. In this context, the implementation of multipurpose AI-based systems capable of detecting pathological changes across multiple anatomical regions concurrently appears highly relevant. Such systems not only have the potential to reduce diagnostic error rates but may also contribute to mitigating the increasing clinical workload of radiologists [6]. However, the diagnostic performance of AI tools requires rigorous evaluation, including systematic assessment of potential sources of error and identification of limitations that may constrain their widespread clinical adoption [2].

2. Materials and Methods

2.1. The Study Design

This single-center retrospective diagnostic accuracy study followed the CLAIM and STARD guidelines [7,8]. The study workflow is illustrated in Figure 1. Prior to the analysis, dataset preparation and radiologist training were performed according to predefined inclusion and exclusion criteria.

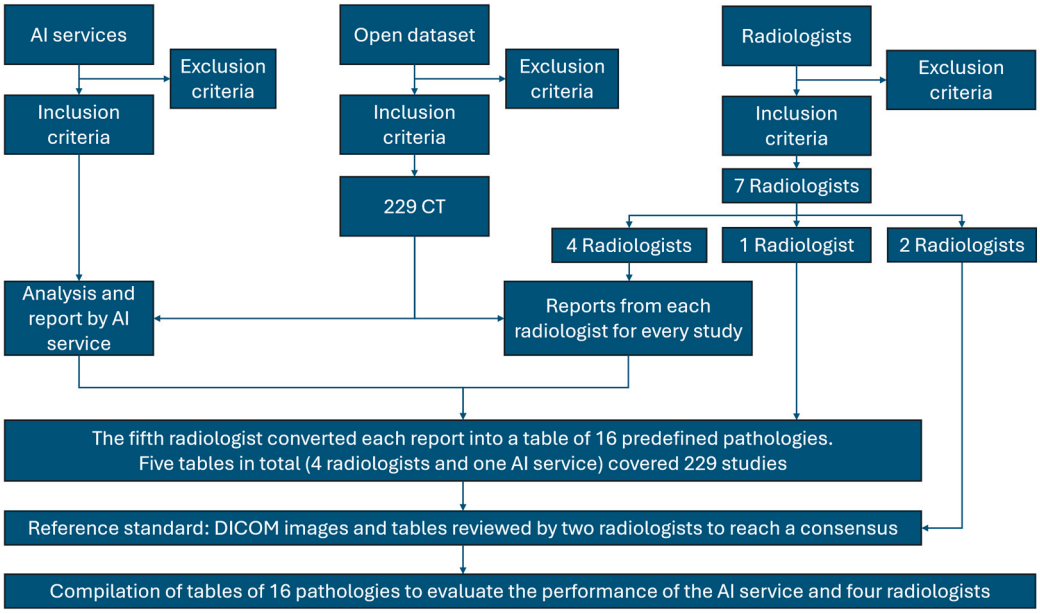


Figure 1. Study Design. From 250 CT examinations screened, 21 were excluded for technical/protocol deviations, yielding 229 studies. Four radiologists independently interpreted each study; one compiler radiologist converted all reports (4 radiologists + AI) into standardized tables; two expert referees reviewed original DICOM images and tables to establish the consensus reference standard. Performance of AI and radiologists was then evaluated against this reference.

2.2. Study Registration

This retrospective diagnostic accuracy study analyzed publicly available, anonymized CT data and therefore did not require separate protocol registration (e.g., ClinicalTrials.gov). Ethical approval was obtained from the local ethics committee of Moscow City Clinical Hospital No. 1. (protocol dated 1 March 2024). The BIMCV-COVID-19+ dataset had prior approval from the Hospital Arnau de Vilanova ethics committee (CElm 12/2020, Valencia,

Spain) and was funded through regional and EU Horizon 2020 grants. All data were anonymized before release, so informed consent was waived.

2.3. Data Source

Anonymized CT scans were acquired in 2020. Retrospective evaluation by radiologists and the AI system was carried out between 14 March 2024, and 2 November 2024. All CT examinations originated from the publicly available BIMCV-COVID-19+ dataset (Valencia Region, Spain), collected in 2020 from 11 public hospitals and standardized to UMLS terminology [9].

2.3.1. Inclusion Criteria

Adult chest and abdominal CT images; slice thickness $\leq$ 1 mm; scan coverage extending from the lung apices to the ischial bones, acquired during deep inspiratory breath-hold.

2.3.2. Exclusion Criteria

CT studies with protocol deviations (slice thickness > 1 mm or incomplete anatomic coverage), severe motion or beam-hardening artifacts, non-standard patient positioning, or upload/parse failures. Studies with corrupted DICOM tags or missing key series were also excluded.

2.4. Data Preprocessing

No additional data preprocessing was applied beyond standard DICOM parsing of the publicly available BIMCV-COVID-19+ dataset; images were analyzed as provided.

2.5. Data Partitions

2.5.1. Assignment of Data to Partitions

All 229 eligible CT examinations were used solely as an independent test set for the locked AI system. No additional training or validation split was performed.

2.5.2. Level of Disjointness Between Partitions

Each examination (one patient) was treated as a single unit of analysis; no overlap existed between cases.

2.6. Intended Sample Size

The required sample size was calculated a priori as 236 examinations to estimate sensitivity and specificity with 95% confidence and $\pm$10% error. To account for possible data loss (artifacts, upload failures, or absent non-contrast series), 250 CTs were selected; after exclusions, 229 remained for final analysis.

2.7. De-Identification Methods

All CT examinations were provided as part of the publicly available BIMCV-COVID-19+ dataset, which had been fully anonymized by the data provider before release. Personal identifiers—including patient name, date of birth, medical record numbers, and examination dates—were removed from DICOM headers. Only anonymous study IDs linking imaging data to accompanying radiology reports were retained for analysis. The de-identification procedure of the BIMCV dataset was reviewed and approved by the local ethics committee (CElm: 12/2020, Valencia, Spain).

2.8. Handling of Missing Data

Studies lacking mandatory series (e.g., contrast-enhanced scans without a corresponding non-contrast series) or with missing annotations/upload failures were treated as tech-

nical exclusions. No additional data imputation or image reconstruction was performed; such cases were not included in endpoint analyses.

2.9. Image Acquisition Protocol

Detailed scanning parameters can be obtained from the original publicly available dataset [9]. Examinations were performed on multislice CT scanners routinely used in hospitals of the Valencia region. Slice thickness was $\leq$1 mm, and coverage extended from the lung apices to the ischial bones. Scans were obtained during deep-inspiration breath-hold. Both non-contrast and contrast-enhanced studies were included; contrast-enhanced examinations lacking a non-contrast series were excluded. Detailed scanner models, reconstruction kernels, and exposure parameters were not available in the public dataset; all scans followed standard clinical protocols for chest–abdomen–pelvis CT in the region.

2.10. Human Readers

Seven radiologists participated in the study. Inclusion criteria comprised at least three years of experience in interpreting chest and abdominal CT scans. Exclusion criteria included failure to complete calibration or training according to the study protocol. The radiologists were assigned to three roles: annotators (n = 4), compiler (n = 1), and referees (n = 2). The annotators, all board-certified radiologists with 5–8 years of experience, independently interpreted all CT examinations using RadiAnt DICOM Viewer 2023.1 outside their routine clinical duties. Prior to the study, all readers attended a short calibration session with sample clinical cases to standardize interpretation criteria. Case order was randomized individually, and readers were blinded to AI outputs, clinical data, and one another's results. For each of the 16 predefined pathologies, findings were recorded for ROC AUC analysis. The compiler (8 years of experience) reviewed all radiologist and AI reports to compile structured tables containing the same 16 pathologies, resulting in five tables (four radiologists and one AI system) covering 229 CT studies. The referees (each with over 8 years of experience) independently reviewed DICOM images and resolved discrepancies by consensus, after which they were granted access to all five result tables for comparative evaluation.

2.11. Annotation Workflow

Each of the six radiologists (annotators, n = 4; referees, n = 2) reviewed every CT slice of all 229 examinations to ensure comprehensive assessment. The compiler (n = 1) subsequently analyzed the outputs from the four annotators and from the AI system (IRA LABS AI service), which processed the same 229 CT studies and produced both DICOM SEG annotations and DICOM SR structured reports. For each study, the compiler registered the presence or absence (binary classification) of all 16 predefined pathologies.

2.12. Reference Standard

The reference standard for performance evaluation was the consensus of two senior radiologists (>8 years' experience) who were not involved in the initial readings. They independently reviewed all CT examinations without access to AI outputs or initial reader reports using RadiAnt DICOM Viewer 2023.1. Disagreements were resolved by consensus or, if unresolved, by a third adjudicator. Formal inter-reader variability was not a primary objective; however, prior work highlights substantial variability in CT reporting [10].

Expert annotations for all 229 examinations were documented in a standardized table covering 16 predefined pathological features. Performance comparisons were made among: (1) initial annotations from four radiologists, (2) AI system outputs, and (3) the expert consensus reference standard.

For exploratory error analysis only, the same experts re-examined cases after reviewing AI outputs to categorize AI detections and errors; these post-AI reviews were not used to generate reference standard metrics.

2.13. Model

2.13.1. Model Description

A multipurpose AI service (IRA LABS, registered medical device RU №2024/22895) was used for simultaneous detection of 16 predefined pathologies on chest–abdominal CT (Table 1).

Table 1. Thresholds for pathologic changes assessed by the multi-target AI service. Thresholds were defined according to institutional methodological guidance [11,12].

Abnormalities	Thresholds
Pulmonary nodules	Presence of at least one pulmonary nodule or lesion larger than 6 mm in short-axis diameter
Airspace opacities (including consolidations/infiltrates)	Presence of any size/volume
Emphysema	Presence of any volume
Aortic dilatation/aneurysm	For aortic dilatation, the threshold diameters were defined as follows: ≥40 mm for the ascending aorta and aortic arch, ≥30 mm for the descending thoracic aorta, ≥25 mm for the abdominal aorta. For aortic aneurysm, the threshold was defined as a diameter > 55 mm in any segment.
Pulmonary artery dilatation	>30 mm in diameter
Coronary artery calcium	Agatston score > 1
Enlarged intrathoracic lymph nodes	Presence of at least one intrathoracic lymph node enlarged to >15 mm in short-axis diameter
Adrenal thickening	Presence of thickening >10 mm
Urolithiasis	Presence of at least one urinary calculus
Rib fractures	Presence of at least one fracture
Low vertebral body density	Attenuation < +150 HU
Vertebral compression fractures	Vertebral body deformity > 25% (Genant grade 2)

The final release version (v6.1, January 2024) identical to the one deployed in clinical practice was applied without retraining or parameter changes. Input consisted of DICOM CT series; output was generated as DICOM SEG annotations and DICOM SR structured reports.

2.13.2. AI Service Inclusion and Exclusion Criteria

Inclusion Criteria

Software registered as a certified medical device (MD) utilizing artificial intelligence (AI) technology in the official national registry of medical software. Software tested and validated within the Moscow Experiment—a large-scale governmental initiative for clinical

deployment of computer vision AI in radiology [13]. Demonstrated diagnostic performance with ROC AUC $\geq$ 0.81 for each target pathology, in accordance with methodological recommendations [11,12]. Capability to analyze both chest and abdominal CT examinations within a single inference pipeline.

Exclusion Criteria

AI products whose participation in the Moscow Experiment was suspended, discontinued, or failed official performance verification [13]. Systems limited to single-region analysis (e.g., chest-only AI) or lacking multiclass pathology detection capability.

The IRA LABS AI service (version 6.1, January 2024) was selected because it fulfilled all inclusion criteria and provided the broadest pathology coverage among eligible AI services participating in the Moscow Experiment [13].

2.13.3. Software and Environment

The proprietary system was executed as an off-the-shelf product. Internal architecture, model parameters, and potential ensemble methods are not publicly disclosed by the developer. Inference was run on a workstation with AMD Ryzen 7 7700, 64 GB RAM, 480 GB SSD, and NVIDIA RTX 4060 (8 GB) GPU, using Ubuntu 22.04.

2.13.4. Initialization of Model Parameters

The pre-trained production version of the AI model was used as released by the developer. No fine-tuning, weight reinitialization, or hyperparameter modification was performed for this study. The operating point (decision thresholds) was the vendor's default, locked a priori and not tuned on the test set.

2.13.5. Training

Details of Training Approach

No additional training or fine-tuning was performed for this study. The AI service was applied as an off-the-shelf production version (v6.1, January 2024) identical to the clinically deployed release.

Method of Selecting the Final Model

The version used was previously chosen and locked by IRA LABS during its clinical validation within the Moscow Experiment on computer-vision technologies. No modifications were made for the present evaluation [13].

Ensembling Techniques

The developer has not disclosed whether internal ensemble methods were used. For this study, a single instance of the AI service was applied for CT analysis, providing DICOM SEG annotations and structured text reports.

2.14. Evaluation

2.14.1. Metrics

Diagnostic performance of the AI system and radiologists was quantified using the area under the ROC curve (AUC, 95% CI via DeLong) [14]. For each of the 16 pathologies, True positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) counts were recorded, and classification errors stratified as minor, intermediate, or major.

2.14.2. Robustness Analysis

AUC was additionally calculated for clinically significant findings only to assess robustness for critical pathologies.

2.14.3. Methods for Explainability

AI outputs were interpreted via DICOM SEG visualizations and structured text reports. Errors were cross-checked against the reference standard and categorized by clinical significance.

2.14.4. Data Independence

BIMCV-COVID-19+ had not been used in model training; all 229 CT studies were independent of the AI development data. Testing was limited to this single external dataset; no further external validation was available.

2.14.5. Comparison and Evaluation Methodology

Comparative analysis was performed using five structured 229 $\times$ 16 matrices (four human readers and one AI system), each containing binary presence/absence determinations for all predefined pathologies. These matrices were compared with the expert consensus reference standard established by the two referees to derive the metrics described above.

For pathologies with quantitative thresholds (Table 1), the AI system applied predefined anatomical cut-offs (e.g., $\geq$40 mm for ascending aorta) based on DICOM SR measurements, while radiologists relied on visual assessment and manual measurements.

2.14.6. Conversion of AI Outputs to Binary Labels

The fifth radiologist (compiler) converted the AI outputs into binary presence/absence labels for all 16 pathologies. Binary classification was based primarily on DICOM SR measurements and text findings, cross-referenced against thresholds specified in Table 1. DICOM SEG masks served as visual confirmation but were not used for independent volumetric analysis.

Example 1: Aortic dilatation/aneurysm The DICOM SR contained a quantitative measurement: "Ascending aorta diameter: 42 mm." According to the predefined threshold ($\geq$40 mm for ascending aorta, Table 1), this measurement was classified as "positive" for aortic dilatation. Cases with measurements below threshold were labeled "negative."

Example 2: Pulmonary nodules The DICOM SEG file displayed a 3D segmentation mask outlining a pulmonary nodule. The accompanying DICOM SR reported: "In the right lung, a 10 $\times$ 8 mm node (average size 9 mm according to Fleischner)." Since the threshold is $\geq$6 mm (Table 1), this case was classified as "positive." The compiler visually verified that the segmentation mask adequately encompassed the nodule before assigning the label.

2.14.7. Handling of Small Segmentations and Borderline Cases

Segmentations that partially extended beyond lesion boundaries were accepted if the DICOM SR measurement remained above the predefined threshold (Table 1) and the core pathology was adequately captured. Cases where segmentation contours substantially misrepresented the pathology (e.g., including adjacent structures or anatomically incorrect regions) were flagged as false positives during expert consensus review. No additional voxel-based or volumetric filtering beyond the diameter thresholds specified in Table 1 was applied.

2.15. Outcomes

2.15.1. Primary Outcome

TP, FP, TN, and FN for AI and human readers in detecting each of the 16 predefined pathologies (Table 1). All errors were stratified by clinical significance:

- Minor—no change in patient management or follow-up needed (examples: missed simple cysts < 5 mm, false-positive osteosclerosis misclassified as rib fracture).

- Intermediate—unlikely to affect primary disease treatment but requiring further testing or follow-up (examples: false-positive enlarged lymph nodes, over-detection of small pulmonary nodules).
- Major—likely to change treatment strategy or primary diagnosis (examples: missed liver/renal masses, missed intrathoracic lymphadenopathy suggestive of metastases).

Expert radiologists assigned these classifications during consensus review based on potential impact on clinical decision-making.

2.15.2. Secondary Outcomes

Area under the ROC curve (AUC) with 95% confidence intervals for each pathology and for the aggregated set, for both AI and radiologists.

Comparative analysis of AI versus radiologists using multi-reader multi-case (MRMC) methods.

Exploratory error review by experts after viewing AI outputs to categorize AI detections (not used as reference standard).

2.16. Sample Size Calculation

The minimum required sample size was estimated using the formula for a single proportion: $n = Z^2 \times P \times (1 - P)/d^2$, where Z = 1.96 (95% confidence), P = 0.81 (expected sensitivity/specificity), and d = 0.10 (margin of error). This yielded approximately 59 cases with pathology. Assuming an average disease prevalence of 25% across the evaluated pathologies, the total required sample size was $N = 59/0.25 \approx 236$ examinations. To compensate for potential data loss (artifacts, annotation errors) and the multifocal nature of the study (chest + abdomen), the planned sample was increased to 250. After exclusions, 229 studies remained for final analysis.

2.17. Statistical Analysis

Statistical analysis was performed using RStudio (version 2025.09.2+418; Posit Software, PBC, Boston, MA, USA) [15] with the *irr* [16] and *pROC* [17] packages. Data visualization was carried out with GraphPad Prism version 10.2.2 (GraphPad Software Inc., San Diego, CA, USA) [18]. Descriptive statistics were reported as absolute numbers (*n*) and proportions (%). Diagnostic performance was assessed using ROC analysis with calculation of AUC and 95% confidence intervals via DeLong's method [14]. Comparisons between AI and radiologists used multi-reader multi-case DBM/OR analysis (*RJafroc*) [19]. To control for multiple testing across 16 pathologies, Benjamini–Hochberg correction (q = 0.05) was applied. A two-sided $p < 0.05$ was considered statistically significant.

3. Results

3.1. Overall Diagnostic Performance

In 229 chest–abdominal CT examinations independently interpreted by four radiologists and the AI system, the AI system produced 423 errors (11.5% of all evaluated features). The prevalence of evaluated pathologies ranged from 9.6% (urolithiasis) to 70.7% (coronary artery calcium) (Table 2).

Table 2. Prevalence of Pathological Features in the Study Dataset (N = 229).

Pathology	Positive Cases (n)	Negative Cases (n)	Prevalence (%)
Coronary artery calcium	162	67	70.7
Renal lesions	148	81	64.6
Airspace opacities	144	85	62.9
Low vertebral density	126	103	55.0
Emphysema	98	131	42.8
Liver lesions	97	132	42.4
Pulmonary nodules	96	133	41.9
Pleural effusion	73	156	31.9
Adrenal thickening	69	160	30.1
Rib fractures	63	166	27.5
Enlarged intrathoracic lymph nodes	55	174	24.0
Pulmonary artery dilatation	54	175	23.6
Vertebral compression fractures	51	178	22.3
Epicardial fat	38	191	16.6
Aortic dilatation/aneurysm	32	197	14.0
Urolithiasis	22	207	9.6

False positives predominated (n = 262; 61.9%) over false negatives (n = 161; 38.1%), whereas radiologists showed more false negatives (470–562) than false positives (5–27) (Figure 2). Contingency tables (TP, FP, TN, FN) for each pathology across all features, 4 radiologists, and the AI are provided in Supplementary Files (Tables S1 and S2).

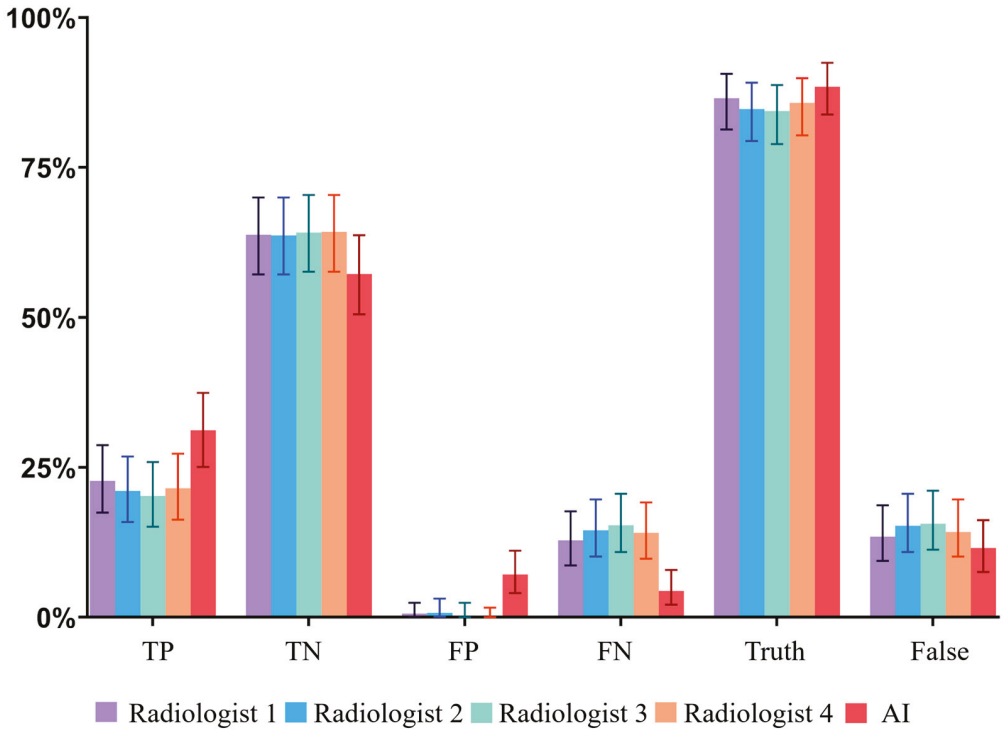


Figure 2. Bar charts showing the relative numbers of TP, TN, FP, FN, true (TP + TN) and false (FP + FN) responses; whiskers = 95% CI. AI exhibited a more liberal operating point (higher FP, lower FN) than individual radiologists.

3.2. Clinical Significance of Errors

All errors were stratified by clinical significance. For both radiologists and AI, minor errors were most frequent, followed by intermediate, with major errors being rare (Figure 3).

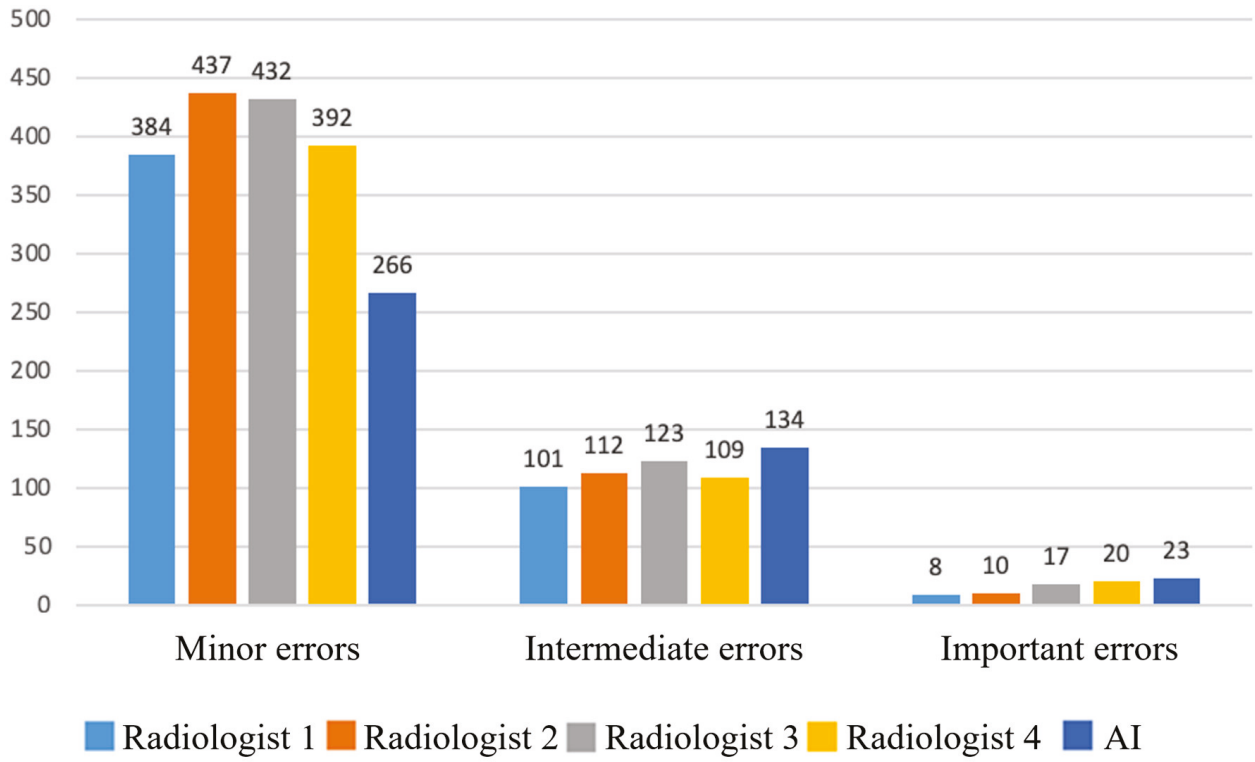


Figure 3. Absolute numbers of false responses (FP + FN) by clinical significance for physicians and AI. Only 5.4% of AI errors were classified as important; ≈0.63% of all assessed instances represented clinically important AI errors.

3.3. Breakdown of Clinically Significant AI Errors

Major errors (false negatives, $n = 20$):

- Liver lesions (8);
- Renal lesions (2);
- Adrenal lesions (2);
- Impaired lung aeration (atelectasis, 2);
- Enlarged intrathoracic lymph nodes (3);
- Pulmonary nodule (1);
- Low vertebral body density (1);
- Urolithiasis (1) (Figure 4).

Intermediate errors (mostly false positives, $n = 91$):

- Intrathoracic lymph nodes (16);
- Pulmonary nodules (15);
- Impaired aeration (15);
- Aortic dilatation/aneurysm (10);
- Adrenal thickening (10) (Figure 5).

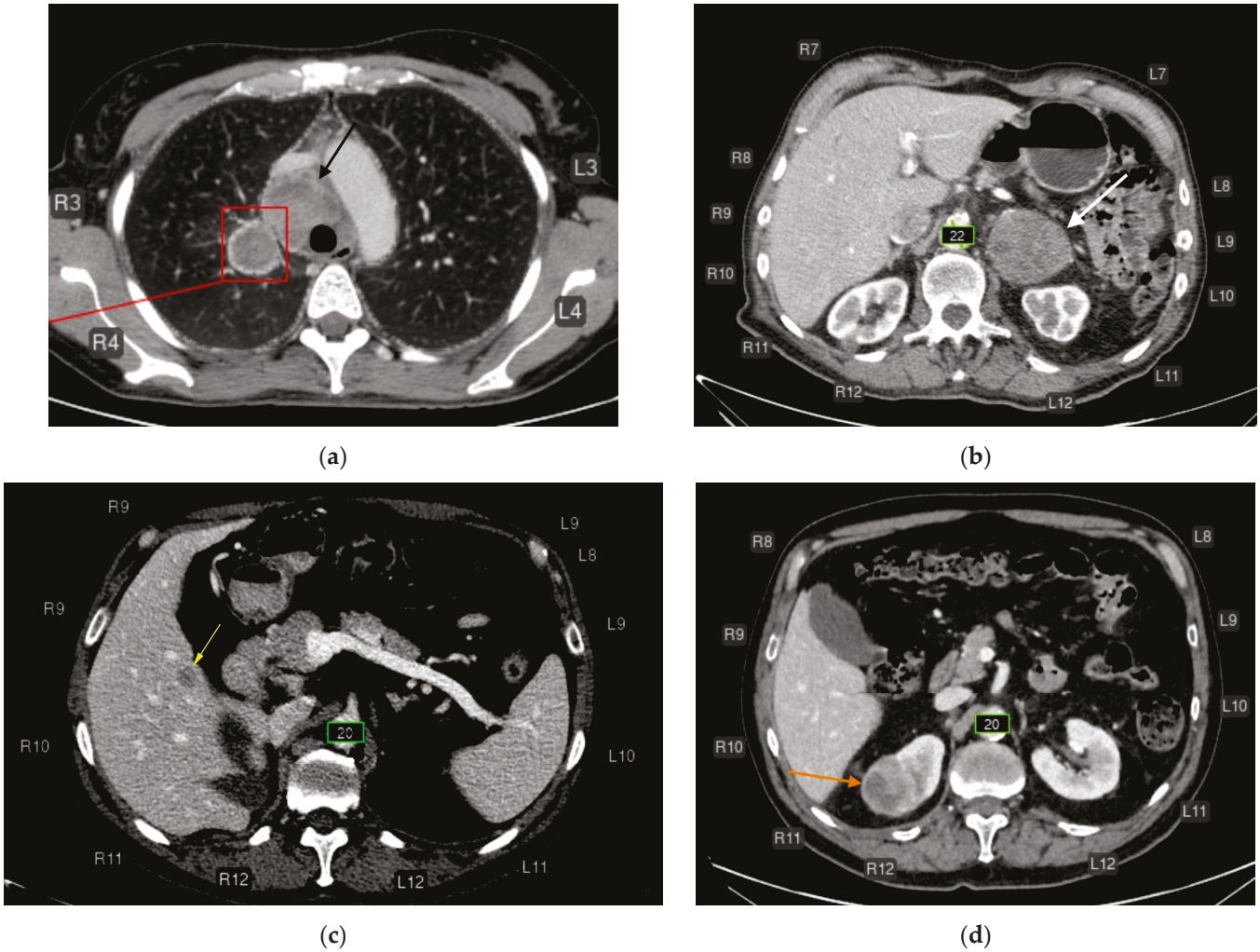


Figure 4. CT of the chest (**a**) and abdomen (**b**–**d**) after AI processing. Clinically significant findings missed by the AI that could substantially affect patient outcomes (false-negative errors). No segmentation (missed pathology): (**a**) Enlarged intrathoracic lymph nodes (black arrow), possibly representing metastases. False-negative error by the AI service. In addition, the AI correctly labeled the pulmonary lesion and rib numbering. (**b**) Left adrenal mass (white arrow), possibly a tumor or metastasis. False-negative error by the AI service. The AI correctly measured the abdominal aortic diameter and labeled rib numbering. (**c**) Hypovascular liver mass (yellow arrow), possibly a metastasis. False-negative error by the AI service. The AI correctly measured the abdominal aortic diameter and labeled rib numbering. (**d**) Right renal mass (orange arrow), possibly a tumor. False-negative error by the AI service. The AI correctly measured the abdominal aortic diameter and labeled rib numbering.

Minor errors (n = 266):

- Missed small simple cysts (<5 mm) in kidney or liver (70 cases) (Figure 6a);
- Incorrect segmentation of rib fractures (32 cases) misclassified from bone islands, artifacts, or costal cartilage transitions (Figure 6b).

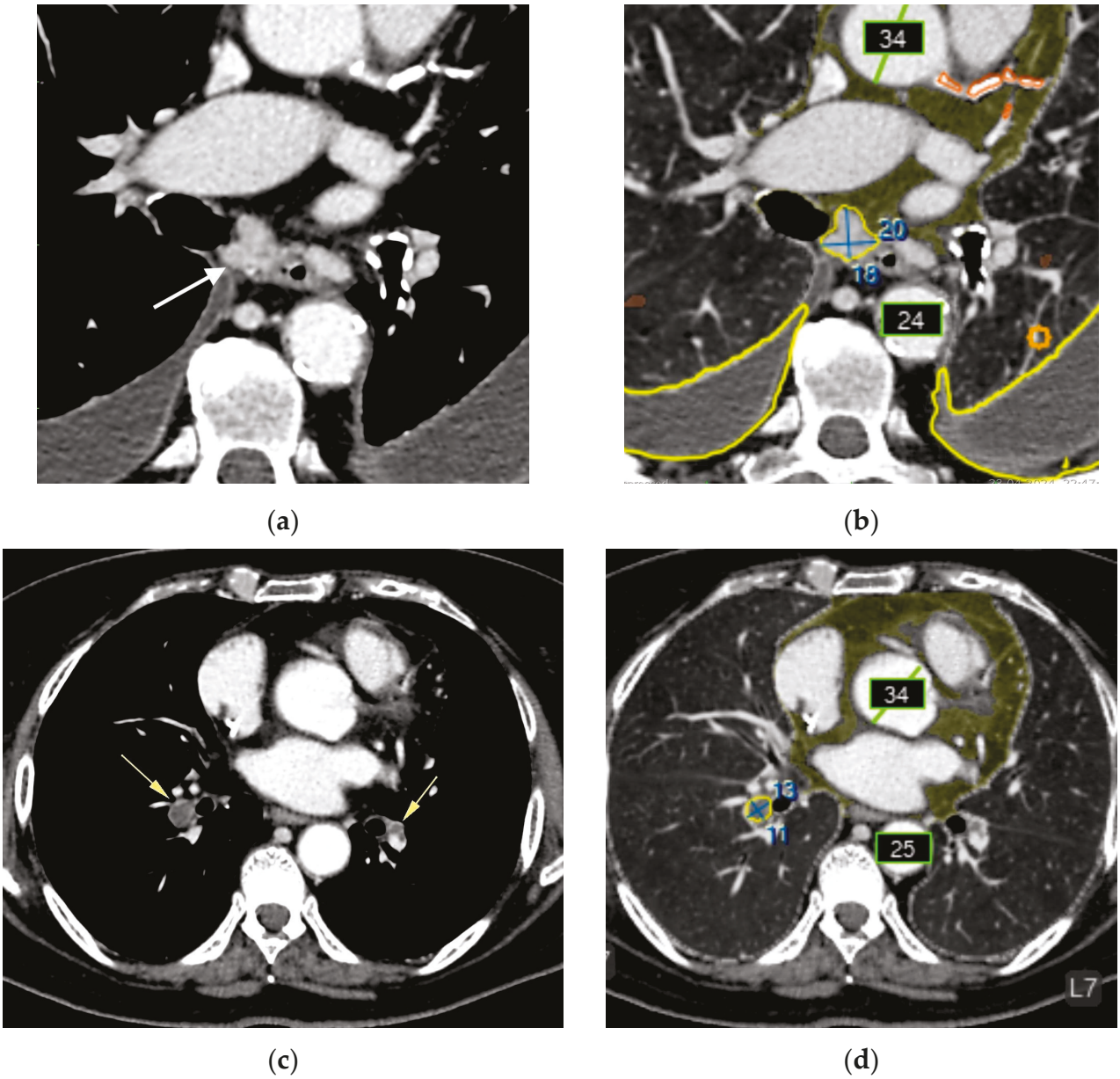
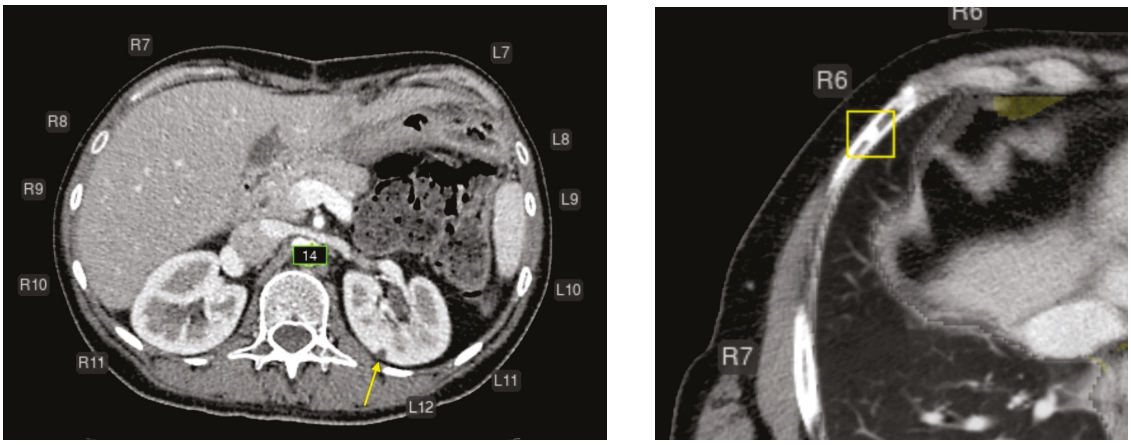


Figure 5. Chest CT before (**a** ,**c**) and after AI processing (**b** ,**d**). Intermediate errors. (**a** ,**c**) Two adjacent, non-enlarged bifurcation lymph nodes (white arrow) are segmented by the AI as a single enlarged node (false-positive error). Although a radiologist can recognize this error, it may be challenging for less experienced readers. The AI service correctly identified bilateral pleural effusion (blue contour), measured the diameters of the ascending and descending thoracic aorta (green), coronary calcifications (orange), paracardial fat (dark green), emphysema (brown), and a small focus of pulmonary infiltration associated with COVID-19. (**b** ,**d**) A thrombus in the right pulmonary artery (yellow arrow) is misclassified as enlarged intrathoracic lymph nodes (false-positive error). The AI was not trained to detect pulmonary artery thrombi but was trained to identify enlarged lymph nodes in this region. Due to limited training examples containing both lymphadenopathy and thrombi, the model misinterpreted the lesion as lymph node enlargement. Nevertheless, the detected abnormality could prompt a radiologist to recognize a potential thrombus. The AI correctly measured the diameters of the ascending and descending thoracic aorta (green), coronary calcifications (orange), and paracardial fat (dark green).



(**a**) (**b**)

Figure 6. Axial CT of the abdomen (**a**) and chest (**b**) after AI processing. Missed findings without clinical significance. (**a**) A simple cyst in the left kidney measuring approximately 4–5 mm (yellow arrow) was not segmented (false-negative error). This is most likely a simple cyst that does not require follow-up. The AI correctly measured the abdominal aortic diameter (green) and labeled rib numbering. (**b**) A focal area of osteosclerosis in the anterior segment of the right 6th rib was highlighted as a consolidated fracture (yellow box; false-positive error). This finding does not require follow-up. The AI correctly labeled rib numbering and measured the aortic diameter.

3.4. ROC Analysis and AUC Values

AUCs were calculated for each pathology and in aggregate for radiologists and AI (Table 3).

Table 3. Area under the ROC curve for diagnostic performance of four radiologists and AI.

Pathology (Number of True Positive Findings from 229)	AUC [95% CI]				
	Radiologist 1	**Radiologist 2**	**Radiologist 3**	**Radiologist 4**	**AI**
Enlarged intrathoracic lymph nodes (55)	0.952 [0.916–0.989]	0.964 [0.932–0.996]	0.891 [0.836–0.946]	0.936 [0.892–0.981]	0.854 [0.803–0.904]
Aortic dilatation/aneurysm (32)	0.681 [0.593–0.769]	0.567 [0.505–0.629]	0.567 [0.505–0.629]	0.55 [0.495–0.605]	0.947 [0.926–0.969]
Vertebral compression fractures (51)	0.6 [0.544–0.656]	0.55 [0.508–0.592]	0.56 [0.515–0.605]	0.55 [0.508–0.592]	0.943 [0.913–0.972]
Coronary artery calcification (CAC) (162)	0.838 [0.784–0.893]	0.73 [0.665–0.794]	0.837 [0.788–0.885]	0.864 [0.825–0.903]	0.875 [0.824–0.926]
Lung nodules (96)	0.906 [0.867–0.945]	0.887 [0.845–0.929]	0.849 [0.803–0.895]	0.909 [0.87–0.948]	0.863 [0.818–0.909]
Urolithiasis (22)	0.818 [0.715–0.921]	0.773 [0.666–0.879]	0.795 [0.69–0.901]	0.773 [0.666–0.879]	0.523 [0.478–0.567]
Airspace opacities (infiltrates, consolidations) (144)	0.92 [0.89–0.95]	0.941 [0.915–0.967]	0.948 [0.923–0.973]	0.965 [0.944–0.986]	0.81 [0.757–0.862]
Liver masses (97)	0.969 [0.945–0.993]	0.918 [0.88–0.955]	0.852 [0.806–0.898]	0.928 [0.893–0.963]	0.793 [0.741–0.846]
Renal masses (148)	0.964 [0.94–0.988]	0.949 [0.924–0.974]	0.949 [0.924–0.974]	0.967 [0.945–0.989]	0.778 [0.732–0.824]
Rib fractures (63)	0.574 [0.529–0.619]	0.574 [0.529–0.619]	0.557 [0.517–0.598]	0.541 [0.506–0.576]	0.899 [0.868–0.929]

Table 3. *Cont.*

Pathology (Number of True Positive Findings from 229)	AUC [95% CI]				
	Radiologist 1	Radiologist 2	Radiologist 3	Radiologist 4	AI
Pleural effusion (73)	0.942 [0.905–0.979]	0.952 [0.918–0.986]	0.952 [0.918–0.986]	0.938 [0.9–0.976]	0.929 [0.9–0.959]
Pulmonary artery dilatation (54)	0.574 [0.526–0.622]	0.565 [0.52–0.61]	0.528 [0.497–0.559]	0.519 [0.493–0.544]	0.959 [0.929–0.988]
Low vertebral body density (126)	0.504 [0.496–0.513]	0.513 [0.498–0.528]	0.504 [0.496–0.513]	0.509 [0.497–0.521]	0.9 [0.863–0.937]
Adrenal thickening (69)	0.848 [0.793–0.903]	0.75 [0.691–0.81]	0.726 [0.666–0.786]	0.768 [0.709–0.827]	0.849 [0.795–0.902]
Emphysema (98)	0.738 [0.687–0.79]	0.704 [0.654–0.755]	0.742 [0.691–0.793]	0.78 [0.729–0.83]	0.941 [0.914–0.968]
All pathologies (1328)	0.815 [0.802–0.828]	0.790 [0.777–0.804]	0.782 [0.769–0.796]	0.801 [0.788–0.814]	0.883 [0.872–0.894]

Aggregate AUC: AI = 0.88; highest radiologist = 0.81 (Figure 7).

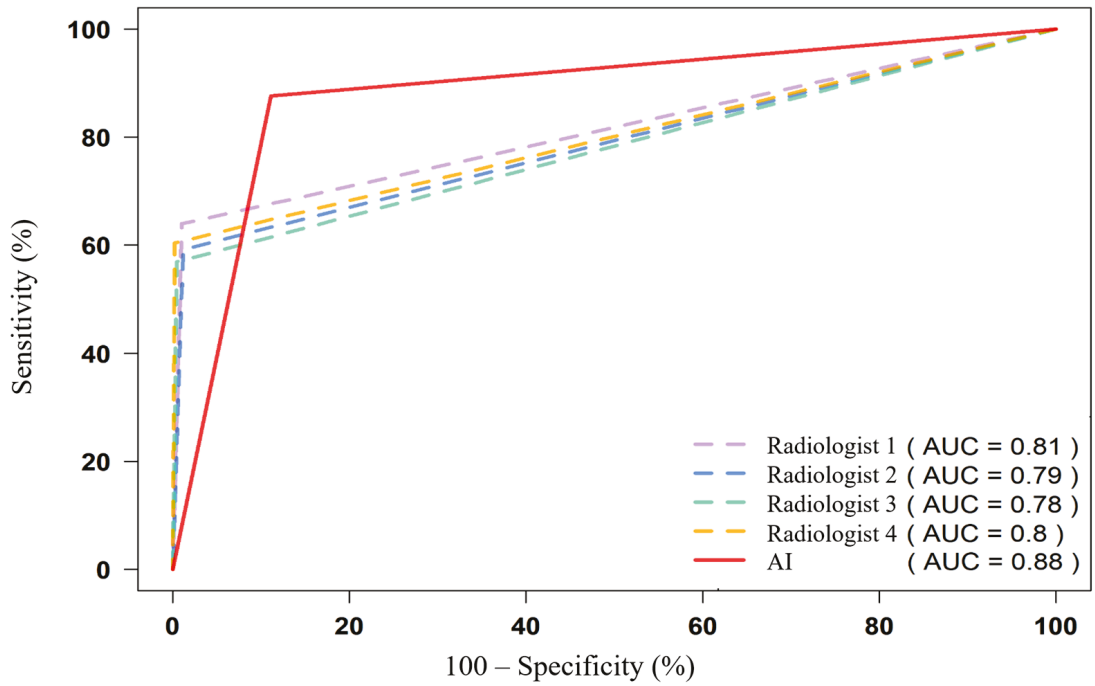


Figure 7. ROC curves comparing radiologists and AI for all features. The AI shows excellent discrimination for vascular, osseous, and emphysema targets, with lower performance on urolithiasis and small solid-organ lesions.

Clinically significant findings only: mean AI AUC = 0.90 (Table 4).

Table 4. AUC analysis for AI on clinically significant findings only.

Pathology	AUC [95% CI]
Airspace opacities (infiltrates, consolidations)	0.90 [0.85–0.95]
Emphysema	1.00 [0.95–1.00]
Lung nodules	0.97 [0.93–1.00]

Table 4. *Cont.*

Pathology	AUC [95% CI]
Enlarged intrathoracic lymph nodes	0.94 [0.89–0.99]
Pleural effusion	1.00 [0.98–1.00]
Aortic dilatation/aneurysm	1.00 [1.00]
Coronary artery calcification (CAC)	0.96 [0.69–1.00]
Adrenal thickening	0.95 [0.93–1.00]
Rib fractures	0.98 [0.89–1.00]
Vertebral compression fractures	0.73 [0.32–1.00]
Liver masses	0.87 [0.79–0.95]
Renal masses	0.86 [0.69–1.00]
Urolithiasis	0.55 [0.00–1.00]
Average value for all pathologies	0.9 [0.71–1.00]

3.5. Diagnostic Performance Categories

According to standard diagnostic categories, AI performance was predominantly excellent or good.

Urolithiasis was the only feature where AI showed inadequate performance.

Radiologists demonstrated poor or inadequate performance for several features, including:

- Aortic dilatation/aneurysm;
- Vertebral compression fractures;
- Rib fractures;
- Pulmonary artery dilatation;
- Low vertebral body density;
- Increased epicardial fat volume here.

4. Discussion

In this multi-reader evaluation of 229 chest and abdominal CT examinations comprising 3664 feature-level assessments, the multi-target AI service achieved an aggregate AUC of 0.88 (95% CI 0.87–0.89), outperforming the four independent radiologists (AUC 0.78–0.81) (Table 2 , Figure 7). Diagnostic performance was good to excellent for most of the 16 predefined targets. The AI demonstrated clear advantages in vascular, osseous, and morphometric findings, while showing relative deficits for solid-organ masses and airspace disease. The only unsatisfactory target was urolithiasis (AUC $\approx$ 0.52), which persisted in sensitivity analysis (AUC 0.55). Across modalities, commercial AI shows target-dependent performance, e.g., for airspace disease on chest radiographs [20].

Although the AI system produced more false positives than false negatives (61.9% vs. 38.1%), clinically important AI errors were rare—only 0.63% of all assessed instances—and were mainly missed focal lesions. Stratification of errors revealed that 94.6% of AI mistakes were minor or intermediate. Intermediate false positives often reflected adjacency or merging artifacts and vascular–nodal confusion, whereas minor errors were primarily tiny cysts and rib fracture overcalls. The AI's superiority in measuring diameters, densities, calcifications, and vertebral deformities likely stems from stable morphometric cues, whereas parenchymal textures and small solid-organ lesions remain more challenging. Its underperformance in urolithiasis plausibly relates to protocol sensitivity: many cases were contrast-enhanced only, whereas optimal stone detection requires non-contrast CT [21].

Although our final sample size (n = 229) was slightly below the a priori calculated requirement (n = 236), the impact on precision was minimal, increasing the margin of error from ±10.0% to ±10.2%. The narrow confidence intervals observed for most AUC estimates suggest adequate statistical power was maintained.

Our aggregate results align with recent meta-analyses [22,23] reporting that state-of-the-art imaging AI reaches AUC ≈ 0.86–0.94 and, in selected tasks, can match or surpass individual radiologists [23,24]. The complementary patterns of liberal AI (more FP) and conservative readers (more FN) suggest that hybrid strategies—such as triage or double-reading—may reduce important misses while managing FP burden [24]. Incorrect AI outputs can influence readers' decisions, underscoring the need for guardrails in hybrid workflows [25]. Similar FP–FN patterns have been synthesized in systematic reviews of AI error characteristics [26].

4.1. Strengths and Limitations

Key strengths include:

Multi-reader design with blinded interpretation and expert consensus reference standard;

Simultaneous evaluation of 16 diverse targets, enabling a comprehensive view of performance;

Stratification of errors by clinical significance, which provides insights beyond raw sensitivity and specificity;

Interpretive pitfalls and potential automation bias necessitate procedural safeguards and reader training [27].

Limitations include:

Single external dataset, limiting generalizability across protocols and institutions;

Protocol heterogeneity within BIMCV-COVID-19+ and incomplete non-contrast phases for some studies;

Lack of transparency regarding the proprietary model's architecture and potential ensembling strategies;

The perfect ROC AUC scores (1.00) reflect the small sample size of clinically significant cases available for analysis, rather than necessarily indicating superior model performance. These findings require validation with substantially larger datasets.

Translation from curated datasets to clinical applicability can vary across modalities and tasks [28];

Absence of formal inter-reader variability analysis and external validation on independent cohorts. Limitations of AI services observed in radiography evaluations further argue for protocol-aware validation [29].

Binary classification prioritized clinical relevance over volumetric precision. No quantitative segmentation assessment (e.g., Dice coefficient) was performed; minor imperfections were tolerated if DICOM SR measurements exceeded thresholds. This pragmatic approach may introduce measurement variability and warrants future voxel-level validation.

4.2. Public Health Implications

Use of multi-target CT AI may improve early detection of significant findings, speed up reporting, and optimize resource use—critical for systems facing radiologist shortages. Such deployment could enhance population-level outcomes and reduce costs from delayed diagnoses, but requires careful monitoring of error profiles, protocol harmonization, and adherence to evidence-based standards to ensure equitable, safe benefits.

4.3. Future Directions

Perform multi-center replication with protocol-aware validation to ensure robustness across scanners and acquisition techniques;

Explore operating point calibration or task-specific thresholds to optimize the FP–FN balance;

Develop targeted refinements for small solid-organ lesions, parenchymal textures, and urolithiasis detection;

Investigate workflow integration strategies, including triage, double reading, or AI-assisted decision support, to translate performance gains into clinical benefit.

Evaluate the economic and public health impact of multi-target AI deployment, including cost-effectiveness analyses, resource allocation, and potential reductions in population-level morbidity and healthcare expenditures.

5. Conclusions

A clinically deployed multi-target AI service demonstrated high diagnostic accuracy on chest and abdominal CT across 16 predefined features, outperforming individual radiologists on several vascular, osseous, and morphometric targets while underperforming on urolithiasis and small solid-organ lesions. Clinically important AI errors were rare and predominantly involved missed focal lesions. These findings support the use of multi-target AI as a complementary second reader, provided protocol alignment is ensured and error profiles are prospectively monitored. Future multi-center validations and workflow studies are warranted to confirm generalizability and define optimal integration strategies in routine practice.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics15212778/s1 , Table S1. Complete raw data from all four radiologists and the AI service across all 229 CT examinations. Contingency table analysis presenting true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), total correct responses (True), and total incorrect responses (False) with percentages and 95% confidence intervals (CI). Data encom-passes all 3,664 individual feature assessments (16 pathological features × 229 examinations). Each row represents the complete diagnostic performance record of each radiologist and the AI service for the entire study dataset, enabling di-rect comparison of detection patterns and error distributions across all readers. Table S2. Detailed raw diagnostic data for each of the 16 individual pathological features, stratified by radiologist and AI service. Contingency table values (TP, TN, FP, FN, True, False) with percentages and 95% CI provided separately for each pathology across all radiologists and AI service. This comprehensive breakdown represents reports at the pathology level, allowing detailed examination of feature-specific accuracy patterns, detection capabilities, and error characteriza-tion for both human readers and the artificial intelligence system across the entire study population.

Author Contributions: Conceptualization, V.A.G. and V.Y.C.; methodology, V.A.N., N.Y.K., E.V.K., U.M.M., V.Y.C. and V.A.G.; validation and formal analysis, V.A.N., N.Y.K., E.V.K. and U.M.M., M.S.G.; investigation and data curation, V.A.N., N.Y.K., E.V.K., U.M.M., A.V.H., V.D.S. and M.I.S.; writing—original draft preparation, V.A.G. and V.A.N.; writing—review and editing, all authors; visualization, V.A.N. and M.S.G.; supervision, V.Y.C. and V.A.G.; project administration, V.A.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by the authors' institutions.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Review Board of Moscow City Clinical Hospital No. 1 (Protocol No.1 on 1 March 2024). The original BIMCV-COVID-19+ dataset had prior ethical approval from the Hospital Arnau de Vilanova, Valencia (CElm: 12/2020).

Informed Consent Statement: Patient consent was waived because only de-identified, publicly available data were analyzed, and no individual could be identified from the materials used in this study.

Data Availability Statement: Publicly available data from the BIMCV-COVID-19+ initiative were analyzed in this study (https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/—accessed on 16 September 2025). Derived, de-identified result tables and R/GraphPad summaries are available from the corresponding author on reasonable request.

Acknowledgments: We thank the **BIMCV consortium** and the Valencia-region hospitals for curating and sharing the dataset, and our colleagues for valuable feedback on early drafts of this manuscript. Editorial assistance was provided using ChatGPT (GPT-5, OpenAI) for language refinement. The authors reviewed and edited all content and are fully responsible for the final version.

Conflicts of Interest: Several authors are employees or advisors of IRA LABS, the producer of the evaluated AI service. To address potential conflicts a priori, we (i) prespecified the evaluation protocol, eligibility criteria, endpoints, and statistical analysis plan; (ii) used a public dataset (BIMCV-COVID-19+) external to IRA LABS for all imaging; (iii) ensured that initial readings, discrepancy resolution, and the consensus reference standard were performed solely by radiologists who are not IRA LABS employees and were blinded to developer identities; (iv) instituted a data lock before manuscript drafting; (v) limited IRA LABS contributors' role to providing the off-the-shelf, commercially released AI service at its default operating point; and (vi) excluded IRA LABS personnel from re-reading or adjudicating AI outputs. The company had no influence on inclusion/exclusion of cases or results. All authors affirm that these safeguards were implemented and that the analyses faithfully represent the findings.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
CT	Computed tomography
AUC	Area under the ROC curve
ROC	Receiver operating characteristic
CI	Confidence interval
DICOM	Digital Imaging and Communications in Medicine
SR	Structured report
IRB	Institutional review board
APC	Article processing charge
MRMC	Multi-reader multi-case
HU	Hounsfield unit
TP	True positive
FP	False positive
TN	True negative
FN	False negative
CAC	Coronary artery calcification
UMLS	Unified Medical Language System
BIMCV-COVID-19+	Biomedical Imaging and COVID-19 dataset (Valencia region)

References

1. Katal, S.; York, B.; Gholamrezanezhad, A. AI in Radiology: From Promise to Practice—A Guide to Effective Integration. *Eur. J. Radiol.* **2024** , *181* , 111798. [CrossRef]
2. Buijs, E.; Maggioni, E.; Mazziotta, F.; Lega, F.; Carrafiello, G. Clinical Impact of AI in Radiology Department Management: A Systematic Review. *Radiol. Med.* **2024** , *129* , 1656–1666. [CrossRef] [PubMed]

3. Pokataev, I.A.; Dudina, I.A.; Kolomiets, L.A.; Morkhov, K.Y.; Nechushkina, V.M.; Rumyantsev, A.A.; Tyulyadin, S.A.; Urmancheeva, A.F.; Khokhlova, S.V. Ovarian Cancer, Primary Peritoneal Cancer, and Fallopian Tube Cancer: Practical Recommendations of RUSSCO, Part 1.2. *Malig. Tumors* **2024** , *14* , 82–101. (In Russian) [CrossRef]
4. Chernina, V.Y.; Belyaev, M.G.; Silin, A.Y.; Avetisov, I.O.; Pyatnitskiy, I.A.; Petrash, E.A.; Basova, M.V.; Sinitsyn, V.E.; Omelyanovskiy, V.V.; Gombolevskiy, V.A.; et al. Diagnostic and Economic Evaluation of a Comprehensive Artificial Intelligence Algorithm for Detecting Ten Pathological Findings on Chest Computed Tomography. *Diagnostics* **2023** , *4* , 105–132. (In Russian) [CrossRef]
5. Wildman-Tobriner, B.; Allen, B.C.; Maxfield, C.M. Common Resident Errors When Interpreting Computed Tomography of the Abdomen and Pelvis: A Review of Types, Pitfalls, and Strategies for Improvement. *Curr. Probl. Diagn. Radiol.* **2019** , *48* , 4–9. [CrossRef]
6. Nechaev, V.A.; Vasiliev, A.Y. Risk Factors for Perception Errors among Radiologists in the Analysis of Imaging Studies. *Vestnik Surgu. Med.* **2024** , *17* , 14–22. (In Russian) [CrossRef]
7. Mongan, J.; Moy, L.; Kahn, C.E., Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol. Artif. Intell.* **2020** , *2* , e200029. [CrossRef]
8. Bossuyt, P.M.; Reitsma, J.B.; E Bruns, D.; A Gatsonis, C.; Glasziou, P.P.; Irwig, L.; Lijmer, J.G.; Moher, D.; Rennie, D.; de Vet, H.C.W.; et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. *BMJ* **2015** , *351* , h5527. [CrossRef]
9. de la Iglesia Vayá, M.; Saborit-Torres, J.M.; Serrano, J.A.M.; Oliver-Garcia, E.; Pertusa, A.; Bustos, A.; Cazorla, M.; Galant, J.; Barber, X.; Orozco-Beltrán, D.; et al. BIMCV COVID-19: A Large Annotated Dataset of RX and CT Images from COVID-19 Patients. *IEEE Dataport* **2021**. [CrossRef]
10. Kulberg, N.S.; Reshetnikov, R.V.; Novik, V.P.; Elizarov, A.B.; Gusev, M.A.; Gombolevskiy, V.A.; Vladzymyrskyy, A.V.; Morozov, S.P. Inter-Observer Variability Between Readers of CT Images: All for One and One for All. *Digit. Diagn.* **2021** , *2* , 105–118. (In Russian) [CrossRef]
11. Gombolevskiy, V.A.; Masri, A.G.; Kim, S.Y.; Morozov, S.P. Manual for Radiology Technicians on Performing CT Examination Protocols. In *Methodological Guidelines No. 17*; SBHI "SPCC for Diagnostics and Telemedicine Tech."; Moscow Healthcare Department: Moscow, Russia, 2017; 56p. (In Russian)
12. Morozov, S.P.; Vladzimirskiy, A.V.; Klyashtorny, V.G. Clinical Trials of Software Based on Intelligent Technologies (Radiology). In *Best Practices in Radiology and Instrumental Diagnostics, Issue 23*; SBHI "SPCC for Diagnostics and Telemedicine Tech."; Moscow Healthcare Department: Moscow, Russia, 2019; 33p. (In Russian)
13. Vasiliev, Y.A. (Ed.) *Computer Vision in Radiology: The First Stage of the Moscow Experiment* , 2nd rev. and exp. ed.; Izdatelskie Resheniya: Moscow, Russia, 2023; 376p. (In Russian)
14. DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the Areas under Two or More Correlated ROC Curves: A Nonparametric Approach. *Biometrics* **1988** , *44* , 837–845. [CrossRef]
15. RStudio Team. RStudio (Posit). 2023. Available online: https://posit.co/ (accessed on 16 September 2025).
16. Gamer, M.; Lemon, J.; Fellows, I.; Singh, P. irr: Various Coefficients of Interrater Reliability and Agreement. R Package Version 0.84.1. Available online: https://CRAN.R-project.org/package=irr (accessed on 16 September 2025).
17. Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.-C.; Müller, M. pROC: An Open-Source Package to Analyze and Compare ROC Curves. *BMC Bioinform.* **2011** , *12* , 77. [CrossRef] [PubMed]
18. GraphPad Software. GraphPad Prism, Version 10.2.2. 2024. Available online: https://www.graphpad.com/ (accessed on 16 September 2025).
19. Chakraborty, D.P. *Observer Performance Methods for Diagnostic Imaging: Foundations, Modeling, and Applications with R-Based Examples*; CRC Press: Boca Raton, FL, USA, 2017.
20. Plesner, L.L.; Müller, F.C.; Brejnebøl, M.W.; Laustrup, L.C.; Rasmussen, F.; Nielsen, O.W.; Boesen, M.; Andersen, M.B. Commercially Available Chest Radiograph AI Tools for Detecting Airspace Disease, Pneumothorax, and Pleural Effusion. *Radiology* **2023** , *308* , e231236. [CrossRef] [PubMed]
21. Räty, P.; Mentula, P.; Lampela, H.; Nykänen, T.; Helanterä, I.; Haapio, M.; Lehtimäki, T.; Skrifvars, M.B.; Vaara, S.T.; Leppäniemi, A.; et al. Intravenous Contrast CT versus Native CT in Acute Abdomen with Impaired Renal Function (INCARO): Study Protocol. *BMJ Open* **2020** , *10* , e037928. [CrossRef]
22. Vasilev, Y.A.; Vladzimirskyy, A.V.; Omelyanskaya, O.V.; Reshetnikov, R.V.; Blokhin, I.A.; Kodenko, M.M.; Nanova, O.G. A Review of Meta-Analyses on the Application of AI in Radiology. *Med. Vis.* **2024** , *28* , 22–41. (In Russian) [CrossRef]
23. Aggarwal, R.; Sounderajah, V.; Martin, G.; Ting, D.S.W.; Karthikesalingam, A.; King, D.; Ashrafian, H.; Darzi, A. Diagnostic Accuracy of Deep Learning in Medical Imaging: A Systematic Review and Meta-Analysis. *npj Digit. Med.* **2021** , *4* , 65. [CrossRef]
24. Seah, J.C.Y.; Tang, C.H.M.; Buchlak, Q.D.; Holt, X.G.; Wardman, J.B.; Aimoldin, A.; Esmaili, N.; Ahmad, H.; Pham, H.; Lambert, J.F.; et al. Effect of a Comprehensive Deep-Learning Model on Chest X-Ray Interpretation Accuracy: A Retrospective MRMC Study. *Lancet Digit. Health* **2021** , *3* , e496–e506. [CrossRef]

25. Bernstein, M.H.; Atalay, M.K.; Dibble, E.H.; Maxwell, A.W.P.; Karam, A.R.; Agarwal, S.; Ward, R.C.; Healey, T.T.; Baird, G.L. Can Incorrect AI Results Impact Radiologists? A Multi-Reader Pilot Study of Lung Cancer Detection with Chest Radiography. *Eur. Radiol.* **2023** , *33* , 8263–8269. [CrossRef]
26. Zeng, A.; Houssami, N.; Noguchi, N.; Nickel, B.; Marinovich, M.L. Frequency and Characteristics of Errors by Artificial Intelligence (AI) in Reading Screening Mammography: A Systematic Review. *Breast Cancer Res. Treat.* **2024** , *207* , 1–13. [CrossRef]
27. Behzad, S.; Tabatabaei, S.M.H.; Lu, M.Y.; Eibschutz, L.S.; Gholamrezanezhad, A. Pitfalls in Interpretive Applications of AI in Radiology. *Am. J. Roentgenol.* **2024** , *223* , e2431493. [CrossRef]
28. Pedrosa, J.; Aresta, G.; Ferreira, C.; Carvalho, C.; Silva, J.; Sousa, P.; Ribeiro, L.; Mendonça, A.M.; Campilho, A. Assessing Clinical Applicability of COVID-19 Detection in Chest Radiography with Deep Learning. *Sci. Rep.* **2022** , *12* , 6596. [CrossRef]
29. Vasilev, Y.A.; Vladzymyrskyy, A.V.; Arzamasov, K.M.; Shulkin, I.M.; Astapenko, E.V.; Pestrenin, L.D. Limitations in the Application of AI Services for Chest Radiograph Analysis. *Digit. Diagn.* **2024** , *5* , 407–420. (In Russian) [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.

diagnostics

MDPI

Article

Feasibility of Artificial Intelligence Models for Longitudinal CT Analysis of Epicardial Adipose Tissue After Immunotherapy

Eliodoro Faiella [1,*], Stefania Lamja [1,*], Rebecca Casati [1], Michele Tondo [2], Raffaele Ragone [1], Adriano Redi [1], Elva Vergantino [2], Bruno Beomonte Zobel [1,2], Francesco Grasso [1,3] and Domiziana Santucci [1]

1 Fondazione Policlinico Campus Biomedico di Roma, 00128 Roma, Italy; rebecca.casati@policlinicocampus.it (R.C.); raffaele.ragone@policlinicocampus.it (R.R.); adriano.redi@policlinicocampus.it (A.R.); b.zobel@policlinicocampus.it (B.B.Z.); r.grasso@policlinicocampus.it (F.G.); d.santucci@policlinicocampus.it (D.S.)
2 Università Campus Biomedico di Roma, 00128 Roma, Italy; michele.tondo@unicampus.it (M.T.); elva.vergantino@unicampus.it (E.V.)
3 Research Unit of Interventional Radiology, Department of Radiology, 00128 Roma, Italy
* Correspondence: e.faiella@policlinicocampus.it (E.F.); s.lamja@policlinicocampus.it (S.L.)

Abstract

Background: Epicardial adipose tissue (EAT) is an imaging-derived biomarker increasingly associated with cardiovascular inflammation and metabolic risk. Computed tomography (CT) allows for accurate volumetric quantification of EAT, but the clinical interpretation of longitudinal changes remains challenging. Artificial Intelligence (AI) may provide additional value by identifying patterns and predictors of EAT variation. **Purpose**: To evaluate longitudinal changes in CT-derived EAT volume and to assess the feasibility and performance of AI-based models in discriminating patients with EAT increase after immunotherapy. **Methods**: In this retrospective single-center study, EAT was volumetrically segmented on baseline and follow-up CT scans. EAT change (ΔEAT) was calculated, and patients were dichotomized according to EAT increase (ΔEAT > 0). Three supervised AI models—Support Vector Machine (SVM), Artificial Neural Network (ANN), and Extreme Gradient Boosting (XGBoost)—were trained using imaging-derived and clinical variables. Given the limited sample size and class imbalance, stratified two-fold cross-validation was adopted. Model performance was assessed using AUC, accuracy, and F1-score, and model interpretability was explored using permutation importance. **Results**: EAT volume showed a statistically significant increase at follow-up. In the AI analysis, SVM and ANN demonstrated good discriminative performance, with ANN achieving the highest AUC (~0.90). XGBoost failed to show meaningful predictive capability. Baseline EAT volume and follow-up duration emerged as the most relevant features. **Conclusions**: AI-based models, particularly SVM and ANN, are feasible tools for the analysis of CT-derived EAT changes, even in small cohorts. These results support the integration of AI-assisted EAT assessment into imaging-based cardio-oncology research.

Keywords: epicardial adipose tissue; artificial intelligence; cardiovascular imaging

1. Introduction

Epicardial adipose tissue (EAT) is a visceral fat depot located between the myocardium and the visceral pericardium, sharing the same microcirculation with the underlying myocardium and lacking a fascial boundary. This close anatomical and functional relationship allows for a direct paracrine and vasocrine interaction between EAT and the heart. Unlike

subcutaneous fat, EAT exhibits high metabolic activity and secretes a wide range of bioactive molecules, including pro-inflammatory cytokines, adipokines, and pro-atherogenic mediators, which may influence myocardial function, coronary atherosclerosis, and overall cardiovascular risk [1–3]. Increasing evidence suggests that EAT acts not merely as a passive fat reservoir but as an active organ involved in cardiovascular pathophysiology [4,5].

From an imaging perspective, EAT has gained increasing attention as a quantitative and reproducible biomarker. Computed tomography (CT) enables accurate, non-invasive volumetric assessment of EAT with high spatial resolution and standardized attenuation thresholds. CT-based EAT quantification is widely available and can be derived not only from dedicated cardiac CT but also from routinely acquired oncologic or chest CT scans, enhancing its potential applicability in real-world clinical settings [6]. While cross-sectional EAT measurements have been extensively investigated and associated with cardiovascular outcomes, metabolic disorders, and systemic inflammation, the clinical interpretation of longitudinal changes in EAT volume remains poorly explored [7]. In particular, the dynamic behavior of EAT over time and its potential modulation by systemic therapies represent an emerging and insufficiently characterized area of research.

Importantly, epicardial adipose tissue has been shown to provide incremental risk information beyond traditional cardiovascular risk factors, reflecting localized inflammatory and metabolic activity that may not be fully captured by systemic biomarkers [8,9].

Immunotherapy is associated with profound immune activation and systemic inflammatory responses, often accompanied by metabolic alterations. These therapies may influence adipose tissue biology, potentially affecting both the quantity and functional properties of EAT [10]. In this context, longitudinal changes in EAT may reflect complex interactions between immune activation, inflammation, metabolism, and cardiovascular homeostasis. However, disentangling these interactions using conventional statistical methods may be challenging, as traditional approaches typically focus on linear associations and average effects. Increased EAT volume has been associated with ischemic heart disease, atrial arrhythmias—particularly atrial fibrillation—and left ventricular diastolic dysfunction, supporting its role as a clinically relevant cardiometabolic biomarker [11].

Artificial intelligence (AI), and machine learning techniques in particular, offer powerful tools to model complex, non-linear, and multivariate relationships among imaging-derived and clinical variables [12–14] Importantly, certain AI models are well suited to small and heterogeneous datasets, which are common in early-stage imaging studies and real-world clinical research [15]. By leveraging these methods, AI may provide complementary insights beyond conventional analyses and support the identification of subtle longitudinal patterns in imaging biomarkers [16].

Despite growing interest in EAT quantification, standardized approaches for interpreting longitudinal changes are still lacking, and most available evidence is derived from cross-sectional analyses. This gap limits the translation of EAT measurements into longitudinal risk assessment strategies and highlights the need for methodological frameworks capable of capturing temporal dynamics in imaging biomarkers.

Unlike previous AI-based studies focusing primarily on cross-sectional EAT quantification, the present work specifically investigates longitudinal changes and their modeling in a real-world immunotherapy cohort, thereby addressing a currently underexplored dynamic dimension of EAT assessment.

In this study, we investigate the feasibility of applying AI-based approaches to the analysis of longitudinal CT-derived EAT changes in a real-world cohort of limited size, with the aim of exploring the potential diagnostic value of AI-enhanced imaging biomarkers in this clinical setting.

https://doi.org/10.3390/diagnostics16060852

2. Materials and Methods

2.1. Study Design and CT Imaging

This retrospective study included consecutive patients with available baseline and follow-up computed tomography examinations acquired after the initiation of immunotherapy. Baseline CT was defined as the examination performed closest to treatment initiation, while follow-up CT corresponded to the first available examination after therapy onset suitable for longitudinal comparison. CT scans were performed according to institutional imaging protocols and included contrast-enhanced acquisitions routinely used for oncologic assessment.

All CT examinations were reviewed to ensure adequate coverage of the cardiac region and sufficient image quality for epicardial adipose tissue (EAT) analysis. Only patients with technically adequate scans at both time points and complete EAT segmentation were included in the final analysis. Exclusion criteria comprised severe motion artifacts, incomplete cardiac coverage, or insufficient image quality precluding reliable adipose tissue delineation. The study was conducted in accordance with institutional standards for retrospective imaging research.

CT acquisition parameters were consistent with institutional oncologic imaging protocols and did not substantially vary across patients in terms of slice thickness and reconstruction algorithms. All examinations provided adequate cardiac coverage suitable for epicardial adipose tissue quantification.

2.2. Epicardial Adipose Tissue Quantification

EAT was segmented and quantified volumetrically on non-ECG-synchronized CT scans using attenuation-based thresholds consistent with adipose tissue. Adipose tissue was defined using an attenuation threshold range between −190 and −30 Hounsfield Units. Segmentation included all voxels within the typical fat attenuation range and encompassed the entire adipose compartment enclosed by the visceral pericardium along the full cardiac silhouette, from the pulmonary artery bifurcation to the cardiac apex.

Semi-automated segmentation was performed, followed by manual review and correction when necessary to ensure anatomical accuracy, particularly at the pericardial boundaries. EAT volume was recorded in milliliters at baseline (EAT_pre) and at follow-up (EAT_post). Longitudinal change was calculated as ΔEAT = EAT_post − EAT_pre. To account for interindividual variability in follow-up duration, a normalized metric representing EAT change per month (ΔEAT/month) was additionally derived and used for descriptive analysis.

2.3. Feature Definition and Outcome

The primary outcome of the artificial intelligence analysis was binary classification of EAT increase, defined as a positive longitudinal change in EAT volume ($\Delta EAT > 0$). Input features were selected a priori based on clinical relevance and included imaging-derived, temporal, and clinical variables. Imaging-derived features comprised baseline EAT volume, while temporal information included the duration of follow-up between CT examinations. Clinical covariates included age, sex, body mass index, and the presence of cardiometabolic comorbidities.

Missing data were handled within the machine learning pipeline using median imputation to minimize bias while preserving the limited sample size.

A total of 42 patients were included in the artificial intelligence modeling analysis. Patients were dichotomized according to longitudinal EAT increase ($\Delta EAT > 0$ vs. $\Delta EAT \leq 0$), and class distribution was explicitly reported to ensure transparency regarding dataset balance.

https://doi.org/10.3390/diagnostics16060852

To minimize dimensionality relative to sample size and reduce overfitting risk, only baseline EAT volume was included as imaging-derived input. The total number of input variables was kept low relative to the number of observations to maintain an acceptable feature-to-sample-size ratio and preserve model stability.

Input variables were selected based on established clinical associations between epicardial adipose tissue and cardiometabolic risk factors reported in prior literature. Multicollinearity among clinical variables (e.g., BMI, dyslipidemia, diabetes mellitus) was assessed prior to modeling and did not reveal critical redundancy that would compromise model stability.

2.4. Artificial Intelligence Analysis

Three supervised machine learning models were evaluated to explore different modeling strategies. A Support Vector Machine (SVM) with radial basis function kernel was selected for its robustness in small datasets and its ability to capture non-linear decision boundaries. An Artificial Neural Network (ANN) was implemented as a compact multilayer perceptron using an LBFGS solver to enhance numerical stability and reduce overfitting in limited sample sizes. Extreme Gradient Boosting (XGBoost) was included as a representative ensemble tree-based method capable of modeling complex feature interactions.

Feature standardization was applied for SVM and ANN models, while class imbalance was addressed using class weighting strategies. Given the limited number of observations and the imbalance between outcome classes, stratified two-fold cross-validation was employed to maximize data usage while preserving class distribution. Model performance was assessed using the area under the receiver operating characteristic curve (AUC), accuracy, and F1-score. Model interpretability was explored through permutation importance analysis applied to the SVM classifier to identify the relative contribution of input features.

Given the limited sample size, stratified two-fold cross-validation was selected to maximize the number of observations available for training while preserving class distribution. Although this approach may increase variability in performance estimates, it reduces the risk of excessive data fragmentation. Model performance metrics are therefore reported as mean $\pm$ standard deviation across folds to reflect variability.

Hyperparameter tuning was limited and predefined to reduce overfitting risk. Feature standardization, median imputation, and class weighting were performed within the cross-validation pipeline to prevent data leakage. No information from validation folds was used during model training or preprocessing steps.

Model performance metrics are reported as mean $\pm$ standard deviation across folds. Area under the curve (AUC) values are presented with 95% confidence intervals.

3. Results

3.1. Pre–Post Comparisons (Expanded Description)

Epicardial adipose tissue (EAT) volume demonstrated a consistent and statistically significant increase between baseline and follow-up CT examinations across the study population. Both mean and median ΔEAT values indicated a positive longitudinal trend, suggesting that EAT remodeling occurred in the majority of patients during the observation period. The concordance between parametric and non-parametric statistical tests supports the robustness of this finding and indicates that the observed change was not driven by outliers or non-normal data distribution.

In contrast, changes in visceral fat volume showed greater dispersion. While the mean increase reached statistical significance using parametric testing, the borderline result observed with the Wilcoxon signed-rank test suggests substantial inter-individual

variability. This discrepancy highlights potential differences in the temporal dynamics and biological behavior of epicardial versus visceral adipose tissue compartments.

Pre–post comparisons for epicardial and visceral adipose tissue are reported in Table 1.

Table 1. Pre–Post Comparison of Adipose Tissue Compartments.

Variable	Mean Δ	Paired *t*-Test (*p*)	Wilcoxon Test (*p*)
Epicardial adipose tissue (mL)	+23.8	0.0010	0.0011
Visceral fat volume (mL)	+318.5	0.0374	0.0512

3.2. Regression Analysis: Absolute vs. Normalized EAT Change

To further detail the determinants of epicardial adipose tissue change, results of the linear regression analysis for absolute ΔEAT are summarized in Table 2.

Table 2. Linear Regression Analysis for Absolute ΔEAT (n = 42).

Variable	Direction of Association	Relative Contribution
Follow-up duration	Positive	Highest
Dyslipidemia	Positive	Moderate
Baseline EAT volume	Negative	Low
Sex	Negative	Low
Diabetes mellitus	Negative	Low
Age	Negative	Low
Body mass index	Neutral	Minimal

Linear regression analysis revealed that different determinants emerged depending on whether absolute or time-normalized EAT change was considered. In the model evaluating absolute ΔEAT, follow-up duration represented the dominant explanatory variable, indicating that longer exposure time was associated with greater EAT remodeling. Dyslipidemia status also contributed meaningfully, suggesting a potential interaction between lipid metabolism and epicardial fat dynamics.

When ΔEAT was normalized per month, the explanatory power of the model increased substantially, underscoring the importance of accounting for heterogeneous follow-up intervals. Regression results for normalized EAT change per month are summarized in Table 3.

Table 3. Linear Regression Analysis for Normalized ΔEAT per Month (n = 13).

Variable	Direction of Association	Relative Contribution
Baseline EAT volume	Positive	Highest
Arterial hypertension	Positive	High
Body mass index	Negative	Moderate
Diabetes mellitus	Negative	Moderate
Dyslipidemia	Negative	Moderate
Follow-up duration	Neutral	Minimal

In this normalized framework, baseline EAT volume and arterial hypertension emerged as the strongest positive contributors, while metabolic comorbidities such as diabetes mellitus and dyslipidemia exhibited inverse associations. These findings suggest that baseline cardiometabolic status may modulate the rate of EAT change rather than its absolute magnitude.

https://doi.org/10.3390/diagnostics16060852

The normalized ΔEAT/month analysis was limited to 13 patients due to incomplete follow-up interval data in the remaining cohort. Given the reduced sample size, these results should be interpreted cautiously, and statistical power is limited.

3.3. Artificial Intelligence Performance Across Models

AI-based classification analyses further emphasized differences among modeling strategies. The SVM and ANN models demonstrated consistent and clinically meaningful discriminative performance despite the small and imbalanced dataset. Notably, the ANN implemented with an LBFGS solver achieved the highest overall performance metrics, suggesting superior capacity to capture non-linear feature interactions under constrained conditions.

In contrast, the XGBoost model failed to achieve discriminative performance beyond chance level. This finding underscores the importance of aligning algorithm complexity with dataset size and supports the preferential use of simpler, well-regularized models in small imaging cohorts. Performance metrics of the evaluated AI models are reported in Table 4.

Table 4. Performance metrics of AI models for classification of EAT increase (ΔEAT > 0).

Model	AUC	Accuracy	F1-Score
SVM	0.817	0.762	0.862
ANN (LBFGS)	0.900	0.917	0.944
XGBOOST	0.500	0.488	0.455

AUC values are reported with 95% confidence intervals calculated across cross-validation folds.

A consolidated summary of longitudinal epicardial adipose tissue changes and artificial intelligence model performance is presented in Table 5.

Table 5. Summary of longitudinal changes in epicardial adipose tissue and model performance.

Domain	Metric	Result
EAT longitudinal change	Mean ΔEAT (mL)	+23.8
	Median ΔEAT (mL)	+26.0
	95% CI of mean ΔEAT (mL)	10.2–37.4
	Paired *t*-test (*p*-value)	0.0010
	Wilcoxon test (*p*-value)	0.0011
Visceral fat change	Mean Δ visceral fat (mL)	+318.5
	Paired *t*-test (*p*-value)	0.0374
	Wilcoxon test (*p*-value)	0.0512
AI classification	Best-performing model	ANN (LBFGS)
	AUC (best model)	0.90
	Accuracy (best model)	0.92
	F1-score (best model)	0.94

4. Discussion

The present study explores the application of artificial intelligence (AI) techniques to the longitudinal assessment of epicardial adipose tissue quantified on computed tomography, with the aim of identifying patients exhibiting EAT increase after immunotherapy. The results demonstrate two main findings: first, EAT behaves as a dynamic imaging biomarker, showing measurable longitudinal variation; second, AI-based models—particularly Support Vector Machines (SVM) and compact Artificial Neural

Networks (ANN)—are capable of discriminating EAT increase even in a small and imbalanced cohort.

4.1. Epicardial Adipose Tissue as a Dynamic Imaging Biomarker

EAT has traditionally been investigated as a static parameter, often measured at a single time point and correlated with cardiovascular risk factors or clinical outcomes. However, growing evidence supports the concept of EAT as a metabolically active and plastic tissue, capable of responding to systemic inflammatory, metabolic, and pharmacologic stimuli. From this perspective, longitudinal changes in EAT volume may provide complementary information beyond baseline measurements alone, reflecting dynamic cardiovascular and systemic processes [1].

CT-based volumetric quantification offers a robust and reproducible method to capture these changes over time. In our study, the marked inter-individual variability in ΔEAT highlights the heterogeneous response of epicardial fat to immunotherapy-related systemic effects. Such heterogeneity may be masked by conventional population-level analyses and reinforces the need for analytical approaches capable of identifying subject-specific patterns [8]. This characteristic makes longitudinal EAT an ideal target for AI-driven analysis, which is inherently designed to model complex, multivariate, and non-linear relationships.

4.2. Added Value of Artificial Intelligence in EAT Analysis

Traditional statistical methods, including paired comparisons and linear regression models, are effective for detecting overall trends and identifying linear associations. However, they may be limited in capturing higher-order interactions among imaging-derived and clinical variables, particularly in small and heterogeneous datasets. AI-based approaches address this limitation by enabling multivariate pattern recognition without prespecified assumptions regarding data distribution or linearity [8].

In this study, AI models were tasked with a clinically intuitive objective: classification of EAT increase versus non-increase [17,18]. Despite the limited sample size, both SVM and ANN achieved good discriminative performance, suggesting that meaningful patterns are embedded within the available features [19]. These findings support the concept that EAT-related information derived from CT imaging can be effectively leveraged by AI even in early-phase, exploratory imaging studies.

4.3. Model Selection in Small Imaging Datasets

A key methodological insight emerging from this work concerns the relationship between dataset size and algorithm choice. Ensemble tree-based methods such as XGBoost are known to perform exceptionally well in large and diverse datasets but may be prone to instability and overfitting in small cohorts. In contrast, SVM and ANN—when appropriately constrained—demonstrated superior robustness in the present setting [20–22].

While the limited sample size likely contributed to the suboptimal performance of XGBoost, additional factors may have played a role. Tree-based ensemble methods often require greater heterogeneity and larger datasets to fully exploit feature interactions. Furthermore, pronounced class imbalance and limited feature dimensionality may have constrained the representational advantage typically observed with boosting algorithms.

The strong performance of the ANN implemented with an LBFGS solver is particularly noteworthy. This optimization strategy avoids stochastic gradient descent and internal validation splits, both of which can introduce variability in limited datasets. Similarly, class-weighted SVM mitigated the impact of class imbalance, a frequent challenge in longitudinal imaging research. The near-random performance observed with XGBoost further underscores the importance of tailoring algorithm selection to dataset characteristics rather than relying on model complexity alone.

https://doi.org/10.3390/diagnostics16060852

4.4. Interpretability and Feature Relevance

Interpretability remains a critical requirement for the clinical translation of AI models in medical imaging. In this study, permutation importance analysis applied to the SVM model identified baseline EAT volume and follow-up duration as the most influential predictors of EAT increase. This finding is biologically plausible, as baseline EAT reflects the underlying metabolic and inflammatory milieu, while follow-up duration captures the temporal window over which EAT remodeling may occur. The prominence of imaging-derived features further reinforces the central role of CT-based EAT quantification in AI-driven analysis [19,23].

4.5. Clinical Implications and Future Perspectives

From a clinical imaging perspective, AI-assisted longitudinal EAT analysis may enhance cardiovascular risk stratification and patient monitoring, particularly in populations exposed to systemic therapies such as immunotherapy. Automated or semi-automated pipelines integrating CT-based segmentation with AI prediction could support early identification of patients undergoing unfavorable cardiometabolic remodeling [23].

From a research standpoint, this study provides a methodological framework for future investigations. It demonstrates that meaningful AI analyses can be conducted even in small cohorts, provided that algorithm selection, validation strategy, and interpretability tools are carefully optimized. This is particularly relevant in emerging fields such as cardio-oncology, where large datasets are often not readily available.

In addition, the use of artificial intelligence for longitudinal imaging biomarker analysis may have relevant implications for personalized medicine. Rather than relying on population-level thresholds, AI-based approaches enable patient-specific pattern recognition, potentially allowing for an individualized assessment of cardiometabolic risk trajectories. In the context of immunotherapy, where treatment-related inflammatory responses may vary substantially among patients, such personalized imaging biomarkers could provide complementary information to traditional clinical and laboratory parameters.

Furthermore, the ability to extract meaningful information from routinely acquired CT scans supports the concept of opportunistic imaging. The integration of AI-assisted EAT analysis into standard oncologic imaging workflows could be achieved without additional radiation exposure or dedicated cardiac imaging protocols, thereby enhancing the clinical value of existing imaging data.

Although classification of ΔEAT > 0 may represent a potential imaging signal of cardiometabolic remodeling in patients undergoing immunotherapy, the present findings should be interpreted as exploratory and hypothesis-generating. The current study does not establish direct clinical decision thresholds but rather demonstrates the feasibility of AI-assisted longitudinal modeling in a real-world cohort. Future larger-scale and outcome-driven investigations are required to determine whether such modeling approaches can meaningfully inform risk stratification or therapeutic strategies in cardio-oncology practice.

4.6. Limitations

Several limitations should be acknowledged. The retrospective and single-center design may limit generalizability. The small sample size, although representative of real-world imaging cohorts, precluded external validation and restricted the complexity of the models that could be reliably applied. In addition, the lack of hard cardiovascular outcomes prevents direct assessment of the prognostic implications of EAT increase. Future multicenter studies with larger populations and outcome data are warranted to confirm and extend these findings.

Another limitation relates to the use of non-ECG-synchronized CT scans, which may introduce variability in cardiac phase-dependent measurements. However, previous studies have demonstrated acceptable reproducibility of EAT volumetric quantification on non-gated CT, particularly when semi-automated segmentation and consistent anatomical landmarks are applied. This supports the reliability of the approach adopted in the present study and its applicability in routine clinical practice.

Importantly, no independent test set or external validation cohort was available. Therefore, model performance should be interpreted as exploratory and hypothesis-generating rather than definitive. External validation in larger multicenter cohorts is necessary before clinical translation.

5. Conclusions

AI-driven analysis of CT-derived epicardial adipose tissue is feasible and informative, even in small real-world cohorts. In this study, Support Vector Machines and compact Artificial Neural Networks demonstrated robust performance in modeling longitudinal EAT changes, supporting their suitability for early-phase, exploratory imaging research.

While these findings demonstrate the feasibility of AI-based modeling of longitudinal CT-derived epicardial adipose tissue changes, their clinical implications remain exploratory. The results should be interpreted as hypothesis-generating, supporting further validation in larger, multicenter cohorts before integration into routine clinical workflows.

Future studies involving larger, multicenter cohorts and clinical outcome data are warranted to validate these results and to further explore the clinical and diagnostic implications of AI-assisted longitudinal EAT assessment.

Author Contributions: Methodology, A.R.; Software, R.R.; Resources, M.T.; Writing—original draft, S.L. and R.C.; Writing—review and editing, S.L. and D.S.; Visualization, E.V.; Supervision, E.F., B.B.Z. and F.G.; Funding acquisition, S.L. 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: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. Iacobellis, G. Epicardial adipose tissue in endocrine and metabolic diseases. *Endocrine* **2014**, *46*, 8–15. [CrossRef] [PubMed]
2. Matloch, Z.; Cinkajzlova, A.; Mraz, M.; Haluzik, M. The Role of Inflammation in Epicardial Adipose Tissue in Heart Diseases. *Curr. Pharm. Des.* **2018**, *24*, 297–309. [CrossRef] [PubMed]
3. Ansaldo, A.M.; Montecucco, F.; Sahebkar, A.; Dallegri, F.; Carbone, F. Epicardial adipose tissue and cardiovascular diseases. *Int. J. Cardiol.* **2019**, *278*, 263–264. [CrossRef]
4. Dhore-Patil, A.; Urina-Jassir, D.; Samson, R.; Le Jemtel, T.H.; Oparil, S. Epicardial Adipose Tissue Thickness and Preserved Ejection Fraction Heart Failure. *Curr. Hypertens. Rep.* **2024**, *26*, 381–388. [CrossRef] [PubMed]
5. Brandt, V.; Bekeredjian, R.; Schoepf, U.J.; Varga-Szemes, A.; Emrich, T.; Aquino, G.J.; Decker, J.; Bayer, R.R.; Ellis, L.; Ebersberger, U.; et al. Prognostic value of epicardial adipose tissue volume in combination with coronary plaque and flow assessment for the prediction of major adverse cardiac events. *Eur. J. Radiol.* **2022**, *148*, 110157. [CrossRef] [PubMed]
6. Nerlekar, N.; Thakur, U.; Lin, A.; Koh, J.Q.S.; Potter, E.; Liu, D.; Muthalaly, R.G.; Rashid, H.N.; Cameron, J.D.; Dey, D.; et al. The Natural history of Epicardial Adipose Tissue Volume and Attenuation: A long-term prospective cohort follow-up study. *Sci. Rep.* **2020**, *10*, 7109. [CrossRef] [PubMed]

7. Thuny, F.; Naidoo, J.; Neilan, T.G. Cardiovascular complications of immune checkpoint inhibitors for cancer. *Eur. Heart J.* **2022**, *43*, 4458–4468. [CrossRef] [PubMed] [PubMed Central]
8. Greco, F.; Salgado, R.; Van Hecke, W.; Del Buono, R.; Parizel, P.M.; Mallio, C.A. Epicardial and pericardial fat analysis on CT images and artificial intelligence: A literature review. *Quant. Imaging Med. Surg.* **2022**, *12*, 2075–2089. [CrossRef] [PubMed] [PubMed Central]
9. Goeller, M.; Achenbach, S.; Marwan, M.; Doris, M.K.; Cadet, S.; Commandeur, F.; Chen, X.; Slomka, P.J.; Gransar, H.; Cao, J.J.; et al. Epicardial adipose tissue density and volume are related to subclinical atherosclerosis, inflammation and major adverse cardiac events in asymptomatic subjects. *J. Cardiovasc. Comput. Tomogr.* **2018**, *12*, 67–73. [CrossRef] [PubMed] [PubMed Central]
10. Foldyna, B.; Hadzic, I.; Zeleznik, R.; Langenbach, M.C.; Raghu, V.K.; Mayrhofer, T.; Lu, M.T.; Aerts, H.J.W.L. Deep learning analysis of epicardial adipose tissue to predict cardiovascular risk in heavy smokers. *Commun. Med.* **2024**, *4*, 44. [CrossRef]
11. Miller, R.J.H.; Shanbhag, A.; Killekar, A.; Lemley, M.; Bednarski, B.; Van Kriekinge, S.D.; Kavanagh, P.B.; Feher, A.; Miller, E.J.; Einstein, A.J.; et al. AI-derived epicardial fat measurements improve cardiovascular risk prediction from myocardial perfusion imaging. *npj Digit. Med.* **2024**, *7*, 24. [CrossRef]
12. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
13. Ben-Hur, A.; Weston, J. A user's guide to support vector machines. *Methods Mol. Biol.* **2010**, *609*, 223–239. [CrossRef] [PubMed]
14. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
15. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347. [CrossRef] [PubMed]
16. Domingos, P. A Few Useful Things to Know About Machine Learning. *Commun. ACM* **2012**, *55*, 78–87. [CrossRef]
17. Cundari, G.; Marchitelli, L.; Pambianchi, G.; Catapano, F.; Conia, L.; Stancanelli, G.; Catalano, C.; Galea, N. Imaging biomarkers in cardiac CT: Moving beyond simple coronary anatomical assessment. *Radiol. Med.* **2024**, *129*, 380–400. [CrossRef] [PubMed]
18. Tesche, C.; Gray, H.N. Machine learning and deep neural networks applications in coronary flow assessment. *J. Thorac. Imaging* **2020**, *35*, S66–S71. [CrossRef]
19. Feng, Y.; Xu, Z.; Zhang, L.; Zhang, Y.; Xu, H.; Zhuang, X.; Zhang, H.; Xie, X. Machine-learning-derived radiomics signature of pericoronary tissue in coronary CT angiography associates with functional ischemia. *Front. Physiol.* **2022**, *13*, 980996. [CrossRef]
20. Sagris, M.; Antonopoulos, A.S.; Simantiris, S.; Oikonomou, E.; Siasos, G.; Tsioufis, K.; Tousoulis, D. Pericoronary fat attenuation index—A new imaging biomarker and its diagnostic and prognostic utility: A systematic review and meta-analysis. *Eur. Heart J.-Cardiovasc. Imaging* **2022**, *23*, e526–e536. [CrossRef]
21. Nichols, J.H.; Samy, B.; Nasir, K.; Fox, C.S.; Schulze, P.C.; Bamberg, F.; Hoffmann, U. Volumetric measurement of pericardial adipose tissue from contrast-enhanced coronary computed tomography angiography: A reproducibility study. *J. Cardiovasc. Comput. Tomogr.* **2008**, *2*, 288–295. [CrossRef]
22. Rebelo, A.F.; Ferreira, A.M.; Fonseca, J.M. Automatic epicardial fat segmentation and volume quantification on non-contrast cardiac Computed Tomography. *Comput. Methods Programs Biomed. Update* **2022**, *2*, 100079. [CrossRef]
23. Ding, X.; Terzopoulos, D.; Diaz-Zamudio, M.; Berman, D.S.; Slomka, P.J.; Dey, D. Automated pericardium delineation and epicardial fat volume quantification from noncontrast CT. *Med. Phys.* **2015**, *42*, 5015–5026. [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.

diagnostics

MDPI

Article

Nomogram for Osteoporosis Risk Using LDCT Trabecular Parameters

Pin-Chieh Wu [1,2], Yun-Ju Wu [3], Chiao-Lin Hsu [1,4], Hsien-Chung Yu [1], Chi-Shen Chen [1] and Fu-Zong Wu [3,5,6,*]

1 Health Management Center, Kaohsiung Veterans General Hospital, Kaohsiung 813414, Taiwan
2 Department of Pharmacy, Chia Nan University of Pharmacy and Science, Tainan 717301, Taiwan
3 Department of Radiology, Kaohsiung Veterans General Hospital, Kaohsiung 813414, Taiwan
4 Center for Geriatrics and Gerontology, Kaohsiung Veterans General Hospital, Kaohsiung 813414, Taiwan
5 Faculty of Medicine, School of Medicine, National Yang Ming Chiao Tung University, Taipei 112304, Taiwan
6 Faculty of Clinical Medicine, National Yang Ming Chiao Tung University, Taipei 112304, Taiwan
* Correspondence: cmvwu1029@gmail.com

Abstract

Background: Low-dose computed tomography (LDCT) offers a unique opportunity to assess osteoporosis risk during routine lung cancer screening. This study aims to develop an integrated prediction model using trabecular bone features from LDCT and clinical factors to identify high-risk individuals early. **Methods:** This retrospective observational cohort study included 429 adults who underwent both DEXA and LDCT scans within one week at Kaohsiung Veterans General Hospital (2018–2022). Clinical data, including demographics, lifestyle factors, and comorbidities, were extracted from electronic records. Participants were categorized into osteoporotic (T-score ≤ -2.5) and non-osteoporotic groups. Trabecular bone morphometry was assessed at the T12 vertebra using QUIBIM Precision® software, analyzing parameters such as BV/TV, Tb.Th, Tb.N, Tb.Sp, D2D, D3D, and QTS. ROI placement and measurements followed standardized protocols. Ethical approval was obtained, and informed consent was waived. Statistical analyses included t-tests, ROC curves, logistic regression, and Delong tests to compare clinical and trabecular predictors of osteoporosis using SPSS v22. **Results:** In this study of 429 individuals, osteoporosis was significantly associated with female gender, older age, lower BMI, and smaller waist circumference. Trabecular bone morphometry revealed that osteoporotic individuals had significantly thinner trabeculae (lower Tb.Th), higher trabecular number (Tb.N), and more complex trabecular architecture (higher D2D/D3D), with lower QTS. Logistic regression showed that Model 1—the model combining clinical factors and Tb.N—showed a slightly higher predictive performance (AUC 0.738) than Model 2 (AUC 0.711), although the improvement was modest (p = 0.022). A nomogram based on age, sex, BMI, waist circumference, and Tb.N effectively estimated osteoporosis probability, providing a clinically useful tool for risk stratification. **Conclusions:** In conclusion, combining trabecular bone morphometry (Tb.N) from routine LDCT with age, sex, BMI, and waist circumference enhances osteoporosis risk prediction, enabling personalized assessment without extra radiation during standard lung cancer screening.

Keywords: osteoporosis; nomogram; LDCT; trabecular bone

1. Introduction

Low-dose computed tomography (LDCT) is vital for lung cancer screening, enabling early detection and intervention [1]. In the context of holistic care, LDCT screening's

widespread use also offers opportunities to assess other health conditions, including osteoporosis and cardiovascular risk, through detailed bone structure and coronary calcium score analysis [2]. Although dual-energy X-ray absorptiometry (DEXA) is a reliable tool for diagnosing osteoporosis, its accessibility remains limited [3]. Therefore, identifying high-risk populations for osteoporosis and providing them with subsequent diagnostic confirmation and care has become an important public health issue in aging societies.

In a recent study, we identified age, height, weight, and BMI as key variables for our osteoporosis prediction model and assessed its performance using the area under the receiver operating characteristic curve (AUC). The artificial neural network model significantly outperformed the Osteoporosis Self-Assessment Tool for Asians model and other machine learning algorithms in both men and women (AUC: 0.67 for men; 0.77 for women), and demonstrated consistent performance across different populations [4]. However, bone mechanical competence is also governed by trabecular micro-architecture; its integrity depends on several interrelated attributes, including trabecular thickness, number, connectivity, and overall volume [5]. Research that integrates such micro-architectural metrics with clinical variables in osteoporosis risk models remains sparse, and the predictive performance of these combined models has yet to be thoroughly evaluated.

The rationale for the present study is to address this gap by developing an integrated prediction model that incorporates trabecular bone morphometric features extracted from LDCT scans alongside clinical factors. By using data from LDCT lung cancer screening, this integrated approach aims to facilitate the early identification and management of individuals at high risk for osteoporosis within a population already undergoing routine imaging.

2. Materials and Methods

2.1. Study Design and Participants

This retrospective observational cohort study examined individuals aged 20 years or older who underwent DEXA at the Kaohsiung Veterans General Hospital in Taiwan. Participants were included if both a DEXA scan and an LDCT scan were performed within one week. All chest LDCT scans were performed using a 256-slice CT scanner (Revolution CT, GE Healthcare, Milwaukee, WI, USA), covering the region from the lung apex to the lung base. Scans were obtained during full inspiration without the administration of contrast agents. The scanning protocols across different vendors were generally comparable and included a tube voltage of 120 kVp, a slice thickness of 1–2.5 mm, and reconstruction using a soft-tissue kernel.

Inclusion criteria were as follows: participants aged $\geq$18 years who underwent both DEXA and LDCT within a one-week interval, with complete clinical and demographic data available. Exclusion criteria included patients with a history of metabolic bone disorders other than osteoporosis and individuals undergoing active osteoporosis treatment at the time of evaluation. Between January 2016 and December 2022, a total of 510 individuals were enrolled. Of these, 57 were excluded due to incomplete or missing key data, and 24 were excluded for having Z-scores instead of T-scores, leaving a final cohort of 429 participants for analysis, shown in Figure 1. The study drew upon clinical and demographic data obtained from electronic medical records. The collected information included demographic characteristics such as age, gender, body mass index (BMI), body fat percentage, and waist circumference. Renal function markers, including blood urea nitrogen (BUN), serum creatinine, and estimated glomerular filtration rate (eGFR), were also recorded. Additionally, lifestyle factors such as smoking status (categorized as never/passive, former smoker, or current smoker), alcohol consumption, and physical activity levels were assessed. Relevant comorbidities, including cardiovascular disease, thyroid disease, cancer history, autoimmune disease, and diabetes mellitus, were also

https://doi.org/10.3390/diagnostics16101429

documented. A total of 429 eligible participants were categorized into two groups based on osteoporosis status: non-osteoporotic (Osteoporosis [−], N = 334) and osteoporotic (Osteoporosis [+], N = 95). Osteoporosis was diagnosed based on T-scores from DEXA scans, with a T-score of ≤ -2.5 used to define osteoporosis according to World Health Organization (WHO) criteria [6]. This study was reviewed and approved by the Institutional Review Board of Kaohsiung Veterans General Hospital (approval number: KSVGH22-CT11-17). As it involved retrospective analysis of anonymized data, the requirement for informed consent was waived. The study was carried out in accordance with the ethical principles outlined in the Declaration of Helsinki.

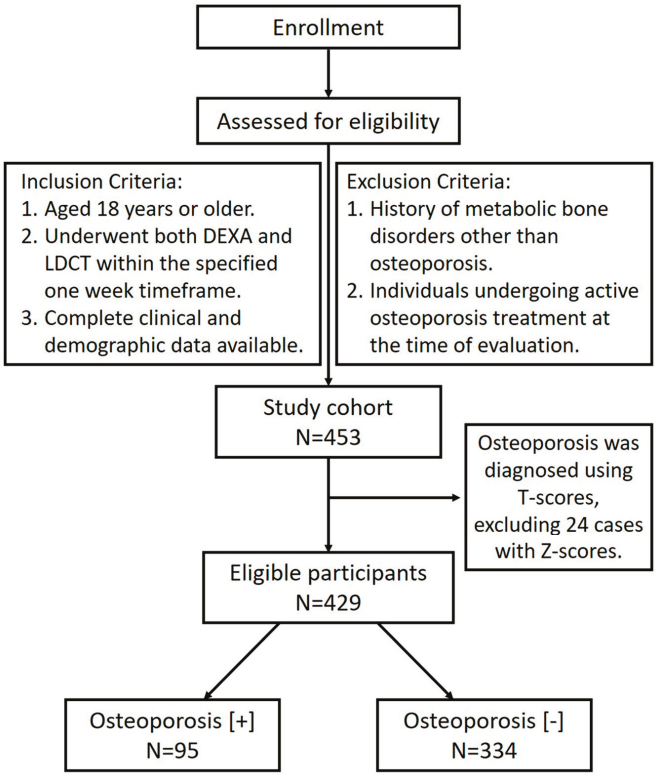


Figure 1. Flowchart of LDCT Participants Based on Enrollment Criteria and Osteoporosis Status.

2.2. Measurement of Trabecular Bone Morphometry

For trabecular bone analysis, the T12 vertebral body was selected as the region of interest (ROI). A circular ROI with an area of 1.5 cm^2 was placed in the axial view of the vertebral body (central portion). T12 was selected because it is consistently included in routine low-dose chest CT (LDCT) examinations, providing a practical and reproducible level for opportunistic assessment. Its trabecular structure can be clearly visualized, enabling reliable ROI placement and quantitative analysis, thereby supporting its use as a standardized vertebral level for trabecular evaluation. Interobserver reproducibility was assessed in 30 randomly selected cases, yielding an intraclass correlation coefficient (ICC) of 0.910, indicating excellent agreement. Consistent with our findings, previous studies have also reported high reproducibility for vertebral trabecular measurements, with ICC values ranging from 0.933 to 0.994 (all $p < 0.001$) [7].

The analysis covered a defined volume of interest (VOI) with a total slice thickness of 1 cm, consisting of five slices, each 2 mm thick, shown in Figure 2. Trabecular bone microstructure was quantitatively analyzed using QUIBIM Precision® software (version 2.8, QUIBIM S.L., Valencia, Spain). The following parameters were extracted: bone volume fraction (BV/TV, %), trabecular thickness (Tb.Th, mean, μm), trabecular number (Tb.N, arbitrary units [au]), trabecular separation (Tb.Sp, mean, μm), two-dimensional

and three-dimensional fractal dimensions (D2D and D3D), and the quantitative trabecular score (QTS).

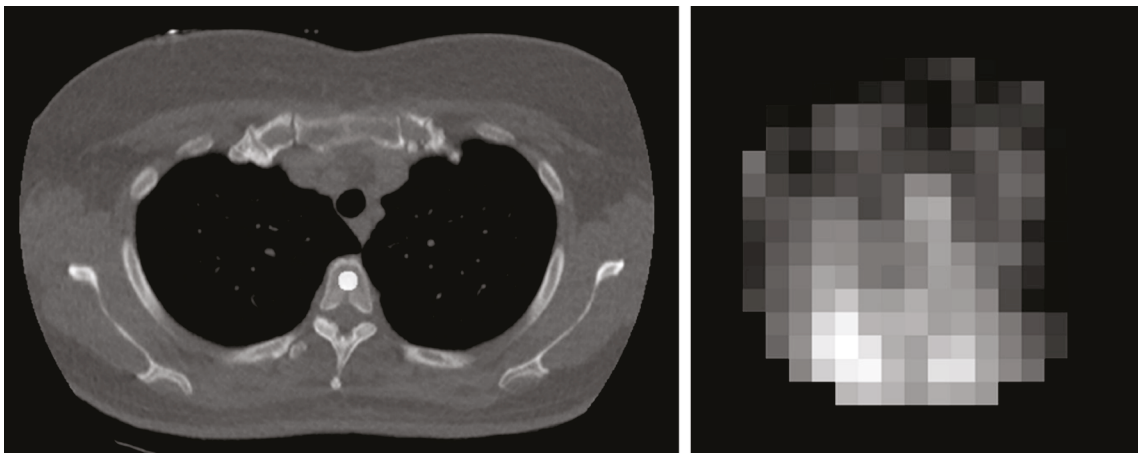


Figure 2. Trabecular bone analysis was performed on the T12 vertebral body, which was selected as the region of interest (ROI). A circular ROI of 1.5 cm^2 was placed centrally in the axial view. The volume of interest (VOI) encompassed five consecutive slices, each 2 mm thick, totaling 1 cm in thickness.

BV/TV represents the proportion of bone volume relative to the total volume of the region of interest. In QUIBIM, the reconstructed trabecular bone is composed of numerous small voxels; accordingly, BV/TV is calculated as the ratio of voxels identified as bone (n_voxel of BV) to the total number of voxels within the volume of interest (n_voxel of TV). Tb.Th is evaluated by analyzing axial slices, in which each trabecular structure is depicted by multiple contours forming a tree-like branching pattern composed of numerous skeletal lines. Tb.Th is calculated by first measuring the minimum two-dimensional width (2D width) of each skeletal line, followed by averaging these widths across all skeletal lines and contours. Tb.N is derived as the ratio of BV/TV to Tb.Th, estimating the number of trabeculae per unit length. Tb.Sp reflects the average distance between trabeculae, while its standard deviation indicates the variability in spacing. Additionally, the D2D and D3D provide measures of the complexity and structural irregularity of the trabecular network in 2D and 3D space, respectively. The QTS, a built-in parameter in the QUIBIM trabecular analysis module, serves as an overall index of trabecular bone quality. All measurements were performed in a standardized manner to ensure consistency in ROI placement and parameter extraction.

2.3. Statistical Analysis

The statistical analysis was performed using SPSS Statistics Version 22.0 (IBM Corp., Armonk, NY, USA). Descriptive statistics were calculated for all variables, and continuous variables were presented as means ± standard deviations (SD), while categorical variables were expressed as frequencies and percentages.

An independent t-test was used to compare clinical and trabecular bone morphometry parameters between patients with and without osteoporosis. To assess multicollinearity among the parameters, a Variance Inflation Factor (VIF) test was conducted. Receiver operating characteristic (ROC) curve analysis was conducted to evaluate the discriminatory ability of individual trabecular bone parameters for osteoporosis. Two multivariable logistic regression models were built. Model 1 combined the screened clinical covariates with the trabecular bone morphometric parameter that exhibited the strongest univariable discriminatory power, whereas Model 2 contained only the screened clinical covariates. Model performance will be quantified by the AUC. Statistical significance is defined as

https://doi.org/10.3390/diagnostics16101429

$p < 0.05$. Finally, ROC curve analysis was performed, and the AUC was calculated to compare the predictive performance of the two logistic regression models. The Delong test assessed AUC differences to determine significant predictive differences between models for osteoporosis status. With a study cohort of 429 individuals, post hoc analysis indicated sufficient statistical power. Missing data accounted for ≤4% of all observations and appeared to occur at random.

3. Results

3.1. Study Population Characteristic

Table 1 presents the clinical characteristics of a study cohort of 429 individuals, categorized based on osteoporosis status: osteoporosis-negative (N = 334) and osteoporosis-positive (N = 95). Among the cohort, 31.2% are female and 68.8% are male, with osteoporosis significantly more common in females (32.1% vs. 17.6%, $p = 0.001$). The osteoporosis group has a higher mean age (68.13 ± 8.030 vs. 63.83 ± 8.101 years, $p < 0.001$). BMI (23.51 ± 3.1 vs. 25.67 ± 3.48 kg/m^2, $p < 0.001$) and waist circumference (85.19 ± 10.359 vs. 89.99 ± 9.174 cm, $p < 0.001$) are lower in individuals with osteoporosis, while body fat percentage does not differ significantly ($p = 0.287$). Biochemical markers, including BUN ($p = 0.543$), creatinine ($p = 0.683$), and eGFR ($p = 0.373$), show no significant differences. Lifestyle factors such as smoking ($p = 0.574$), alcohol consumption ($p = 0.428$), and exercise habits ($p = 0.748$) also do not significantly differ between groups. Regarding medical history, there are no significant differences in the prevalence of cardiovascular disease ($p = 0.197$), thyroid disease ($p = 0.311$), cancer history ($p = 0.572$), autoimmune disease ($p = 0.291$), or diabetes mellitus ($p = 0.086$). Overall, osteoporosis is associated with female gender, older age, lower BMI, and lower waist circumference, while other lifestyle and medical factors appear unrelated.

Table 1. Patient clinical characteristics in the study cohort according to osteoporosis status (N = 429).

	Total, N = 429	**Osteoporosis (−), N = 334**	**Osteoporosis (+), N = 95**	***p*-Value**
Gender				0.001
Female	134 (31.2%)	91 (27.2%)	43 (45.3%)	
Male	295 (68.8%)	243 (72.8%)	52 (54.7%)	
Age (years)	64.78 ± 8.271	63.83 ± 8.101	68.13 ± 8.030	<0.001
BMI (kg/m^2)	25.187 ± 3.688	25.67 ± 3.567	23.51 ± 3.628	<0.001
Body fat percentage (%)	24.808 ± 7.156	25.01 ± 7.058	24.12 ± 7.482	0.287
Waist circumference (cm)	88.91 ± 9.652	89.99 ± 9.174	85.19 ± 10.359	<0.001
Blood urea nitrogen (mg/dL)	19.03 ± 9.264	18.88 ± 8.747	19.54 ± 10.941	0.543
Creatinine (mg/dL)	1.415 ± 0.662	1.42 ± 0.681	1.39 ± 0.595	0.683
eGFR (mL/min/1.73 m^2)	51.649 ± 9.275	51.86 ± 8.919	50.90 ± 10.448	0.373
Smoking, n (%)				0.574
No or passive smoking	292 (68.1%)	223 (66.8%)	69 (72.6%)	
Quit smoking (at least one year)	70 (16.3%)	55 (16.5%)	15 (15.8%)	
Smoking	65 (15.2%)	54 (16.2%)	11 (11.6%)	
Alcohol drinking, n (%)				0.428
Alcohol drinking (−)	408 (95.3%)	316 (94.9%)	92 (96.8%)	
Alcohol drinking (+)	20 (4.7%)	17 (5.1%)	3 (4.4%)	

https://doi.org/10.3390/diagnostics16101429

Table 1. *Cont.*

	Total, N = 429	Osteoporosis (−), N = 334	Osteoporosis (+), N = 95	*p*-Value
Exercise, n (%)				0.748
Exercise (−)	260 (63.1%)	202 (63.5%)	58 (61.7%)	
Exercise (+)	152 (36.9%)	116 (36.5%)	36 (38.3%)	
Cardiovascular disease, n (%)				0.197
No	387 (90.2%)	298 (89.2%)	89 (93.7%)	
Yes	42 (9.8%)	36 (10.8%)	6 (6.3%)	
Thyroid disease, n (%)				0.311
No	410 (95.6%)	321 (96.1%)	89 (93.7%)	
Yes	19 (4.4%)	13 (3.9%)	6 (6.3%)	
Cancer history, n (%)				0.572
No	401 (93.5%)	311 (93.1%)	90 (94.7%)	
Yes	28 (6.5%)	23 (6.9%)	5 (5.3%)	
Autoimmune disease, n (%)				0.291
No	421 (98.1%)	329 (98.5%)	92 (96.8%)	
Yes	8 (1.9%)	5 (1.5%)	3 (3.2%)	
Diabetes mellitus, n (%)				0.086
No	299 (69.7%)	226 (67.7%)	73 (76.8%)	
Yes	130 (30.3%)	108 (32.3%)	22 (23.2%)	

Missing data: Smoking (total: 2 cases, 0.5%; osteoporosis (−) group: 2 cases, 0.6%); Alcohol drinking (total: 1 case, 0.2%; osteoporosis (−) group: 1 case, 0.3%); Exercise (total: 17 cases, 4.0%; osteoporosis (−) group: 16 cases, 4.8%; osteoporosis (+) group: 1 cases, 1.1%). Abbreviations: BMI, body mass index; eGFR, estimated glomerular filtration rate.

3.2. Trabecular Bone Morphometry in the Study Population

Table 2 shows that osteoporotic individuals have significantly reduced trabecular thickness (Tb.Th, $p = 0.004$) but an increased trabecular number (Tb.N, $p = 0.001$), indicating an increased count of thinner struts. Both 2D (D2D, $p = 0.015$) and 3D fractal dimensions (D3D, $p = 0.003$) are significantly elevated, suggesting greater structural complexity. Additionally, the QTS is notably lower in the osteoporosis group ($p = 0.002$), reflecting a decline in overall trabecular integrity. However, bone volume fraction (BV/TV, $p = 0.418$) and trabecular separation (Tb.Sp, $p = 0.222$) do not show significant differences, indicating comparable bone volume and trabecular spacing between groups.

Table 2. Trabecular bone morphometry analysis according to osteoporosis status.

Variable	Osteoporosis (−), N = 334	Osteoporosis (+), N = 95	*p*-Value
BV/TV	41.03 ± 4.287	41.41 ± 3.237	0.418
Tb.Th	1.53 ± 0.278	1.44 ± 0.194	0.004
Tb.Sp	1.89 ± 1.315	1.72 ± 0.246	0.222
Tb.N	0.27 ± 0.045	0.29 ± 0.036	0.001
D2D	1.65 ± 0.087	1.67 ± 0.068	0.015
D3D	1.83 ± 0.100	1.87 ± 0.085	0.003
QTS	19.17 ± 3.066	18.10 ± 2.756	0.002

Abbreviations: BV/TV, bone volume to total volume; Tb.Th, trabecular thickness; Tb.Sp, trabecular separation; Tb.N, trabecular number; D2D, 2D fractal dimension; D3D, 3D fractal dimension; QTS, quality of trabecular structure.

3.3. Logistic Regression Model for Osteoporosis Prediction

The logistic regression analysis in Table 3 evaluates two models for predicting osteoporosis: Model 1, which integrates both clinical parameters and Tb.N—the variable with the highest univariable discriminatory power, as indicated by its AUC of 0.612 (Table 4)—was compared with Model 2, which relies solely on clinical variables. In Model 1, age is a significant predictor (OR = 1.07, 95% CI: 1.037–1.104, $p < 0.001$), indicating an increasing likelihood of osteoporosis with age. Gender (OR = 0.571, 95% CI: 0.325–1.002, $p = 0.051$) shows a trend toward significance, suggesting males may have a lower risk, though not statistically significant. BMI (OR = 0.829, 95% CI: 0.722–0.951, $p = 0.008$) is negatively associated with osteoporosis, implying a protective effect, while waist circumference (OR = 1.019, 95% CI: 0.967–1.075, $p = 0.478$) is not significantly associated. Tb.N, a trabecular bone morphometry parameter with the highest value, was highly significant (OR = 19,062.087, 95% CI: 22.233–16,343,266.88, $p = 0.004$), indicating a detrimental micro-architectural shift attributable to significant trabecular thinning despite no significant change in BV/TV. The overall fit of Model 1 is strong, with an overall odds ratio (OR = 299.455, 95% CI: 63.133–1420.388, $p < 0.001$), suggesting that integrating trabecular bone morphometry significantly enhances predictive performance. In Model 2, age remains a significant predictor (OR = 1.06, 95% CI: 1.029–1.092, $p < 0.001$), reinforcing its association with osteoporosis risk. Gender (OR = 0.517, 95% CI: 0.298–0.898, $p = 0.019$) is statistically significant, indicating females have a higher osteoporosis risk. BMI (OR = 0.832, 95% CI: 0.727–0.952, $p = 0.007$) remains protective, while waist circumference (OR = 1.01, 95% CI: 0.959–1.063, $p = 0.714$) does not show a significant effect. The overall performance of Model 2 (OR = 366.701, 95% CI: 70.006–1920.821, $p < 0.001$) suggests strong predictive power but slightly lower robustness compared to Model 1.

Table 3. Logistic regression models for prediction of osteoporosis.

Variable	Coefficient	OR	95% CI	*p*-Value
Model 1: Combined model (clinical and trabecular bone morphometry profiles)				
Age	0.068	1.07	1.037–1.104	<0.001
Gender	−0.561	0.571	0.325–1.002	0.051
BMI	−0.188	0.829	0.722–0.951	0.008
Waist circumference	0.019	1.019	0.967–1.075	0.478
Tb.N	9.855	19,062.087	22.233–16,343,266.88	0.004
Overall				
Model 1	5.702	299.455	63.133–1420.388	<0.001
Model 2: Clinical model				
Age	0.059	1.06	1.029–1.092	<0.001
Gender	−0.659	0.517	0.298–0.898	0.019
BMI	−0.184	0.832	0.727–0.952	0.007
Waist circumference	0.01	1.01	0.959–1.063	0.714
Overall				
Model 2	5.905	366.701	70.006–1920.821	<0.001

Abbreviations: BMI, body mass index; Tb.N, trabecular number; OR, odds ratio; CI, confidence interval.

Table 4. Evaluating osteoporosis predictive performance with models and trabecular bone morphometry.

Variable	**AUC**	**95% CI**	***p*-Value**	
Tb.Th	0.362	0.299–0.425	<0.001	
Tb.N	0.612	0.549–0.675	0.001	
D2D	0.585	0.521–0.650	0.011	
D3D	0.609	0.546–0.671	0.001	
QTS	0.367	0.303–0.430	<0.001	
Model 1	0.738	0.680–0.796	<0.001	
Model 2	0.711	0.647–0.774	<0.001	
Comparison	Difference between areas	SE	95% CI	*p*-value
Model 1 vs. Model 2	0.0273	0.0119	0.00393–0.0506	0.022

Abbreviations: Tb.Th, trabecular thickness; Tb.N, trabecular number; D2D, 2D fractal dimension; D3D, 3D fractal dimension; QTS, quality of trabecular structure; AUC, area under the ROC curve; CI, confidence interval; SE, standard error.

3.4. Assessment of the Performance of the Different Models for Osteoporosis

Table 4 evaluates the predictive performance of different models and trabecular bone morphometry variables for osteoporosis, including statistical metrics such as the AUC, 95% confidence intervals (CI), and *p*-values. The results show that Tb.Th has the lowest AUC (0.362), indicating poor predictive performance, while Tb.N has an AUC of 0.612 with statistical significance ($p = 0.001$). The 2D and 3D fractal dimensions (D2D, D3D) have AUCs of 0.585 and 0.609, respectively, indicating moderate predictive ability, whereas the QTS has an AUC of 0.367, showing weak predictive performance. Regarding predictive models, Model 1 has the highest AUC (0.738), demonstrating the best predictive capability, while Model 2 has a slightly lower AUC (0.711) but remains statistically significant. Comparing Model 1 and Model 2, the AUC difference is 0.0273, with a standard error (SE) of 0.0119, a confidence interval of 0.00393–0.0506, and a *p*-value of 0.022, indicating that Model 1 is significantly superior to Model 2. Overall, among trabecular bone morphometry variables, Tb.N shows better predictive performance, while predictive models outperform individual trabecular parameters, with Model 1 demonstrating the highest predictive ability. Sensitivity analysis using randomly sampled balanced (95/95) and imbalanced (95/334) datasets showed consistent performance, indicating model robustness to class imbalance. The balanced model achieved an AUC of 0.751 (95% CI: 0.683–0.820, $p < 0.001$), demonstrating significant discriminative ability for osteoporosis diagnosis.

3.5. Logistic-Based Nomogram for Osteoporosis Prediction

The probability of osteoporosis in the study cohort was determined using a multivariable logistic regression model incorporating five potential predictive factors. Each significant variable was assigned a score based on a point scale, and the total probability of osteoporosis was estimated by summing these scores and mapping them onto a total point scale. As shown in Figure 3, this nomogram visually represents the predictive model, integrating clinical variables such as Tb.N value, waist circumference, BMI, sex, and age. Each variable is assigned a corresponding score based on its value, and the cumulative score is then matched to a probability scale at the bottom to estimate the likelihood of osteoporosis. For example, a 79-year-old male with a Tb.N value of 0.08, a waist circumference of 90 cm, and a BMI of 28.05 would receive approximately 1.1 point for Tb.N (Tb.N value = 0.08), 2.2 for waist circumference (90 cm), 6.9 for BMI (BMI = 28.5), 0 for sex (male), and 8 for age (79 year-old). The total score of 18.2 points corresponds to an estimated osteoporosis probability of approximately 45%.

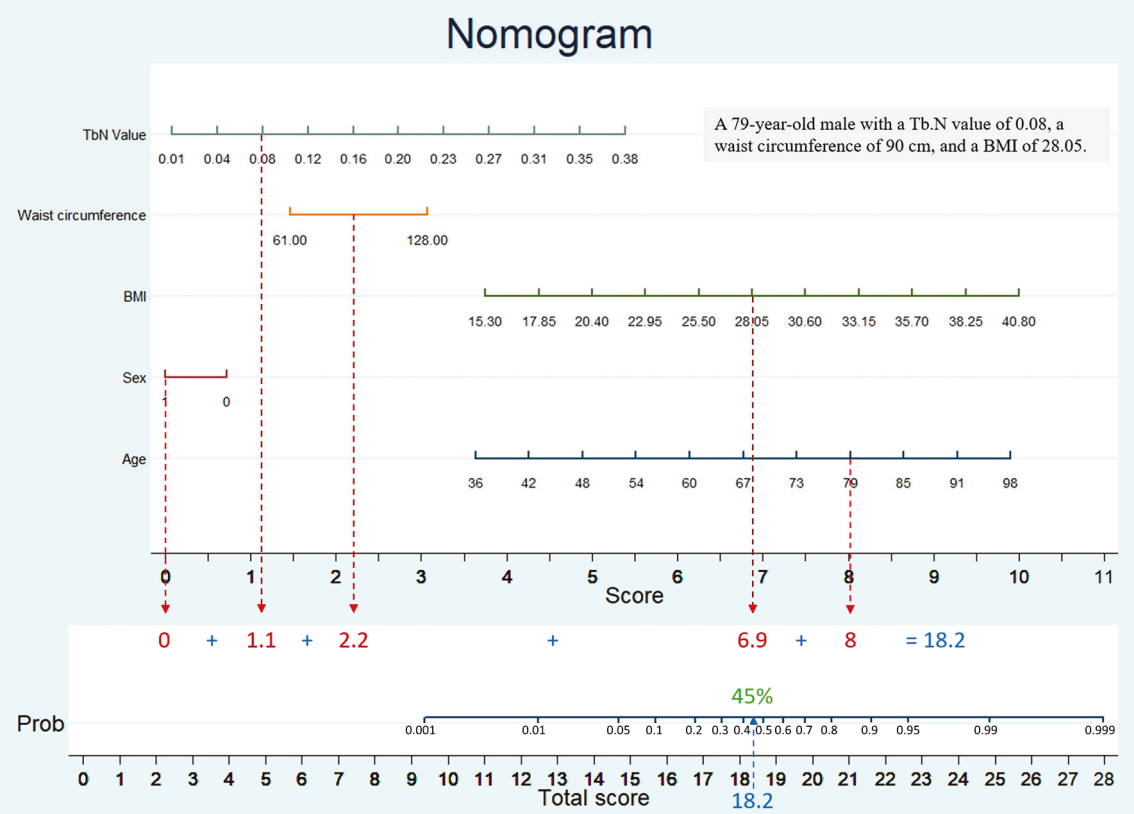


Figure 3. Nomogram for Predicting Osteoporosis Risk Based on TbN, Waist Circumference, BMI, Sex, and Age.

4. Discussion

Osteoporosis is a systemic disease defined as a reduction in bone mass associated with an impaired bone architecture, increased bone fragility and increased fracture risk. The micro-architecture of trabecular bone is acknowledged as a key determinant of bone quality [8]. In healthy bone, trabeculae form a dense, interlacing lattice that efficiently distributes mechanical loads. Conversely, reductions in trabecular volume or thickness weaken this internal scaffold, diminish load-bearing capacity, and ultimately heighten fracture risk. This study shows that adding a single trabecular morphometric variable to a clinical model (age, sex, waist circumference and BMI) significantly improves discriminatory performance, with the area under the curve rising from 0.711 to 0.738 ($p < 0.05$). This result demonstrates that trabecular morphometry provides meaningful, incremental information for osteoporosis risk assessment.

Previous investigations have shown that osteoporosis is accompanied by changes in trabecular micro-architecture. Jin et al. analysed femoral specimens from an ovariectomised (OVX) mouse model of osteoporosis with micro-CT, comparing a normal control group (n = 10) with an OVX group (n = 46). BV/TV, Tb.Th, Tb.N, Tb.Sp and fractal dimension were calculated; all four morphometric indices differed significantly between groups ($p < 0.05$) [9]. In a clinical study of 108 men with lumbar osteopenia (mean age 52.1 years, T-score < −2.5), Legrand et al. found that, although BV/TV and Tb.Th did not have significant differences between 62 subjects with at least one vertebral compression fracture and 46 fracture-free controls, both Tb.Sp and Tb.N were significantly altered [10]. Majumdar et al. employed high-resolution magnetic resonance imaging at 1.5 T to acquire distal radius images in three female groups—premenopausal normal (n = 10), postmenopausal normal (n = 9) and postmenopausal fracture (n = 11). Compared with the postmenopausal-normal group, the postmenopausal fracture group showed significant differences in BV/TV, Tb.Sp

and Tb.N. In addition, both Tb.Sp and Tb.N exhibited moderate correlations with trabecular BMD [11]. These studies highlight the diagnostic relevance of trabecular morphometry, yet they remain limited to animal models or small human cohorts.

More recently, LDCT—acquired for lung-cancer screening—has emerged as an opportunistic source of skeletal information. Pan et al. demonstrated that simple vertebral Hounsfield-unit (HU) thresholds at T11–L2 could identify low BMD and osteoporosis, and Marque et al. showed that serial LDCT HU values sensitively track two-year changes in vertebral trabecular density. These HU-based approaches demonstrated that LDCT images can be a tool for bone assessment [12,13]; however, they only quantify bone quantity and do not interrogate micro-architectural quality.

The present study analyzed a cohort of more than 400 participants, providing a larger-scale evaluation of trabecular morphometry in the context of osteoporosis. Our micro-architectural analysis showed no significant difference in BV/TV between osteoporosis and non-osteoporosis groups, whereas Tb.Th was lower in the osteoporotic group—an observation consistent with earlier reports [9]. In the present software, Tb.N is calculated according to the previously published formula [14,15]—as BV/TV divided by Tb.Th, a BV/TV that shows no significant change combined with markedly thinner trabeculae inevitably drives the ratio upward, producing an apparent increase in Tb.N. Rather than indicating a true gain in trabeculae, this elevation reflects the deterioration of the trabecular network—namely, thinning plates. For that reason, Tb.N was retained as a model variable: when interpreted in conjunction with BV/TV and Tb.Th, it serves as an integrated quantitative marker of micro-architectural compromise and adds discriminatory power to the multivariable analysis, with its Tb.N value yielding the highest AUC of 0.623. Fractal dimension captures the degree of irregularity and fragmentation in trabecular bone texture. Previous studies have shown that individuals with osteoporosis or osteopenia exhibit higher FD values than healthy controls, reflecting a more fragmented and complex trabecular architecture [16]. Consistent with these findings, the present study likewise observed a significantly higher FD in the osteoporotic cohort compared with the non-osteoporotic group.

We translated age, sex, BMI and LDCT-derived Tb.N into a nomogram that supplies an individualized probability of osteoporosis without extra imaging or radiation, thereby extending LDCT from HU-based bone-quantity screening to micro-architectural evaluation. Since 1 July 2022, Taiwan's National Lung Cancer Early Detection Program has provided biennial LDCT to adults aged 45–74 years who are heavy smokers or have a lung cancer family history—an age range that overlaps the prime diagnostic window for osteoporosis [17]. In regions where osteoporosis remains under-detected, such as Taiwan [4,18], our LDCT nomogram delivers simultaneous, actionable information on lung cancer and skeletal health during the same visit.

A recent narrative review highlights that, with the aid of artificial intelligence, LDCT possesses considerable multifunctional potential [2]. By applying artificial intelligence algorithms to LDCT images, it is possible to not only predict the risk of lung cancer but also assess the risks of smoking-related comorbidities, such as cardiovascular disease and mortality. The present study provides evidence that the same LDCT scan can be used to extract Tb.N, which offers additional discriminatory value for assessing osteoporosis risk. With the aid of emerging artificial intelligence tools, LDCT holds forward-looking potential as an opportunistic screening modality that enables simultaneous evaluation of pulmonary, cardiovascular, and skeletal health in a single scan.

Our study has some limitations. First, this was a single-centre retrospective investigation, which may limit the generalizability of the findings to the broader population. Nevertheless, correlations between trabecular bone morphometric features and osteoporo-

sis were identified in the relatively larger cohort. Second, several potentially relevant covariates—such as duration of comorbid conditions, medication history (e.g., glucocorticoids, hormone-replacement therapy)—were not systematically recorded. These factors can modulate bone density and micro-architecture and may have introduced residual confounding.

In addition, no subgroup analyses by sex or age groups were performed. These factors may influence the observed results and should be considered in future studies. Further studies with a comprehensive capture of relative variables are therefore warranted. Third, we acknowledge the computational nature of Tb.N and therefore avoid interpreting it as a direct biological indicator of the actual number of trabeculae. The observed increase in Tb.N should be interpreted with caution, as it may largely reflect reductions in trabecular thickness (Tb.Th) and related structural alterations rather than a true increase in trabecular elements. Although BV/TV was not statistically significant in our cohort, this may reflect the limited sample size and inherent variability in measurements. All measurement procedures followed standardized protocols to minimize potential bias. Fourth, the model did not undergo internal or external validation, which may limit the generalizability of the findings. Calibration and Decision Curve Analysis were not performed due to limited sample size, and future studies should assess clinical utility. In the limitations, we noted that multi-vertebral analyses in future studies may help validate the robustness and generalizability of the results. Fifth, the extremely large odds ratio for Tb.N (19,062.087) and its wide confidence interval indicate potential model instability, likely influenced by the variable scale and limited number of events.

In conclusion, our findings demonstrate that trabecular bone morphometry provides valuable information for assessing osteoporosis risk. When Tb.N extracted from routine LDCT scans is combined with age, sex, waist circumference and BMI, the ability to discriminate osteoporosis appears to improve modestly. The resulting nomogram offers an individualized probability of osteoporosis during the same LDCT scan that many adults already undergo for lung cancer screening, without requiring additional radiation or imaging time.

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

Funding: This study was funded by Kaohsiung Veterans General Hospital (Grant No. KSVGH112-133, KSVGH115-080, KSVGH115-145, KSVGH115-159).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Kaohsiung Veterans General Hospital (protocol code KSVGH22-CT11-17 and date 01/11/2022).

Informed Consent Statement: As this was a retrospective study for an anonymous medical record, consent to participate was not required. No consent was required for publication of this article.

Data Availability Statement: All data used for this study were obtained from anonymous medical records and are therefore not publicly accessible.

Acknowledgments: The authors thank all radiologists at the hospitals for assisting with the collection of imaging data used in this study.

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

References

1. Meza, R.; Jeon, J.; Toumazis, I.; Ten Haaf, K.; Cao, P.; Bastani, M.; Han, S.S.; Blom, E.F.; Jonas, D.E.; Feuer, E.J.; et al. Evaluation of the Benefits and Harms of Lung Cancer Screening With Low-Dose Computed Tomography: Modeling Study for the US Preventive Services Task Force. *JAMA* **2021**, *325*, 988–997. [CrossRef] [PubMed]
2. De Luca, G.R.; Diciotti, S.; Mascalchi, M. The Pivotal Role of Baseline LDCT for Lung Cancer Screening in the Era of Artificial Intelligence. *Arch. Bronconeumol.* **2025**, *61*, 359–367. [CrossRef] [PubMed]
3. Chanprasertpinyo, W.; Punsawad, C.; Khwanchuea, R.; Sukkriang, N.; Yincharoen, P.; Rerkswattavorn, C. Comparison between calcaneus quantitative ultrasound and the gold standard DXA in the ability to detect osteoporosis in chronic obstructive pulmonary disease patients. *J. Orthop. Surg. Res.* **2023**, *18*, 778. [CrossRef] [PubMed]
4. Huang, W.C.; Chen, I.S.; Yu, H.C.; Chen, C.S.; Wu, F.Z.; Hsu, C.L.; Wu, P.C. A simple and user-friendly machine learning model to detect osteoporosis in health examination populations in Southern Taiwan. *Bone Rep.* **2025**, *24*, 101826. [CrossRef] [PubMed]
5. Liu, X.S.; Sajda, P.; Saha, P.K.; Wehrli, F.W.; Guo, X.E. Quantification of the roles of trabecular microarchitecture and trabecular type in determining the elastic modulus of human trabecular bone. *J. Bone Miner. Res.* **2006**, *21*, 1608–1617. [CrossRef] [PubMed]
6. Ensrud, K.E.; Crandall, C.J. Osteoporosis. *Ann. Intern. Med.* **2017**, *167*, Itc17–Itc32. [CrossRef] [PubMed]
7. Chan, C.H.; Wu, H.Y.; Chen, L.H.; Vy, V.P.T.; Liang, C.H.; Ting, Y.W.; Lu, Y.C.; Chan, W.P. Vertebral trabecular microarchitecture changes in the normally ageing population. *Quant. Imaging Med. Surg.* **2026**, *16*, 79. [CrossRef] [PubMed]
8. Chappard, D.; Baslé, M.F.; Legrand, E.; Audran, M. Trabecular bone microarchitecture: A review. *Morphologie* **2008**, *92*, 162–170. [CrossRef] [PubMed]
9. Jin, S.-C.; Chu, C.-H.; Chen, J.-C. Trabecular Bone Morphological Analysis for Preclinical Osteoporosis Application Using Micro Computed Tomography Scanner. *J. Med. Biol. Eng.* **2016**, *36*, 96–104. [CrossRef]
10. Legrand, E.; Chappard, D.; Pascaretti, C.; Duquenne, M.; Krebs, S.; Rohmer, V.; Basle, M.F.; Audran, M. Trabecular bone microarchitecture, bone mineral density, and vertebral fractures in male osteoporosis. *J. Bone Miner. Res.* **2000**, *15*, 13–19. [CrossRef] [PubMed]
11. Majumdar, S.; Genant, H.K.; Grampp, S.; Newitt, D.C.; Truong, V.H.; Lin, J.C.; Mathur, A. Correlation of trabecular bone structure with age, bone mineral density, and osteoporotic status: In vivo studies in the distal radius using high resolution magnetic resonance imaging. *J. Bone Miner. Res.* **1997**, *12*, 111–118. [CrossRef] [PubMed]
12. Pan, Y.L.; Wu, Y.B.; Wang, H.G.; Yu, T.H.; He, D.; Lu, X.J.; Zhao, F.F.; Ma, H.F.; Wang, Y.J.; Cai, Y.K. Opportunistic use of chest low-dose computed tomography (LDCT) imaging for low bone mineral density and osteoporosis screening: Cutoff thresholds for the attenuation values of the lower thoracic and upper lumbar vertebrae. *Quant. Imaging Med. Surg.* **2024**, *14*, 4792–4803. [CrossRef] [PubMed]
13. Marques, M.L.; da Silva, N.P.; van der Heijde, D.; Reijnierse, M.; Baraliakos, X.; Braun, J.; van Gaalen, F.; Ramiro, S. Hounsfield Units measured in low dose CT reliably assess vertebral trabecular bone density changes over two years in axial spondyloarthritis. *Semin. Arthritis Rheum.* **2023**, *58*, 152144. [CrossRef] [PubMed]
14. Kim, N.; Lee, J.G.; Song, Y.; Kim, H.J.; Jin, S.Y.; Cho, G. Evaluation of MRI resolution affecting trabecular bone parameters: Determination of acceptable resolution. *Magn. Reson. Med.* **2012**, *67*, 218–225. [CrossRef] [PubMed]
15. Pichone, A.; Gomes, C.P.; Lima, L.F.C.; Moreira, C.A.; Paranhos-Neto, F.P.; Madeira, M.; Lopes, R.T.; Farias, M.L.F.; Leite, M., Jr. Assessment of trabecular and cortical parameters using high-resolution peripheral quantitative computed tomography, histomorphometry and microCT of iliac crest bone core in hemodialysis patients. *Bone Rep.* **2022**, *16*, 101173. [CrossRef] [PubMed]
16. Czyz, M.; Kapinas, A.; Holton, J.; Pyzik, R.; Boszczyk, B.M.; Quraishi, N.A. The computed tomography-based fractal analysis of trabecular bone structure may help in detecting decreased quality of bone before urgent spinal procedures. *Spine J.* **2017**, *17*, 1156–1162. [CrossRef] [PubMed]
17. Administration, H.P. The First Country to Provide Lung Cancer Screening for Citizens with a Family History of Lung Cancer or a History of Heavy Smoking: The Lung Cancer Early Detection Program was Launched on July 1, 2022. Available online: https://www.hpa.gov.tw/EngPages/Detail.aspx?nodeid=1051&pid=16553 (accessed on 22 May 2025).
18. Lee, M.T.; Fu, S.H.; Hsu, C.C.; Chen, H.M.; Lin, J.W.; Tsai, K.S.; Hwang, J.S.; Lin, S.C.; Wu, C.H.; Wang, C.Y. Epidemiology and clinical impact of osteoporosis in Taiwan: A 12-year trend of a nationwide population-based study. *J. Formos. Med. Assoc.* **2023**, *122*, S21–S35. [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.

diagnostics

MDPI

Article

Predictors of In-Hospital Mortality after Thrombectomy in Anterior Circulation Large Vessel Occlusion: A Retrospective, Machine Learning Study

Ivan Petrović [1,*], Serena Broggi [2], Monika Killer-Oberpfalzer [3,4], Johannes A. R. Pfaff [5], Christoph J. Griessenauer [6], Isidora Milosavljević [1], Ana Balenović [1], Johannes S. Mutzenbach [3] and Slaven Pikija [3,*]

1 Faculty of Medicine, University of Novi Sad, 21000 Novi Sad, Serbia; 014568@mf.uns.ac.rs (I.M.); 014631@mf.uns.ac.rs (A.B.)
2 Neurology and Stroke Unit, ASST dei Sette Laghi, 21100 Varese, Italy; serena.broggi@gmail.com
3 Department of Neurology, University Hospital Salzburg, Christian Doppler Klinik, Paracelsus Medical University Salzburg, 5020 Salzburg, Austria; m.killer@salk.at (M.K.-O.); j.mutzenbach@salk.at (J.S.M.)
4 Institute of Neurointervention, Christian Doppler Klinik, Paracelsus Medical University Salzburg, 5020 Salzburg, Austria
5 Department of Neuroradiology, University Hospital Salzburg, Christian Doppler Klinik, Paracelsus Medical University Salzburg, 5020 Salzburg, Austria; j.pfaff@salk.at
6 Department of Neurosurgery, University Hospital Salzburg, Christian Doppler Klinik, Paracelsus Medical University, 5020 Salzburg, Austria; c.griessenauer@salk.at
* Correspondence: 014662@mf.uns.ac.rs (I.P.); s.pikija@salk.at (S.P.)

Abstract: Background: Despite the increased use of mechanical thrombectomy (MT) in recent years, there remains a lack of research on in-hospital mortality rates following the procedure, the primary factors influencing these rates, and the potential for predicting them. This study aimed to utilize interpretable machine learning (ML) to help clarify these uncertainties. Methods: This retrospective study involved patients with anterior circulation large vessel occlusion (LVO)-related ischemic stroke who underwent MT. The patient division was made into two groups: (I) the in-hospital death group, referred to as miserable outcome, and (II) the in-hospital survival group, or favorable outcome. Python 3.10.9 was utilized to develop the machine learning models, which consisted of two types based on input features: (I) the Pre-MT model, incorporating baseline features, and (II) the Post-MT model, which included both baseline and MT-related features. After a feature selection process, the models were trained, internally evaluated, and tested, after which interpretation frameworks were employed to clarify the decision-making processes. Results: This study included 602 patients with a median age of 76 years (interquartile range (IQR) 65–83), out of which 54% (n = 328) were female, and 22% (n = 133) had miserable outcomes. Selected baseline features were age, baseline National Institutes of Health Stroke Scale (NIHSS) value, neutrophil-to-lymphocyte ratio (NLR), international normalized ratio (INR), the type of the affected vessel ('Vessel type'), peripheral arterial disease (PAD), baseline glycemia, and premorbid modified Rankin scale (pre-mRS). The highest odds ratio of 4.504 was observed with the presence of peripheral arterial disease (95% confidence interval (CI), 2.120–9.569). The Pre-MT model achieved an area under the curve (AUC) value of around 79% utilizing these features, and the interpretable framework discovered the baseline NIHSS value as the most influential factor. In the second data set, selected features were the same, excluding pre-mRS and including puncture-to-procedure-end time (PET) and onset-to-puncture time (OPT). The AUC value of the Post-MT model was around 84% with age being the highest-ranked feature. Conclusions: This study demonstrates the moderate to strong effectiveness of interpretable machine learning models in predicting in-hospital mortality following mechanical thrombectomy for ischemic stroke, with AUCs of 0.792 for the Pre-MT model and 0.837 for the Post-MT model. Key predictors included patient age, baseline NIHSS, NLR, INR, occluded vessel type, PAD, baseline glycemia, pre-mRS, PET, and OPT. These findings provide valuable insights into risk factors and could improve post-procedural patient management.

Keywords: ischemic stroke; machine learning; mechanical thrombectomy; in-hospital mortality

1. Introduction

Despite major advances in emergent treatment, stroke is still a devastating disease identified as one of the leading causes of death and disability on a global level [1–8]. Acute ischemic stroke (AIS), caused by the culprit cessation of brain circulation leading to brain infarction, accounts for more than two-thirds of strokes [8–10]. Current state-of-the-art therapy includes intravenous thrombolysis (IVT) with alteplase or tenecteplase and mechanical thrombectomy (MT), alone or in combination with so-called bridging IVT [10–13]. Following large clinical trials, MT, particularly when combined with IVT, has become the new standard of care in patients with large vessel occlusion (LVO)-related acute stroke, as it results in better functional outcomes and a higher degree of angiographic revascularization, without a significant difference in the incidence of symptomatic intracerebral hemorrhage (sICH) or death, compared to IVT alone [14–19]. Nevertheless, it is imperative to acknowledge that certain patients may exhibit unfavorable prognoses or succumb shortly after the intervention. Therefore, it is of the highest interest to investigate whether there are a priori features of patients that should be assessed, treated, or modified before the procedure, which could be relevant to later adverse outcomes.

In previously published work, the estimated rate of in-hospital mortality in stroke ranges from 11 to 15% [20]. The percentage of deaths escalate further if 90-day mortality is included, reaching as high as 27%, with 52% of all death cases occurring during the first week after the MT treatment [21]. About 2/3 of in-hospital deaths are attributed to non-modifiable, baseline factors such as age and baseline stroke severity, while post-stroke complications determine the remaining third [20,22]. In practice, a degree of independence prior to the stroke measured with the premorbid modified Rankin scale (pre-mRS) is an important determining factor in whether MT is going to be performed. It is a 7-point scale where 0 signifies no symptoms, 4 indicates mobility only with assistance from another person, and 6 denotes death [23]. Stroke patients with a pre-mRS score of 4 or 5 before the onset of symptoms are generally not considered for mechanical thrombectomy and are excluded from most trials on endovascular treatment (EVT) [24]. However, for stroke patients with a pre-mRS score of 0–3, mechanical thrombectomy should not be withheld based solely on other features. Nevertheless, in an ideal scenario, having a reliable predictive tool before intervention could enable stakeholders to anticipate potential outcomes better.

Machine learning (ML) has become apparent as a competent predictive tool in many medical fields since, through modeling, it can handle complex relationships in data, uncover subtle information, and use these insights to automatically generate and summarize new knowledge [25]. This cannot be said for classical statistical approaches, as the increase in data complexity may make classical statistical inference less tractable [26,27]. In the currently available literature, ML models have shown superiority when compared to the standard statistical and clinical prediction models [28], and since there is increasing complexity, numerosity, and multifactoriality of available data sets [29], interest in their usage is rapidly increasing. On the other hand, clinicians can still feel confused and uncertain about the use of these models in everyday practice. The ML's "black box" represents a major barrier to this progress since an understanding of a model's decision-making process is crucial for its implementation in the clinical environment [30]. This concern could play an important role in daily interactions with patients, because, without an explanation of how a diagnosis, prognosis, or treatment plan was made, a lack of trust can be expected [31]. Fortunately, significant progress has been accomplished within the last years by developing interpretation methods, such as the Sharpley Additive Explanation (SHAP) and Local Interpretable Model-Agnostic Explanations (LIMEs), which show the potential in helping with the interpretational riddle on both feature and individual levels. To the best of our knowledge, interpretable machine learning (IML) models and their role

in the prediction of in-hospital death after anterior circulation MT in AIS patients are still an unexplored research area.

In this research, we aimed to develop and internally validate IML models that would predict the probability of in-hospital death after anterior circulation MT in LVO-AIS patients. We hypothesized that ML would generate an easily understandable predictive model, highlighting the key features contributing to the unfavorable outcome and offering a better understanding of this complexity. The built model could be used to identify AIS patients less likely to benefit from recanalization and patients with a high risk of dying after the procedure.

2. Materials and Methods

2.1. Analyzed Group and Data Processing

This retrospective study analyzed the clinical data of consecutive patients with anterior circulation LVO-related ischemic stroke, treated with mechanical thrombectomy (MT) at the tertiary university middle-volume endovascular treatment center, Christian-Doppler-Klinik (Salzburg, Austria). The local ethics committee approved the retrospective collection of data. The data were collected from the clinical information system during the analyzed period of 9 years (2012–2020), and the inclusion criteria were that (I) the patient was older than 18 years; (II) MT was used as a therapeutic approach; (III) occluded vessel(s) were in the anterior circulation of the brain; and additionally to be included into the in-hospital death (IHD) group, (IV) if a patient died, it occurred due to the nature of stroke or stroke-related complications [22,32] during the time of hospitalization. On the contrary, patients with occlusion of the posterior circulation were excluded from the study, and patients who died after the period of in-hospital stay were not included in the IHD group.

The gathered data were divided into several subcategories, based on their similarities (Table 1), and were analyzed on three levels: (I) the whole group; (II) the in-hospital death group (miserable outcome); and (III) the in-hospital survival group (favorable outcome). Python 3.10.9, provided by Anaconda, Inc. (Austin, TX, USA), was the programming language used in this research. In the first steps of preprocessing, variables from the data were screened for missing values, and if this number went beyond 20%, they were excluded from any further analysis. A categorization of the variables followed, in which the division was made into categorical and continuous groups. Missing values were imputed using the most common value method for categorical variables and the median value for continuous ones.

Table 1. Whole sample analysis.

Variables	All Patients (N = 602)	Patients with the Miserable Outcome (N = 133)	Patients with the Favorable Outcome (N = 469)	*p* Value
Sociodemographic data				
Age, median (IQR)	76 (65–83)	81 (74–87)	74 (62–82)	<0.001
Sex, *n* (%)				0.113
Female	328/602 (54%)	81/133 (61%)	247/469 (53%)	
Male	274/602 (46%)	52/133 (39%)	222/469 (47%)	
Days to In-hospital death (IQR)		6 (2.5–14)		
Vascular risk factors and comorbidities				
Previous stroke, *n* (%)	92/602 (15%)	32/133 (24%)	60/469 (13%)	0.002
Peripheral arterial disease (PAD), *n* (%)	33/602 (5%)	19/133 (14%)	14/469 (3%)	<0.001
Tobacco smoking *n* (%)	381 (63%)	81 (61%)	300 (64%)	0.393
Atrial fibrillation, *n* (%)	237/602 (39%)	62/133 (47%)	175/469 (37%)	0.066

Table 1. *Cont.*

Variables	All Patients (N = 602)	Patients with the Miserable Outcome (N = 133)	Patients with the Favorable Outcome (N = 469)	*p* Value
Diabetes mellitus, *n* (%)	90/602 (15%)	26/133 (20%)	64/469 (14%)	0.122
Arterial hypertension, *n* (%)	433/602 (72%)	103/133 (77%)	330/469 (70%)	0.135
Internal carotid artery (ICA) stenosis, *n* (%)	86/602 (14%)	24/133 (18%)	62/469 (13%)	0.206
Ischemic heart disease, *n* (%)	110/602 (18%)	31/133 (23%)	79/469 (17%)	0.115
Kidney failure, *n* (%)	53/602 (9%)	22/133 (17%)	31/469 (6%)	0.001
Heart failure, *n* (%)	75/602 (12%)	22/133 (17%)	53/469 (11%)	0.142
Statins usage, *n* (%)	149/602 (25%)	31/133 (23%)	118/469 (25%)	0.747
Antiplatelets drug usage, *n* (%)	158/602 (26%)	38/133 (29%)	120/469 (26%)	0.563
Anticoagulants usage, *n* (%)	58/602 (10%)	24/133 (18%)	34/469 (7%)	<0.001
Direct oral anticoagulants usage, *n* (%)	57/602 (9%)	20/133 (15%)	37/469 (8%)	0.02
Baseline laboratory values				
C-reactive protein (CRP) (mg/dL), median (IQR)	0.39 (0.18–0.91)	0.42 (0.19–1.48)	0.39 (0.17–0.81)	0.008
Glycemia (mg/dL), median (IQR)	120 (107–143)	124 (112–157)	119 (106–141)	0.004
HbA1c (%), median (IQR)	5.7 (5.5–5.9)	5.7 (5.5–5.9)	5.7 (5.4–5.9)	0.167
Creatinine (mg/dL), median (IQR)	0.89 (0.77–1.08)	0.93 (0.82–1.25)	0.88 (0.76–1.05)	0.001
Cholesterol (mg/dL), median (IQR)	153 (134–177)	153 (132–158)	153 (134–181)	0.025
Triglycerides (mg/dL), median (IQR)	100 (79–129)	100 (81–122)	100 (78–133)	0.405
High-density lipoprotein (HDL) (mg/dL), median (IQR)	46 (39–53)	46 (40–49)	46 (39–54)	0.708
Low-density lipoprotein (LDL) (mg/dL), median (IQR)	95 (78–112)	95 (73–98)	95 (78–116)	0.002
White Blood Cells (WBC) ($\times 10^9$/L), median (IQR)	8.59 (7.00–10.8)	8.59 (7.08–11.89)	8.59 (6.99–10.76)	0.107
Neutrophils ($\times 10^9$/L), median (IQR)	6.27 (4.58–8.23)	6.27 (4.98–8.88)	6.02 (4.52–8.11)	0.005
Lymphocytes ($\times 10^9$/L), median (IQR)	1.55 (1.10–2.10)	1.40 (0.91–1.97)	1.63 (1.15–2.11)	0.001
Red Blood Cells (RBC) ($\times 10^{12}$/L), median (IQR)	4.41 (4.07–4.74)	4.41 (4.01–4.81)	4.41 (4.09–4.72)	0.447
Hematocrit (HCT) (%), median (IQR)	39.2 (36.0–42.2)	39.20 (35.3–42.4)	39.20 (36.20–42.20)	0.566
Platelets (PLT) ($\times 10^9$/L), median (IQR)	224 (188–273)	224 (187–284)	224 (188–269)	0.368
Fibrinogen (mg/dL), median (IQR)	341 (286–399)	359 (315–446)	341 (277–394)	0.002
International normalized ratio (INR), median (IQR)	1.20 (1.20–1.20)	1.20 (1.20–1.21)	1.20 (1.20–1.20)	0.009
Neutrophil-to-lymphocyte ratio (NLR), median (IQR)	4.06 (2.41–6.50)	3.95 (2.86–8.67)	3.95 (2.29–5.86)	0.001
Systemic inflammatory index (SII), median (IQR)	908 (511–1522)	919 (618–1951)	918 (494–1395)	0.006
Clinical data				
pre-mRS $\geq$ 2, *n* (%)	91/602 (15%)	32/133 (24%)	59/469 (13%)	0.002
Baseline NIHSS, median (IQR)	18 (12–22)	20 (16–24)	17 (11–21)	<0.001

Table 1. *Cont.*

Variables	All Patients (N = 602)	Patients with the Miserable Outcome (N = 133)	Patients with the Favorable Outcome (N = 469)	*p* Value
Wake-up stroke, *n* (%)	114/602 (19%)	32/133 (24%)	82/469 (17%)	0.113
ASPECTS > 6, *n* (%)	536/602 (89%)	116/133 (87%)	420/469 (90%)	0.546
TOAST				
Cardioembolic cause (CE), *n* (%)	331/602 (55%)	86/133 (65%)	245/469 (52%)	0.015
Large artery atherosclerosis (LAA), *n* (%)	81/602 (13%)	12/133 (9%)	69/469 (15%)	0.120
Other, or unknown cause, *n* (%)	190/602 (32%)	35/133 (26%)	155/469 (33%)	0.171
Thrombolysis, *n* (%)	361/602 (60%)	64/133 (48%)	297/469 (63%)	0.002
Occluded vessel and leptomeningeal collaterals				
Vessel type, *n* (%)				
Middle Cerebral Artery-M1 segment	335/602 (56%)	55/133 (41%)	280/469 (60%)	<0.001
Middle Cerebral Artery-M2 segment	47/602 (8%)	11/133 (8%)	36/469 (8%)	0.966
Internal Carotid Artery-I type	93/602 (15%)	20/133 (15%)	73/469 (16%)	0.990
Internal Carotid Artery-L type	33/602 (5%)	12/133 (9%)	21/469 (4%)	0.069
Internal Carotid Artery-T type	94/602 (16%)	35/133 (26%)	59/469 (13%)	<0.001
Vessel side, *n* (%)				0.567
Right	287/602 (48%)	60/133 (45%)	227/469 (48%)	
Left	315/602 (52%)	73/133 (55%)	242/469 (52%)	
Leptomeningeal collaterals (LC), *n* (%)				0.001
Good or Equal to the unaffected side	454/602 (75%)	85/133 (64%)	369/469 (79%)	
Absent	148/602 (25%)	48/133 (36%)	100/469 (21%)	
Endovascular procedure information				
Onset-to-puncture time (OPT)	192 (154–270)	200 (152–329)	195.0 (140–285)	0.181
Puncture-to-end time (PET), median (IQR)	52 (38–66)	62 (54–90)	55 (33–79)	0.358
Onset to the procedure end time (OPET)	264 (219–318)	276 (237–389)	269 (205–343)	0.269
Number of steps, median (IQR)	2 (1–3)	3 (2–4)	2 (1–4)	0.011
Procedure type, *n* (%)				
Failed/Abandoned thrombectomy, *n* (%)	55/602 (9%)	11/133 (8%)	44/469 (9%)	0.824
Aspiration only	229/602 (38%)	41/133 (31%)	188/469 (40%)	0.066
Stent-retriever only	24/602 (4%)	5/133 (4%)	19/469 (4%)	1.000
Aspiration and stent-retriever	294/602 (49%)	76/133 (57%)	218/469 (46%)	0.038
Procedure complications, *n* (%)				
No complications	586/602 (97%)	129/133 (97%)	457/469 (97%)	1.000
Downstream complications	11/602 (2%)	3/133 (2%)	8/469 (2%)	0.959
Distal complications	5/602 (1%)	1/133 (1%)	4/469 (1%)	1.000
Vessel perforation	45/602 (7%)	8/133 (6%)	37/469 (8%)	0.590
Bleeding type, *n* (%)				
No bleeding	415/602 (69%)	85/133 (64%)	330/469 (70%)	0.189

Table 1. *Cont.*

Variables	All Patients (N = 602)	Patients with the Miserable Outcome (N = 133)	Patients with the Favorable Outcome (N = 469)	*p* Value
Symptomatic bleeding	18/602 (3%)	1/133 (1%)	17/469 (4%)	0.153
Any other bleeding	169/602 (28%)	47/133 (35%)	122/469 (26%)	0.045
Thrombolysis in Cerebral Infarction (TICI), *n* (%)				0.040
Incomplete recanalization (TICI = 0-2a)	152/602 (25%)	24/133 (18%)	128/469 (27%)	
Complete recanalization (TICI = 2b-3)	450/602 (75%)	109/133 (82%)	341/469 (73%)	
Internal carotid artery stenting, *n* (%)	34/602 (6%)	11/133 (8%)	23/469 (5%)	0.203
Osteoclastic decompressive craniectomy, *n* (%)	29/602 (5%)	8/133 (6%)	21/469 (4%)	0.616

IQR, interquartile range; mRS, modified Rankin scale; NIHSS, the National Institutes of Health Stroke Scale; ASPECTS, Alberta stroke program early CT score; TOAST, Trial of ORG 10,172 in Acute Stroke Treatment.

Following a prior procedure, the data were divided into the outcome (target variable) and the input features. Outcome was represented by the parameter 'in-hospital death', which included two options: (I) *miserable outcome*, if a patient died during an in-hospital stay, and (II) *favorable outcome* if a patient survived. In this research, the terms in-hospital death and miserable outcome were alternately used to describe the same result, as we wanted to differentiate it from the unfavorable, or poor outcome, which could refer to different degrees of disability and therefore diminish the significance of the issue. The input features consisted of 53 variables, which were statistically compared between the two possible outcomes (Supplementary Table S1). Categorical features were expressed as numbers (percentages), and the chi-squared test was used to determine differences between the two outcome groups. Continuous features were presented as medians (interquartile range) and analyzed by the Student *t*-test or Mann–Whitney U test, based on sample normality. Two-tailed tests were used, and statistical significance was observed at level $p < 0.05$ for every variable. To address the collinearity between features, a correlation analysis was performed. We used the chi-squared test of independence for categorical features, presenting results as *p* values, with a statistically significant correlation being $p < 0.05$. The Pearson and Spearman Correlation coefficients were calculated for continuous features, based on their distribution, representing results as coefficients ranging from −1 to +1, presenting the strength and direction of the linear relationship between two continuous variables.

Data normalization followed, which is one of the first preprocessing tasks to be performed during analysis [33], and it implies that the data are transformed or scaled so that an equal contribution of each feature is achieved. This was carried out by Z-score normalization, in which the values of a feature are normalized based on the mean (average) value and standard deviation, as this process reduces numerical instabilities between the analyzed features [34,35].

2.2. Two-Step Feature Selection Process

After the previous steps, variables were divided into two categories: (I) *Baseline features* ($n = 42$) that were gathered or present during the patient's admission, and (II) *Intervention-related features* ($n = 11$) collected or calculated after the endovascular treatment. As the input features included 53 variables, the feature selection technique was applied as it enables the selection of those features that contribute the most to the prediction variable or the outcome in machine learning (ML) algorithms. It is primarily carried out to remove unimportant features, improve analysis efficacy, and adapt the data set to best suit ML classifiers [36]. First, irrelevant features were removed after the univariate analysis, and then, redundant and useless features were excluded by a wrapper approach [37]. Univariate analysis was based on the features' type. The Shapiro–Wilk normality test was conducted

before the continuous variables analysis [38], and as asymmetric distribution was observed, the Mann–Whitney-Wilcoxon test was applied [39], while the chi-squared test was used in the categorical variables analysis [40]. After the removal of the irrelevant features, Recursive Feature Elimination (RFE) was used, as a method of wrapper models [41]. This algorithm eliminates one backward feature during one iteration and prevents information loss while showing high classification performances when compared to many other feature reduction methods [42]. This algorithm was used with four different classifiers—logistic regression (LG), random forest (RF), gradient boosting (GB), and extreme gradient boosting (XGB)—and the number of possible features to be included in the classification was not limited, and it solely depended on the model. The RFE process was performed for every classifier separately, and two features' sets were gained: (I) *Pre-MT set*, including only baseline features, and (II) *Post-MT set*, which included both baseline and intervention-related variables.

To statistically address the influence of chosen variables on the miserable outcome, binomial regression analysis was performed, calculating the odds ratios and plotting the estimated marginal means plots. The most optimal cut-off values to predict the miserable outcomes were determined based on sensitivity and specificity.

2.3. Data Sampling and Construction of Two Predictive Models

As the outcome group was unbalanced (miserable outcome, n = 133 vs. favorable outcome, n = 469), we performed a data sampling technique to reduce the disparity. There are two basic sampling techniques, including random oversampling (ROS) and random undersampling (RUS). Despite the lack of conclusive evidence concerning the superiority of any approach, it has been reported that oversampling may lead to overfitting of the model [43–46], which is why we chose the RUS as more appropriate. In this approach, the majority classes are randomly eliminated, so that equal distribution can be achieved [47].

By random splitting, data were divided into three groups: (I) a *Training set* that included 70% of the input data, (II) an *Evaluation set* consisting of 15%, and (III) a *Test set* comprising 15% of the original data. Using Python's Scikit-learn library, four classifiers were trained, internally evaluated, and tested: logistic regression (LR), random forest (RF), gradient boosting (GB), and extreme gradient boosting (XGB). As previously mentioned, two predictive models were created for every classifier based on the input features' sets: (I) *Pre-MT model*, and (II) *Post-MT model*. During the training, each classifier used RFE-preselected feature sets, and the training included 10-fold cross-validation, during which the training set was divided into 1 validating and 9 training subsets (Figure 1). After the internal evaluation and testing, results were expressed through evaluation metrics such as accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic (AUC-ROC) curve. Since the AUC has been proven to be a better measure than accuracy in the learning algorithm evaluation [48], it was used as a measure of discrimination [49], while the model calibration was expressed through calibration curves and Brier scores, where lower scores reflected better model calibration [50].

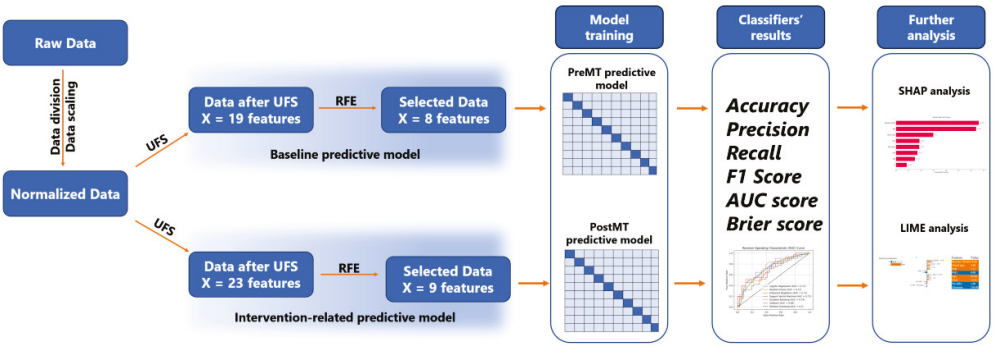


Figure 1. Summarized steps of the analysis. After a two-step feature selection process, with univariate feature selection (UFS) as the first step and Recursive Feature Elimination (RFE) as the second, models

were trained, evaluated, and tested. After obtaining the classifiers' results, features and individual level interpretations were performed for the best-performing classifiers in both models, using SHAP and LIME.

2.4. Interpretable Framework-Sharpley Additive Explanation (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME)

Best-performing classifiers, based on the AUC value, were more thoroughly analyzed through Sharpley Additive Explanation (SHAP) values. This locally interpretable explanatory method can rationalize machine learning (ML) algorithm predictions, regardless of their complexity, as it identifies and maps features that reduce or increase the probability of the predicted outcome [51]. Additionally, SHAP can help with the understanding of features' importance and, therefore, enhance their clinical interpretability [25]. Features were presented based on their order of importance (Absolute Mean SHAP Plot) in the predictive model.

To better understand the classification-making process on an individual level, Local Interpretable Model-Agnostic Explanations (LIMEs) were applied, as they can extract the contribution of key features during it [52]. This analysis was performed for both models, randomly choosing a patient from both groups. This figure (Figure 5)/Figure 5. shows the overall predicted probability of in-hospital death and favorable outcome on the left, classification details in the middle, and feature values and categories on the right. Continuous features were presented with their true values, while the categorical features were labeled based on the included category.

3. Results

3.1. Analyzed Group

A total of 602 patients fulfilled the inclusion criteria. In our cohort, around 54% (n = 328) of all patients were women, 133 (22%) had miserable outcomes, and the median age was 76 years (interquartile range (IQR) 65–83 years). Statistical feature comparison is presented in Table 1, and the correlation matrices are stored in Supplementary Table S2 for categorical features and Supplementary Table S3 for continuous ones. Out of all baseline variables, age (81 (IQR 74–87) vs. 74 (IQR 62–82), $p < 0.001$), baseline NIHSS value (20 (IQR 16–24) vs. 17 (IQR 11–21), $p < 0.001$), and peripheral artery disease (19/133 (14%) vs. 14/469 (3%), $p < 0.001$) showed the most statistically significant disparity between the two analyzed groups, while the most significant intervention-related feature was the 'Number of steps' (3 (IQR 2–4) vs. 2 (IQR 1–4), $p = 0.011$). Correlation analysis showed high collinearity between the previous stroke, the presence of peripheral arterial disease, and higher pre-mRS values.

3.2. The Results of a Two-Step Feature Selection Process

During a two-step process, variables for the predictive models were chosen as a product of univariate feature selection (UFS) and Recursive Feature Elimination (RFE). In the baseline set, those 8 features were age, baseline NIHSS value, neutrophil-to-lymphocyte ratio (NLR), international normalized ratio (INR), the type of the affected vessel ('Vessel type'), peripheral arterial disease (PAD), baseline glycemia, and premorbid modified Rankin scale (pre-mRS). In the intervention-related set, the RFE method did not include pre-mRS, but it added two features: puncture-to-the-procedure end time (PET), and onset-to-puncture time (OPT). The RFE-selected sets were used in further analysis. The whole process is summarized in Figure 1.

Binomial logistic regression results are presented in Table 2, showing the influence of chosen features on in-hospital death (IHD) occurrence. Out of continuous features, INR had the highest impact with the most optimal cut-off value of 1.29, and an increase of 0.2 led to 37.2% higher chances of IHD. One unit increase in baseline NIHSS increased the risk of IHD by approximately 9%, with the optimal cut-off value of 20, one unit

of NLR (the optimal cut-off = 3.8) as well as 10 mg/dL of blood glucose (the optimal cut-off = 120 mg/dL) by 7%, and one year of age (the optimal cut-off value = 78 years) by 6%. Selected time frames showed a lower influence on IHD, where every 10 min of OPT (the optimal cut-off value = 285 min) increased risk by 1% and 10 min of PET (the optimal cut-off value = 54 min) by 2%. The cut-off values tables are stored in Supplementary Table S4. Categorical variables showed high influence, as the presence of peripheral arterial disease (PAD) increased the chances of developing IHD 3.5 times and a premorbid mRS value of 2 or higher by 80%. The results are plotted in Figure 2.

Table 2. Binomial logistic regression results of the chosen parameters associated with in-hospital death after mechanical thrombectomy for anterior large vessel occlusion stroke, N = 602.

Predictor	Estimate	95% Confidence Interval		SE	z	*p* Value	Odds Ratio	95% Confidence Interval	
		Lower	Upper					Lower	Upper
Age	0.053	0.034	0.074	0.010	5.289	<0.001	1.061	1.034	1.076
Baseline NIHSS	0.071	0.037	0.105	0.017	4.104	<0.001	1.086	1.038	1.111
Glycemia	0.005	0.000	0.009	0.002	2.317	0.020	1.007	1.001	1.009
NLR	0.055	0.016	0.095	0.020	2.736	0.006	1.071	1.016	1.100
INR	1.050	0.398	1.710	0.334	3.150	<0.001	2.866	1.488	5.517
PAD	1.505	0.751	2.259	0.384	3.910	<0.001	4.504	2.120	9.569
pre-mRS	0.590	0.074	1.105	0.263	2.240	0.025	1.803	1.077	3.019
OPT	0.000	0.000	0.001	0.000	0.258	0.796	1.001	0.999	1.001
PET	0.002	0.001	0.006	0.002	1.049	0.294	1.002	0.998	1.006

SE, standard error; NIHSS, National Institutes of Health Stroke Scale; NLR, Neutrophil-to-lymphocyte ratio; INR, international normalized ratio; PAD, peripheral arterial disease; pre-mRS, premorbid modified Rankin scale; OPT, onset-to-puncture time; PET, puncture-to-end time.

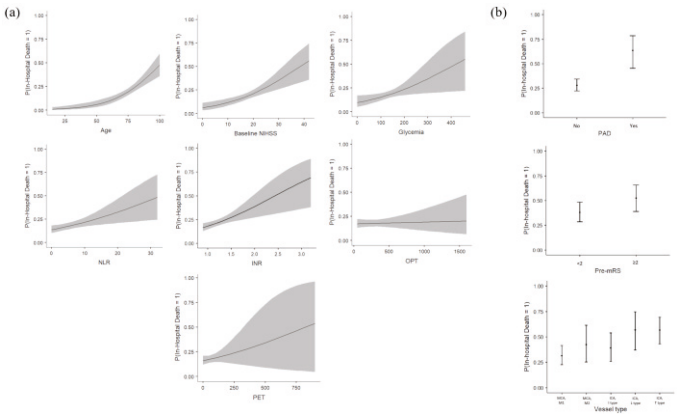


Figure 2. Estimated marginal means graphs of chosen features. Continuous (**a**) and categorical (**b**) features have been presented in the graph. Continuous features are represented with their true values, and categorical based on the analyzed categories. Estimated marginal means of continuous features are presented with a solid line, representing the predicted probability of in-hospital death based on the respective variable shown on the x-axis. The gray shaded areas depict the 95% confidence intervals around the estimated marginal means, with a wider gray area indicating more uncertainty and a narrower gray area indicating less uncertainty about the estimate. NIHSS, National Institutes of Health Stroke Scale; NLR, neutrophil-to-lymphocyte ratio; INR, international normalized ratio; OPT, onset-to-puncture time; PET, puncture-to-end time; PAD, peripheral arterial disease; pre-mRS, premorbid modified Rankin scale; MCA, middle cerebral artery; ICA, internal carotid artery.

3.3. Predictive Models Evaluation

After the training, internal evaluation, and testing, the performances of the classifiers were expressed through the previously mentioned parameters (Table 3). Their ROC curves

and the calibration plot with reliability curves are visualized in Figure 3. For the Pre-MT model, gradient boosting (GB) showed itself as the best-performing classifier, during all phases, with an AUC of almost 80% (0.7903) and the lowest Brier score of 0.2009. The post-MT classifier evaluation showed that the extreme gradient boosting (XGB) had the highest performance with an AUC score of 84% (0.8372) and a Brier score of 0.1194.

Table 3. Evaluation of the classifiers associated with in-hospital death after mechanical thrombectomy for anterior large vessel occlusion stroke, N = 602.

Classifier		Accuracy	Precision	Recall	F1-Score	AUC-ROC	Brier Score
Pre-MT Model	LR	0.7222	0.7083	0.6800	0.6939	0.7821	0.2205
	RF	0.7222	0.7273	0.6400	0.6809	0.7648	0.2058
	GB	0.6667	0.6207	0.7200	0.6667	0.7917	0.2009
	XGB	0.7037	0.6800	0.6800	0.6800	0.7503	0.2182
Post-MT Model	LR	0.7222	0.7273	0.6400	0.6809	0.7848	0.1328
	RF	0.7073	0.6800	0.6800	0.6800	0.8145	0.1192
	GB	0.6852	0.6667	0.6400	0.6531	0.7931	0.1197
	XGB	0.7407	0.7200	0.7200	0.7200	0.8372	0.1194

MT, mechanical thrombectomy; LR, logistic regression; RF, random forest; GB, gradient boosting; XGB, extreme gradient boosting; AUC-ROC, the area under the receiver operating characteristic curve.

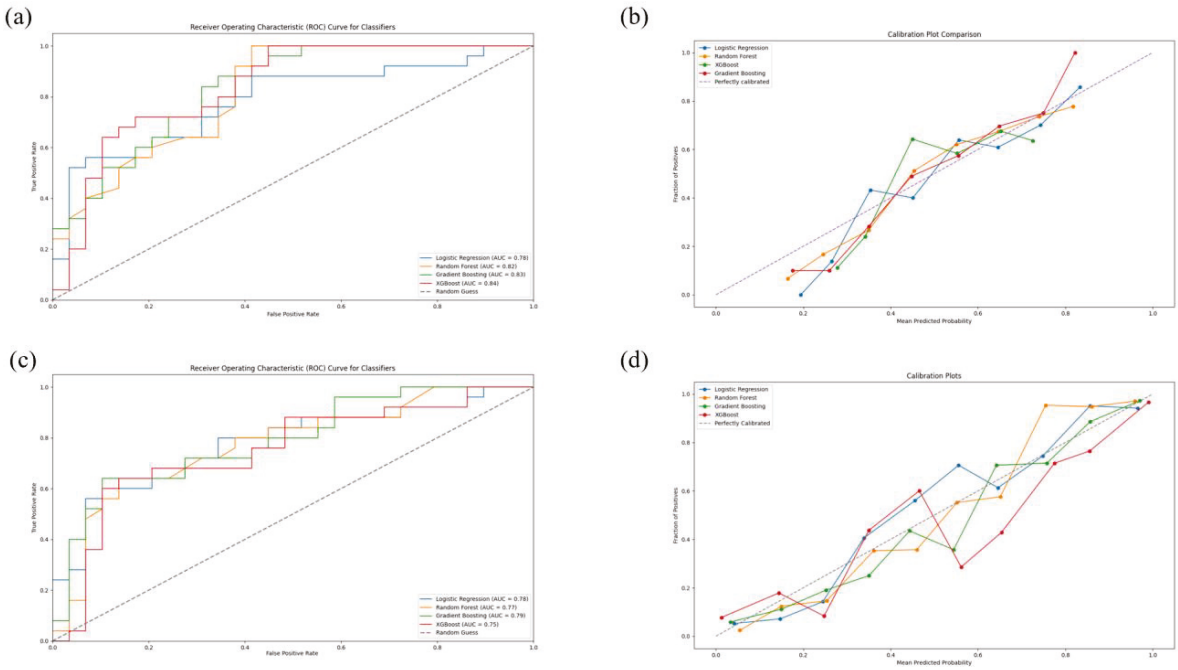


Figure 3. Evaluation of the predictive models. The receiver operating characteristic curves (ROCs) for pre-MT (**a**) and post-MT (**c**) models. Random guessing is represented by the black dashed line, with an AUC of 50% (0.5). The calibration plots with reliability curves for pre-MT (**b**) and post-MT (**d**) classifiers are also visualized, with the black dashed line representing perfect calibration.

3.3.1. SHAP Analysis–Feature Level Interpretation

For the best-performing classifier, in both models, SHAP values were calculated, and the order of importance is presented in Figure 4. Based on the SHAP analysis of the Pre-MT model (Figure 4a), the most influential features that increase the risk of developing a miserable outcome were baseline NIHSS, age, NLR, glycemia, and INR. In addition, pre-mRS of two or higher, internal carotid artery occlusions (type L and T ICA terminus occlusions),

and present peripheral arterial disease were also important contributing factors. The SHAP analysis of the XGB classifier (Figure 4b) shows that the same factors as previously mentioned influence the outcome prediction, with the addition of longer 'Puncture-to-end time' (PET) and 'Onset-to-the-puncture time' (OPT).

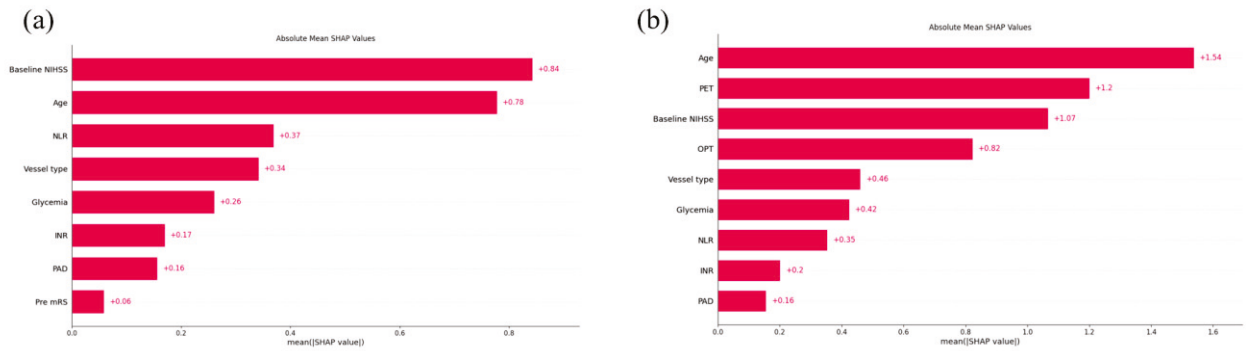


Figure 4. The SHAP analysis. The SHAP analysis graphs for in-hospital death prediction for the Pre-MT (**a**) and the Post-MT (**b**) models were visualized as absolute mean SHAP values. NIHSS, National Institutes of Health Stroke Scale; NLR, neutrophil-to-lymphocyte ratio; INR, international normalized ratio; PAD, peripheral arterial disease; pre-mRS, premorbid modified Rankin scale; OPT, onset-to-puncture time; PET, puncture-to-end time.

3.3.2. LIME Analysis–Individual Level Interpretation

The Pre-MT model's LIME analysis is visualized in Figure 5a, simplifying the GB classifier's decision-making algorithm, and Figure 5b visualizes the post-MT model, based on the prediction of XGB. Continuous features are presented with their true values, while the categorical features are labeled as follows. pre-mRS category 0 corresponds to a score of 2 or higher, peripheral arterial disease (PAD) category 1 represents the presence of the disease, while the vessel type is associated with the division in Table 1, within the 5 categories (1, M1 segment of the middle cerebral artery; 2, M2 segment of the middle cerebral artery; 3, internal carotid artery occlusion–type I; 4, internal carotid artery occlusion–type L; 5, internal carotid artery occlusion–type T). In chosen patients, probability scores, based on eight features in the Pre-MT model, which is nine in the Post-MT one, predict both favorable and miserable outcomes.

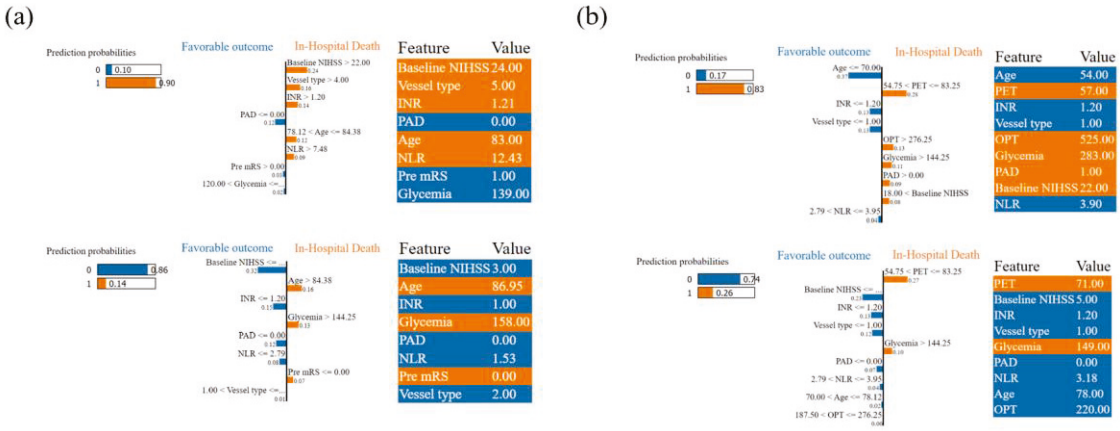


Figure 5. The LIME analysis. Local Interpretable Model-Agnostic Explanations (LIME) plot from the testing set of the GB (**a**) and XGB (**b**) models. True positive (upper examples) and true negative (lower examples) patients were chosen for the interpretation. NIHSS, National Institutes of Health Stroke Scale; INR, international normalized ratio; PAD, peripheral arterial disease; NLR, neutrophil-to-lymphocyte ratio; pre-mRS, premorbid modified Rankin scale; OPT, onset-to-puncture time; PET, puncture-to-end time.

4. Discussion

To the best of our knowledge, this is the first study that utilizes interpretable machine learning models for the in-hospital mortality analysis of patients who underwent thrombectomy for anterior circulation occlusive stroke. Four classifiers were trained, internally evaluated, and tested for both models. The two best-performing classifiers were the gradient boosting (GB) for the Pre-MT model and the extreme gradient boosting (XGB) for the Post-MT predictive model, which included both baseline and intervention-related parameters. Based on eight variables, GB achieved an AUC score of almost 80%, with a Brier score of 0.2, while the XGB accomplished better performances, an AUC score of 84%, and a Brier score of 0.12, using nine features. In addition to seven identical features that were incorporated into both models (age, baseline NIHSS value, neutrophil-to-lymphocyte ratio (NLR), international normalized ratio (INR), the type of the affected vessel, baseline glycemia, and the presence of peripheral arterial disease), the baseline model also included premorbid modified Rankin scale (pre-mRS), while the second model involved puncture-to-the-procedure-end time (PET) and onset-to-puncture time (OPT). We aimed to more thoroughly explain the "black box" of ML models through Sharpley Additive Explanation (SHAP), which interpreted the model at the feature level, and Local Interpretable Model-Agnostic Explanations (LIMEs), used for the interpretation at the individual level.

4.1. Baseline Features Prior to Thrombectomy

In our baseline, Pre-MT predictive model, baseline NIHSS value and age stood out as the most significant features, which is similar to the previous research [20,22]. Out of all analyzed laboratory findings, the neutrophil-to-lymphocyte ratio followed by baseline glycemia and INR value have proven themselves as the most appropriate for mortality prediction.

Hyperglycemia was linked to unfavorable clinical outcomes and mortality, in patients with large vessel occlusion (LVO) that were treated with the MT [53], which probably occurs as a consequence of metabolic alterations, such as intracellular acidosis, decreased mitochondrial function, and increased levels of reactive oxygen species (ROS), leading to neuronal damage [54]. In a previously published study, in which interpretable machine learning was used to predict the development of 24 h early neurological deterioration (END) in MT-treated ischemic stroke patients, hyperglycemia was the most important feature that contributed to the unfavorable outcome [25]. Even though hypoglycemia was proven to predict higher 90-day mortality in MT-treated patients [54], it did not demonstrate itself as significant in this analysis. A second laboratory feature, the neutrophil-to-lymphocyte ratio (NLR), was also linked with the stroke outcome. Higher NLR was associated with a more common death, unfavorable, or poor 90-day outcome in ischemic stroke patients who underwent reperfusion therapy [55,56]. Accompanied by hyperglycemia, which is an important factor for the development of thromboinflammation [57], elevated NLR represents a response to the inflammatory nature of stroke, which is a major contributor to its pathobiology and outcome [58]. Although MT can be safely and effectively performed in emergent large vessel occlusion (ELVO) patients with high INR [59], these patients are more prone to worse clinical outcomes and higher mortality risk [60], which was also found in our study.

Taking into consideration occluded vessels, our analysis has shown that the T-type internal carotid occlusion was mostly associated with in-hospital death, while the isolated middle cerebral artery (MCA) occlusion, specifically in the M1 segment, was more abundant in the favorable outcome group. In the available literature, no research has explicitly analyzed the connection between the occlusion type and in-hospital mortality. However, it was found that the internal carotid artery (ICA) occlusion is a significant predictor of 90-day mortality [61] and that the L- and T-types of terminal carotid occlusion were more frequent in the 90-day mortality group [62]. Outcomes of the thrombectomy of the MCA M1 and M2 segments showed that the 90-day functional outcomes did not differ between segments [63]. Although peripheral artery disease showed no association with

the 90-day functional or safety outcomes in a large cohort study [64], it certainly is a strong independent predictor for stroke [65], and in our analysis, it was an important contributing factor to the miserable outcome.

It is estimated that only around 20% of the pre-stroke-disabled patients (mRS > 2) return to their premorbid values after the MT, and, additionally, a good outcome is less frequent, while the mortality and risk of symptomatic intracranial hemorrhage (sICH) are higher [66–69]. Based on the previous information, higher premorbid mRS could be classified as an important unfavorable factor, which was also chosen by our feature selection process.

4.2. Features Pertinent to Mechanical Thrombectomy

Similar to the previous model, in previously published works, most of the intervention-related features have shown significance in the 90-day functional outcome and survival prediction, but their role in the context of in-hospital mortality has not been analyzed. It is shown that shorter onset-to-puncture (OPT) is associated with a lower degree of 3-month disability and higher independence at discharge [70–73], and besides long-term outcomes, in this research, it was shown that the longer OPT is an important predictor of in-hospital death. Other than this, the feature selection model also included PET, which showed even higher predictive capabilities, compared to the OPT. In the available literature, this parameter is associated with 3-month mortality, especially if it exceeded 60 min, and post-interventional complications if this period went beyond 35 min. Possible explanations for this occurrence might be that the longer procedure duration usually happens when there is an occlusive material resistant to extraction. Besides this, an extended intervention timeframe can burden compensatory collaterals and lead to irreversible injuries [72–74]. Even though these factors individually demonstrated low predictive power, they collectively contributed to the model's overall predictions. This highlights the importance of analyzing multiple features simultaneously rather than isolated, a task that is effectively performed by machine learning models.

There are several study limitations. First, all included patients died due to stroke, or its complications, but the precise cause of death was not more thoroughly differentiated in the research. As some of the most common conditions that contribute to in-hospital mortality are respiratory infections and brain edema [20,22], future studies may focus on these as an outcome and not the overall in-hospital mortality of ischemic stroke. Second, in our study, we used a hybrid feature selection approach, which is just one of many techniques that can be used during this preprocessing phase. It means that a variety of different methods could potentially be applied to this phase in the forthcoming studies, so that their influence on the results may become a subject of the analysis. Third, regardless of the high predictive power, for the evaluation and implementation of tested models, a higher number of included patients is necessary. Finally, this study only included patients with anterior circulation ischemic stroke, so it does not apply to the posterior vasculature.

5. Conclusions

In this study, two supervised ML models were constructed to predict in-hospital death after the anterior circulation MT, in large-vessel-occlusion-related acute ischemic stroke patients. The constructed models achieved good predictive performances, and the prediction-making process was more thoroughly understood with the usage of SHAP and LIME. The baseline parameters showed the strongest impact on the miserable outcome prediction, but the highest predictive power was obtained when time frames, OPT and PET, were also included in the model. We found that baseline predictors NIHSS and age, neutrophil-to-lymphocyte ratio, followed by baseline glycemia and INR value prove to be the most appropriate for mortality prediction. In addition, a longer time elapsed from arterial puncture to the end of the intervention was also associated with higher mortality. The majority of features related to endovascular treatment did not show an essential role in the miserable outcome prediction, since factors like the success of recanalization (TICI) and

intervention-related complications demonstrated low or almost no impact on in-hospital death prediction.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics14141531/s1, Table S1: Breakdown of features used in the research; Table S2: Categorical features correlation; Table S3: Continuous features correlation; Table S4: Cut-off values for continuous features.

Author Contributions: Conceptualization, I.P., S.B. and S.P.; methodology, I.P.; software, I.P.; formal analysis, I.P., S.B., I.M. and A.B.; resources, M.K.-O., J.A.R.P. and C.J.G.; data curation, M.K.-O., J.A.R.P. and C.J.G.; writing—original draft preparation, I.P., S.B., I.M. and A.B.; writing—review and editing, S.P.; visualization, I.P.; supervision, J.A.R.P., J.S.M. and S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study protocol was per the guidance of the Christian Doppler Klinik (Salzburg, Austria) committee for the protection of human subjects (protocol UN 2553). The local ethics committee obtained written approval for the retrospective study of patients with acute ischemic stroke (415-EP/73/750-2017).

Informed Consent Statement: According to Austrian regulations, informed consent is not required for routinely collected clinical and radiological data as used in this study.

Data Availability Statement: All data generated or analyzed during this study are included in this published article.

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

References

1. Yang, C.; Zhu, C.; Sui, Y.; Guo, Y.; Wu, N.; Dong, Q.; Xu, R.; Qian, J.; Li, J. Synergetic impact of lipoprotein(a) and fibrinogen on stroke in coronary artery disease patients. *Eur. J. Clin. Investig.* **2024**, *54*, e14179. [CrossRef] [PubMed]
2. Kamel, H.; Healey, J.S. Cardioembolic Stroke. *Circ. Res.* **2017**, *120*, 514–526. [CrossRef]
3. Feske, S.K. Ischemic Stroke. *Am. J. Med.* **2021**, *134*, 1457–1464. [CrossRef]
4. Krishna, J.T.V.; Kumar, P. *Risk Factor Assessment, Etiology, Clinico-Radiological Profile and Prognosis in Cva*; NaN, 2023; pp. 37–39. Available online: https://www.citedrive.com/en/discovery/risk-factor-assessment-etiologyclinico-radiological-profile-and-prognosis-in-cva/ (accessed on 17 June 2024).
5. Deng, G.; Chu, Y.-H.; Xiao, J.; Shang, K.; Zhou, L.-Q.; Qin, C.; Tian, D.-S. Risk Factors, Pathophysiologic Mechanisms, and Potential Treatment Strategies of Futile Recanalization after Endovascular Therapy in Acute Ischemic Stroke. *Aging Dis.* **2023**, *14*, 2096–2112. [CrossRef] [PubMed]
6. Tsalta-Mladenov, M.E.; Andonova, S.P. Quality of life after ischaemic stroke-accent on patients with thrombolytic therapy. *Egypt. J. Neurol. Psychiatry Neurosurg.* **2021**, *57*, 163. [CrossRef]
7. Matsumoto, K.; Nohara, Y.; Soejima, H.; Yonehara, T.; Nakashima, N.; Kamouchi, M. Stroke Prognostic Scores and Data-Driven Prediction of Clinical Outcomes After Acute Ischemic Stroke. *Stroke* **2020**, *51*, 1477–1483. [CrossRef]
8. Phipps, M.S.; Cronin, C.A. Management of acute ischemic stroke. *BMJ* **2020**, *368*, l6983. [CrossRef] [PubMed]
9. Abdu, H.; Tadese, F.; Seyoum, G. Comparison of Ischemic and Hemorrhagic Stroke in the Medical Ward of Dessie Referral Hospital, Northeast Ethiopia: A Retrospective Study. *Neurol. Res. Int.* **2021**, *2021*, 9996958. [CrossRef] [PubMed]
10. Berge, E.; Whiteley, W.; Audebert, H.; De Marchis, G.M.; Fonseca, A.C.; Padiglioni, C.; de la Ossa, N.P.; Strbian, D.; Tsivgoulis, G.; Turc, G. European Stroke Organisation (ESO) guidelines on intravenous thrombolysis for acute ischaemic stroke. *Eur. Stroke J.* **2021**, *6*, I–LXII. [CrossRef]
11. Turc, G.; Turc, G.; Bhogal, P.; Fischer, U.; Khatri, P.; Lobotesis, K.; Mazighi, M.; Schellinger, P.D.; Toni, D.; De Vries, J.; et al. European Stroke Organisation (ESO)-European Society for Minimally Invasive Neurological Therapy (ESMINT) Guidelines on Mechanical Thrombectomy in Acute Ischemic Stroke. *J. Neurointerv. Surg.* **2023**, *15*, e8. [CrossRef]
12. Powers William, J. Acute Ischemic Stroke. *N. Engl. J. Med.* **2020**, *383*, 252–260. [CrossRef]
13. Wang, R.P.C.; Li, S.; Hao, L.; Wang, Z.B.; Ge, Z.B.; Yang, S. A meta-analysis of intravenous thrombolysis versus bridging therapy for ischemic stroke. *Medicine* **2022**, *101*, e30879. [CrossRef]
14. Berkhemer, O.A.; Fransen, P.S.S.; Beumer, D.; Berg, L.A.V.D.; Lingsma, H.F.; Yoo, A.J.; Schonewille, W.J.; Vos, J.A.; Nederkoorn, P.J.; Wermer, M.J.H.; et al. A randomized trial of intraarterial treatment for acute ischemic stroke. *N. Engl. J. Med.* **2015**, *372*, 11–20. [CrossRef] [PubMed]

15. Campbell, B.C.; Mitchell, P.J.; Kleinig, T.J.; Dewey, H.M.; Churilov, L.; Yassi, N.; Yan, B.; Dowling, R.J.; Parsons, M.W.; Oxley, T.J.; et al. Endovascular therapy for ischemic stroke with perfusion-imaging selection. *N. Engl. J. Med.* **2015**, *372*, 1009–1018. [CrossRef] [PubMed]
16. Goyal, M.; Demchuk, A.M.; Menon, B.K.; Eesa, M.; Rempel, J.L.; Thornton, J.; Roy, D.; Jovin, T.G.; Willinsky, R.A.; Sapkota, B.L.; et al. Randomized assessment of rapid endovascular treatment of ischemic stroke. *N. Engl. J. Med.* **2015**, *372*, 1019–1030. [CrossRef]
17. Jovin, T.G.; Chamorro, A.; Cobo, E.; De Miquel, M.A.; Molina, C.A.; Rovira, A.; Román, L.S.; Serena, J.; Abilleira, S.; Ribo, M.; et al. Thrombectomy within 8 hours after symptom onset in ischemic stroke. *N. Engl. J. Med.* **2015**, *372*, 2296–2306. [CrossRef]
18. Stent-Retriever Thrombectomy after Intravenous t-PA vs. t-PA Alone in Stroke | NEJM [Internet]. Available online: https://www.nejm.org/doi/full/10.1056/nejmoa1415061 (accessed on 14 February 2024).
19. Goyal, M.; Menon, B.K.; Van Zwam, W.H.; Dippel, D.W.J.; Mitchell, P.J.; Demchuk, A.M.; Dávalos, A.; Majoie, C.B.L.M.; Van Der Lugt, A.; De Miquel, M.A.; et al. Endovascular thrombectomy after large-vessel ischaemic stroke: A meta-analysis of individual patient data from five randomised trials. *Lancet* **2016**, *387*, 1723–1731. [CrossRef] [PubMed]
20. Johnston, S.C.; Mendis, S.; Mathers, C.D. Global variation in stroke burden and mortality: Estimates from monitoring, surveillance, and modelling. *Lancet Neurol.* **2009**, *8*, 345–354. [CrossRef]
21. Sluis, W.M.; Hinsenveld, W.H.; Goldhoorn, R.-J.B.; Potters, L.H.; Bruggeman, A.A.; van der Hoorn, A.; Bot, J.C.; van Oostenbrugge, R.J.; Lingsma, H.F.; Hofmeijer, J.; et al. Timing and causes of death after endovascular thrombectomy in patients with acute ischemic stroke. *Eur. Stroke J.* **2023**, *8*, 215–223. [CrossRef]
22. Bustamante, A.; García-Berrocoso, T.; Rodriguez, N.; Llombart, V.; Ribó, M.; Molina, C.; Montaner, J. Ischemic stroke outcome: A review of the influence of post-stroke complications within the different scenarios of stroke care. *Eur. J. Intern. Med.* **2016**, *29*, 9–21. [CrossRef]
23. Broderick, J.P.; Adeoye, O.; Elm, J. Evolution of the Modified Rankin Scale and Its Use in Future Stroke Trials. *Stroke* **2017**, *48*, 2007–2012. [CrossRef]
24. Benali, F.; Kappelhof, M.; Ospel, J.; Ganesh, A.; McDonough, R.V.; Postma, A.A.; Goldhoorn, R.-J.B.; Majoie, C.B.L.M.; Wijngaard, I.v.D.; Lingsma, H.F.; et al. Benefit of successful reperfusion achieved by endovascular thrombectomy for patients with ischemic stroke and moderate pre-stroke disability (mRS 3): Results from the MR CLEAN Registry. *J. Neurointerv. Surg.* **2023**, *15*, 433–438. [CrossRef]
25. Yang, T.; Hu, Y.; Pan, X.; Lou, S.; Zou, J.; Deng, Q.; Zhang, Q.; Zhou, J.; Zhu, J. Interpretable Machine Learning Model Predicting Early Neurological Deterioration in Ischemic Stroke Patients Treated with Mechanical Thrombectomy: A Retrospective Study. *Brain Sci.* **2023**, *13*, 557. [CrossRef]
26. De Bin, R.; Benner, A.; Ambrogi, F.; Lusa, L.; Boulesteix, A.-L.; Migliavacca, E.; Binder, H.; Michiels, S.; Sauerbrei, W.; McShane, L.; et al. Statistical analysis of high-dimensional biomedical data: A gentle introduction to analytical goals, common approaches and challenges. *BMC Med.* **2023**, *21*, 182. [CrossRef]
27. Statistics for Machine Learning | Packt [Internet]. Available online: https://www.packtpub.com/product/statistics-for-machine-learning/9781788295758 (accessed on 21 May 2024).
28. Hoffman, H.; Wood, J.; Cote, J.R.; Jalal, M.S.; Otite, F.O.; Masoud, H.E.; Gould, G.C. Development and Internal Validation of Machine Learning Models to Predict Mortality and Disability After Mechanical Thrombectomy for Acute Anterior Circulation Large Vessel Occlusion. *World Neurosurg.* **2023**, *182*, E137–E154. [CrossRef] [PubMed]
29. Campagnini, S.; Arienti, C.; Patrini, M.; Liuzzi, P.; Mannini, A.; Carrozza, M.C. Machine learning methods for functional recovery prediction and prognosis in post-stroke rehabilitation: A systematic review. *J. Neuroeng. Rehabil.* **2022**, *19*, 54. [CrossRef]
30. Hu, C.; Li, L.; Huang, W.; Wu, T.; Xu, Q.; Liu, J.; Hu, B. Interpretable Machine Learning for Early Prediction of Prognosis in Sepsis: A Discovery and Validation Study. *Infect. Dis. Ther.* **2022**, *11*, 1117–1132. [CrossRef] [PubMed]
31. Ting Sim, J.Z.; Fong, Q.W.; Huang, W.; Tan, C.H. Machine learning in medicine: What clinicians should know. *Singap. Med. J.* **2023**, *64*, 91–97.
32. Bustamante, A.; Giralt, D.; García-Berrocoso, T.; Rubiera, M.; Álvarez-Sabín, J.; Molina, C.; Serena, J.; Montaner, J. The impact of post-stroke complications on in-hospital mortality depends on stroke severity. *Eur. Stroke J.* **2016**, *2*, 54–63. [CrossRef]
33. Izonin, I.; Tkachenko, R.; Shakhovska, N.; Ilchyshyn, B.; Singh, K.K. A Two-Step Data Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain. *Mathematics* **2022**, *10*, 1942. [CrossRef]
34. Mukhyber, S.J.; DhahirAbdulhadeAbdulah, A.D.M. Effect Z-score Normalization on Accuracy of classification of liver disease. *Turk. J. Comput. Math. Educ. (TURCOMAT)* **2021**, *12*, 658–662.
35. Mihaela, G.; Ruxandra, P.S. The importance of normalization methods for mining medical data. *Int. J. Comput. Technol.* **2015**, *14*, 6014–6020. [CrossRef]
36. Jain, S.; Saha, A. Rank-based univariate feature selection methods on machine learning classifiers for code smell detection. *Evol. Intell.* **2021**, *15*, 609–638. [CrossRef]
37. [PDF] A Two-Phase Feature Selection Method Using Both Filter and Wrapper | Semantic Scholar [Internet]. Available online: https://www.semanticscholar.org/paper/A-two-phase-feature-selection-method-using-both-and-Yuan-Tseng/1d46dc187f99524367948ae7b3683f27f333c36b (accessed on 21 May 2024).
38. Yap, B.W.; Sim, C.H. Comparisons of various types of normality tests. *J. Stat. Comput. Simul.* **2011**, *81*, 2141–2155. [CrossRef]

39. Martínez-Murcia, F.J.; Górriz, J.M.; Ramírez, J.; Puntonet, C.G.; Salas-González, D. Computer Aided Diagnosis tool for Alzheimer's Disease based on Mann–Whitney–Wilcoxon U-Test. *Expert Syst. Appl.* **2012**, *39*, 9676–9685. [CrossRef]
40. Sarkar, S.D.; Goswami, S.; Agarwal, A.; Aktar, J. A Novel Feature Selection Technique for Text Classification Using Naïve Bayes. *Int. Sch. Res. Not.* **2014**, *2014*, e717092. [CrossRef] [PubMed]
41. (PDF) Improving the Classification Accuracy Using Recursive Feature Elimination with Cross-Validation [Internet]. Available online: https://www.researchgate.net/publication/344181117_Improving_the_Classification_Accuracy_using_Recursive_Feature_Elimination_with_Cross-Validation (accessed on 21 May 2024).
42. Lu, X.; Yang, Y.B.; Wu, F.; Gao, M.B.; Xu, Y.; Zhang, Y.B.; Yao, Y.B.; Du, X.B.; Li, C.B.; Wu, L.B.; et al. Discriminative analysis of schizophrenia using support vector machine and recursive feature elimination on structural MRI images. *Medicine* **2016**, *95*, e3973. [CrossRef] [PubMed]
43. Liu, A.Y.C. The Effect of Oversampling and Undersampling on Classifying Imbalanced Text Datasets. Available online: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=cade435c88610820f073a0fb61b73dff8f006760 (accessed on 17 June 2024).
44. Mohammed, R.; Rawashdeh, J.; Abdullah, M. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. In Proceedings of the 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 7–9 April 2020; pp. 243–248. Available online: https://ieeexplore.ieee.org/abstract/document/9078901 (accessed on 22 May 2024).
45. Chawla, N.V.; Japkowicz, N.; Kotcz, A. Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.* **2004**, *6*, 1–6. [CrossRef]
46. Batista, G.E.A.P.A.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.* **2004**, *6*, 20–29. [CrossRef]
47. Yap, B.W.; Rani, K.A.; Rahman, H.A.A.; Fong, S.; Khairudin, Z.; Abdullah, N.N. An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. In Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013), Kuala Lumpur, Malaysia, 16–18 December 2013; Herawan, T., Deris, M.M., Abawajy, J., Eds.; Springer: Singapore, 2014; pp. 13–22.
48. Ling, C.X.; Huang, J.; Zhang, H. AUC: A Better Measure than Accuracy in Comparing Learning Algorithms. In *Advances in Artificial Intelligence*; Xiang, Y., Chaib-Draa, B., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2003; pp. 329–341.
49. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature | Users' Guide to the Medical Literature | JAMA | JAMA Network [Internet]. Available online: https://jamanetwork.com/journals/jama/fullarticle/2656816 (accessed on 22 May 2024).
50. Jelovsek, J.E.; Hill, A.J.; Chagin, K.M.; Kattan, M.W.; Barber, M.D. Predicting Risk of Urinary Incontinence and Adverse Events after Midurethral Sling Surgery in Women. *Obstet. Gynecol.* **2016**, *127*, 330–340. [CrossRef]
51. Rodríguez-Pérez, R.; Bajorath, J. Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. *J. Med. Chem.* **2019**, *63*, 8761–8777. [CrossRef] [PubMed]
52. Katuwal, G.J.; Chen, R. Machine Learning Model Interpretability for Precision Medicine [Internet]. *arXiv* **2016**. Available online: http://arxiv.org/abs/1610.09045 (accessed on 22 May 2024).
53. Perez-Vega, C.; Domingo, R.A.; Tripathi, S.; Ramos-Fresnedo, A.; Kashyap, S.; Quinones-Hinojosa, A.; Lin, M.P.; Fox, W.C.; Tawk, R.G. Influence of glucose levels on clinical outcome after mechanical thrombectomy for large-vessel occlusion: A systematic review and meta-analysis. *J. Neurointerv. Surg.* **2021**, *14*, 17–21. [CrossRef]
54. Wang, L.; Zhou, Z.; Tian, X.; Wang, H.; Yang, D.; Hao, Y.; Shi, Z.; Lin, M.; Wang, Z.; Zheng, D.; et al. Impact of Relative Blood Glucose Changes on Mortality Risk of Patient with Acute Ischemic Stroke and Treated with Mechanical Thrombectomy. *J. Stroke Cerebrovasc. Dis.* **2019**, *28*, 213–219. [CrossRef] [PubMed]
55. Pikija, S.; Sztriha, L.K.; Killer-Oberpfalzer, M.; Weymayr, F.; Hecker, C.; Ramesmayer, C.; Hauer, L.; Sellner, J. Neutrophil to lymphocyte ratio predicts intracranial hemorrhage after endovascular thrombectomy in acute ischemic stroke. *J. Neuroinflamm.* **2018**, *15*, 319. [CrossRef]
56. Sharma, D.; Spring, K.J.; Bhaskar, S.M.M. Role of Neutrophil-Lymphocyte Ratio in the Prognosis of Acute Ischaemic Stroke After Reperfusion Therapy: A Systematic Review and Meta-analysis. *J. Central Nerv. Syst. Dis.* **2022**, *14*, 11795735221092518. [CrossRef]
57. Desilles, J.-P.; Syvannarath, V.; Ollivier, V.; Journé, C.; Delbosc, S.; Ducroux, C.; Boisseau, W.; Louedec, L.; Di Meglio, L.; Loyau, S.; et al. Exacerbation of Thromboinflammation by Hyperglycemia Precipitates Cerebral Infarct Growth and Hemorrhagic Transformation. *Stroke* **2017**, *48*, 1932–1940. [CrossRef] [PubMed]
58. Inflammation and Stroke: An Overview—PubMed [Internet]. Available online: https://pubmed.ncbi.nlm.nih.gov/27730544/ (accessed on 22 May 2024).
59. Pandhi, A.; Tsivgoulis, G.; Ishfaq, M.F.; Katsanos, A.; Magoufis, G.; Malhotra, K.; Krishnan, R.; Arthur, A.; Hoit, D.; Elijovich, L.; et al. Mechanical thrombectomy outcomes in large vessel stroke with high international normalized ratio. *J. Neurol. Sci.* **2019**, *396*, 193–198. [CrossRef]
60. Chen, H.; Ahmad, G.; Colasurdo, M.; Yarbrough, K.; Schrier, C.; Phipps, M.S.; Cronin, C.A.; Mehndiratta, P.; Cole, J.W.; Wozniak, M.; et al. Mildly elevated INR is associated with worse outcomes following mechanical thrombectomy for acute ischemic stroke. *J. Neurointerv. Surg.* **2022**, *15*, e117–e122. [CrossRef]

61. Nogueira, R.G.; Liebeskind, D.S.; Sung, G.; Duckwiler, G.; Smith, W.S. Predictors of good clinical outcomes, mortality, and successful revascularization in patients with acute ischemic stroke undergoing thrombectomy: Pooled analysis of the Mechanical Embolus Removal in Cerebral Ischemia (MERCI) and Multi MERCI Trials. *Stroke* **2009**, *40*, 3777–3783. [CrossRef]
62. Liebeskind, D.S.; Flint, A.C.; Budzik, R.F.; Xiang, B.; Smith, W.S.; Duckwiler, G.R.; Nogueira, R.G. Carotid I's, L's and T's: Collaterals shape the outcome of intracranial carotid occlusion in acute ischemic stroke. *J. Neurointerv. Surg.* **2015**, *7*, 402–407. [CrossRef] [PubMed]
63. Mönch, S.; Boeckh-Behrens, T.; Maegerlein, C.; Berndt, M.; Wunderlich, S.; Zimmer, C.; Friedrich, B. Mechanical Thrombectomy of the Middle Cerebral Artery—Neither Segment nor Diameter Matter. *J. Stroke Cerebrovasc. Dis.* **2019**, *29*, 104542. [CrossRef] [PubMed]
64. Pirson, F.A.; Hinsenveld, W.H.; Staals, J.; de Ridder, I.R.; van Zwam, W.H.; Schreuder, T.H.; Roos, Y.B.; Majoie, C.B.; van der Worp, H.B.; Uyttenboogaart, M.; et al. Peripheral Artery Disease in Acute Ischemic Stroke Patients Treated with Endovascular Thrombectomy; Results from the MR CLEAN Registry. *Front. Neurol.* **2020**, *11*, 560300. [CrossRef] [PubMed]
65. Meves, S.H.; Diehm, C.; Berger, K.; Pittrow, D.; Trampisch, H.-J.; Burghaus, I.; Tepohl, G.; Allenberg, J.-R.; Endres, H.G.; Schwertfeger, M.; et al. Peripheral arterial disease as an independent predictor for excess stroke morbidity and mortality in primary-care patients: 5-year results of the getABI study. *Cerebrovasc. Dis.* **2010**, *29*, 546–554. [CrossRef] [PubMed]
66. Millán, M.; Ramos-Pachón, A.; Dorado, L.; Bustamante, A.; Hernández-Pérez, M.; Rodríguez-Esparragoza, L.; Gomis, M.; Remollo, S.; Castaño, C.; Werner, M.; et al. Predictors of Functional Outcome After Thrombectomy in Patients with Prestroke Disability in Clinical Practice. *Stroke* **2022**, *53*, 845–854. [CrossRef] [PubMed]
67. Seker, F.; Pfaff, J.; Schönenberger, S.; Herweh, C.; Nagel, S.; Ringleb, P.; Bendszus, M.; Möhlenbruch, M. Clinical Outcome after Thrombectomy in Patients with Stroke with Premorbid Modified Rankin Scale Scores of 3 and 4: A Cohort Study with 136 Patients. *Am. J. Neuroradiol.* **2018**, *40*, 283–286. [CrossRef] [PubMed]
68. Brugnara, G.; Neuberger, U.; Mahmutoglu, M.A.; Foltyn, M.; Herweh, C.; Nagel, S.; Schönenberger, S.; Heiland, S.; Ulfert, C.; Ringleb, P.A.; et al. Multimodal Predictive Modeling of Endovascular Treatment Outcome for Acute Ischemic Stroke Using Machine-Learning. *Stroke* **2020**, *51*, 3541–3551. [CrossRef] [PubMed]
69. Asdaghi, N.; Wang, K.; Gardener, H.; Jameson, A.; Rose, D.Z.; Alkhachroum, A.; Gutierrez, C.M.; Ying, H.; Mueller-Kronast, N.; Sur, N.B.; et al. Impact of Time to Treatment on Endovascular Thrombectomy Outcomes in the Early Versus Late Treatment Time Windows. *Stroke* **2023**, *54*, 733–742. [CrossRef] [PubMed]
70. Ota, T.; Nishiyama, Y.; Koizumi, S.; Saito, T.; Ueda, M.; Saito, N. Impact of onset-to-groin puncture time within three hours on functional outcomes in mechanical thrombectomy for acute large-vessel occlusion. *Interv. Neuroradiol.* **2018**, *24*, 162–167. [CrossRef]
71. Saver, J.L.; Goyal, M.; van der Lugt, A.; Menon, B.K.; Majoie, C.B.L.M.; Dippel, D.W.; Campbell, B.C.; Nogueira, R.G.; Demchuk, A.M.; Tomasello, A.; et al. Time to Treatment with Endovascular Thrombectomy and Outcomes from Ischemic Stroke: A Meta-analysis. *JAMA* **2016**, *316*, 1279–1289. [CrossRef]
72. Farghaly, W.M.; Ahmed, M.A.; El-Bassiouny, A.; Hamed, A.A.; Shehata, G.A. Predictors of three months mortality after endovascular mechanical thrombectomy for acute ischemic stroke. *Egypt. J. Neurol. Psychiatry Neurosurg.* **2022**, *58*, 96. [CrossRef]
73. Spiotta, A.M.; Vargas, J.; Turner, R.; Chaudry, M.I.; Battenhouse, H.; Turk, A.S. The golden hour of stroke intervention: Effect of thrombectomy procedural time in acute ischemic stroke on outcome. *J. Neurointerv. Surg.* **2014**, *6*, 511–516. [CrossRef] [PubMed]
74. Alawieh, A.; Pierce, A.K.; Vargas, J.; Turk, A.S.; Turner, R.D.; Chaudry, M.I.; Spiotta, A.M. The golden 35 min of stroke intervention with ADAPT: Effect of thrombectomy procedural time in acute ischemic stroke on outcome. *J. Neurointerv. Surg.* **2018**, *10*, 213–220. [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.

diagnostics

MDPI

Article

Using Machine Learning and Feature Importance to Identify Risk Factors for Mortality in Pediatric Heart Surgery

Lorenz A. Kapsner [1,2], Manuel Feißt [3], Ariawan Purbojo [4], Hans-Ulrich Prokosch [1], Thomas Ganslandt [1], Sven Dittrich [5], Jonathan M. Mang [6,*,†] and Wolfgang Wällisch [5,†]

1 Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91058 Erlangen, Germany
2 Institute of Radiology, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91054 Erlangen, Germany
3 Institute of Medical Biometry, University of Heidelberg, 69117 Heidelberg, Germany
4 Department of Paediatric Cardiac Surgery, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91054 Erlangen, Germany
5 Department of Pediatric Cardiology, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91054 Erlangen, Germany
6 Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, 91054 Erlangen, Germany
* Correspondence: jonathan.mang@uk-erlangen.de
† These authors contributed equally to this work.

Abstract: Background: The objective of this IRB-approved retrospective monocentric study was to identify risk factors for mortality after surgery for congenital heart defects (CHDs) in pediatric patients using machine learning (ML). CHD belongs to the most common congenital malformations, and remains the leading mortality cause from birth defects. **Methods:** The most recent available hospital encounter for each patient with an age <18 years hospitalized for CHD-related cardiac surgery between the years 2011 and 2020 was included in this study. The cohort consisted of 1302 eligible patients (mean age [SD]: 402.92 [±562.31] days), who were categorized into four disease groups. A random survival forest (RSF) and the 'eXtreme Gradient Boosting' algorithm (XGB) were applied to model mortality (incidence: 5.6% [n = 73 events]). All models were then applied to predict the outcome in an independent holdout test dataset (40% of the cohort). **Results:** RSF and XGB achieved average C-indices of 0.85 (±0.01) and 0.79 (±0.03), respectively. Feature importance was assessed with 'SHapley Additive exPlanations' (SHAP) and 'Time-dependent explanations of machine learning survival models' (SurvSHAP(t)), both of which revealed high importance of the maximum values of *serum creatinine* observed within 72 h post-surgery for both ML methods. **Conclusions:** ML methods, along with model explainability tools, can reveal interesting insights into mortality risk after surgery for CHD. The proposed analytical workflow can serve as a blueprint for translating the analysis into a federated setting that builds upon the infrastructure of the German Medical Informatics Initiative.

Keywords: risk factors; mortality; congenital heart defects (CHDs); machine learning (ML); random survival forest (RSF); eXtreme Gradient Boosting (XGB); feature importance

1. Introduction

With a prevalence of approximately 1.08% of all newborns in Germany [1], congenital heart defects (CHDs) are some of the most common congenital malformations. Many different diseases with different etiologies, including genetic [2] and nongenetic [3] causes, comprise CHD, however, therapy often requires interventional or surgical treatment [4]. Despite advances in the survival of cardiac patients, CHD is still the leading cause of mortality from birth defects [5] and imposes a substantial long-term burden [6]. According to the report of the German Registry for Cardiac Operations and Interventions in Patients with CHD, the overall in-hospital mortality for all surgical cases (Norwood type operations

only registered since 2018) within a 6-year period (2013–2018) was 2.5%, rising to 6–10% in complex cases with multiple treatments [4]. International studies based on centers or registries have shown similar overall mortality rates ranging from 2.5 to 4.8% [7,8]. Globally, CHD-related mortality and morbidity remain significantly higher in low- and middle-income countries than in high-income countries [9]. Analysis of surgical outcomes and associated risk factors can therefore lead to shorter hospital stays [10] and reduce the long-term financial burden on healthcare systems [11,12]. The development of accurate risk prediction models is critical to improving outcomes.

While classical statistical methods such as the semiparametric Cox proportional hazards (CPH) regression for survival analysis have a well-founded theoretical–mathematical background and are well interpretable, clinicians and researchers strive to achieve various improvements by applying artificial intelligence (AI) methods in the healthcare domain, e.g., better detection and diagnosis of diseases [13]. However, a disadvantage of many AI algorithms is limited transparency regarding the underlying decision-making processes, which has led to the term 'black box algorithms' being coined for these systems [14]. This has recently led to the development of new statistical methods to ensure the comprehensibility and interpretability of the results generated by these AI methods. These efforts are referred to as *explainable artificial intelligence* (xAI) [15].

During the last decade, several machine learning (ML) algorithms have been adapted for modeling survival data. Moncanda-Torres et al. performed a survival analysis with two ML methods—a random survival forest (RSF) and the eXtreme Gradient Boosting algorithm 'xgboost' (XGB) [16]—to predict breast cancer mortality risk in a large cohort [17], applying XBG's implementation to perform Cox regression for survival data. For a better understanding of the models' results, they applied the xAI method *SHapley Additive exPlanations* (SHAP) [18], which revealed interesting insights into how the different features contributed to the ML models' decisions [17].

With respect to CHD, Du et al. recently applied XGB to predict the probability of in-hospital mortality after surgery in a large Chinese cohort of over 24,000 patients [19]. They were able to show that the ML method performed better in predicting the probability of in-hospital mortality after surgery compared to the RACHS-1 and STS-EACTS risk stratification scores [19]. Given the large number of different heart centers in Germany [4] and Central Europe, it would be interesting to link these data across institutions, hospitals, and federal states as well and apply these ML technologies in order to gain more insight into risk factors for CHD. The German Medical Informatics Initiative (MII) [20] established an infrastructure during the last 6 years with the goal of performing multicentric studies in a privacy-preserving manner in Germany.

As preparatory ground work, the aim of this feasibility study was to implement a data collection strategy and identify risk factors for mortality after surgery for CHD in pediatric patients at one German pediatric heart center. We applied ML methods and xAI to model post-surgery mortality risk and to allow for better interpretability of the results. Furthermore, the established analysis workflow could serve as a blueprint and benchmark for a future translation into a federated setting.

We start by describing the study sample (Section 2.1). Then, we explain the data collection (Section 2.2) and the data preprocessing and feature engineering steps (Section 2.3) and elaborate in detail on the ML experiments (Section 2.4), the application of the xAI methods (Section 2.5), and the statistical analyses (Section 2.6). In the Results section, we first provide details on the study sample using descriptive statistics (Section 3.1) and present the 'feature-engineered' dataset (Section 3.2) used to carry out the ML experiments (Section 3.3). The latter include computation of xAI measures (Section 3.3.1) and the application of a Cox proportional hazards regression model (Section 3.3.2) to compare the ML results with a standard statistical approach. We discuss our findings, relate them to the current literature, discuss the application of the utilized ML methods (Section 4), and elaborate on the limitations of the work (Section 4.1). Finally, we outline future efforts in

the identification of risk factors for mortality after surgery for CHD in pediatric patients (Section 5).

2. Material and Methods

After designing the study and defining the inclusion and exclusion criteria, data associated with eligible patient encounters were extracted from the clinical data warehouse of the University Hospital Erlangen (UHE), and each case was allocated to one of four disease groups—*univentricular heart disease 1 (UVHD I)*, *univentricular heart disease 2 (UVHD II)*, *biventricular complex heart disease (BVHD cmplx.)*, and *biventricular simple heart disease (BVHD smpl.)*—using dedicated combinations of procedure and diagnosis codes (ICD), as defined in Supplemental Table S1. The survival outcome was determined using the status information, if a patient was discharged alive or died during the hospital stay, and the post-surgery follow-up time, i.e., the duration in days between surgery and discharge from hospital. Data extraction was performed using Structured Query Language (SQL), and the data were temporarily saved in a text file (comma-separated file format) before being imported into R software (version 4.4.2) [21], in which all further steps of this work (data preprocessing, ML experiments, computation of the feature importance, data analysis) were carried out, as described below.

2.1. Ethics Statement and Study Sample

This monocentric retrospective study was approved by the ethics committee of the Friedrich-Alexander-University Erlangen-Nürnberg, waiving the need for informed consent. The authors declare that this research was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects. The study period was between the years 2011 and 2020. Inclusion criteria were hospitalization for cardiac surgery for CHD at the Pediatric Heart Center of the UHE within the study period and age at the time of the surgery < 18 years. The procedure codes (OPS-codes, the German modification of the International Classification of Procedures in Medicine) that define eligible surgical corrections of congenital heart disease are provided in Supplemental Table S1. In brief, univentricular hearts (UVHs) were characterized by the presence of one rudimentary and one dominant ventricle (left or right), regardless of whether they had double-inlet left or right ventricles or a single atrioventricular inlet. To retain a sufficient sample size for the analysis, the group *UVHD I* comprised both patients undergoing the Norwood procedure, either as primary procedure or in a comprehensive stage II surgery, and single ventricle patients with PA-Banding or Shunt as palliative surgery stage I (sub-groups *Ia* and *Ib* in Supplemental Table S1). Likewise, the *UVHD II* group represented patients undergoing stage II and III of univentricular pathway palliation or BCPC, Glenn and TCPC, or Fontan surgery (sub-groups *IIa* and *IIb* in Supplemental Table S1).

Assignment of cases to the cardiac diagnosis groups comprising simple and complex biventricular heart diseases was performed according to Erikssen et al. [22] and further adjusted by the corresponding mortality group described by Jacobs et al. [23]. The key element of risk stratification, and thus classification between these two diagnostic groups was based on the corresponding in-hospital mortality risk estimate. For example, we classified a patient with surgical correction of AVSD as having simple biventricular heart disease (*BVHD smpl.*) due to the low in-hospital mortality rate of 2.5%, but AVSD with TOF was classified as complex biventricular disease (*BVHD cmplx.*) [23].

Premature newborns and newborns with a birth-weight <2500 g with persistent ductus arteriosus (PDA) ligation as the only cardiac procedure were excluded from the dataset (newborns with birth-weights <2500 g not matching the before-mentioned criteria were not excluded). Furthermore, to ensure a complete heart surgery history for all patients, in the present study we included only the most recent available hospital encounters (cases) for patients who were born within the study period (date of birth $\geq$1 January 2011). The next subsection lists the data elements that were included in this analysis.

2.2. Data Collection

The dataset included patient-related (sex, weight, height), encounter-related (year of the hospital admission, days until discharge after surgery), disease-related (heart disease group, deceased status [censored/deceased], associated malformations, syndrome association, presence of chromosomal alterations, presence of pulmonary hypertension), surgery-related (duration of the surgery, duration of bypassing with a heart lung machine [HLM], duration of a circulatory arrest, hypothermia, duration of using an aortic cross clamp, if the thorax was left open after surgery), and post-surgery related data elements. The latter group comprises measurements of laboratory values observed within 72 h after surgery. Further details on the definitions of disease-related features are reported in Supplemental Table S2. In order to prepare the collected data for the ML experiments and the statistical modeling, data preprocessing and feature engineering steps were required, which are described in the following subsection.

2.3. Data Preprocessing and Feature Engineering

For the experiments in this work, the dataset was partitioned by 60% to 40% into a training dataset and a holdout test dataset. Missing values were imputed with multivariate imputation by chained equations using the 'mice' R package [24], with further details given in the supplemental methods. The engineering of new features was performed after imputation of missing values.

To incorporate information on each patient's hospitalization history, (A) the number of previous CHD-related hospital admissions with heart surgery-related procedure codes was recorded for each case in a new ordinal feature (*No. of previous hospitalizations*), and (B) the most severe heart disease group (as defined in the Methods section) that was available in the hospitalization history was recorded likewise in the feature *Heart disease history*. If previous hospitalizations for CHD surgery were present in a patient's history, the heart disease groups were ranked by severity (in decreasing order), as follows: *UVHD I*, *BVHD cmplx.*, *UVHD II*, and *BVHD smpl*. Due to a high correlation of the weight variable and the age at surgery, body weight below <2500 g was coded in a binary manner, and the continuous weight variable was removed from the dataset. Furthermore, the height variable was removed from the analysis, as it was also strongly correlated with age at surgery. Age at surgery was kept in the dataset in favor of age at hospital admission, as they were also highly correlated with each other. Furthermore, the duration of the hospital stay was also removed from the dataset, as it is a linear combination of the days between hospital admission and surgery and the days until discharge after surgery, the latter being the time component of the survival outcome. For each laboratory value observed within a 72 h post-surgery period, the parameter that is typically associated with a more severe clinical condition was kept in the dataset, i.e., the maximum levels for serum creatinine, urea, and C-reactive protein. In contrast, for leukocytes, the minimum values were kept for further analyses as an indicator for leukopenia. For the remaining continuous features, the absolute Pearson's correlation coefficient was <0.7. Finally, n = 19 independent variables were used to train the ML methods. The survival outcome was formed by the status variable and the post-surgery follow-up time, i.e., the duration in days between surgery and discharge from hospital (which equals the duration between surgery and death for deceased patients). Outlier cases observed with regard to the post-surgery follow-up time were censored at the 99.5% quantile of the post-surgery follow-up time, which was computed using the whole cohort. The as-such prepared data elements then served as input data for the ML experiments, which are described in the next subsection.

2.4. Machine Learning Experiments

Two ML algorithms that are known for their outstanding predictive abilities in modeling survival data were used throughout this analysis, namely an RSF using the implementation from the 'ranger' [25] R package (version 0.16.0) and XGB using the implementation from the R package 'xgboost' (version 1.7.8.1) [26]. Especially in the medical domain,

transparency is required for clinical decision support systems so that users are able to understand the systems' suggestions [27]. Both algorithms, RSF and XGB, belong to the family of 'tree-based models', which (at least to some extent) can be more interpretable as compared to other ML methods [28]. To make the results of the applied ML methods comparable with each other, predictions were obtained as risk scores. These scores can be used to rank observations by their mortality risk, even if predictions from different algorithms are on different scales. Predicting risk scores further allowed the use of Harrell's concordance index (C-index) [29] as an evaluation metric, which is commonly used to assess prognostic models in survival analyses.

First, the hyperparameters were optimized on the training dataset with three-fold cross validation (CV). A Bayesian hyperparameter optimization was employed using the R package 'ParBayesianOptimization' [30]. For all experiments, Bayesian optimization was parameterized with 128 sampling runs and with the parameter Kappa of the upper confidence bound set to $\kappa = 3.5$. Its default value is $\kappa = 2.576$, corresponding to the ~99th percentile of the upper confidence bound with higher values allow to increase the unexplored search space. For each ML method, the Bayesian process was initialized with 50 randomly chosen parameter settings from a pre-computed parameter grid. For XGB, the parameters optimized with the Bayesian process in this study included the maximum depth of a tree ('max_depth'), the step size shrinkage used in each update ('learning_rate'), the subsample ratio of the training instances ('subsample'), the subsample ratio of columns when constructing each tree ('colsample_bytree'), and the minimum sum of instance weight that is required in a child ('min_child_weight'). The number of rounds for boosting ('num_round') were optimized using 'early stopping', a technique that stops the learning process if no further improvement of the validation metric is observed for a pre-defined number of training iterations. Early stopping was set to 500 iterations in this study. The parameters optimized for the RSF were the number of trees ('num.trees'), the maximal tree depth ('max.depth'), the number of variables available for splitting in each node ('mtry'), the minimum node size ('min.node.size'), and the fraction of the training instances ('sample.fraction'). For comparability, the hyperparameters of the two ML methods were optimized on identical CV folds, which were generated in a stratified manner based on the time variable, the status variable, and the disease group using the 'splitTools' [31] R package.

The as-such identified hyperparameter setting with the highest cross-validated C-index of each ML method was then validated in a 10 times repeated 10-fold CV on the training dataset. The folds of the repeated CV were also generated in a stratified manner and provided to all validation experiments. All of the trained 10×10 repeated CV models per ML method were then finally applied to predict the outcome in the holdout test dataset.

Similar to Moncanda-Torres et al. [17], we here also computed a CPH regression to compare the results of the ML methods with a standard statistical method for survival analysis. Therefore, the CPH models were fitted using the 'survival' R package [32] following the same experimental setup as outlined above. A 10×10 repeated CV was employed using the previously computed validation CV folds, and finally, all resulting CPH models were applied to predict the outcome in the holdout test dataset. Due to the rather small number of events in the dataset, only a subset of the available features were used as independent variables for the CPH regression, namely the union set of the previously identified $n = 5$ most important features of each ML method. Additionally, one single CPH model was fitted with all training data observations to compare the CPH results of the 10×10 repeated CV with the statistical standard approach for modeling survival data. In order to further explore the results of the ML methods, model explainability algorithms were employed, which are described below.

2.5. Model Explainability Using SHAP

The *SHapley Additive exPlanations* (SHAP) framework for interpreting model predictions was published in 2017 by Lundberg and Lee [18]. SHAP has a solid theoretical

foundation based on Shapley values that were originally proposed in a game theory context by Lloyd S. Shapley [33]. These values show the magnitude of a feature's influence on the model prediction with respect to the base-level average. When applied in a local explanation setting, SHAP values can provide insights into how a specific observation's feature values contribute to the model's prediction with respect to a base-level prediction. By combining the local explanations of many observations in a dataset, global insights into the functioning of a model can be gained [34]. SHAP values can be computed efficiently for tree-based models using the TreeSHAP algorithm [34]. While this algorithm is already implemented in 'xgboost', SHAP values for 'ranger' models can be computed with the 'treeshap' [35] R package.

When performing ML experiments, the application of resampling strategies such as CV or repeated CV is indispensable for decent model evaluation [36]. For a global model explanation in this study, SHAP values were computed for all observations in the independent holdout test dataset using each of the 10×10 repeated CV models per ML method. The results were then aggregated by averaging the SHAP values for each observation and feature across all 100 repeated CV models per ML method in the following manner:

- Let $i : \{i = 1, 2, ..., n\}$ be the i-th observation in the survival dataset.
- Let $d_j : \{j = 1, 2, ..., p\}$ be the j-th feature in the dataset.
- Let $\phi(i_*, d_j)$ be the SHAP value of the j-th feature of observation i_*.
- Let $M_k : \{k = 1, 2, ..., l\}$ be the k-th repeated CV model.

Then, the for each feature and observation, the aggregated SHAP values ϕ_{agg} across all repeated CV models were computed as:

$$\phi_{agg}(i_*, d_j) = \frac{1}{l}\sum_{k=1}^{l} \phi(i_*, d_j) \tag{1}$$

This approach was used in favor of arbitrarily selecting one out of the 100 models to extract the feature importance values, e.g., by selecting the model with the highest evaluation metric or—more robust—by selecting the model whose mean or median evaluation metric was closest to the overall mean or median evaluation metric of the repeated CV experiment. The aim of aggregating SHAP values across all repeated CV models was to ensure that all information generated by the repeated CV experiments were incorporated into the feature importance measures.

Furthermore, we applied the algorithm *Time-dependent explanations of machine learning survival models* (SurvSHAP (t)), a generalization of SHAP to survival models proposed by Krzyzinski et al., to compute time-dependent feature importance for survival models [37]. This algorithm overcomes the limitation that applying explanation methods, which were originally intended for standard regression and classification tasks (such as SHAP) to survival models (such as CPH), results in losing some of the importance information originating from the survival function. As this algorithm was only available for local explanations [37], we proposed an extension to aggregate SurvSHAP(t) values for each feature and time point across multiple observations in order to derive global SurvSHAP(t) values:

- Let $t \in \{t_1, ..., t_m\}$ be the times to the event of interest, where $t_{i1} < t_{i2} < ... < t_{im}$.
- As defined by Krzyzinski et al. [37], $\phi_t(i_*, d_j)$ is the SurvSHAP(t) value of the j-th feature of observation i_* at time point t.

Then, global SurvSHAP(t) values ϕ_{gt} for a dataset with multiple observations can be computed as:

$$\phi_{gt}(i_*, d_j) = \frac{1}{n}\sum_{i=1}^{n} \phi_t(i_*, d_j) \tag{2}$$

Likewise as above for SHAP, SurvSHAP(t) values were computed for all observations in the independent holdout test dataset using each of the 10×10 repeated CV models per ML method. Similarly, these results were also aggregated by averaging the SurvSHAP(t)

values for each observation and feature across all 100 repeated CV models per ML method, resulting in aggregated global SurvSHAP(t) values ϕ_{agg_t}:

$$\phi_{agg_t}(i_*, d_j) = \frac{1}{l}\sum_{k=1}^{l} \phi_{gt}(i_*, d_j) \quad (3)$$

These aggregated global SurvSHAP(t) values were then used to visualize the time-dependent importance of the variables with regard to survival probability. In this study, SurvSHAP(t) values were computed only for the RSF, as in contrast to the R implementation of 'xgboost', the 'ranger' R package already allowed prediction of the survival function, which is a necessary prerequisite for the application of the SurvSHAP(t) algorithm.

2.6. Statistical Analysis

All experiments and statistical analyses were performed in the statistical programming language R, version 4.3.2 [21]. The experiments were computed on a 64-bit Windows 10 desktop PC with an Intel Core i7-6700 CPU @ 3.40GHz with 8 logical CPUs and 24 GB RAM. Summary statistics were computed in base R [21] and with the R package 'DescrTab2' [38]. Descriptive statistics include mean and standard deviation for continuous variables and relative and absolute frequencies for categorical variables. Stratified splitting of the survival dataset by means of the time variable, the status variable, and the disease group variable was performed with the 'splitTools' R package [31] (version 1.0.1) for both data partitioning and generating the CV folds. SHAP values were visualized using the 'shapviz' R package [39]. SurvSHAP(t) values were computed with the 'survex' R package [40].

3. Results

3.1. Sample Characteristics

A total of 1302 eligible patients (median age at hospital admission [IQR]: 159 [63 to 502.25] days) were included in this analysis, of which 50 patients (0.50 [0 to 17.25] days) were included in the disease group *univentricular heart disease 1 (UVHD I)*, 111 patients (1097 [219 to 1332] days) in the disease group *univentricular heart disease 2 (UVHD II)*, 291 patients (20 [0 to 232.50] days) in the disease group *biventricular complex heart disease (BVHD cmplx.)*, and 850 patients (169 [112 to 504.75] days) in the disease group *biventricular simple heart disease (BVHD smpl.)*. The median number of days of admission to the hospital prior to cardiac surgery was 1 (1 to 3) days. While the median duration of the whole hospital stay was 25 (14 to 55.50) days for patients with *UVHD I*, the patients in the *UVHD II* and *BVHD* groups were hospitalized for a shorter period, at 18 (9.50 to 32.50) and 8 (6 to 15) days, respectively. As expected, the *BVHD smpl.* group had the shortest median hospital stay, at 7 (6 to 11) days. Given the compelling necessity of surgical intervention in the neonatal and infant period, the median age at admission was lowest for the most severe heart disease categories, at 0.50 (0 to 17.25) days in the *UVHD I* group and 20 (0 to 232.50) days in the *BVDH cmplx.* cohort. The median time between admission and surgery varied considerably between patients with *UVHD I* and patients with *UVHD II*, at 6 (2.25 to 9.50) days and 1 (1 to 1) days, respectively. Patients with *BVHD smpl.* were admitted 1 (1 to 1) days on median before surgery, compared with 3 (1 to 7) days in the *BVDH cmplx.* cohort.

A total of 73 patients (5.6%) died during the post-surgery follow-up period (*UVHD I*: 62% [31/50 cases]; *UVHD II*: 7.2% [8/111 cases]; *BVHD cmplx.*: 9.6% [28/291 cases]; *BVHD smpl.*: 0.7% [6/850 cases]). The median post-surgery follow-up period (i.e., days between surgery and discharge from hospital) was 17 (11 to 48.75) days for patients with *UVHD I* and 14 (8 to 30.50) days for patients with *UVHD II*, whereas the follow-up periods for patients with *BVHD cmplx.* and *BVHD smpl.* were 10 (7 to 17) and 6 (5 to 8) days, respectively. Further patient characteristics and additional parameters, including disease-related information, are given in Table 1.

Table 1. Patient characteristics. State of the dataset before feature engineering. UVHD I: disease group univentricular heart defect (HD) 1; UVHD II: disease group univentricular HD 2; BVHD cmplx.: disease group biventricular HD complex; BVHD smpl.: disease group biventricular HD simple.

Category	Variable	UVHD I	UVHD II	BVHD Cmplx.	BVHD Smpl.	Total Cohort
Demographics	*N* (%)	50 (3.84)	111 (8.53)	291 (22.35)	850 (65.28)	1302 (100.00)
	Median age at adm. (IQR) [days]	0.50 (0 to 17.25)	1097 (219 to 1332)	20 (0 to 232.50)	169 (112 to 504.75)	159 (63 to 502.25)
	Median height (IQR) [cm]	51.50 (49.25 to 53.75)	92 (67.50 to 100)	54 (50 to 67.50)	65 (59 to 78)	64 (55 to 78)
	Missing height (%)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.08)	1 (0.08)
	Median weight (IQR) [kg]	3.30 (2.96 to 3.88)	13 (6.93 to 15)	3.96 (3.29 to 6.93)	69 (4.69 to 9)	5.81 (3.96 to 9.10)
	Missing weight (%)	0 (0.00)	5 (0.38)	3 (0.23)	21 (1.61)	29 (2.23)
	Sex: male (%)	30 (60.00)	73 (65.77)	187 (64.26)	441 (51.88)	731 (56.14)
	Sex: female (%)	20 (40.00)	38 (34.23)	104 (35.74)	409 (48.12)	571 (43.86)
Encounter-related	Median days duration of stay (IQR)	25 (14 to 55.50)	18 (9.50 to 32.50)	15 (9 to 23.50)	7 (6 to 11)	9 (6 to 17)
	Median days between adm. and surgery (IQR)	6 (2.25 to 9.50)	1 (1 to 1)	3 (1 to 7)	1 (1 to 1)	1 (1 to 3)
	Median days disch. after surgery (IQR)	17 (11 to 48.75)	14 (8 to 30.50)	10 (7 to 17)	6 (5 to 8)	7 (5 to 12)
Disease-related	Median no. of previous adm. (IQR)	0 (0 to 0)	2 (1 to 2)	0 (0 to 1)	0 (0 to 0)	0 (0 to 0)
	Deceased status (%)	31 (62.00)	8 (7.21)	28 (9.62)	6 (0.71)	73 (5.61)
	Malformations (%)	16 (32.00)	22 (19.82)	39 (13.40)	88 (10.35)	165 (12.67)
	Chrom. alterations (%)	3 (6.00)	2 (1.80)	26 (8.93)	180 (21.18)	211 (16)
	Pulm. hypertension (%)	0 (0.00)	2 (1.80)	4 (1.37)	9 (1.06)	15 (1)

As some cases in our cohort had very long in-hospital follow-up times, with one patient being discharged 260 days after the surgery, to address potential biasing of the results, outliers with follow-up times above the 99.5% percentile, which corresponded to 130 days until discharge after surgery, were censored at that time point (see also Supplemental Figure S1). This affected a total of seven outlier cases, including $n = 1$ event.

As outlined in the methods section, only the most recent CHD-related hospitalization for each patient was included in this analysis. However, to incorporate information on each patient's hospitalization history, the most severe heart disease group that was recorded in prior hospitalizations (which were available from the hospital information system) was added as a new feature *heart disease history* in this analysis. Supplemental Table S3 shows the number of patients of each heart defect group stratified by *heart disease history* and deceased status. It can be seen that there were 96 patients ($n = 5$ events) in the heart disease groups *BVHD cmplx.*, *BVHD smpl.*, and *UVHD II*, which were categorized as *UVHD Ia* or *UVHD Ib* groups in previous hospitalizations.

The partitioning of the as-such prepared dataset resulted in $n = 780$ patients (60% of the cohort) that were allocated to the training dataset, of which 31 patients ($n = 18$ events) belonged to the *UVHD I* group, 65 patients ($n = 3$ events) belonged to the *UVHD II* group, 172 patients ($n = 17$ events) belonged to the *BVHD cmplx.* group, and 512 patients ($n = 4$ events) belonged to the *BVHD smpl.* group (Supplemental Table S4). The median post-surgery follow-up periods were 15 days (UVHD I), 14 days (UVHD II), 10 days (BVHD cmplx.), and 6 days (BVHD smpl.). The independent holdout test dataset consisted of $n = 522$ observations (40% of the cohort), of which 19 observations ($n = 13$ events) belonged to the disease group with *UVHD I*, 46 observations ($n = 4$ events) belonged to the disease group with *UVHD II*, 119 observations ($n = 11$ events) belonged to the disease group with *BVHD cmplx.*, and 338 observations ($n = 2$ events) belonged to the disease group with *BVHD smpl.* (Supplemental Table S4). The median post-surgery follow-up periods were 33 days (*UVHD I*), 14.5 days (*UVHD II*), 10 days (*BVHD cmplx.*), and 6 days (*BVHD smpl.*).

3.2. Feature Engineering and Feature Selection

As a preparation step for the ML experiments, features were engineered and selected after missing values had been imputed following the approach outlined in the Methods section. A total of 21 features were finally selected for the ML experiments, comprising 19 predictor variables and the survival outcome, which was formed by the status variable and the post-surgery follow-up time. A summary of the final dataset used for the ML experiments is given in Table 2.

Table 2. Machine learning dataset. State of the dataset after feature engineering. [1]: Chi-squared test. [2]: Wilcoxon rank sum test. UVHD 1: disease group univentricular heart defect (HD) 1; UVHD 2: disease group univentricular HD 2; BVHD cmplx.: disease group biventricular HD complex; BVHD smpl.: disease group biventricular HD simple.

Category	Variables		Censored	Deceased	Total	*p*
Demographics	Sex	m	686 (56%)	45 (62%)	731 (56%)	0.263 [1]
		w	544 (44%)	27 (38%)	571 (44%)	
	Weight <2500 g	No	1201 (98%)	62 (86%)	1263 (97%)	<0.001 [1]
		Yes	29 (2%)	10 (14%)	39 (3%)	
Disease-related	Chrom. alterations	No	1027 (83%)	64 (89%)	1091 (84%)	0.227 [1]
		Yes	203 (17%)	8 (11%)	211 (16%)	
	Disease group	BVHD cmplx.	263 (21%)	28 (39%)	291 (22%)	
		BVHD smpl.	844 (69%)	6 (8%)	850 (65%)	
		UVHD I	19 (2%)	31 (43%)	50 (4%)	<0.001 [1]
		UVHD II	104 (8%)	7 (10%)	111 (9%)	
	Heart disease history	No previous hospitalization	956 (78%)	56 (78%)	1012 (78%)	0.360 [1]
		BVHD smpl.	52 (4%)	0 (0%)	52 (4%)	
		UVHD II/UVHD III	5 (0%)	0 (0%)	5 (0%)	
		BVHD cmplx.	123 (10%)	9 (12%)	132 (10%)	
		UVHD Ib	35 (3%)	4 (6%)	39 (3%)	
		UVHD Ia	59 (5%)	3 (4%)	62 (5%)	
	Malformations	No	1092 (89%)	45 (62%)	1137 (87%)	<0.001 [1]
		Yes	138 (11%)	27 (38%)	165 (13%)	
	Pulm. hypertension	No	1215 (99%)	72 (100%)	1287 (99%)	0.346 [1]
		Yes	15 (1%)	0 (0%)	15 (1%)	
Encounter-related	Days between admission and surgery	mean ± sd	3.5 ± 9.8	6.9 ± 7.5	3.7 ± 9.7	<0.001 [2]
		min−max	0–182	0–34	0–182	
	Days until discharge after surgery	mean ± sd	12 ± 16	27 ± 28	13 ± 18	<0.001 [2]
		min − max	1–130	0–122	0–130	
	No. of previous admissions	mean ± sd	0.32 ± 0.66	0.32 ± 0.69	0.32 ± 0.66	0.967 [2]
		min − max	0–4	0–3	0–4	
Laboratory analytes	C-reactive protein (maximum)	mean ± sd	58 ± 44	68 ± 56	58 ± 45	0.326 [2]
		min − max	1–356	0.1–259	0.1–356	
	Leukocytes (minimum)	mean ± sd	9.7 ± 3.5	6.9 ± 3.6	9.6 ± 3.6	<0.001 [2]
		min − max	1.5–26	1.6–16	1.5–26	
	Serum creatinine (maximum)	mean ± sd	0.44 ± 0.26	0.85 ± 0.39	0.47 ± 0.28	<0.001 [2]
		min − max	0.17–3.2	0.25–2.2	0.17–3.2	
	Urea (maximum)	mean ± sd	32 ± 17	48 ± 21	33 ± 17	<0.001 [2]
		min − max	5–152	12–107	5–152	

Table 2. *Cont.*

Category	Variables		Censored	Deceased	Total	*p*
Surgery-related	Age at surgery	mean ± sd	422 ± 568	152 ± 344	407 ± 561	<0.001 [2]
		min − max	0–3466	0–1597	0–3466	
	Aortic cross clamp time	mean ± sd	72 ± 57	68 ± 61	72 ± 57	0.437 [2]
		min − max	0–280	0–284	0–284	
	Circulatory arrest during surgery	No	1091 (89%)	44 (61%)	1135 (87%)	<0.001 [1]
		Yes	139 (11%)	28 (39%)	167 (13%)	
	Heart lung machine during surgery	0 min.	165 (13%)	10 (14%)	175 (13%)	0.282 [1]
		≥1 min–<90 min	207 (17%)	7 (10%)	214 (16%)	
		≥90 min	858 (70%)	55 (76%)	913 (70%)	
	Hypothermia during surgery	>32 °C	308 (25%)	19 (26%)	327 (25%)	<0.001 [1]
		≥28 °C–≤32 °C	455 (37%)	12 (17%)	467 (36%)	
		<28 °C	467 (38%)	41 (57%)	508 (39%)	
	Open thorax	No	1162 (94%)	31 (43%)	1193 (92%)	<0.001 [1]
		Yes	68 (6%)	41 (57%)	109 (8%)	

3.3. Machine Learning Experiments

Details on the configuration of the Bayesian hyperparameter optimization, as well as the identified hyperparameter settings for XGB and RSF, are given in Table 3. The optimal number of boosting iterations for XGB was 22. On the independent holdout test dataset, the RSF models trained with a 10 × 10 repeated CV achieved a C-index of 0.85 (±0.01) on average, whereas the XGB models achieved a C-index of 0.79 (±0.03) on average (Figure 1). The prediction performance of each algorithm computed with the full independent holdout test dataset, as well as stratified by disease groups, is given in Supplemental Figure S2. It can be seen that RSF and XGB performed best in *BVHD smpl.* cases, with average C-indices of 0.92 (±0.02) and 0.85 (±0.07), respectively, whereas CPH showed the best results in *UVHD II* cases, with an average C-index of 0.82 (±0.06).

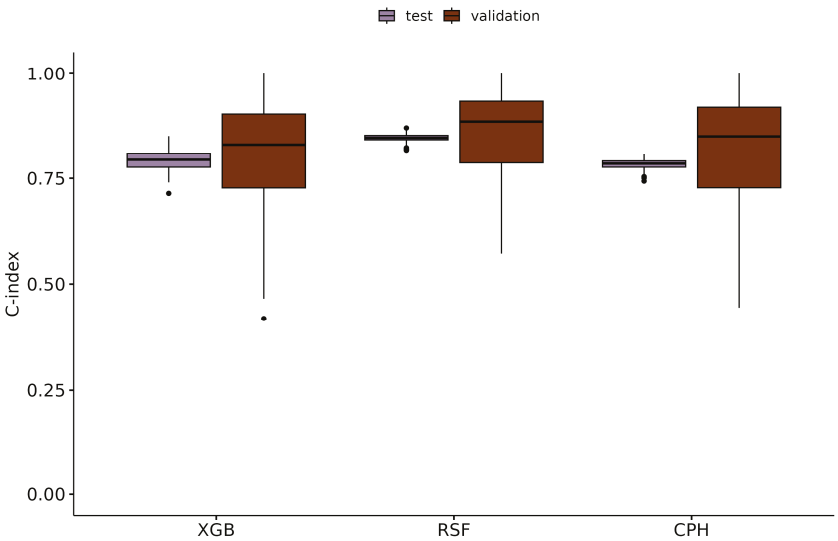


Figure 1. Performance: boxplots to visualize the performance of the applied models during validation and when predicting the outcome in the independent holdout test dataset. The underlying data for each boxplot are the performance of the 100 models from the repeated CV during validation and when applying these 100 models to predict the outcome in the holdout test dataset. XGB: xgboost; RSF: random survival forest; CPH: Cox proportional hazards regression.

Table 3. Optimized hyperparameters. XGB: xgboost; RSF: random survival forest.

Algorithm	Hyperparameter	Minimum	Maximum	Step Size	Bounds	Optimized Value
XGB	colsample_bytree	0.5	0.8	0.3	[0.3, 1]	0.8
	learning_rate	0.01	0.11	0.05	[0.001, 0.2]	0.11
	max_depth	1	9	4	[1, 40]	5
	min_child_weight	1	9	4	[0, 10]	1
	subsample	0.5	0.8	0.3	[0.3, 1]	0.5
RSF	max.depth	1	9	4	[1, 40]	40
	min.node.size	1	9	4	[1, 20]	20
	mtry	2	6	2	[2, 9]	2
	num.trees	500	1000	500	[100, 1000]	100
	sample.fraction	0.5	0.8	0.3	[0.3, 1]	0.63

3.3.1. Feature Importance

For each ML method, SHAP values were computed for all observations in the holdout test dataset using all trained models from the respective 10 × 10 repeated CV. The global SHAP values averaged by feature and observation across all of the respective 100 CV models are visualized with beeswarm plots in Figure 2 for XGB (left) and RSF (right). The corresponding mean absolute SHAP values are given in in Supplemental Table S5. For the two ML methods XGB and RSF, SHAP identified the maximum values of *serum creatinine* observed within 72 h post-surgery as the most important feature (Figure 2). Also, the post-surgery observed maximum values of *urea* and the *age at surgery* were of high importance for both algorithms (Figure 2). The beeswarm plots further indicate that, in addition to the disease group, the *aortic cross clamp time*, the maximum post-surgery values of *C-reactive protein*, *days between admission and surgery*, and if *circulatory arrest* was applied during surgery were ranked by SHAP among the ten most important features for both ML methods to predict the outcome in the holdout test dataset (Figure 2). The binary variable if the thorax was left open after the surgery was also a notable feature ranked fourth for XGB and sixth for RSF. In comparison, for RSF, Figure 3 shows the global SurvSHAP(t) values averaged by feature, observation, and time-point across all repeated CV models. These results also take the time dependency of the feature importance into account. The union set of the five most important features of XGB and RSF according to SHAP comprises the data elements *age at surgery*, *aortic cross clamp time*, *days between admission and surgery*, *disease group*, *open thorax*, *serum creatinine (maximum)*, and *urea (maximum)* (see Table 4).

Table 4. Top features. The table shows the n = 5 most important features (by means of the mean absolute SHAP values) for xgboost (XGB) and ranger (RSF), respectively.

Top Features	XGB	RSF
Age at surgery	0.12	0.18
Aortic cross clamp time	n/a	0.19
Days between admission and surgery	n/a	0.24
Disease group	0.20	0.19
Open thorax	0.13	n/a
Serum creatinine (maximum)	0.38	0.31
Urea (maximum)	0.15	n/a

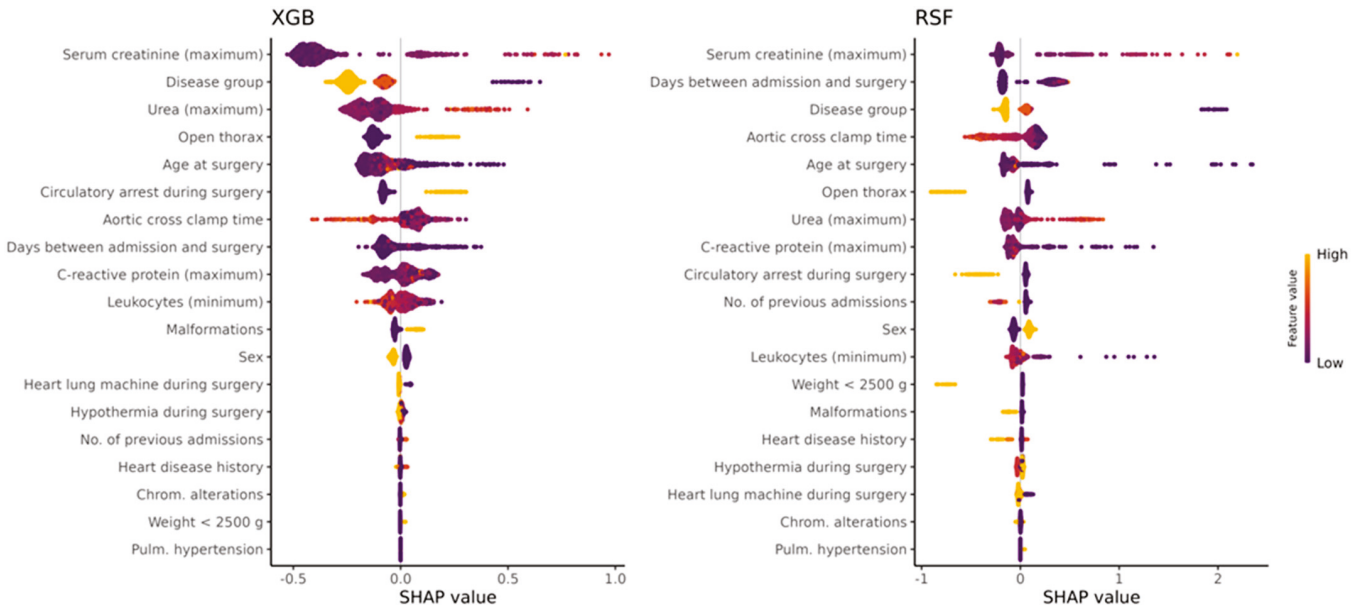


Figure 2. Feature importance: SHAP beeswarm plots. XGB: xgboost; RSF: random survival forest.

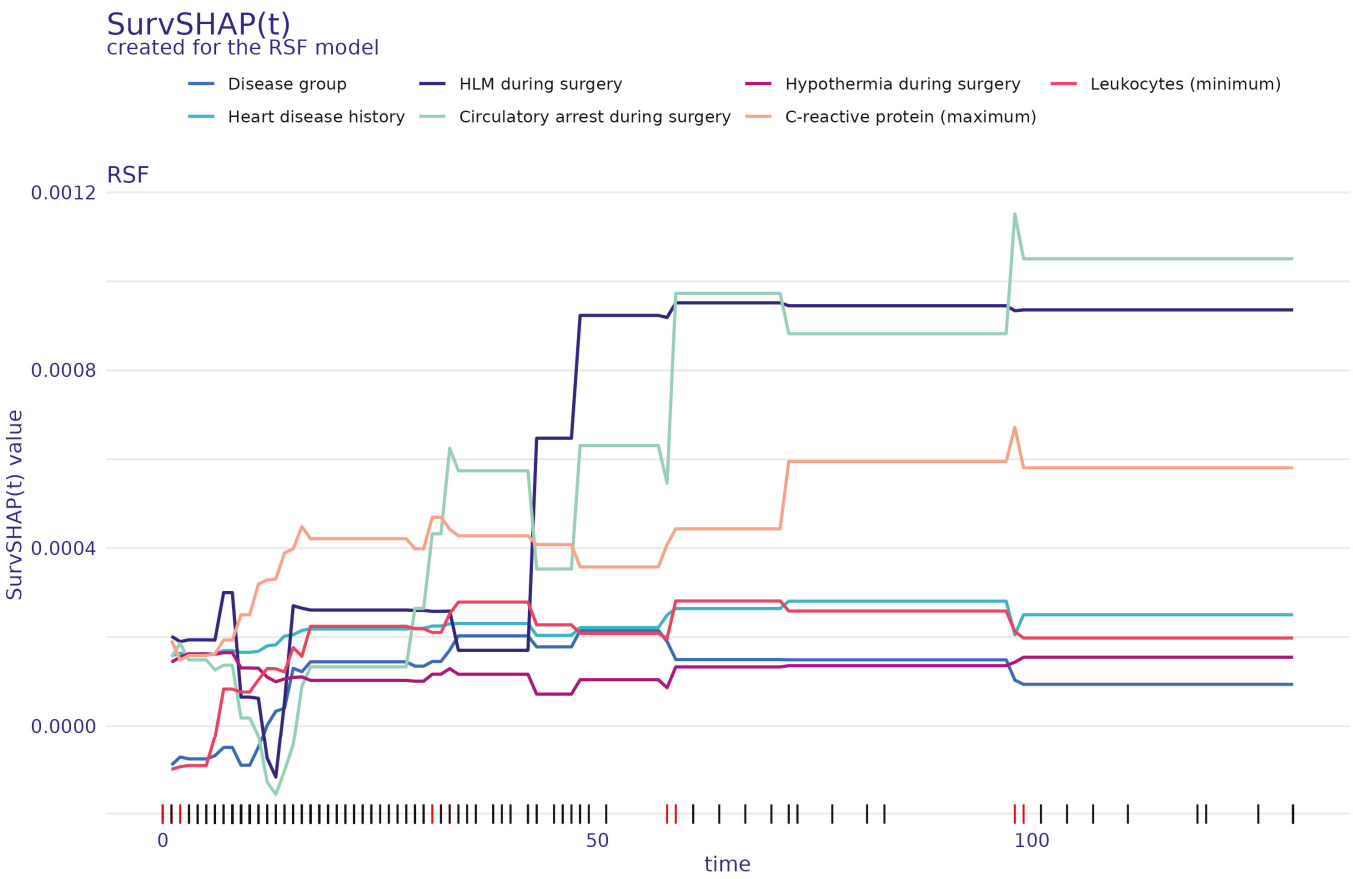


Figure 3. Feature importance: global SurvSHAP(t) values showing feature importance as a function of the survival time for the random survival forest (RSF). The SurvSHAP(t) values of the seven most important features identified as such were averaged by feature and evaluation time-point across all 100 repeated CV models. On the x-axis, red ticks mark event time points and black ticks mark censored time points.

To provide more insights on the feature importance variability among the repeated CV models of each ML method, Figure 4 visualizes the number of CV models in which a certain feature occurred within the five most important features according to SHAP for XGB (top left) and RSF (top right), as well as for RSF according to SurvSHAP(t) (bottom left). This visualization shows that the maximum values of *serum creatinine* observed within 72 h post-surgery was ranked according to SHAP the most important feature in 86% of all XGB models and in 59% of all RSF models. Likewise, *serum creatinine* was also ranked as the most important feature according to SurvSHAP(t) in 44% of all repeated CV models, which was on par with the *days between admission and surgery* (Figure 4). Similar to SHAP, the *disease group*, *age at surgery*, and *open thorax*, but also the post-surgery minimum values of *leukocytes* and the maximum values of *urea*, were ranked amongst the five most important features in many of the repeated CV models when computing SurvSHAP(t) values.

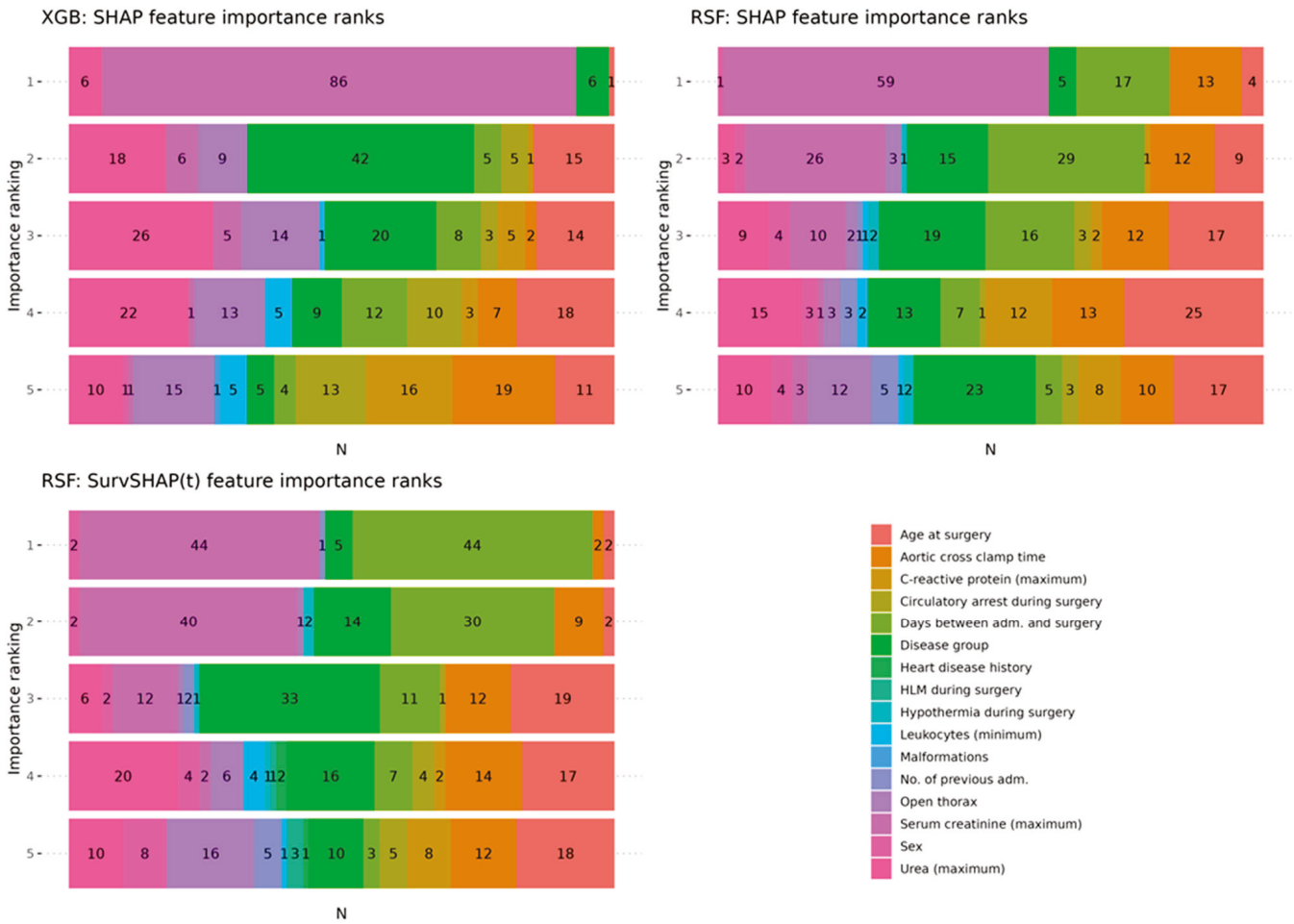
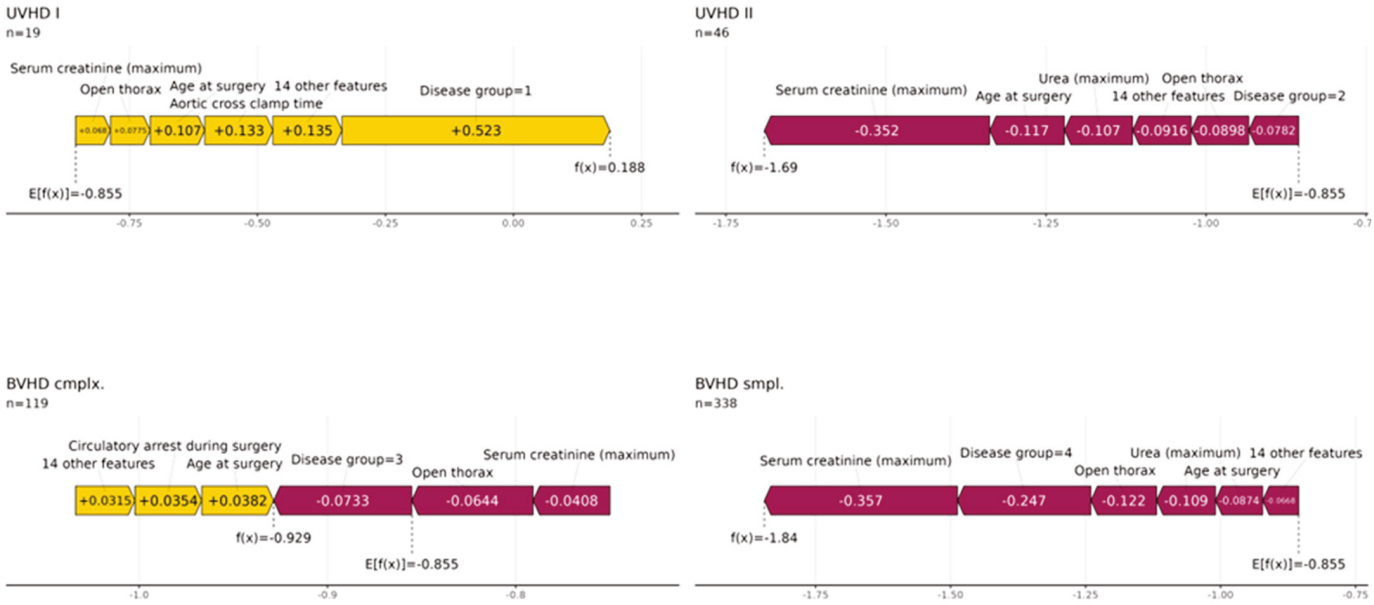


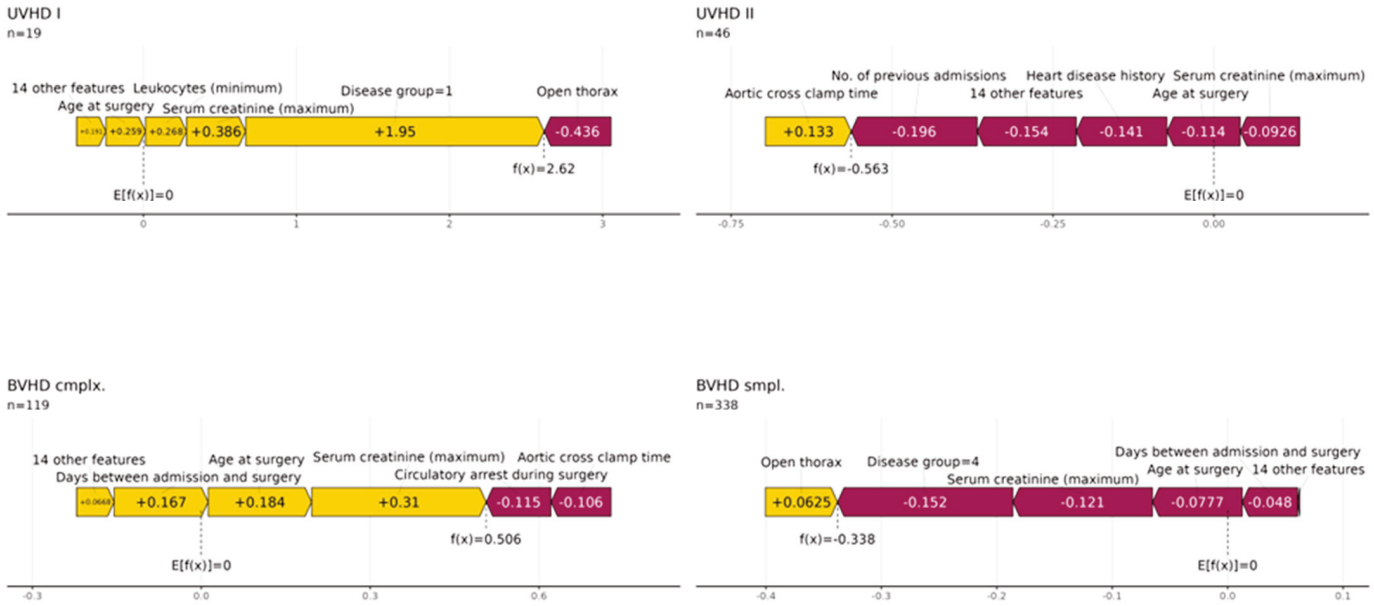
Figure 4. Feature importance ranks: counts of the occurrence of variables within the five most important features as defined by their mean absolute SHAP values and SurvSHAP(t) values, respectively, for each repeated CV model. XGB: xgboost; RSF: random survival forest.

Furthermore, force plots of the aggregated global SHAP values as shown in Figure 5 for XGB and RSF are a visualization to provide more insights into the underlying data structure. Using these plots, it can be visualized how the specific observations' feature values in the independent holdout test dataset influenced the models' predictions stratified by the four disease groups. It is noteworthy that, according to the RSF force plots, the maximum values of *serum creatinine* noticeably seemed to increase the mortality risk on average for the disease groups *UVHD I* and *BVHD cmplx.*, whereas a negative effect of this variable was observed in the two other groups. The XGB force plots instead show an

increased mortality risk of higher post-surgery maximum values of *serum creatinine* only for *UVHD I* cases, whereas in all other disease groups higher levels were associated with decreased mortality risk. Furthermore, it can be seen that the central tendency of the *serum creatinine* values is similar between the training dataset and the test dataset for the censored cases in all four disease groups. In contrast, the deceased cases in the CHD groups *UVHD II* and *BVHD smpl.* in the test dataset exhibited noticeably lower maximum values of *serum creatinine* than the corresponding training dataset (see Figure 6). Nevertheless, Figure 5 also reveals that inclusion in the disease group *UVHD I* was identified as the most important factor for increased mortality risk by both ML methods (Figure 5).



(**a**) XGB



(**b**) RSF.

Figure 5. Feature importance: SHAP force plots for XGB and RSF. UVHD I: disease group univentricular heart failure (HF) 1; UVHD II: disease group univentricular HF 2; BVHD cmplx.: disease group biventricular HF complex; BVHD smpl.: disease group biventricular HF simple; XGB: xgboost; RSF: random survival forest. (**a**) XGB. (**b**) RSF.

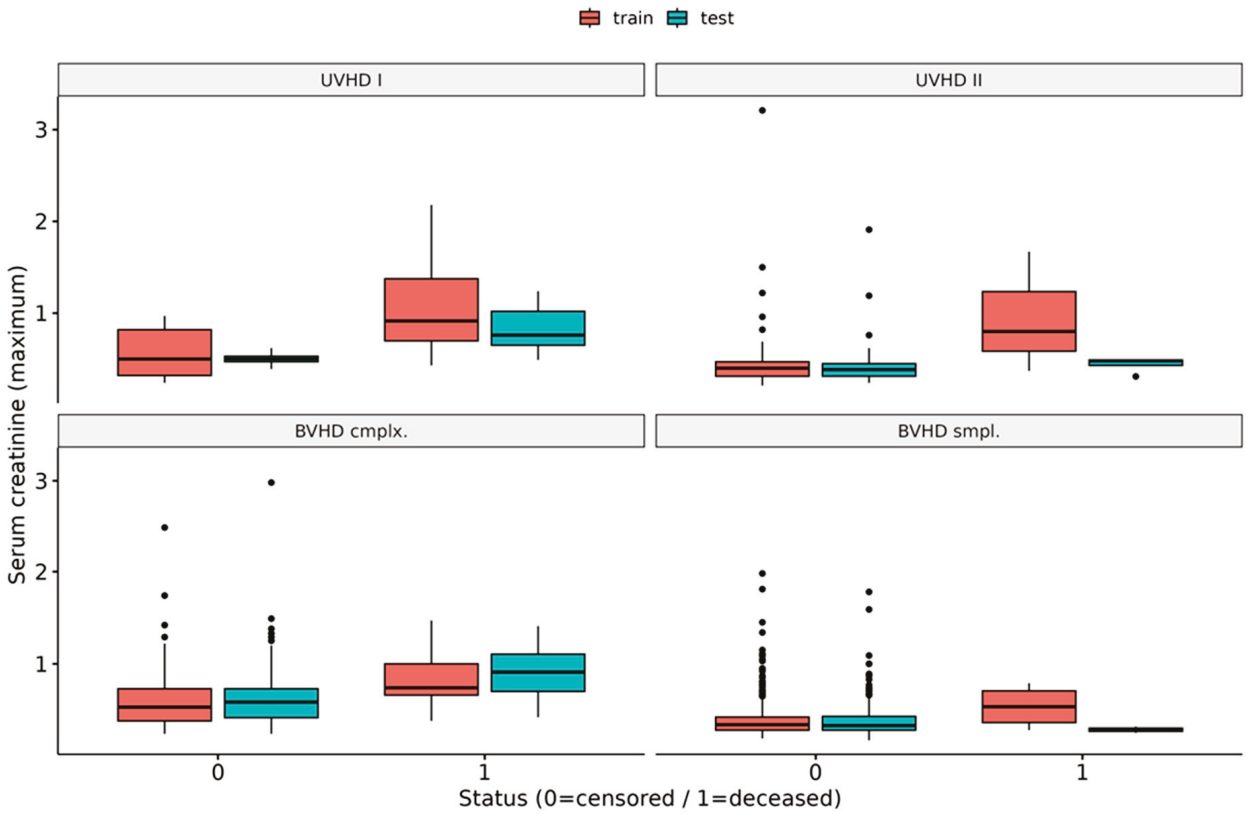


Figure 6. Distribution of the maximum values of serum creatinine observed within 72 h post-surgery for the heart disease groups, stratified by the training dataset and the test dataset. UVHD I: disease group univentricular heart failure (HF) 1; UVHD II: disease group univentricular HF 2; BVHD cmplx.: disease group biventricular HF complex; BVHD smpl.: disease group biventricular HF simple.

3.3.2. Comparison with CPH

For comparison with a standard statistical approach, the CPH regression models were fitted in the same experimental setup as the ML methods but using only a subset of the available features. This feature subset was formed by the union set of the five most important features, as previously identified by XGB and RSF, which were used as independent variables for the CPH regression models. The CPH models achieved an average C-index of 0.78 (±0.01) in predicting the outcome in the holdout test dataset (Figure 1; for further details see supplemental results and Table S6).

4. Discussion

In this study, we applied a random survival forest and the 'eXtreme Gradient Boosting' algorithm to identify risk factors for mortality after surgery for CHD in pediatric patients at one German pediatric heart center. Both ML methods performed well in predicting the mortality risk scores in the holdout test dataset, with C-indices of 0.85 (±0.01) and 0.79 (±0.03) on average for RSF and XGB, respectively. In comparison, conventional Cox regression also performed very well, with an average C-index of 0.78 (±0.01) (Figure 1).

This is in line with other recently published studies using machine learning models for analysis of quality assessment in pediatric patients with CHD [8,19,41]. The ML methods had greater predictive power than the classical statistic methods and standard risk categories STAT and RACHS-1 in estimating in-hospital mortality, with AUCs between 0.83 and 0.88. These innovative ML algorithms appear to be ideal for complex multidimensional data and individualized risk prediction, owing to enhanced capturing of complex and nonlinear relationships [42,43].

Treatment strategies for CHD have advanced over the past decade, leading to low and consistent CHD-related overall mortality rates around 2.5% to 4.8% [7,8,44,45], depending

on the patient composition of the studies. These refined surgical techniques and postoperative management have also led to paradigm shift in neonatal CHD approaches, aiming for earlier correction of complex biventricular heart defects rather than relying on palliative surgery [22]. This approach has been adopted in our center as well, which is reflected by the distribution of cardiac disease categories in this study, with about one-third of all cases being classified as complex patients, including those with *BVHD cmplx.* and *UVHD I* conditions.

We know from the literature that certain non-modifiable factors such as younger age, lower weight, and procedure type are linked to increased hospital mortality [8,46,47]. Not surprisingly, the pre-surgery risk allocation into disease groups turned out to be particularly relevant with regard to in-hospital mortality risk. This was confirmed by our xAI results, as well as by the CPH regression. In this regard it is to be noted that, at 56%, the mortality rate in the *UVHD I* group appears to be quite high in our analysis compared to the literature. However, this observed mortality rate is directly related to the cohort definition used in this analysis, where only the most recent available hospitalization for each patient was included. *UVHD* cases that were successfully discharged from hospital in the past and readmitted later, but included in a different heart defect group (due to a different surgery being performed), are consequently missing from the calculation of the *UVHD I* mortality rates, which leads to the observed higher mortality rates for this particular group in our analysis. In the study by Du et al., CHD risk categories were also identified as one of the most important features for their XGB model's predictions [19]. While in our analysis disease groups were built by combining dedicated diagnosis and surgical procedure codes, the risk categories used in their study were based on surgical procedures and generated by clustering them into groups with homogeneous estimated mortality risks according to a dedicated statistical method [19]. Furthermore, a peculiarity in the statistical modeling of survival data is that the outcome is composed of a time variable and a status variable. The time variable indicates the total follow-up time for an observation, whereas the status indicates if an event was either observed at the end of the follow-up time frame or not, which is then referred to as *censoring*. If such data is available, it can, for example, be used for visualizing Kaplan-Meier plots and to model semiparametric CPH regression models. Hence, in our study, we have implemented the XGB algorithm using a Cox regression, which utilizes the information provided by the survival function and accounts for censoring. In contrast, Du et al. applied XGB with a binary classification [19], which is why our results are not fully comparable to theirs.

Research has highlighted the importance of an individual's ability to handle the stress of surgery [43], which resulted in the inclusion of more patient specific pre-and postoperative variables respective complications in mortality and morbidity prediction models in cardiac surgery [45].

As part of the evolution of mortality risk models of STS–CHSD, Jacobs et al. identified pre-procedural renal dysfunction as an important contributor to increased mortality after pediatric heart surgery [23]. Cardiac surgery is a significant stressor on the kidneys, leading to pediatric acute kidney injury (AKI) in 33–43% of surgical cases [48,49], although most patients recover in a short period of time [50]. The TRIBE-AKI consortium demonstrated the importance of assessing renal function, as acute kidney injury was associated with longer mechanical ventilation, extended stays in intensive care unit (ICU), and a trend towards higher mortality [49].

Interestingly, in our analysis, the maximum creatinine level (within 72 h postoperatively) was the strongest predictor of mortality. In fact, it was ranked as the top feature in both applied ML methods, with higher feature values indicating higher risk scores (Figure 3). Global SHAP values also indicated, albeit to a lesser extent, high importance for maximum urea levels. Likewise, the influence of the maximum *creatinine* values on the mortality risk can also be seen in the SurvSHAP(t) values, with an even greater increased risk for patients with longer follow-up times (Figure 3). The importance of the post-surgery maximum *creatinine* value is further underlined by the fact that the majority of repeated CV

models of both algorithms ranked it as the most important feature (Figure 4). In a subgroup analysis, the increasing effect of the maximum *serum creatinine* value on mortality risk could be shown for *UVHD I* cases and *BVHD cmplx.* cases for the RSFmodels, whereas the XGB models showed an increased risk for mortality of the higher post-surgery maximum *serum creatinine* values only in *UVHD I* cases (Figure 5). These results may be better explainable when looking at the distribution of *serum creatinine* between the different disease groups: it is noteworthy that, for the censored cases, outliers with very high values were present in all diseases groups but *UVHD I* in the training dataset and in the test dataset (Figure 6). Thus, XGB might interpret the presence of high outlier values as a protective marker depending on the heart disease group. In contrast, the figure also shows that the differences in the central tendency between censored and deceased cases in the training dataset are obvious for all disease groups. In the test dataset however, for the deceased cases, the central tendency differs from the training dataset in *UVHD II* and *BVHD smpl.* cases (Figure 6). This observation suggests that RSF may have associated this central tendency in these subgroups with a decreased mortality risk.

It should be noted that creatinine (and/or urea) levels were not a significant factor in the large database and mortality studies performed by Bertsimas et al. and Du et al. [8,19]. This, however, is different for dialysis-dependent AKI. Brown et al. demonstrated an association between this postoperative complication and increased 6-month mortality [51]. However, 30-day mortality was not affected by this complication. As the serum creatinine level was the most significant variable in risk prediction in our study, we suggest the inclusion of postoperative renal function monitoring in the form of creatinine and urea levels in future mortality prediction models.

In our study, the union set of the five most important features of the two applied ML methods XGB and RSF according to SHAP comprises the data elements *age at surgery*, *aortic cross-clamp time*, *days between admission and surgery*, *disease group*, *open thorax*, *serum creatinine (maximum)*, and *urea (maximum)*. The days between admission and surgery could reflect a worse preoperative status or admission in a decompensated cardiovascular state with the need for recovery, but this is speculative, as we did not further investigate the reason for the delayed time to surgery. The importance of preoperative status in predicting mortality was also emphasized by Bertsimas et al. in their large database study [8]. A short time (<12 days) since last admission and any common preoperative risk factor (i.e., shock, sepsis, mechanical ventilator) were associated with worse outcomes [8]. Consistent with previous studies using machine learning models for risk prediction, aortic cross-clamp or bypass time is a significant and meaningful risk factor [44,52,53]. Assessment of inflammatory state as an expression of the individual response to perioperative and postoperative factors (such as prolonged CPB) and complications demonstrated a correlation between lower postoperative leukocyte minimums and higher risk of mortality, as identified by SHAP and SurvSHAP(t) (Figures 2 and 3). This laboratory value could serve as a surrogate parameter for a pronounced or abnormal inflammatory response to surgery and CPB. Systemic inflammatory response syndrome (SIRS) is common after cardiac surgery, with a rate of 20 to 32.5% [54,55] and is related to increased mortality [56]. Most commonly, however, SIRS is associated with an elevated white blood cell count [56]. We did not measure SIRS in our setting because it is difficult to accurately define and measure SIRS in the neonatal and pediatric postoperative course [57], and it ideally needs to be established prospectively. Therefore, it cannot be concluded whether this result might be related to an increased SIRS reaction.

The results of the CPH models on the independent holdout test dataset are good and almost on par with XGB. Compared to the ML methods, a big advantage when using the CPH regression is the good interpretability of its results. However, it has to be noted that, in our experimental setup, the input features provided to the CPH regression were already known to be important for the ML methods. Since these features obviously contained information relevant for the ML methods, the probability that they could also be important for the CPH regression was higher from the outset. Thus, the application of the ML methods

and subsequent computation of the feature importance could be considered as kind of a (computationally very expensive) feature selection strategy for the CPH regression in our setup. Hence, the results of the CPH may not be fully comparable to those of the ML methods. Thus, using CPH might not serve well for the most fair comparison of a standard statistical method with ML methods. Therefore, for example, elastic net regression (ENR) would be more suitable, as it also includes variable selection. However, as the interpretation of CPH is more comprehensive and straightforward compared to penalized regression methods such as ENR, here we only used CPH as a base level for comparison with RSF and XGB.

4.1. Limitations

This study had several limitations. The major limitation is the rather low sample size, and especially the relatively low number of events, which hampered statistical modeling, especially when using ML algorithms that usually require much more data than standard statistical methods. This low number of events may have influenced the application of resampling strategies such as the CV techniques utilized here, potentially producing fold configurations without any events. The not-optimal splitting becomes obvious in the differing data distributions between the training dataset and the holdout test dataset, for example, with regard to the follow-up time periods or to the distribution of the post-surgery *creatinine* values among the different disease groups. On the one hand, we addressed this issue by partitioning the data by 60% to 40% into a training dataset and a holdout test dataset in order to ensure that enough events were available for each disease group in both data partitions. Furthermore, generation of the CV folds was stratified by disease group, time- variable, and status variable to avoid the accumulation of edge cases in a few folds. On the other hand, validation of the hyperparameters was implemented with a 10× repeated 10-fold CV to take the variability of the differently configured folds into account. Indeed, the boxplots in Figure 1 show high variability of the respective model performance among the 100 folds during the hyperparameter validation of each algorithm.

Despite all the adjustments that were made regarding the probably insufficient sample size in this study, it would be beneficial to overcome this limitation by analyzing data from multiple German pediatric heart centers jointly. One goal of this feasibility study was to establish an analytical workflow that can be used as a blueprint for designing a multicentric study in the future. During the last 6 years, the German Medical Informatics Initiative (MII) [20] established an infrastructure in German university hospitals to conduct such multicentric studies in a privacy-preserving manner. Based on this infrastructure, the Bavarian Cancer Research Center, for example, is currently establishing approaches for federated machine learning (fML).

However, we have identified several challenges that still need to be addressed in order to conduct our analysis in a federated multicentric setting. The statistical learning methods applied in this study, namely RSF, XGB, and CPH, cannot directly be applied in a federated setting. The adaption of these algorithms for fML is part of ongoing research efforts, both for RSF [58–60] and for XGB [61–63]. Similarly, current research aims at translating the statistical modeling of survival data into a federated setting [64–66], including federated survival forests [67], as well as an extension of CPH, so called 'discrete-time Cox models' [64]. To also make the outputs of the fML models better understandable and interpretable, xAI methods such as SHAP need to be translated to the federated learning setting as well, which is also currently being investigated by many research groups [68–73]. Another important aspect is the handling of missing values in the federated setting. Here, we used a multiple imputation approach, which could, for example, be implemented at each participating site separately.

A limitation of the cardiology diagnosis groups is their reliance on ICD and OPS codes, which can be misleading if not accurately coded. To test the accuracy of the algorithm, the dataset was compared to two other registries, the German national registry and the ECHSA database, which regularly receive data on surgical and interventional procedures from

our center. The results showed similar numbers of cases and a comparable proportion of complex operations, confirming the accuracy of the algorithm. The wide variety of diseases, diagnoses, procedures, and combinations in the field of pediatric cardiology and congenital heart surgery makes categorization challenging. Categorization frequently fails to consider the interaction between surgical and patient-specific factors. Moreover, even within the same category, there is considerable variation in severity, which makes accurate prognosis and risk stratification challenging.

5. Conclusions

In this work, we applied advanced ML methods to identify risk factors for mortality after surgery for CHD in pediatric patients at one German Pediatric Heart Center. We demonstrated that ML methods can be applied in combination with dedicated model explainability tools to reveal interesting insights into mortality risk after surgery for CHD, such as the seemingly high importance of the *maximum values of serum creatinine* observed within 72 h post-surgery. With respect to clinical relevance, future efforts are needed to validate the findings and to investigate potential countermeasures. Furthermore, building upon this preparatory groundwork, future work needs to investigate if the analysis workflow established here can be translated into a federated setting in order to analyze a larger cohort.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics14222587/s1, Table S1. Inclusion criteria. If groups are defined by both OPS and ICD codes, the respective codes are concatenated with a logical 'AND' between code systems (ICD/OPS) and 'OR' within one code system. OPS: German procedure coding system; ICD: internationial classification of diseases; UVHD I: disease group univentricular heart defect (HD) 1; UVHD II: disease group univentricular HD 2; BVHD cmplx.: disease group biventricular HD complex; BVHD smpl.: disease group biventricular HD simple. Table S2. Group definitions. Destinct groups are defined by both OPS and ICD codes, the respective codes are concatenated with a logical 'OR'. OPS: German procedure coding system; ICD: internationial classification of diseases. Table S3. Number of patients per heart defect group stratified by the heart disease history and the deceased status. UVHD I: disease group univentricular heart defect (HD) 1; UVHD II: disease group univentricular HD 2; BVHD cmplx.: disease group biventricular HD complex; BVHD smpl.: disease group biventricular HD simple. N/A: not applicable, i.e. no previous hospitalization. Table S4. Dataset partitions into a training dataset (60%) and an independent holdout test dataset (40%). UVHD I: disease group univentricular heart defect (HD) 1; UVHD II: disease group univentricular HD 2; BVHD cmplx.: disease group biventricular HD complex; BVHD smpl.: disease group biventricular HD simple. Table S5. Feature importance. Mean absolute SHAP values for XGB and RSF. The feature importance was computed from the SHAP values that were averaged by feature and observation across all 100 repeated CV models for each ML method. The rank of each variable according to the computed SHAP values are given in brackets next to the SHAP value. Table S6. Results of the Cox Proportional Hazards model. Columns 2 and 3 present the median values and 95% conficence intervals (CI) for the hazard ratio (HR) and *p*-values of the 100 models from the repeated CV (weighted by the number of samples in the training folds). Columns 4 and 5 present the model output of the single Cox PH model fitted with the complete training dataset. The reference group of the variable 'Disease group' was univentricular heart defect (UVHD) I. HR: harzard ratio. CI: confidence interval. UVHD II: disease group univentricular heart defect (HD) 2; BVHD cmplx.: disease group biventricular HD complex; BVHD smpl.: disease group biventricular HD simple; Inf: infinite. Figure S1. Boxplot of the days until discharge after surgery stratified by disease group and deceased status. Dashed line: 99.5% percentile of the days until discharge computed over the whole sample. 0: censored; 1: deceased. UVHD I: disease group univentricular heart defect (HD) 1; UVHD II: disease group univentricular HD 2; BVHD cmplx.: disease group biventricular HD complex; BVHD smpl.: disease group biventricular HD simple. Figure S2. Performance: Boxplots to visualize the performance on the independent holdout test dataset per disease group for each algorithm. The underlying data for each boxplot is the performance of the 100 models from the repeated CV during validation and when applying these 100 models to predict the outcome in the holdout test dataset subsetted to the respective disease group. UVHD I: disease group univentricular heart defect (HD) 1; UVHD II:

disease group univentricular HD 2; BVHD cmplx.: disease group biventricular HD complex; BVHD smpl.: disease group biventricular HD simple.

Author Contributions: Conceptualization, L.A.K. and J.M.M.; data curation, L.A.K., W.W. and J.M.M.; formal analysis, L.A.K.; investigation, L.A.K.; methodology, L.A.K. and M.F.; software, L.A.K.; visualization, L.A.K.; writing—original draft, L.A.K. and J.M.M.; statistics, M.F.; writing—review and editing, W.W., M.F., H.-U.P., T.G., S.D. and A.P.; validation, A.P., W.W. and S.D.; funding acquisition, H.-U.P. and T.G.; resources, H.-U.P. and T.G.; data extraction, J.M.M.; project administration, W.W.; supervision, W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded in part by the German Federal Ministry of Education and Research (BMBF) within the Medical Informatics Initiative (MIRACUM Module 1b: FKZ 01ZZ2301A; NUM-DIZ: FKZ 01KX2121).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of "Ethikkommission der Friedrich-Alexander-Universität Erlangen-Nürnberg" (protocol code 297_21 Bc as of 21 July 2021).

Informed Consent Statement: Patient consent was waived due to the retrospective design of this study.

Data Availability Statement: The datasets generated and/or analyzed during the current study are not publicly available, due to internal data transfer policies, but are available from the corresponding author on reasonable request. The code used to perform the experiments is available at GitHub: https://github.com/kapsner/CHD-risk-factors (as of 14 November 2024). As no suitable R package existed for performing the desired analyses with the algorithms implemented in 'ranger' [25], 'xgboost' [26], and 'survival' [32], we developed a software framework consisting of the R packages 'mlexperiments' to perform the hyperparameter optimization, as well as the (repeated) CV, and 'mlsurvlrnrs' to provide some learner algorithms for survival data. The R package 'mlexperiments' is publicly available under open source license at The Comprehensive R Archive Network (CRAN): https://CRAN.R-project.org/package=mlexperiments (as of 14 November 2024). The R package 'mlsurvlrnrs' is publicly available under open source license at CRAN: https://CRAN.R-project.org/package=mlsurvlrnrs (as of 14 November 2024). Throughout this work, the computation of SHAP values for random survival models with the 'ranger' R package has been contributed as a new feature to the R package 'treeshap', which is publicly available at CRAN: https://CRAN.R-project.org/package=treeshap (as of 14 November 2024). Furthermore, the computation of SurvSHAP(t) values using the 'treeshap' R package as well as the computation of global SurvSHAP(t) values has been contributed as a new feature in the R package 'survex', which is publicly available at CRAN: https://CRAN.R-project.org/package=survex (as of 14 November 2024).

Conflicts of Interest: All authors declare that they have no conflicts of interest.

References

1. Lindinger, A.; Schwedler, G.; Hense, H.-W. Prevalence of congenital heart defects in newborns in germany: Results of the first registration year of the PAN study (July 2006 to June 2007). *Klin. Padiatr.* **2010**, *222*, 321–326. [CrossRef]
2. Sifrim, A.; Hitz, M.P.; Wilsdon, A.; Breckpot, J.; Turki, S.H.; Thienpont, B.; McRae, J.; Fitzgerald, T.W.; Singh, T.; Swaminathan, G.J.; et al. Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat. Genet.* **2016**, *48*, 1060–1065. [CrossRef]
3. Patel, S.S.; Burns, T.L. Nongenetic risk factors and congenital heart defects. *Pediatr. Cardiol.* **2013**, *34*, 1535–1555. [CrossRef]
4. Dittrich, S.; Arenz, C.; Krogmann, O.; Tengler, A.; Meyer, R.; Bauer, U.; Hofbeck, M.; Beckmann, A.; Horke, A. German registry for cardiac operations and interventions in patients with congenital heart disease: Report 2021 and 9 years' longitudinal observations on fallot and coarctation patients. *Thorac. Cardiovasc. Surg.* **2022**, *70*, e21–e33. [CrossRef]
5. Gilboa, S.M.; Salemi, J.L.; Nembhard, W.N.; Fixler, D.E.; Correa, A. Mortality resulting from congenital heart disease among children and adults in the united states, 1999 to 2006. *Circulation* **2010**, *122*, 2254–2263. [CrossRef]
6. Marino, B.S.; Cassedy, A.; Drotar, D.; Wray, J. The Impact of Neurodevelopmental and Psychosocial Outcomes on Health-Related Quality of Life in Survivors of Congenital Heart Disease. *J. Pediatr.* **2016**, *174*, 11–22.e2. [CrossRef]
7. Tsao, C.W.; Aday, A.W.; Almarzooq, Z.I.; Alonso, A.; Beaton, A.Z.; Bittencourt, M.S.; Boehme, A.K.; Buxton, A.E.; Carson, A.P.; Commodore-Mensah, Y.; et al. Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association. *Circulation* **2022**, *145*, e153–e639.

8. Bertsimas, D.; Zhuo, D.; Dunn, J.; Levine, J.; Zuccarelli, E.; Smyrnakis, N.; Tobota, Z.; Maruszewski, B.; Fragata, J.; Sarris, G.E. Adverse Outcomes Prediction for Congenital Heart Surgery: A Machine Learning Approach. *World J. Pediatr. Congenit. Heart Surg.* **2021**, *12*, 453–460. [CrossRef]
9. GBD 2017 Congenital Heart Disease Collaborators. Global, regional, and national burden of congenital heart disease, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet Child Adolesc. Health* **2020**, *4*, 185–200. [CrossRef]
10. Jacobs, M.L.; O'brien, S.M.; Jacobs, J.P.; Mavroudis, C.; Lacour-Gayet, F.; Pasquali, S.K.; Welke, K.; Pizarro, C.; Tsai, F.; Clarke, D.R. An empirically based tool for analyzing morbidity associated with operations for congenital heart disease. *J. Thorac. Cardiovasc. Surg.* **2013**, *145*, 1046–1057.e1. [CrossRef]
11. Hickey, P.A.; Connor, J.A.; Cherian, K.M.; Jenkins, K.; Doherty, K.; Zhang, H.; Gaies, M.; Pasquali, S.; Tabbutt, S.; St. Louis, J.D.; et al. International quality improvement initiatives. *Cardiol. Young* **2017**, *27*, S61–S68. [CrossRef] [PubMed]
12. Pace, N.D.; Oster, M.E.; Forestieri, N.E.; Enright, D.; Knight, J.; Meyer, R.E. Sociodemographic Factors and Survival of Infants with Congenital Heart Defects. *Pediatrics* **2018**, *142*, e20180302. [CrossRef]
13. Fogel, A.L.; Kvedar, J.C. Artificial intelligence powers digital medicine. *NPJ Digital Med.* **2018**, *1*, 5. [CrossRef] [PubMed]
14. Bruckert, S.; Finzel, B.; Schmid, U. The next generation of medical decision support: A roadmap toward transparent expert companions. *Front. Artif. Intell.* **2020**, *3*, 507973. [CrossRef] [PubMed]
15. Holzinger, A.; Goebel, R.; Fong, R.; Moon, T.; Müller, K.-R.; Samek, W. (Eds.) *xxAI—Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, 18 July 2020, Vienna, Austria, Revised and Extended Papers*; Springer International Publishing: Cham, Switzerland, 2022; Volume 13200. [CrossRef]
16. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
17. Moncada-Torres, A.; van Maaren, M.C.; Hendriks, M.P.; Siesling, S.; Geleijnse, G. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Sci. Rep.* **2021**, *11*, 6968. [CrossRef]
18. Lundberg, S.; Lee, S.-I. A unified approach to interpreting model predictions. *arXiv* **2017**. [CrossRef]
19. Du, X.; Wang, H.; Wang, S.; He, Y.; Zheng, J.; Zhang, H.; Hao, Z.; Chen, Y.; Xu, Z.; Lu, Z. Machine Learning Model for Predicting Risk of In-Hospital Mortality after Surgery in Congenital Heart Disease Patients. *Rev. Cardiovasc. Med.* **2022**, *23*, 376. [CrossRef]
20. Semler, S.; Wissing, F.; Heyder, R. German medical informatics initiative: A national approach to integrating health data from patient care and medical research. *Methods Inf. Med.* **2018**, *57*, e50–e56. [CrossRef]
21. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2024.
22. Erikssen, G.; Liestøl, K.; Seem, E.; Birkeland, S.; Saatvedt, K.J.; Hoel, T.N.; Døhlen, G.; Skulstad, H.; Svennevig, J.L.; Thaulow, E.; et al. Achievements in Congenital Heart Defect Surgery. *Circulation* **2015**, *131*, 337–346. [CrossRef]
23. Jacobs, J.P.; O'brien, S.M.; Pasquali, S.K.; Kim, S.; Gaynor, J.W.; Tchervenkov, C.I.; Karamlou, T.; Welke, K.F.; Lacour-Gayet, F.; Mavroudis, C.; et al. The Importance of Patient-Specific Preoperative Factors: An Analysis of The Society of Thoracic Surgeons Congenital Heart Surgery Database. *Ann. Thorac. Surg.* **2014**, *98*, 1653–1659. [CrossRef]
24. Van Buuren, S.; Groothuis-Oudshoorn, K. Mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **2011**, *45*, 1–67. [CrossRef]
25. Wright, M.N.; Ziegler, A. Ranger: A fast impleentation of random forests for high dimensional data in C++ and R. *arXiv* **2017**, arXiv:1508.04409. [CrossRef]
26. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou , T.; et al. Xgboost: Extreme Gradient Boosting. R Package Version 1.7.8.1. 2022. Available online: https://CRAN.R-project.org/package=xgboost (accessed on 25 September 2024).
27. Shortliffe, E.H.; Sepúlveda, M.J. Clinical decision support in the era of artificial intelligence. *JAMA* **2018**, *320*, 2199–2200. [CrossRef] [PubMed]
28. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. Tree-Based Methods. In *An Introduction to Statistical Learning. Springer Texts in Statistics*; Springer: New York, NY, USA, 2021; pp. 327–352. [CrossRef]
29. Harrell, F.E.; Lee, K.L.; Mark, D.B. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **1996**, *15*, 361–387. [CrossRef]
30. Wilson, S. ParBayesianOptimization: Parallel Bayesian Optimization of Hyperparameters. 2022. Available online: https://cran.r-project.org/web/packages/ParBayesianOptimization/index.html (accessed on 25 September 2024).
31. Mayer, M. SplitTools: Tools for Data Splitting. 2022. Available online: https://cran.r-project.org/web/packages/splitTools/index.html (accessed on 25 September 2024).
32. Therneau, T.M. A Package for Survival Analysis in R. 2022. Available online: https://cran.r-project.org/web/packages/survival/index.html (accessed on 25 September 2024).
33. Shapley, L.S. A VALUE FOR n-PERSON GAMES. *Contrib. Theory Games* **1953**, *2*, 307–317.
34. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef]
35. Komisarczyk, K.; Kozminski, P.; Maksymiuk, S.; Biecek, P. Treeshap: Fast SHAP Values Computation for Tree Ensemble Models. *arXiv* **2023**, arXiv:2109.09847. [CrossRef]

36. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. Resampling Methods. In *An Introduction to Statistical Learning. Springer Texts in Statistics*; Springer: New York, NY, USA, 2021; pp. 197–223. [CrossRef]
37. Krzyziński, M.; Spytek, M.; Baniecki, H.; Biecek, P. SurvSHAP(t): Time-dependent explanations of machine learning survival models. *Knowl.-Based Syst.* **2023**, *262*, 110234. [CrossRef]
38. Meis, J.; Baumann, L.; Pilz, M.; Sauer, L. DescrTab2: Publication Quality Descriptive Statistics Tables. 2022. Available online: https://cran.r-project.org/web/packages/DescrTab2/index.html (accessed on 25 September 2024).
39. Mayer, M. Shapviz: SHAP Visualizations. 2022. Available online: https://cran.r-project.org/web/packages/shapviz/index.html (accessed on 25 September 2024).
40. Spytek, M.; Krzyziński, M.; Baniecki, H.; Biecek, P. Survex: Explainable Machine Learning in Survival Analysis. 2022. Available online: https://cran.r-project.org/web/packages/survex/index.html (accessed on 25 September 2024).
41. Jalali, A.; Lonsdale, H.; Do, N.; Peck, J.; Gupta, M.; Kutty, S.; Ghazarian, S.R.; Jacobs, J.P.; Rehman, M.; Ahumada, L.M. Deep Learning for Improved Risk Prediction in Surgical Outcomes. *Sci. Rep.* **2020**, *10*, 9289. [CrossRef]
42. Volkova, A.; Ruggles, K.V. Predictive Metagenomic Analysis of Autoimmune Disease Identifies Robust Autoimmunity and Disease Specific Microbial Signatures. *Front. Microbiol.* **2021**, *12*, 621310. [CrossRef]
43. Triedman, J.K.; Newburger, J.W. Trends in Congenital Heart Disease: The Next Decade. *Circulation* **2016**, *133*, 2716–2733. [CrossRef]
44. Beckmann, A.; Dittrich, S.; Arenz, C.; Krogmann, O.; Horke, A.; Tengler, A.; Meyer, R.; Bauer, U.M.M.; Hofbeck, M.; German Quality Assurance/Competence Network for Congenital Heart Defects Investigators. German Registry for Cardiac Operations and Interventions in Patients with Congenital Heart Disease: Report 2020-Comprehensive Data from 6 Years of Experience. *Thorac. Cardiovasc. Surg.* **2021**, *69*, e21–e31. [CrossRef]
45. O'brien, S.M.; Jacobs, J.P.; Pasquali, S.K.; Gaynor, J.W.; Karamlou, T.; Welke, K.F.; Filardo, G.; Han, J.M.; Kim, S.; Shahian, D.M.; et al. The Society of Thoracic Surgeons Congenital Heart Surgery Database Mortality Risk Model: Part 1-Statistical Methodology. *Ann. Thorac. Surg.* **2015**, *100*, 1054–1062. [CrossRef]
46. Kalfa, D.; Krishnamurthy, G.; Duchon, J.; Najjar, M.; Levasseur, S.; Chai, P.; Chen, J.; Quaegebeur, J.; Bacha, E. Outcomes of cardiac surgery in patients weighing <2.5 kg: Affect of patient-dependent and -independent variables. *J. Thorac. Cardiovasc. Surg.* **2014**, *148*, 2499–2506.e1.
47. Kempny, A.; Dimopoulos, K.; Uebing, A.; Diller, G.-P.; Rosendahl, U.; Belitsis, G.; Gatzoulis, M.A.; Wort, S.J. Outcome of cardiac surgery in patients with congenital heart disease in England between 1997 and 2015. *PLoS ONE* **2017**, *12*, e0178963. [CrossRef]
48. Gritti, M.N.; Farid, P.; Manlhiot, C.; Noone, D.; Sakha, S.; Ali, S.; Bernknopf, B.; McCrindle, B.W. Factors Associated with Acute Kidney Injury After Cardiopulmonary Bypass in Children. *CJC Pediatr. Congenit. Heart Dis.* **2023**, *2*, 20–29. [CrossRef]
49. Li, S.; Krawczeski, C.D.; Zappitelli, M.; Devarajan, P.; Thiessen-Philbrook, H.; Coca, S.G.; Kim, R.W.; Parikh, C.R. Incidence, risk factors, and outcomes of acute kidney injury after pediatric cardiac surgery: A prospective multicenter study*. *Crit. Care Med.* **2011**, *39*, 1493. [CrossRef]
50. Zappitelli, M.; Bernier, P.-L.; Saczkowski, R.S.; Tchervenkov, C.I.; Gottesman, R.; Dancea, A.; Hyder, A.; Alkandari, O. A small post-operative rise in serum creatinine predicts acute kidney injury in children undergoing cardiac surgery. *Kidney Int.* **2009**, *76*, 885–892. [CrossRef]
51. Brown, K.L.; Ridout, D.; Pagel, C.; Wray, J.; Anderson, D.; Barron, D.J.; Cassidy, J.; Davis, P.J.; Rodrigues, W.; Stoica, S.; et al. Incidence and risk factors for important early morbidities associated with pediatric cardiac surgery in a UK population. *J. Thorac. Cardiovasc. Surg.* **2019**, *158*, 1185–1196.e7. [CrossRef] [PubMed]
52. Zürn, C.; Hübner, D.; Ziesenitz, V.C.; Höhn, R.; Schuler, L.; Schlange, T.; Gorenflo, M.; A Kari, F.; Kroll, J.; Loukanov, T.; et al. Model-driven survival prediction after congenital heart surgery. *Interdiscip. CardioVascular Thorac. Surg.* **2023**, *37*, ivad089. [CrossRef] [PubMed]
53. Agarwal, H.S.; Wolfram, K.B.; Saville, B.R.; Donahue, B.S.; Bichell, D.P. Postoperative complications and association with outcomes in pediatric cardiac surgery. *J. Thorac. Cardiovasc. Surg.* **2014**, *148*, 609–616.e1. [CrossRef] [PubMed]
54. Boehne, M.; Sasse, M.; Karch, A.; Dziuba, F.; Horke, A.; Kaussen, T.; Mikolajczyk, R.; Beerbaum, P.; Jack, T. Systemic inflammatory response syndrome after pediatric congenital heart surgery: Incidence, risk factors, and clinical outcome. *J. Card. Surg.* **2017**, *32*, 116–125. [CrossRef]
55. Soares, L.C.d.C.; Ribas, D.; Spring, R.; Silva, J.M.F.d.; Miyague, N.I. Clinical profile of systemic inflammatory response after pediatric cardiac surgery with cardiopulmonary bypass. *Arq. Bras. Cardiol.* **2010**, *94*, 127–133. [CrossRef]
56. Güvener, M.; Korun, O.; Demirtürk, O.S. Risk Factors for Systemic Inflammatory Response After Congenital Cardiac Surgery. *J. Card. Surg.* **2015**, *30*, 92–96. [CrossRef]
57. MacCallum, N.S.; Finney, S.J.; Gordon, S.E.; Quinlan, G.J.; Evans, T.W. Modified Criteria for the Systemic Inflammatory Response Syndrome Improves Their Utility Following Cardiac Surgery. *Chest* **2014**, *145*, 1197–1203. [CrossRef]
58. Maglogiannis, I.; Iliadis, L.; Macintyre, J.; Cortez, P. (Eds.) *Artificial Intelligence Applications and Innovations: 18th IFIP WG 12.5 International Conference, AIAI 2022, Hersonissos, Crete, Greece, 17–20 June 2022, Proceedings, Part I*; Springer International Publishing: Cham, Switzerland, 2022; Volume 646.
59. Liu, Y.; Liu, Y.; Liu, Z.; Liang, Y.; Meng, C.; Zhang, J.; Zheng, Y. Federated forest. *IEEE Trans. Big Data* **2022**, *8*, 843–854. [CrossRef]

60. Hauschild, A.-C.; Lemanczyk, M.; Matschinske, J.; Frisch, T.; Zolotareva, O.; Holzinger, A.; Baumbach, J.; Heider, D. Federated random forests can improve local performance of predictive models for various healthcare applications. *Bioinformatics* **2022**, *38*, 2278–2286. [CrossRef]
61. Leung, C.; Law, A.; Sima, O. *Towards Privacy-Preserving Collaborative Gradient Boosted Decision Trees*; UC Berkeley: Berkeley, CA, USA, 2019.
62. Le, N.K.; Liu, Y.; Nguyen, Q.M.; Liu, Q.; Liu, F.; Cai, Q.; Hirche, S. FedXGBoost: Privacy-Preserving XGBoost for Federated Learning. *arXiv* **2021**, arXiv:2106.10662. [CrossRef]
63. Jones, K.; Ong, Y.J.; Zhou, Y.; Baracaldo, N. Federated XGBoost on Sample-Wise Non-IID Data. *arXiv* **2022**, arXiv:2209.01340. [CrossRef]
64. Andreux, M.; Manoel, A.; Menuet, R.; Saillard, C.; Simpson, C. Federated Survival Analysis with Discrete-Time Cox Models. *arXiv* **2020**, arXiv:2006.08997. [CrossRef]
65. Wang, X.; Zhang, H.G.; Xiong, X.; Hong, C.; Weber, G.M.; Brat, G.A.; Bonzel, C.-L.; Luo, Y.; Duan, R.; Palmer, N.P.; et al. SurvMaximin: Robust federated approach to transporting survival risk prediction models. *J. Biomed. Inform.* **2022**, *134*, 104176. [CrossRef]
66. Rahimian, S.; Kerkouche, R.; Kurth, I.; Fritz, M. Practical challenges in differentially-private federated survival analysis of medical data. In Proceedings of the Machine Learning Research Conference on Health, Inference, and Learning (CHIL), Inference, Virtual, 7–8 April 2022; pp. 411–425.
67. Archetti, A.; Matteucci, M. Federated Survival Forests. *arXiv* **2023**, arXiv:2302.02807. [CrossRef]
68. Ben Saad, S.; Brik, B.; Ksentini, A. A trust and explainable federated deep learning framework in zero touch B5G networks. In Proceedings of the GLOBECOM 2022—2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil, 4–8 December 2022; IEEE: Rio de Janeiro, Brazil, 2022; pp. 1037–1042. [CrossRef]
69. Rahman, A.; Hossain, S.; Muhammad, G.; Kundu, D.; Debnath, T.; Rahman, M.; Khan, S.I.; Tiwari, P.; Band, S.S. Federated learning-based AI approaches in smart healthcare: Concepts, taxonomies, challenges and open issues. *Cluster Comput.* **2022**, *26*, 2271–2311. [CrossRef]
70. Renda, A.; Ducange, P.; Marcelloni, F.; Sabella, D.; Filippou, M.C.; Nardini, G.; Stea, G.; Virdis, A.; Micheli, D.; Rapone, D.; et al. Federated learning of explainable AI models in 6G systems: Towards secure and automated vehicle networking. *Information* **2022**, *13*, 395. [CrossRef]
71. Bárcena JL, C.; Daole, M.; Ducange, P.; Marcelloni, F.; Renda, A.; Ruffini, F.; Schiavo, A. Fed-XAI: Federated learning of explainable artificial intelligence models. In Proceedings of the XAI.it 2022: 3rd Italian Workshop on Explainable Artificial Intelligence, Udine, Italy, 28 November–2 December 2022.
72. Rumesh, Y.; Senevirathna, T.; Porambage, P.; Liyanage, M.; Ylianttila, M. Comprehensive Analysis over Centralized and Federated Learning-Based Anomaly Detection in Networks with Explainable AI (XAI). 2023. Available online: https://cris.vtt.fi/en/publications/comprehensive-analysis-over-centralized-and-federated-learning-ba (accessed on 25 September 2024).
73. Bogdanova, A.; Imakura, A.; Sakurai, T. DC-SHAP method for consistent explainability in privacy-preserving distributed machine learning. *Hum.-Cent. Intell. Syst.* **2023**, *3*, 197–210. [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.

diagnostics

MDPI

Article

Supervised Machine Learning-Based Prediction of In-Hospital Mortality Following Hip Fracture in Older Adults

Eduardo Guzmán-Muñoz [1,2], Manuel Vásquez-Muñoz [3,4,*], Yeny Concha-Cisternas [5,*], Rodrigo Olivares-Ordenes [6], Vicente Clemente-Suárez [7,8], Antonio Castillo-Paredes [9] and Rodrigo Yáñez-Sepúlveda [10,11]

1 Escuela de Kinesiología, Facultad de Salud, Universidad Santo Tomás, Talca 3460000, Chile; eguzmanm@santotomas.cl
2 Escuela de Pedagogía en Educación Física, Facultad de Educación, Universidad Autónoma de Chile, Talca 3460000, Chile
3 Center for Health Data Observation and Analysis (CADS), School of Medicine and Health Sciences, Universidad Mayor, Santiago 8580745, Chile
4 Escuela de Medicina, Facultad de Medicina y Ciencias de la Salud, Universidad Mayor, Santiago 8580745, Chile
5 Vicerrectoría de Investigación e Innovación, Universidad Arturo Prat, Iquique 1100000, Chile
6 Escuela de Ingeniería Informática, Universidad de Valparaíso, Valparaíso 2340000, Chile; rodrigo.olivares@uv.cl
7 Faculty of Medicine, Universidad Europea de Madrid, 28670 Madrid, Spain; vctxente@yahoo.es
8 Grupo de Investigación en Cultura, Educación y Sociedad, Universidad de la Costa, Barranquilla 080002, Colombia
9 Grupo AFySE (Actividad Física y Salud Escolar), Escuela de Pedagogía en Educación Física, Facultad de Educación, Universidad de Las Américas, Santiago 8370040, Chile; acastillop85@gmail.com
10 Facultad de Educación y Ciencias Sociales, Universidad Andrés Bello, Viña del Mar 2200055, Chile; rodrigo.yanez.s@unab.cl
11 School of Medicine, Universidad Espíritu Santo, Samborondón 092301, Ecuador
* Correspondence: manuel.vasquez@umayor.cl (M.V.-M.); yeconcha@unap.cl (Y.C.-C.)

Abstract

Background/Objectives: Hip fractures in older adults are associated with substantial morbidity, functional decline, and high in-hospital mortality. Early identification of patients at increased risk of death may improve clinical decision-making and resource allocation. This study aimed to develop and internally validate supervised machine learning models to predict in-hospital mortality among older adults hospitalized for hip fracture using nationwide administrative data from Chile. **Methods:** A retrospective cohort study was conducted using anonymized hospital discharge records from the Chilean National Health Fund (FONASA), covering admissions between 1 January 2019 and 31 December 2024, across 72 public hospitals. Demographic, clinical, and care-related variables were included as predictors. Multiple supervised machine learning algorithms were trained and evaluated using stratified train–test partitioning. Model performance was assessed using AUC-ROC, precision, recall, and F1-score. Model interpretability was explored using SHapley Additive exPlanations (SHAP). **Results:** A total of 40,253 hospitalization episodes were analyzed. The Gradient Boosting model achieved the best overall performance, with an AUC-ROC of 0.885 and a favorable balance between precision and recall. SHAP analysis identified age, comorbidity burden, and surgical treatment as the most influential predictors, revealing nonlinear and clinically meaningful contributions to mortality risk. **Conclusions:** Supervised machine learning models based on routinely collected administrative data demonstrated strong predictive performance for in-hospital mortality after hip fracture. Interpretable models may support early risk stratification and clinical decision-making at a national healthcare level.

Keywords: hip fracture; older adults; in-hospital mortality; machine learning; Gradient Boosting; SHAP analysis; predictive modeling; explainable AI

1. Introduction

Hip fracture among older adults remains a pressing global health issue, affecting approximately 1.5 million people each year, with projections suggesting this figure could rise to 2.6 million by 2025 and 4.5 million by 2050 [1]. In 2019, the global incidence in individuals aged 55 years and older was 681.35 per 100,000 population, with a prevalence of 1191.39 and a burden of 130.78 years lived with disability (YLDs) per 100,000 [2]. Over the past three decades, incidence rates have declined among those under 60 but increased sharply in older adults, particularly in women. Although rates remain consistently higher in females, the male-to-female incidence ratio rose modestly from 0.577 to 0.612 between 1990 and 2019 [2]. Falls continue to be the predominant cause across all age groups. These trends highlight the escalating burden of hip fractures in aging populations and the urgent need for improved prevention and prognostic strategies.

These injuries have profound clinical and functional consequences. Hip fractures often lead to loss of mobility, reduced independence, and a substantial decline in quality of life [3,4]. Mortality rates remain high despite advances in surgical and perioperative care—approximately 2–15% of patients die during hospitalization [5–8], and 10–30% within the first year after the event [9,10]. Although advances in surgical management and perioperative care have improved clinical practice, in-hospital mortality after hip fracture remains substantial. Identifying individuals at increased risk of in-hospital death is therefore essential for optimizing clinical pathways and ensuring the most efficient allocation of healthcare resources. As populations age and the incidence of hip fractures continues to rise, the development of accurate, transparent, and clinically meaningful prognostic models becomes increasingly important to support evidence-based decision-making [11].

This challenge is particularly relevant in Chile, where the burden of hip fractures among older adults has grown sharply over recent years. National hospital discharge data indicate that admissions increased by nearly 50% between 2006 and 2017, reaching about 9583 cases—equivalent to 149 per 100,000 adults aged 45 years and older [12]. The direct cost of managing these injuries exceeds 34 million USD annually, placing a considerable strain on the national healthcare system [12]. Hip fractures, therefore, represent a major clinical and public health concern in Chile. Predictive modeling based on routinely collected hospital information may help clarify mortality patterns and inform more effective prevention and management strategies.

In recent years, the use of machine learning in clinical research has expanded rapidly, offering new possibilities for predicting outcomes in complex and heterogeneous patient populations. In contrast to traditional regression models that rely on predefined linear assumptions, machine learning algorithms are capable of capturing nonlinear interactions and identifying multidimensional patterns within data, thereby improving the accuracy of prognostic estimations [13]. A wide range of supervised machine learning techniques—including tree-based ensemble methods, support vector machines, neural networks, and others—have been applied to prognosis in surgical and geriatric medicine [11,14]. Importantly, recent advances in geriatric risk prediction have increasingly emphasized not only predictive performance but also interpretability and clinical usability through explainable artificial intelligence approaches (e.g., SHAP) and user-oriented risk tools, as illustrated by a recent cohort study that developed a visualized machine learning system for sarcopenia risk prediction in older adults [15].

In the field of hip fracture research, several studies have begun incorporating machine learning–based models to estimate mortality risk [16–18]. Some of these approaches have achieved predictive performance comparable to, or even exceeding, that of conventional regression models. For example, a machine learning algorithm based on gradient-boosted decision trees achieved an AUC of about 0.79 for predicting 30-day mortality using preoperative clinical variables in a large dataset [13]. Similarly, random forest and naïve Bayes classifiers have shown comparable accuracy to logistic regression in predicting in-hospital and short-term mortality following hip fracture [11]. However, most prior studies focus on a single modeling technique; there is a relative paucity of research directly comparing multiple machine learning algorithms in this setting [19].

To address this gap, the present study aimed to develop and internally validate several supervised machine learning models for predicting in-hospital mortality among older adults with hip fracture, using routinely collected national hospital data from Chile. We hypothesized that supervised machine learning models—particularly tree-based ensemble methods—would provide accurate and clinically interpretable predictions of in-hospital mortality after hip fracture using routinely collected hospital data, outperforming traditional linear approaches.

2. Materials and Methods

2.1. Study Design and Data Source

This retrospective cohort study developed and internally validated supervised machine learning models to predict in-hospital mortality among older adults hospitalized for hip fracture in Chile. The analysis used anonymized hospital discharge data from the Chilean National Health Fund (Fondo Nacional de Salud, FONASA), which consolidates standardized administrative records from the 72 public hospitals forming the national healthcare network. These institutions deliver moderate- to high-complexity care, including orthopedic trauma and surgical management. The database contains detailed information on patient demographics, diagnostic and procedural codes, and discharge outcomes. Data were extracted from the open-access FONASA repository (https://datosabiertos.fonasa.cl, accessed on 9 February 2026) and included all hospitalizations between 1 January 2019, and 31 December 2024. Each individual was identified using an encrypted unique patient identifier, which allowed linkage of multiple hospitalizations belonging to the same patient. To avoid within-patient clustering and ensure independence of observations, only the first hospitalization for hip fracture per individual during the study period was retained for analysis.

FONASA covers roughly 85% of Chileans aged 60 years and older, representing the public healthcare sector where most hip fractures occur. The dataset undergoes regular audits by the Ministry of Health to ensure internal consistency and coding accuracy. All stages of model development and validation adhered to the TRIPOD-AI reporting guideline, ensuring transparency and reproducibility in the construction of AI-based clinical prediction models.

2.2. Participants

All hospitalizations of adults aged 60 years and older with a principal diagnosis of hip fracture were included. Cases were identified according to the International Classification of Diseases, Tenth Revision (ICD-10) codes: S72.0 (femoral neck fracture), S72.1 (pertrochanteric fracture), and S72.2 (subtrochanteric fracture). Exclusion criteria comprised: (i) missing information on age, sex, or discharge status; (ii) diagnostic codes not falling within the predefined range; (iii) transfer cases or duplicate admissions; and (iv) records with all predictor variables absent. Exclusions were intentionally restricted

to data quality–related issues in order to preserve the representativeness of the national cohort. Clinical variables not captured in administrative records—such as physiological parameters, laboratory values, time to surgery, or frailty scales—were not excluded but were unavailable in the dataset.

To ensure data integrity, validation procedures involved cross-checking unique admission and discharge identifiers to confirm internal consistency and remove duplicates. The final analytic cohort included 40,253 hospitalization episodes of older adults treated across the 72 public hospitals in Chile.

2.3. Outcome

The primary outcome was in-hospital mortality, defined as death occurring during the same hospitalization period. Mortality status was extracted from the standardized discharge condition field within the FONASA database, which documents each patient's vital status at discharge.

During data preprocessing, the positive class (death) was identified automatically through string matching for terms such as "falle", "muer", "defun", and "dead" appearing in the discharge condition variable. When no direct match was detected, the final label was assigned lexicographically to preserve consistency across the dataset. For clinical coherence and interpretability, death was pre-specified as the positive outcome class in both sensitivity and main analyses.

Only deaths verified at the time of discharge were included; post-discharge mortality data were not available. Each record corresponded to a distinct hospitalization episode, confirmed through cross-referencing of admission and discharge identifiers to avoid duplication.

2.4. Predictors

Predictor variables were selected a priori based on three predefined criteria: (i) consistent availability across the national administrative hospital discharge database, (ii) early availability during hospitalization to allow potential clinical applicability, and (iii) prior evidence supporting their association with in-hospital mortality after hip fracture. These criteria were applied to ensure both methodological robustness and clinical relevance while preserving the generalizability of the model to real-world inpatient settings. Based on these criteria, seven routinely collected variables were selected for model development and organized into three conceptual domains:

1. Demographic variables: Sex (male/female), ethnicity (categorical), and age (continuous, in years) [5–8].
2. Disease burden and treatment: Number of comorbidities (count of secondary diagnoses recorded at discharge) and surgical treatment (yes/no) [5–8].
3. Care process indicators: Transferring service or first hospital ward (categorical; e.g., orthopedics, internal medicine, intensive care) and length of hospital stay (continuous, in days) [5–8].

Categorical variables were encoded using one-hot encoding, while continuous variables were standardized before model training. This preprocessing step facilitated comparability among predictors and minimized potential bias related to variable scaling during model optimization. All selected predictors were available for all patients prior to discharge, allowing model application in early inpatient settings.

2.5. Data Management and Preprocessing

To ensure robustness and reproducibility across heterogeneous CSV exports, a standardized data management and preprocessing workflow was established. The datasets were imported through an automated parser capable of recognizing multiple delimiters

(;, „ \t, |) and character encodings (UTF-8 or Latin-1). Lines containing formatting inconsistencies were automatically omitted to preserve dataset integrity. Column headers were normalized to lower-case ASCII and converted to snake_case notation, while synonym mapping was applied to standardize variable names. For instance, variables such as "cirugia_(0 = no/1 = si)" were unified as "cirugia," and all variations referring to "days of hospitalization" were consolidated under a single label.

Before model training, variable types were harmonized to ensure compatibility across models. The variable representing surgical treatment was recoded to binary format (0/1, yes/no, or true/false), and both age and length of stay were converted to numerical type. Records lacking outcome data or missing all predictor values were removed. Missing data were handled by imputing the median for continuous variables and the mode for categorical ones. Numerical features were scaled, and categorical variables were encoded using a OneHotEncoder configured to ignore unseen categories. Both transformations were implemented within a ColumnTransformer to ensure consistent preprocessing across datasets.

To prevent data leakage, all preprocessing steps were embedded within the machine learning pipeline using scikit-learn and fitted exclusively on the training data. This design safeguarded methodological transparency, reproducibility, and the validity of model evaluation.

2.6. Data Partition and Validation Strategy

The dataset was randomly divided into training and testing subsets using an 80/20 stratified split to maintain the original distribution of the outcome classes (alive vs. deceased). Randomization was performed with a fixed seed (random_state = 42) to allow full reproducibility of the results. Each hospitalization episode was assigned to only one subset, ensuring that no records overlapped between the training and testing datasets and thereby preserving a strict separation throughout model development and evaluation.

2.7. Model Development

Eight supervised classification algorithms were developed and trained under a standardized scikit-learn framework to enable consistent comparison across models. The classifiers evaluated were:

1. Naïve Bayes (GaussianNB),
2. Multilayer Perceptron (MLP) neural network with two hidden layers (64 and 32 neurons, ReLU activation, $\alpha = 1 \times 10^{-3}$, max_iter = 500, random_state = 42),
3. Random Forest (400 estimators, class_weight = "balanced_subsample", n_jobs = –1),
4. Gradient Boosting (default parameters, random_state = 42),
5. k-Nearest Neighbors (kNN) (k = 7, distance-weighted voting),
6. Logistic Regression (solver = LBFGS, max_iter = 1000, class_weight = "balanced"),
7. Support Vector Machine (SVM) with radial basis function (RBF) kernel (C = 2.0, gamma = "scale", probability = True, class_weight = "balanced", random_state = 42), and
8. Decision Tree (class_weight = "balanced").

All algorithms were implemented as scikit-learn Pipelines, integrating data preprocessing and model fitting into a unified workflow. This approach minimized the risk of data leakage and ensured full reproducibility of the analytical process. Model training was carried out exclusively on the training subset, and performance was subsequently assessed using the independent test data.

2.8. Performance Metrics

Model performance was evaluated solely on the independent test set to ensure an unbiased assessment of predictive ability. Discriminative power was quantified using

https://doi.org/10.3390/diagnostics16040612

the area under the receiver operating characteristic curve (AUC-ROC) for binary classification, and the macro-averaged One-vs-Rest AUC for potential multiclass scenarios. To provide a comprehensive overview of model behavior, additional performance metrics were computed, including accuracy, precision, recall (sensitivity), and F1-score. These complementary indicators allowed for a balanced interpretation of predictive performance across classes.

For each algorithm, individual ROC curves were generated and exported in high-resolution format (600 dpi). A summary figure comparing all ROC curves was also created to visually contrast model performance. When a classifier produced a decision function rather than direct probabilities, a sigmoid transformation was applied to enable ROC curve computation and visualization. This adjustment improved interpretability through probability scaling without affecting the underlying AUC values.

2.9. Model Explainability

Model interpretability was assessed using SHapley Additive exPlanations (SHAP) to estimate the contribution of each predictor to the model's output. For tree-based algorithms, including Random Forest, Gradient Boosting, and Decision Tree models, the TreeExplainer method was employed. In the case of linear models such as Logistic Regression and linear SVMs, the LinearExplainer with interventional perturbation was applied. For nonlinear algorithms lacking native SHAP integration—namely MLP, RBF SVM, kNN, and Naïve Bayes—the KernelExplainer was implemented using a background sample of 100 randomly selected training observations and 200 test instances for visualization purposes.

The interpretability outputs included a series of global and local explanation plots: bar charts of mean absolute SHAP values, beeswarm plots, and dependence plots for the three most influential predictors. Additionally, individual waterfall plots were generated to illustrate subject-specific feature contributions. All figures were exported at 600 dpi resolution. SHAP values were expressed in log-odds, where negative values indicated a reduced probability of survival—reflecting an increased predicted risk of in-hospital mortality, given that the positive class corresponded to survival.

2.10. SHAP-Based Risk Equation

For the best-performing model according to AUC, an interpretable risk equation was derived using SHAP values. The model prediction for each individual can be expressed as:

$$\mathrm{logit}(p) = \beta_0 + \sum_i \phi_i,\ p = \frac{1}{(1+\exp(-\mathrm{logit}(p)))},\ RR = \frac{p}{p^0}$$

where β_0 represents the SHAP base value, corresponding to the expected model output (in log-odds) across the entire dataset, and ϕ_i denotes the individual contribution of predictor *i* to the predicted outcome. The baseline probability can be obtained as $p_0 = \sigma(\beta_0)$, where σ is the logistic (sigmoid) function.

This formulation allows the decomposition of each patient's predicted risk into additive contributions from all predictors, providing an individualized and transparent estimation of mortality risk directly interpretable in clinical terms.

2.11. Handling of Class Imbalance

Given the relatively low prevalence of in-hospital deaths, class imbalance was managed by setting the parameter class_weight = "balanced" in all algorithms that supported this option. This adjustment automatically reweighted the contribution of each class in inverse proportion to its frequency, reducing bias toward the majority (survival) class. In

https://doi.org/10.3390/diagnostics16040612

addition, a stratified train–test split was used to maintain the original outcome distribution across both subsets.

To preserve the empirical characteristics and natural class proportions of the dataset, no oversampling or synthetic data generation techniques were applied in the main analysis. This decision ensured that model training reflected the real-world clinical context represented in the national hospitalization data.

2.12. Sensitivity Analyses

Several pre-specified sensitivity analyses were performed to evaluate the robustness and clinical interpretability of the developed models. First, the orientation of the positive class was reversed—defining death instead of survival as the reference outcome—to verify the consistency of SHAP values and predicted probabilities. Second, the stability of categorical encoding was examined by consolidating low-frequency categories within the transferring service variable to reduce sparsity effects. Third, the reliability of complex algorithms was tested by varying key hyperparameters, including the regularization parameter C in the Support Vector Classifier and the number of estimators in the Random Forest model.

In addition, several methodological extensions were identified as recommended for future work. These include probability calibration using Platt or isotonic scaling, combined with calibration plots and Brier scores; estimation of 95% confidence intervals for the AUC using bootstrap or DeLong methods; and external validation—temporal or geographic—to assess model generalizability. Decision-curve analysis was also proposed to evaluate clinical usefulness, along with fairness auditing across sex, age, and ethnicity to ensure equitable model performance.

2.13. Statistical Considerations

The unit of analysis was the hospitalization episode, with each record corresponding to a distinct patient admission. Records were excluded only when essential variables required for cohort definition or outcome ascertainment, such as age, sex, or discharge status, were missing; missing values in predictor variables were handled using median imputation for continuous variables and mode imputation for categorical variables. Model outputs were primarily descriptive and comparative, and no correction for multiplicity was applied, as all algorithms evaluated the same predictive construct under equivalent analytical conditions. Hospital-level effects were not explicitly modeled, because the analytical focus was placed on patient-level risk prediction and on maximizing generalizability across the national public hospital network. To ensure full reproducibility, the random seed parameter (random_state = 42) was fixed across all stages of the workflow, including data partitioning, model training, and evaluation. Confidence intervals for the AUC were not computed in the main analysis to optimize computational efficiency; however, these can be readily obtained through stratified bootstrap resampling, which yields nonparametric 95% confidence intervals for model discrimination.

2.14. Software and Reproducibility

All analyses were conducted in Python 3.12.2 using the Jupyter Notebook environment (version 7.1.0). The computational workflow was built with widely adopted libraries, including pandas for data manipulation, scikit-learn for machine learning pipelines, matplotlib for visualization, SHAP for interpretability, and python-docx for automated report generation.

The analytical script was designed to automatically log diagnostic information—such as delimiter and encoding detection or column mapping—while executing all preprocessing and model training steps. It also computed the full set of performance metrics and exported

publication-ready outputs, including metric tables (CSV), high-resolution ROC curves and SHAP visualizations (600 dpi), as well as a structured summary report in DOCX format.

To further strengthen reproducibility, the workflow can be containerized (for instance, using Docker 27) and accompanied by the public release of both the source code repository and a synthetic dataset. The latter should retain the statistical structure of the original data while safeguarding patient privacy.

3. Results

The study included 40,253 hospitalization episodes of older adults with hip fracture. The mean age of the cohort was 81.9 years (SD 9.1), and women accounted for 76.8% of the sample. Regarding disease burden and treatment, the mean number of comorbidities was 6.6 (SD 4.2), with the majority of patients presenting three or more comorbid conditions. Most patients underwent surgical treatment during hospitalization (88.0%), while 12.0% were managed non-operatively. The mean length of hospital stay was 12.7 days (SD 13.9).

The Gradient Boosting model demonstrated the best overall performance, achieving an AUC-ROC of 0.885, an F1-score of 0.791, precision of 0.738, and recall of 0.849, indicating a strong balance between sensitivity and specificity (Table 1).

Table 1. Performance metrics of supervised machine learning models for in-hospital mortality prediction.

Model	Accuracy	Precision	Recall (Sensitivity)	F1-Score	AUC-ROC
Gradient Boosting	0.965	0.738	0.849	0.791	0.885
Multilayer Perceptron (Neural Network)	0.965	0.486	0.060	0.107	0.884
Logistic Regression	0.825	0.141	0.777	0.238	0.880
Linear Support Vector Machine	0.832	0.145	0.770	0.244	0.879
Random Forest	0.957	0.218	0.085	0.122	0.824
k-Nearest Neighbors	0.960	0.253	0.078	0.119	0.697
Decision Tree	0.935	0.156	0.194	0.173	0.579
Gaussian Naïve Bayes	0.043	0.035	0.989	0.068	0.499

The MLP neural network and linear approaches—Logistic Regression and linear SVM—reached comparable AUC values (approximately 0.88). However, their calibration between precision and recall was less optimal, which may limit their interpretability in clinical settings. In contrast, algorithms such as Random Forest and k-Nearest Neighbors achieved very high overall accuracies ($\geq$0.96) but markedly low F1-scores ($\leq$0.12), suggesting a strong bias toward the majority (survival) class and poor discrimination of deaths. The Naïve Bayes classifier yielded the highest recall (0.989) but an AUC of only 0.499, indicating near-random predictive performance.

Taken together, the Gradient Boosting algorithm provided the most favorable compromise between discriminative capacity and clinical relevance. Consequently, it was selected for interpretability assessment and individualized mortality risk analysis (Figure 1).

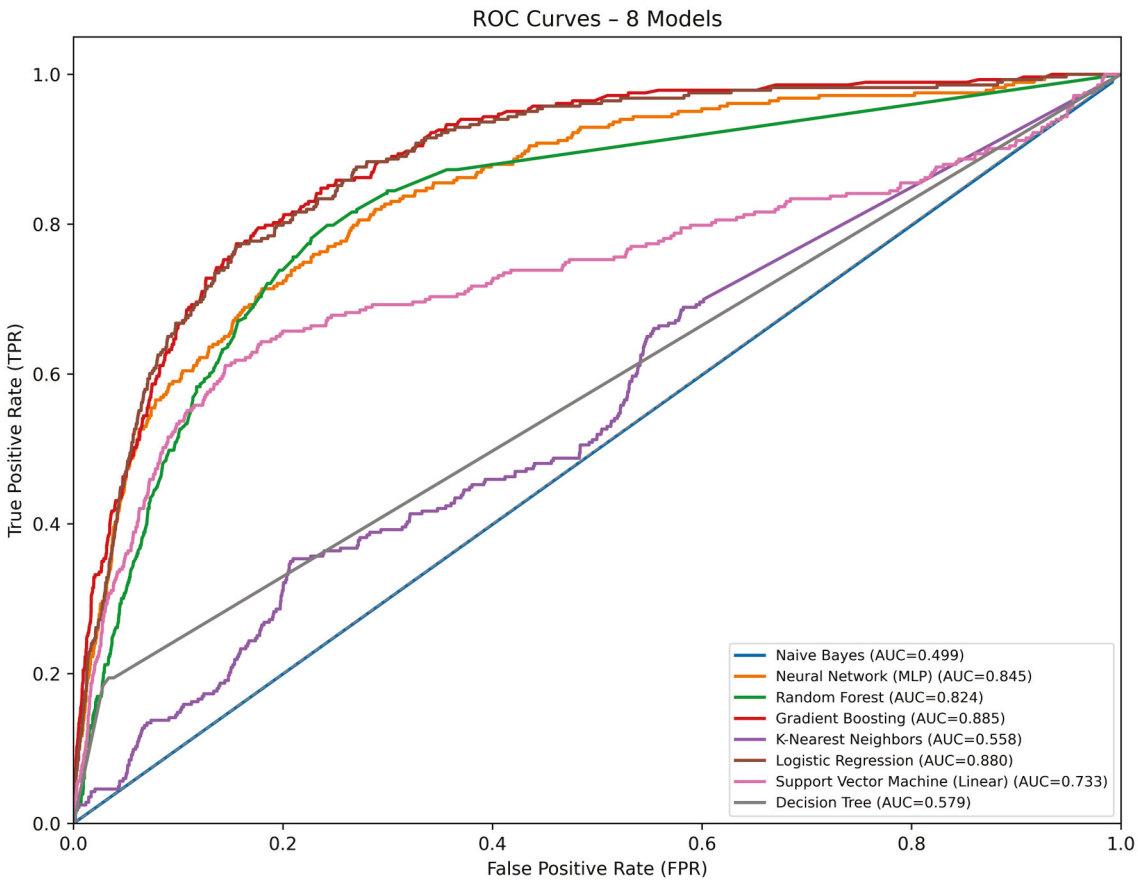


Figure 1. Receiver operating characteristic (ROC) curves for the eight supervised machine learning models. Interpretability analysis was performed using SHapley Additive exPlanations (SHAP) on the Gradient Boosting model. The global importance ranking (Figure 2) revealed that comorbidities had the largest contribution to model output ($|\text{SHAP}| \approx 0.52$), followed by surgical treatment ($|\text{SHAP}| \approx 0.47$) and age ($|\text{SHAP}| \approx 0.27$). Process-related variables such as first hospital service ($\approx$0.08) and length of stay ($\approx$0.07) showed smaller but informative effects, while demographic factors (e.g., sex and ethnicity) had minimal impact (<0.02).

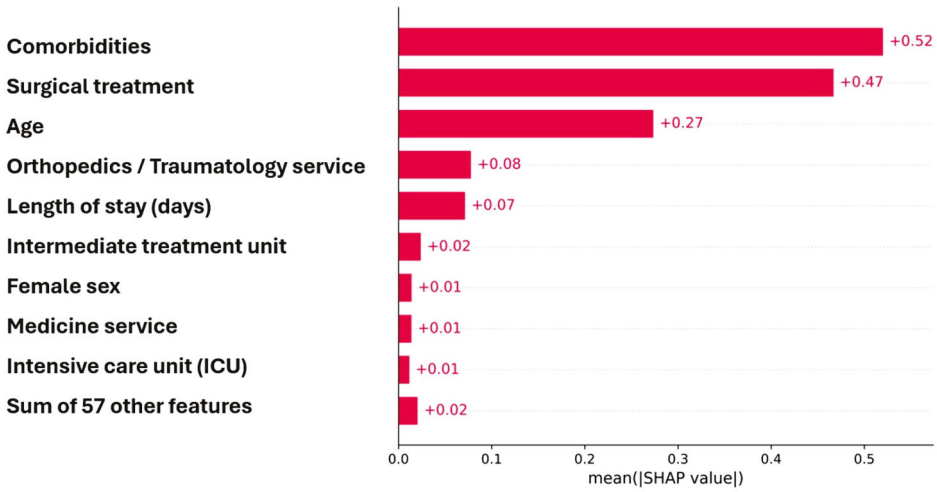


Figure 2. Mean absolute SHAP value ($|\text{SHAP}|$) contribution of predictors for in-hospital mortality after hip fracture.

The beeswarm plot (Figure 3) illustrated that higher comorbidity counts consistently shifted predictions toward lower survival probability (negative SHAP values), whereas surgical treatment was associated with positive SHAP contributions, indicating a protective

effect. Age and prolonged hospitalization also produced predominantly negative shifts, aligning with poorer outcomes. Categories corresponding to intensive or high-complexity care (e.g., ICU or internal medicine) showed small but directionally coherent negative effects.

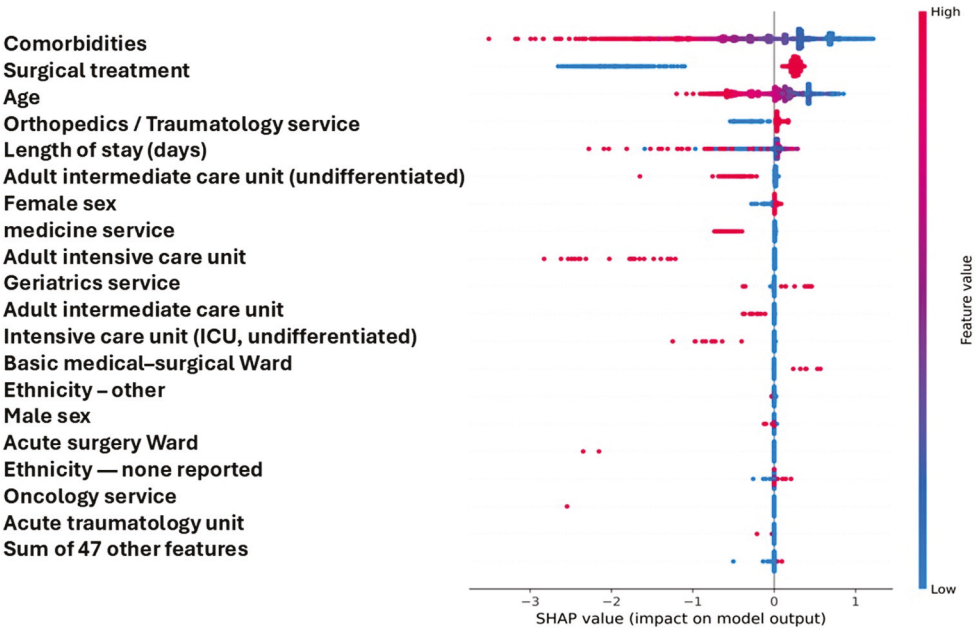


Figure 3. SHAP summary plot showing the impact of each predictor on model output for in-hospital mortality after hip fracture.

Dependence plots confirmed these trends:

- Comorbidities showed a monotonic negative gradient, with risk increasing sharply beyond three comorbidities.
- Surgery displayed two distinct clusters consistent with a binary variable, with surgery exerting a stable protective effect across age groups.
- Age demonstrated a nonlinear negative relationship with predicted survival, with steeper declines in older patients.

Collectively, these results indicate that the model's predictions are driven primarily by disease burden, surgical status, and age—factors that are both clinically plausible and interpretable within the inpatient context.

Individual SHAP waterfall plots (Figure 4) illustrated how the contribution of each predictor combined to shape the model's prediction for individual patients. In a representative case with a high predicted probability of survival, the largest positive contributions were associated with surgical treatment (+0.27), a low number of comorbidities (+0.14), and younger age (+0.11), while a minor negative influence was attributed to a longer hospital stay. This additive breakdown shows how SHAP values can be interpreted as shifts in log-odds, providing transparent insight into the model's reasoning at the individual patient level.

Dependence plots were created for the three most influential predictors—surgical treatment, number of comorbidities, and age—to illustrate their directional influence on the model's log-odds predictions (Figures 5–7). In the case of surgical treatment (Figure 5), two distinct clusters were observed: non-operated patients exhibited strongly negative SHAP values, reflecting a higher predicted risk of mortality, while operated patients showed positive SHAP contributions, consistent with a protective effect. The color gradient by age revealed only a minor interaction, suggesting that the survival benefit associated with surgery remained stable across different age groups.

The comorbidity burden exhibited a monotonic downward trend in SHAP values (Figure 6). As the number of comorbidities increased, the predicted log-odds of survival decreased, particularly beyond three or four comorbidities, suggesting a threshold effect. This variable showed the strongest and most consistent negative influence, independent of surgical status.

The age variable exhibited a clear nonlinear relationship with the model's output (Figure 7). Younger patients were associated with positive SHAP values, reflecting a lower predicted risk of in-hospital mortality, whereas advancing age corresponded to progressively negative SHAP contributions. The interaction coloring by surgical treatment showed that the protective effect of surgery persisted across all age groups, although it tended to diminish slightly among the oldest patients.

Taken together, these dependence plots reinforce the clinical plausibility of the model's predictions, underscoring age, comorbidities, and surgical intervention as the key determinants of in-hospital mortality risk.

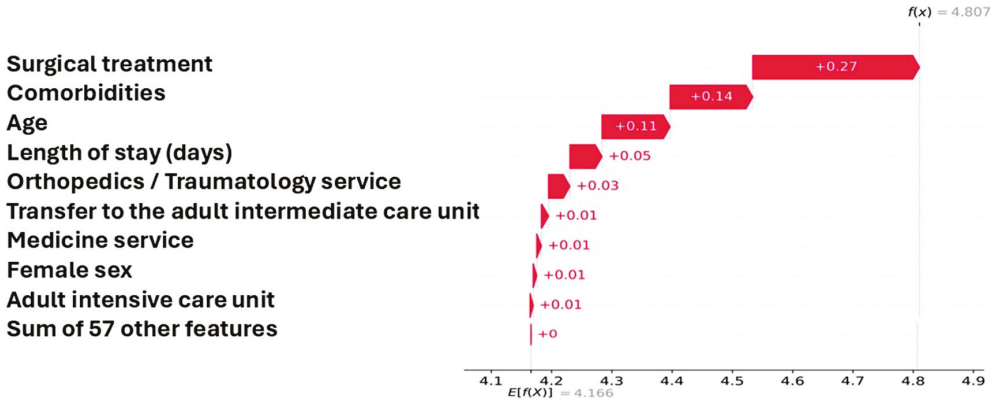


Figure 4. SHAP dependence plot showing the directional impact of top predictors on in-hospital mortality after hip fracture.

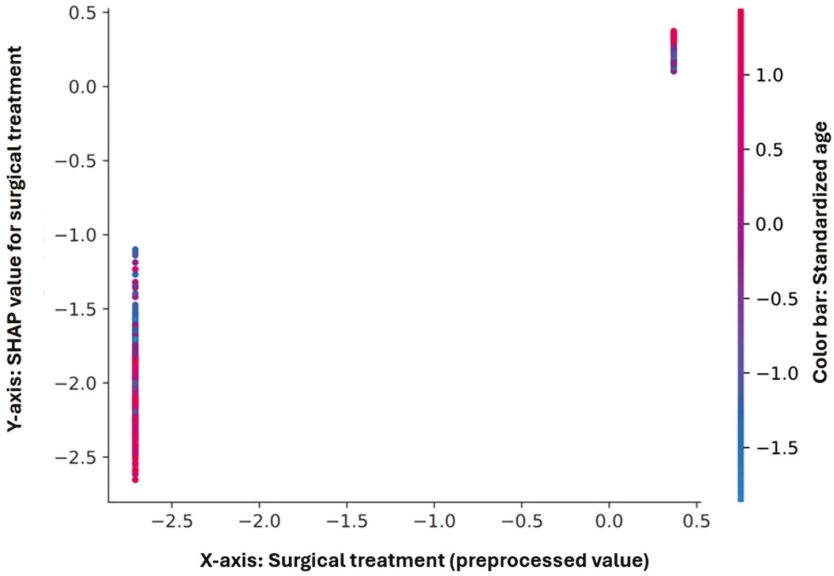


Figure 5. SHAP dependence plot for surgical treatment. The plot depicts the relationship between surgical treatment (x-axis, after preprocessing) and its SHAP contribution to the model output (y-axis). Two well-defined clusters are observed, corresponding to non-operated and operated patients. Negative SHAP values (left cluster) indicate increased predicted mortality, whereas positive SHAP values (right cluster) correspond to lower mortality risk (protective effect). The color gradient represents standardized age, showing minimal interaction with surgical status. The low within-cluster dispersion suggests a stable and robust influence of surgical treatment on model predictions.

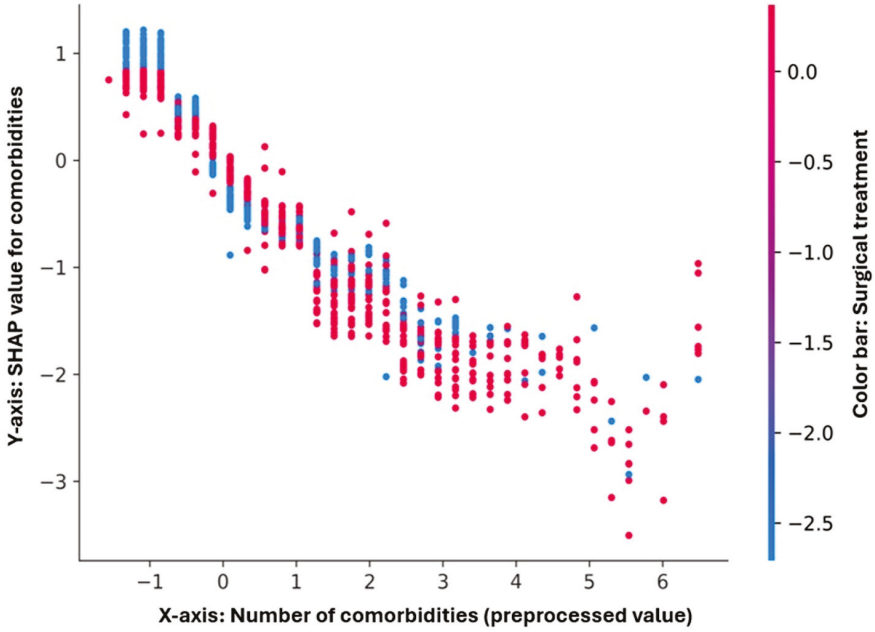


Figure 6. SHAP dependence plot for comorbidities. This plot shows the relationship between the number of comorbidities (x-axis, after preprocessing) and their SHAP contribution to the model output (y-axis). A clear monotonic downward gradient is observed: higher comorbidity counts are associated with more negative SHAP values, indicating increased predicted mortality risk. The color gradient represents surgical treatment, showing that both operated and non-operated patients follow a similar negative pattern, confirming that comorbidity burden exerts a dominant and consistent adverse effect on model predictions.

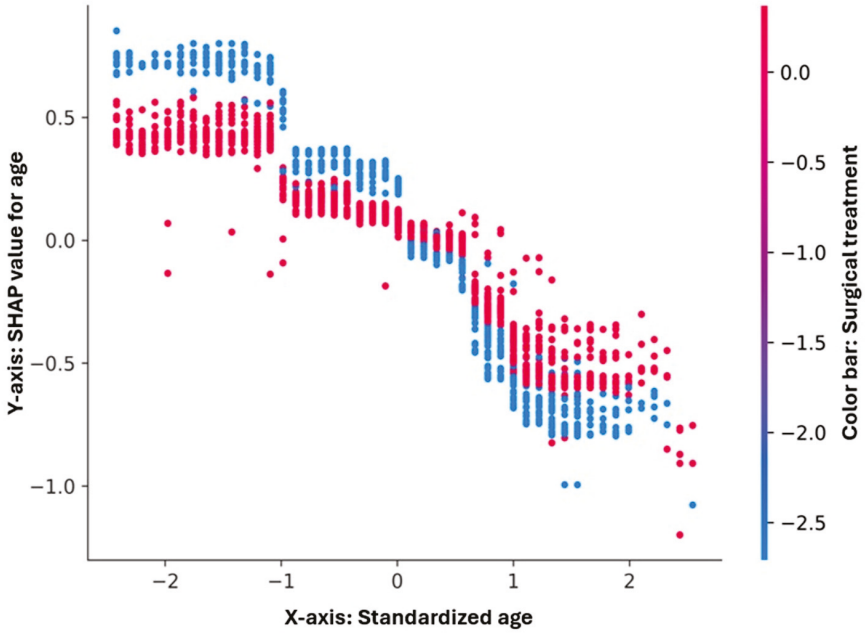


Figure 7. SHAP dependence plot for age. The plot illustrates the relationship between standardized age (x-axis) and its SHAP contribution to the model output (y-axis). A clear negative gradient is observed, indicating that older patients contribute negatively to the log-odds of survival (higher predicted mortality risk). Younger patients show positive SHAP values, suggesting a protective effect. The color gradient represents surgical treatment, revealing a mild interaction: at similar age levels, operated patients tend to cluster slightly higher on the SHAP scale, consistent with a modest protective influence of surgery. Overall, age emerges as a strong and nonlinear determinant of individual mortality risk predictions.

Finally, to provide an interpretable quantitative expression of predicted risk, a SHAP-based risk equation was derived from the Gradient Boosting model—the best-performing algorithm according to AUC. The model output can be expressed as a sum of additive feature contributions (ϕ_i) on the log-odds scale:

$$logit(p) = \beta_0 + \sum_i \phi_i$$

where β_0 represents the model baseline (expected log-odds), and ϕ_i denotes the SHAP value for predictor *i*. The probability of in-hospital survival or mortality is then obtained as:

$$p = \frac{1}{1 + \exp(-logit(p))}$$

The relative risk (RR) for an individual patient can be expressed as:

$$RR = \frac{p}{p^0}, \quad p^0 = \sigma\left(\beta^0\right)$$

For the Gradient Boosting model, the estimated baseline value was $\beta_0 \approx 4.1660$, corresponding to a baseline relative risk $RR_0 \approx 1.0073$. This representation enables the direct computation of individualized mortality risk from the SHAP decomposition, thereby combining interpretability with quantitative precision.

4. Discussion

The main finding of this nationwide study is that a Gradient Boosting machine learning model achieved excellent discrimination for predicting in-hospital mortality among older adults hospitalized with hip fracture (AUC-ROC = 0.885), outperforming the other evaluated algorithms. In addition to its predictive performance, the model provided clinically meaningful interpretability through SHAP analysis, identifying comorbidity burden, surgical treatment, and age as the most influential determinants of mortality risk. These results demonstrate that interpretable, data-driven models based on routinely collected hospital data can effectively support early risk stratification in real-world inpatient settings.

Our findings are consistent with previous research that has applied machine learning techniques to predict clinical outcomes in patients with hip fractures. For example, a recent 2024 study from China reported that an Extreme Gradient Boosting (XGBoost) model achieved an AUC of approximately 0.91 for predicting in-hospital mortality after hip surgery, identifying age, comorbidities, and surgical intervention as the most influential variables. These results closely mirror the patterns observed in our study, reinforcing the robustness and generalizability of tree-based ensemble methods for mortality prediction in this population [20]. Similarly, a Thai study using multiple models reported AUCs ranging from 0.81 to 0.99 for one-year mortality prediction, also highlighting the utility of Random Forest and Gradient Boosting techniques [21]. In contrast, a retrospective analysis in Spain with a smaller sample (~500 patients) failed to surpass an AUC of 0.65, underscoring the need for large, well-curated datasets to build robust models [22]. Additionally, a Swedish registry study involving over 124,000 cases found that regularized logistic regression achieved an AUC of ~0.74, showing that in some contexts, traditional statistical models remain competitive, though with more limited interpretability at the individual level [19].

The SHAP analysis provides important interpretative insights into how established clinical predictors contribute to mortality risk after hip fracture. Although age, comorbidity burden, and surgical treatment are well-known determinants of prognosis, SHAP analysis adds value by quantifying their relative and nonlinear contributions at the individual patient level. In particular, nonlinear age effects, threshold patterns in comorbidity burden,

and the consistently protective contribution of surgical treatment across age strata were observed. This patient-level risk decomposition extends beyond traditional summary statistics or conventional mortality scores and supports transparent, individualized risk stratification in real-world inpatient settings. Although SHAP values were used to identify the most influential predictors, the models were not refitted using only SHAP-ranked features, as SHAP was applied for interpretability rather than post hoc feature selection. Restricting the feature set based solely on importance rankings may lead to information leakage or overly optimistic performance estimates and may overlook complex interactions captured by the full set of clinically plausible predictors.

Several mortality risk scores have been developed to support prognostic assessment after hip fracture, including the Nottingham Hip Fracture Score and indices based on comorbidity burden such as the Charlson Comorbidity Index. These tools have demonstrated clinical utility and remain widely used in routine practice. Rather than replacing established mortality scores, the present machine learning approach should be viewed as complementary. By integrating routinely collected hospital variables and allowing nonlinear interactions, the Gradient Boosting model achieved higher discriminative performance while providing individualized risk decomposition through SHAP-based explanations. This feature enables transparent interpretation of how traditional risk factors—such as age, comorbidities, and surgical treatment—contribute to predicted mortality risk at the patient level, thereby extending the interpretability of conventional scoring systems.

From a clinical perspective, one of the most relevant findings is the negative impact of the number of comorbidities on survival after hip fracture. This is consistent with the geriatric literature, which has consistently shown that a higher burden of chronic diseases worsens prognosis. For example, each one-point increase in the age-adjusted Charlson Comorbidity Index is associated with approximately a 31% increase in mortality risk among hip fracture patients [23]. Likewise, patients with more than three comorbidities experience substantially higher rates of infectious complications and death within the first year after the fracture compared to those with fewer comorbidities [24,25]. Our results confirm the number of comorbidities as the most important predictor, reflecting how preexisting conditions such as heart disease, renal failure, neurological disorders (e.g., dementia), or cancer reduce the patient's physiological reserve and hinder recovery after trauma. In fact, in a study focused on octogenarian patients with hip fracture, the Charlson Index was found to be a better predictor of in-hospital mortality than the ASA scale (BMC Surg), underscoring the importance of quantifying multimorbidity [26]. Our findings reaffirm that high comorbidity translates into greater vulnerability and a higher risk of fatal outcomes; therefore, risk stratification should carefully consider this factor and encourage perioperative interventions aimed at optimizing preexisting medical conditions.

Another notable finding is the protective effect of surgical treatment on in-hospital mortality. In our model, patients who underwent surgery had a lower predicted probability of death during hospitalization, a result that aligns with well-established clinical evidence. Most hip fractures in older adults require surgical repair, and numerous studies have shown that early surgical intervention significantly improves survival outcomes. International guidelines, such as those from the American Academy of Orthopaedic Surgeons (AAOS), recommend performing surgery within 48 h of the fracture, as timely intervention reduces both mortality and postoperative complications [27,28]. Conversely, the non-surgical option is typically reserved for extremely frail patients or those with terminal comorbidities for whom surgical risk is prohibitive. Previous studies have shown that non-surgically managed patients have markedly higher mortality rates than those who undergo surgery, although this difference is influenced by the poorer baseline condition of the non-operated group [29]. In our cohort, the higher mortality observed among patients who did not

undergo surgery is likely attributable to their older age, a greater proportion of very elderly men, and a higher comorbidity burden compared with those who were operated on, factors that are well recognized to increase mortality risk following hip fracture. Consequently, the apparent protective effect of surgery reflects not only the intrinsic benefits of fracture repair—such as enabling early mobilization, effective pain control, and prevention of complications like thrombosis or pressure ulcers—but also a degree of selection bias, since only patients deemed clinically stable are typically eligible for surgery. Nevertheless, our findings reinforce current clinical recommendations to provide surgical treatment to all appropriate candidates, accompanied by adequate preoperative optimization. Timely surgery remains a modifiable determinant of survival and should be prioritized as a cornerstone of hip fracture management in older adults.

Advanced age also emerged as an important prognostic factor, in line with clinical experience and epidemiological studies. As patient age increases, so does the risk of mortality after a hip fracture; it has been estimated that mortality rates may rise by up to 30% in the oldest groups compared with younger adults [30]. This effect is attributed to the biological frailty associated with aging: very old adults often present sarcopenia, reduced cardiac and respiratory reserve, immune alterations, and higher susceptibility to adverse events such as delirium or infections, all of which contribute to worse outcomes. It is therefore unsurprising that age is included in traditional prognostic scales (such as the Nottingham Hip Fracture Score) and in indices like Charlson, and that it consistently appears among mortality predictors identified in modern analyses [19,31]. In the study by Forssten et al., for instance, advanced age was one of the main factors associated with increased one-year mortality [19]. In our cohort, we observed the same trend: older patients had lower in-hospital survival rates. This highlights the need for comprehensive geriatric approaches for very old hip fracture patients, including full geriatric assessment, nutritional optimization, delirium prevention, and aggressive management of acute comorbidities, with the goal of mitigating the negative impact of age on prognosis.

Finally, other findings of our study deserve mention for their clinical relevance, even if they carried a smaller weight in the predictive model. We observed that prolonged hospitalization tends to be associated with worse outcomes. This is consistent with previous reports linking longer hospital stays to higher short-term post-discharge mortality [32]. A plausible explanation is that prolonged hospital stays often arise from medical complications occurring during hospitalization—such as nosocomial infections, pneumonia, thromboembolism, or delirium—or from delayed functional recovery. These complications, in turn, substantially increase the risk of death. Consequently, reducing the length of stay through early rehabilitation, infection prevention, and proactive management of comorbid conditions could have a meaningful impact on survival outcomes. Another relevant aspect is the type of hospital service and care model. The literature has shown that implementing shared-care models between orthopedic surgeons and geriatricians (orthogeriatric units) is associated with lower in-hospital mortality among hip fracture patients [29]. Orthogeriatric collaboration ensures more holistic care: it optimizes comorbidity management upon admission, tailors surgical indication and timing to patient conditions, and promotes preventive interventions (early mobilization, nutritional support, avoidance of inappropriate polypharmacy) that impact recovery. In our setting, adopting a similar multidisciplinary approach could partly explain the favorable outcomes observed and represents an opportunity for improvement where such models are not yet implemented. In summary, these findings reaffirm that mortality following hip fracture in older adults is a multifactorial phenomenon determined by the interaction between patient baseline characteristics (age, comorbidities, functional status), treatment factors (performance and timing of surgery), and the quality of hospital care (prevention of complications, interdisci-

plinary management). Addressing each of these components comprehensively is essential to reduce mortality and improve prognosis in this vulnerable population [29,32].

From a clinical perspective, the proposed Gradient Boosting model may have several practical applications in the inpatient management of older adults with hip fracture. By leveraging routinely available hospital data, the model could support early risk stratification at admission, helping clinicians identify patients at higher risk of in-hospital mortality who may benefit from closer monitoring, multidisciplinary care, or early optimization of comorbid conditions. In particular, the strong contribution of comorbidity burden, surgical treatment, and age aligns with current orthogeriatric principles and reinforces the importance of timely surgical decision-making and comprehensive medical management. Moreover, the use of SHAP-based explanations enables transparent, patient-level interpretation of risk, facilitating clinical communication and supporting shared decision-making without replacing established clinical judgment or existing mortality risk scores.

This study has several limitations that warrant consideration. First, it relied on a retrospective cohort derived from a single national health system (FONASA), which may limit the generalizability of the findings to other healthcare settings. Although the dataset encompasses a large proportion of the Chilean population, external validation—either temporal or geographic—was not performed in the present analysis due to reliance on a single nationwide administrative database. Second, the set of predictors was restricted to administrative and care-related variables routinely available in hospital discharge records. The absence of physiological measures (e.g., vital signs), laboratory results, and detailed perioperative indicators (e.g., ASA score, time to surgery) may have introduced residual confounding and reduced the precision of individual risk estimation. Third, the outcome definition—in-hospital death at discharge—does not account for deaths occurring shortly after discharge. As a result, time-to-event or competing-risk models could not be applied, potentially leading to an underestimation of early post-discharge mortality. Finally, it is important to note that SHAP-based explanations describe model-derived associations, not causal relationships. Although SHAP enhances model transparency and interpretability, its results should be understood within the framework of predictive inference rather than causal explanation.

5. Conclusions

In this nationwide cohort of older adults hospitalized with hip fracture, a Gradient Boosting machine learning model demonstrated excellent discrimination for predicting in-hospital mortality using routinely collected hospital data. Comorbidity burden, surgical treatment, and age emerged as the most influential predictors of mortality risk. The integration of SHAP analysis enhanced model transparency by providing clinically interpretable, patient-level explanations. These findings support the potential role of interpretable machine learning models as complementary tools for early risk stratification in real-world inpatient settings. Future studies should focus on external validation and evaluation of clinical impact.

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

Funding: The Article Processing Charge (APC) was funded by Universidad Arturo Prat.

Institutional Review Board Statement: The study exclusively used fully anonymized and publicly available administrative data released by the Chilean National Health Fund (FONASA) and the Ministry of Health through official open-data platforms. According to Chilean Law No. 19.628 (Ley sobre Protección de la Vida Privada), anonymized information that cannot be linked to identifiable individuals is not considered personal data and therefore does not require ethical review or informed consent. Additionally, Law No. 20.285 (Ley de Transparencia de la Función Pública y Acceso a la Información de la Administración del Estado) regulates the public availability and lawful reuse of governmental open datasets.

Informed Consent Statement: Patient consent was waived due to the use of publicly available and fully anonymized data from the National Health Fund (FONASA, Chile). No individual or identifiable information was collected or accessed.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors thank the National Health Fund (FONASA, Chile) for making the hospital discharge data publicly available through its open-access database.

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

Abbreviations

The following abbreviations are used in this manuscript:

AAOS	American Academy of Orthopaedic Surgeons
AI	Artificial intelligence
ASA	American Society of Anesthesiologists physical status classification
AUC-ROC	Area under the receiver operating characteristic curve
FONASA	Fondo Nacional de Salud (National Health Fund, Chile)
ICD-10	International Classification of Diseases, Tenth Revision
ICU	Intensive care unit
kNN	k-Nearest Neighbors
MLP	Multilayer Perceptron
RBF	Radial basis function
ROC	Receiver operating characteristic
RR	Relative risk
SHAP	Shapley Additive Explanations
SVM	Support Vector Machine
YLDs	Years lived with disability

References

1. Gullberg, B.; Johnell, O.; Kanis, J.A. World-Wide Projections for Hip Fracture. *Osteoporos. Int.* **1997**, *7*, 407–413. [CrossRef]
2. Feng, J.; Zhang, C.; Li, B.; Zhan, S.; Wang, S.; Song, C. Global Burden of Hip Fracture: The Global Burden of Disease Study. *Osteoporos. Int.* **2024**, *35*, 41–52. [CrossRef] [PubMed]
3. Dyer, S.M.; Crotty, M.; Fairhall, N.; Magaziner, J.; Beaupre, L.A.; Cameron, I.D.; Sherrington, C. A Critical Review of the Long-Term Disability Outcomes Following Hip Fracture. *BMC Geriatr.* **2016**, *16*, 158. [CrossRef] [PubMed]
4. Barry, M.L.; Maday, K.R. Reviewing Acute Hip Fractures in Adults. *J. Am. Acad. Physician Assist.* **2021**, *34*, 1–10. [CrossRef] [PubMed]
5. Guzmán-Muñoz, E.; Concha-Cisternas, Y.; Vásquez-Muñoz, M.; Yañez-Sepúlveda, R.; Núñez-Espinosa, C.; Bittelman Saporte, S.; Nemtala Urquiza, F.; Morales Araneda, R. In-Hospital Mortality Among 40,253 Older Adults with Hip Fracture: Survival Outcomes and Multivariate Analysis in a Chilean Cohort. *J. Clin. Med.* **2025**, *14*, 7717. [CrossRef]
6. Hu, F.; Jiang, C.; Shen, J.; Tang, P.; Wang, Y. Preoperative Predictors for Mortality Following Hip Fracture Surgery: A Systematic Review and Meta-Analysis. *Injury* **2012**, *43*, 676–685. [CrossRef]
7. Sheehan, K.J.; Sobolev, B.; Guy, P.; Kuramoto, L.; Morin, S.N.; Sutherland, J.M.; Beaupre, L.; Griesdale, D.; Dunbar, M.; Bohm, E.; et al. In-Hospital Mortality after Hip Fracture by Treatment Setting. *Can. Assoc. J.* **2016**, *188*, 1219–1225. [CrossRef]

8. Warren, M.; Bretherton, C.; Parker, M. Delay to Surgery beyond 12 Hours Is Associated with Increased Hip Fracture Mortality. *Eur. J. Orthop. Surg. Traumatol.* **2024**, *34*, 2973–2980. [CrossRef]
9. Nunes, B.K.G.; Lima, B.R.D.; Guimarães, L.C.D.C.; Guimarães, R.A.; Rosso, C.F.W.; Felipe, L.D.A.; Pagotto, V. Survival Analysis of Hospitalized Elderly People with Fractures in Brazil Over One Year. *Geriatrics* **2020**, *5*, 10. [CrossRef]
10. Downey, C.; Kelly, M.; Quinlan, J.F. Changing Trends in the Mortality Rate at 1-Year Post Hip Fracture—A Systematic Review. *World J. Orthop.* **2019**, *10*, 166–175. [CrossRef]
11. Wang, J.-W.D. Naïve Bayes Is an Interpretable and Predictive Machine Learning Algorithm in Predicting Osteoporotic Hip Fracture In-Hospital Mortality Compared to Other Machine Learning Algorithms. *PLoS Digit. Health* **2025**, *4*, e0000529. [CrossRef]
12. Quevedo, I.; Ormeño, J.C.; Weissglas, B.; Opazo, C. Epidemiology and Direct Medical Cost of Osteoporotic Hip Fracture in Chile. *J. Osteoporos.* **2020**, *2020*, 5360467. [CrossRef] [PubMed]
13. Trad, F.; Isber, B.; Yammine, R.; Hatoum, K.; Obeid, D.; Chahine, M.; Haidar, R.; El-Hajj Fuleihan, G.; Chehab, A. Parsimonious and Explainable Machine Learning for Predicting Mortality in Patients Post Hip Fracture Surgery. *Sci. Rep.* **2025**, *15*, 22922. [CrossRef] [PubMed]
14. Das, A.; Dhillon, P. Application of Machine Learning in Measurement of Ageing and Geriatric Diseases: A Systematic Review. *BMC Geriatr.* **2023**, *23*, 841. [CrossRef] [PubMed]
15. Du, J.; Tao, X.; Zhu, L.; Wang, H.; Qi, W.; Min, X.; Wei, S.; Zhang, X.; Liu, Q. Development of a Visualized Risk Prediction System for Sarcopenia in Older Adults Using Machine Learning: A Cohort Study Based on CHARLS. *Front. Public Health* **2025**, *13*, 1544894. [CrossRef]
16. Li, Y.; Chen, M.; Lv, H.; Yin, P.; Zhang, L.; Tang, P. A Novel Machine-Learning Algorithm for Predicting Mortality Risk after Hip Fracture Surgery. *Injury* **2021**, *52*, 1487–1493. [CrossRef]
17. Asrian, G.; Suri, A.; Rajapakse, C. Machine Learning-based Mortality Prediction in Hip Fracture Patients Using Biomarkers. *J. Orthop. Res.* **2024**, *42*, 395–403. [CrossRef]
18. Lei, M.; Han, Z.; Wang, S.; Han, T.; Fang, S.; Lin, F.; Huang, T. A Machine Learning-Based Prediction Model for in-Hospital Mortality among Critically Ill Patients with Hip Fracture: An Internal and External Validated Study. *Injury* **2023**, *54*, 636–644. [CrossRef]
19. Forssten, M.P.; Bass, G.A.; Ismail, A.M.; Mohseni, S.; Cao, Y. Predicting 1-Year Mortality after Hip Fracture Surgery: An Evaluation of Multiple Machine Learning Approaches. *J. Pers. Med.* **2021**, *11*, 727. [CrossRef]
20. Lei, M.; Feng, T.; Chen, M.; Shen, J.; Liu, J.; Chang, F.; Chen, J.; Sun, X.; Mao, Z.; Li, Y.; et al. Establishment and Validation of an Artificial Intelligence Web Application for Predicting Postoperative In-Hospital Mortality in Patients with Hip Fracture: A National Cohort Study of 52,707 Cases. *Int. J. Surg.* **2024**, *110*, 4876–4892. [CrossRef]
21. Kitcharanant, N.; Chotiyarnwong, P.; Tanphiriyakun, T.; Vanitcharoenkul, E.; Mahaisavariya, C.; Boonyaprapa, W.; Unnanuntana, A. Development and Internal Validation of a Machine-Learning-Developed Model for Predicting 1-Year Mortality after Fragility Hip Fracture. *BMC Geriatr.* **2022**, *22*, 451. [CrossRef]
22. Calvo Lorenzo, I.; Uriarte Llano, I.; Mateo Citores, M.R.; Rojo Maza, Y.; Agirregoitia Enzunza, U. Análisis de modelos algorítmicos de aprendizaje automático para la predicción del estado vital a los seis meses tras fractura de cadera en pacientes mayores de 74 años. *Rev. Esp. Cir. Ortop. Traumatol.* **2025**, *69*, 47–54. [CrossRef] [PubMed]
23. Zhang, D.-L.; Cong, Y.-X.; Zhuang, Y.; Xu, X.; Zhang, B.-F. Age-Adjusted Charlson Comorbidity Index Predicts Postoperative Mortality in Elderly Patients with Hip Fracture: A Prospective Cohort. *Front. Med.* **2023**, *10*, 1066145. [CrossRef] [PubMed]
24. Bekeris, J.; Wilson, L.A.; Bekere, D.; Liu, J.; Poeran, J.; Zubizarreta, N.; Fiasconaro, M.; Memtsoudis, S.G. Trends in Comorbidities and Complications Among Patients Undergoing Hip Fracture Repair. *Anesth. Analg.* **2021**, *132*, 475–484. [CrossRef] [PubMed]
25. Gadgaard, N.R.; Varnum, C.; Nelissen, R.G.H.H.; Vandenbroucke-Grauls, C.; Sørensen, H.T.; Pedersen, A.B. Comorbidity and Risk of Infection among Patients with Hip Fracture: A Danish Population-Based Cohort Study. *Osteoporos. Int.* **2023**, *34*, 1739–1749. [CrossRef]
26. Zhang, H.; Ma, L.; Yu, X. Risk Factors of Postoperative Complications and In-Hospital Mortality after Hip Fracture among Patients Older than 80 Years Old: A Retrospective Study. *BMC Surg.* **2025**, *25*, 122. [CrossRef]
27. Brox, W.T.; Roberts, K.C.; Taksali, S.; Wright, D.G.; Wixted, J.J.; Tubb, C.C.; Patt, J.C.; Templeton, K.J.; Dickman, E.; Adler, R.A.; et al. The American Academy of Orthopaedic Surgeons Evidence-Based Guideline on Management of Hip Fractures in the Elderly. *J. Bone Jt. Surg.* **2015**, *97*, 1196–1199. [CrossRef]
28. Maheshwari, K.; Planchard, J.; You, J.; Sakr, W.A.; George, J.; Higuera-Rueda, C.A.; Saager, L.; Turan, A.; Kurz, A. Early Surgery Confers 1-Year Mortality Benefit in Hip-Fracture Patients. *J. Orthop. Trauma* **2018**, *32*, 105–110. [CrossRef]
29. Dinamarca-Montecinos, J.L.; Améstica-Lazcano, G.; Rubio-Herrera, R.; Carrasco-Buvinic, A.; Vásquez, A. Características Epidemiológicas y Clínicas de Las Fracturas de Cadera En Adultos Mayores En Un Hospital Público Chileno: Experience in 647 Chilean Patients Aged 60 Years or More. *Rev. Méd. Chile* **2015**, *143*, 1552–1559. [CrossRef]

30. Cui, Z.; Feng, H.; Meng, X.; Zhuang, S.; Liu, Z.; Ye, K.; Sun, C.; Xing, Y.; Zhou, F.; Tian, Y. Age-Specific 1-Year Mortality Rates after Hip Fracture Based on the Populations in Mainland China between the Years 2000 and 2018: A Systematic Analysis. *Arch. Osteoporos.* **2019**, *14*, 55. [CrossRef]
31. Vitiello, R.; Pesare, E.; Capece, G.; Di Gialleonardo, E.; De Matthaeis, A.; Franceschi, F.; Maccauro, G.; Covino, M. Surgical Timing and Clinical Factor Predicting In-Hospital Mortality in Older Adults with Hip Fractures: A Neuronal Network Analysis. *J. Orthop. Traumatol.* **2025**, *26*, 30. [CrossRef]
32. Salvador Marín, J.; Ferrández Martínez, F.J.; Fuster Such, C.; Seguí Ripoll, J.M.; Orozco Beltrán, D.; Carratalá Munuera, M.C.; Martínez López, J.F.; Marzo Campos, J.C. Factores de riesgo para el ingreso prolongado y mortalidad intrahospitalaria en la fractura del fémur proximal en pacientes mayores de 65 años. *Rev. Esp. Cir. Ortop. Traumatol.* **2021**, *65*, 322–330. [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.

diagnostics

MDPI

Article

Use of Machine Learning to Predict the Incidence of Type 2 Diabetes Among Relatively Healthy Adults: A 10-Year Longitudinal Study in Taiwan

Ying-Qiang Liu [1,†], Tzu-Wei Chang [2,3,†], Lung-Chun Lee [2], Chia-Yu Chen [4], Pi-Shan Hsu [2], Yu-Tse Tsan [5,6], Chao-Tung Yang [7,8] and Wei-Min Chu [2,9,10,11,*]

1 Department of Medical Education, Taichung Veterans General Hospital, Taichung 407219, Taiwan
2 Department of Family Medicine, Taichung Veterans General Hospital, Taichung 407219, Taiwan
3 Division of Family Medicine, Department of Medicine, Taipei Veterans General Hospital Yuanshan Branch, Yilan 264018, Taiwan
4 Department of Application Value-Added Service, SYSTEX Corporation, Taipei 114730, Taiwan
5 Division of Occupational Medicine, Department of Emergency Medicine, Taichung Veterans General Hospital, Taichung 407219, Taiwan
6 School of Medicine, Chung Shan Medical University, Taichung 402306, Taiwan
7 Department of Computer Science, Tunghai University, Taichung 407224, Taiwan
8 Research Center for Smart Sustainable Circular Economy, Tunghai University, Taichung 407224, Taiwan
9 Geriatrics and Gerontology Research Center, College of Medicine, National Chung Hsing University, Taichung 402202, Taiwan
10 Department of Post-Baccalaureate Medicine, College of Medicine, National Chung Hsing University, Taichung 402202, Taiwan
11 School of Medicine, National Yang Ming Chiao Tung University, Taipei 112304, Taiwan
* Correspondence: williamchu0110@gmail.com
† These authors contributed equally to this work.

Abstract: Background: The prevalence of diabetes is increasing worldwide, particularly in the Pacific Ocean island nations. Although machine learning (ML) models and data mining approaches have been applied to diabetes research, there was no study utilizing ML models to predict diabetes incidence in Taiwan. We aimed to predict the onset of diabetes in order to raise health awareness, thereby promoting any necessary lifestyle modifications and help mitigate disease burden. **Methods**: The research dataset used in the study was retrieved from the Clinical Data Center of Taichung Veterans General Hospital. We collected data from the available electronic health records with a total of 33 items being employed for model construction. Individuals with diabetes and those with missing data were excluded. Ultimately, 6687 adults were included in the final analysis, where we implemented three different ML algorithms, including logistic regression (LR), random forest (RF) and extreme gradient boosting (XGBoost) in order to predict diabetes. **Results**: The top five important factors involved in the prediction model were glycated hemoglobin (HbA1c), fasting blood glucose, weight, free thyroxine (fT4), and triglycerides (TG). Notably, random forest, logistic regression, and XGBoost reached 99%, 99%, and 98% accuracy, respectively. fT4 seems to be one of the significant features in predicting the onset of diabetes. Moreover, this would be the first study using machine learning models to predict diabetes that has demonstrated the importance of thyroid hormone. **Conclusions**: A total of 33 items were able to be put into the machine learning model in order to predict diabetes with promising accuracy. In comparison to prior studies on machine learning models, this study not only identified similar key factors for predicting diabetes but also highlighted the significance of thyroid hormones, a factor that was previously overlooked. Moreover, it highlighted the relevance of predicting type 2 diabetes using more affordable methods, which would be useful for clinical healthcare professionals and endocrinologists who apply the models to clinical practice.

Keywords: machine learning models; diabetes; free thyroxine; glycated hemoglobin; fasting blood glucose; weight; triglycerides

1. Introduction

The prevalence of type 2 diabetes has been increasing throughout the world, reflecting a rising burden on financial expenditures and public health [1]. Remarkably, Pacific Ocean island nations have the highest prevalence of type 2 diabetes. Additionally, the condition is also one of the leading causes of human suffering, affecting quality of life and even premature mortality [2]. Of major concern is that the onset of type 2 diabetes may be delayed 9–12 years upon diagnosis, when patients who present themselves with established microvascular complications, such as diabetic retinopathy, can be treated [3]. It remains important to identify patients with undiagnosed type 2 diabetes during the preclinical period [4], as well as to predict the disease among relatively healthy adults.

For asymptomatic individuals, a conventional medical checkup performed by healthcare professionals involving history taking, physical examinations, and laboratory tests and imaging is essential [5]. Additionally, medical checkups are all part of preventive medicine and provide opportunities for medical personnel to educate the patient while also disseminating information regarding this chronic disease [6], therefore, raising health awareness. Regarding issues surrounding society and government, Christian et al. found that the global cost of diabetes will substantially increase by 2030, greatly impacting the world's healthcare systems, necessitating that urgent action be taken in order to mitigate its economic burden [7]. In the coming future, reducing the impact that diabetes has on both global health and its economy will remain a challenging issue.

During the onset of this decade, the COVID-19 pandemic raged worldwide, leading to an unprecedented hazardous effect on mental health due to the policies of social distancing and quarantine [8]. Indeed, the virus caused a deterioration in, and a negative impact on, many individual's mental health [9] and it may have indirectly raised public awareness of mental health partly due to fear caused by the pandemic. Intriguingly, multiple mechanisms have been postulated, which may explain the rising rate of new-onset diabetes in COVID-19 patients. One of these mechanisms is that severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) enters the islet cell through the angiotensin-converting enzyme-2 (ACE-2) receptors [10], which may eventually impede insulin secretion. A meta-analysis found that 19.70% (95% CI: 10.93–32.91) of COVID-19 patients experienced new-onset diabetes [11]. Given the higher risk of newly detected diabetes in the post-acute COVID-19 population, screening remains essential regardless of the severity during the acute phase of COVID-19 [12]. Through advancing technology, Artificial Intelligence (AI) applications are now being seen in the healthcare system and are gradually receiving attention from healthcare professionals as they become more involved in clinical decision-making [13]. One example is how predicting physical function upon discharge among older adults based upon data taken from electronic health records (EHRs) is now being used [14]. The machine learning (ML) model develops algorithms to analyze patterns from the data taken from EHRs. However, there are still certain limitations due to data quality [15].

To date, numerous ML models and data mining approaches have been applied to diabetes research [16]. Not only the prediction of the onset of diabetes complications but also the early diagnosis of type 2 diabetes can be carried out. Dagliati et al. [17] in Italy discovered that applying data mining methods, as well as using demographic data and clinical data from EHRs among nearly 1000 patients suffering from type 2 diabetes in order to develop predictive models for the onset of microvascular complications was a

promising approach. In recently performed research, a novel ML algorithm, called the extreme gradient boosting (XGBoost) classifier, is even applied as a prediction model for type 2 diabetes, where it reaches a high identification rate [18].

However, any relevant literature regarding the use of ML models for predicting diabetes incidence is scarce in the Asia-Pacific region [18,19], and some of that literature was only completed via a questionnaire survey [20]. Therefore, it remains clear that additional research using ML to predict type 2 diabetes among healthy adults is still warranted. Thus, the objectives of this study are:

1. to select appropriate features predicting type 2 diabetes among relatively healthy adults;
2. to establish a prediction model via the use of different ML algorithms and subsequently compare the performance accuracy.

This is the first study utilizing EHR data taken from a tertiary medical center for the purpose of diabetes prediction among healthy adults in Taiwan having a 10-year longitudinal design.

2. Materials and Methods

2.1. Dataset

The research dataset was retrieved from the Clinical Data Center of Taichung Veterans General Hospital. Relatively healthy adults who had undergone at least two self-paid health examinations within an interval of less than four years during the period from January 2011 to June 2021 were included for analysis. We collected data from EHRs regarding general demographics, medical history, and laboratory tests, with a total of 33 items being employed for model construction. Individuals with a medical history of diabetes prior to the study period and those with missing data were excluded. The parameters used included age, gender, height, weight, waist size, pulse rate, respiration rate, total cholesterol (TC), triglycerides (TG), high density lipoprotein cholesterol (HDL-C), low density lipoprotein cholesterol (LDL-C, measured), aspartate transaminase (AST), alanine transaminase (ALT), total bilirubin, direct bilirubin, r-glutamyl transferase (r-GT), total protein, albumin, blood urea nitrogen (BUN), serum creatinine, estimated glomerular filtration rate (eGFR), uric acid, serum sodium, serum calcium, hemoglobin (Hgb), platelets, high-sensitivity c-reactive protein (hsCRP), fasting blood glucose, glycated hemoglobin (HbA1c), thyroid stimulating hormone (TSH), free thyroxine (fT4), urine glucose, and urine ketone. Ultimately, 6687 adults were included in the final analysis. This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Taichung Veterans General Hospital (protocol code: TCVGH-IRB CE21445B-1, date of approval: 5 January 2023). The Patient Consent form was waived by TCVGH-IRB due to the study's retrospective design.

2.2. Data Pre-Processing

Initially, we merged EHRs and excluded adults who had been diagnosed with diabetes. Individuals were classified as potential diabetes cases based on an HbA1C > 6.5% or a fasting glucose >126 mg/dL. The annotated data after organization were categorized using the following four labels: "No diabetes symptoms", "Potential diabetes symptoms", "Confirmed diabetes", and "No data". Meanwhile, the "df.isnull().sum()" method was used to calculate the number of missing values in each column. A DataFrame entitled "Missing" was created, containing column names and the percentage of missing values. Subsequently, a horizontal bar plot depicting the missing ratios was generated, and any columns with missing ratios exceeding 20% were removed. Recognizing that filling in missing values with a specific number may impact the model's prediction ability, the

remaining missing values were completed using "−999" as an indicator that the values were missing values. Finally, the "train_test_split function" from the "scikit-learn version 1.6 (sklearn 1.6)" software machine learning library was used to partition the data into training and testing sets with a split ratio of 60% and 40%. The training set was utilized for model training, while the testing set was employed to evaluate the model's performance.

2.3. Machine Learning and Prediction Model Development

2.3.1. Random Forest (RF)

RF belongs to Ensemble Learning, representing a more sophisticated decision tree. Comprising multiple decision trees, each tree operates independently within the ensemble. In the classification process, every new sample undergoes both evaluation and classification by each decision tree within the forest. Consequently, each decision tree produces an individual classification result, with the Random Forest consolidating the outcomes through a voting mechanism, aggregating the classification results from all the decision trees.

The RF equation is as follows:

$$RF = \sum_{k=1}^{K} P_k(1 - P_k) = 1 - \sum_{k=1}^{K} P_k^2 \tag{1}$$

2.3.2. Logistic Regression (LR)

The LR model serves as a form of linear classifier, mainly used in binary classification scenarios. It is primary for classifying data based on the available information and makes judgments to assign data points to specific categories. Notably, the output values generated by the logistic regression model for classification purposes are constrained within the range of [0, 1].

2.3.3. XGBoost

The full name of XGBoost is Extreme Gradient Boosting. This algorithm maintains the integrity of the original model through each iteration and adds a new function to the model. This process enables subsequent trees to correct errors made by preceding trees. Moreover, XGBoost incorporates random feature sampling during tree generation, ensuring that not all features are considered in decision-making for each tree. The algorithm contributes to the model's diversity and robustness, while also enhancing the predictive performance.

The equation of XGBoost is as follows:

$$XGBoost = \sum_{i=1}^{n} l\left(y_i, y_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(f_i) \tag{2}$$

2.4. Data Analysis Through Machine Learning

Through the aforementioned data pre-processing, we employed various machine learning models during data predictive analysis, including RF, LR, and XGBoost. Furthermore, we used SHAP (Shapley Additive exPlanations) in order to perform a visualization of the model's attention to data features [21,22]. This could help contribute to enhancing the interpretability of the model prediction as well as enable us to better comprehend the reasons and decision-making processes behind the models. Consequently, the predictive outcomes become more trustworthy and interpretable. Such an approach assists us towards reaching a deeper interpretation of the data, thus allowing us to make more valuable predictions and decisions.

3. Results

The final analysis encompassed a total of 6687 adults. Table 1 shows the demographics and clinical characteristics of all participants during their first health exam. The average age was 57.7 years, while the average FBS and HbA1c was 89.6 and 5.5, respectively, suggesting that participants were relatively healthy when they underwent their health exam.

Table 1. Demographic and clinical characteristics of participants.

Parameter	Mean	25%	50%	75%
age	57.7	50	58	65
height (cm)	165.7	159.5	165.9	171.7
weight (kg)	66.2	56.8	65.6	74.0
waist (cm)	80	73	80	86
respiration rate (/min)	16.6	16	16	18
pulse (/min)	70.6	63	70	77
AST (U/L)	24.3	18	22	27
ALT (U/L)	30.1	17	24	35
total bilirubin (mg/dL)	0.88	0.6	0.8	1.0
direct bilirubin (mg/dL)	0.5	0.5	0.5	0.5
r-GT (U/L)	33.9	16	24	36
serum creatinine (mg/dL)	0.84	0.70	0.83	0.99
BUN (mg/dL)	12	10	12	14
eGFR (ml/min/1.73m2)	97.91	83.09	94.74	109.11
FBS (mg/dL)	89.6	84	89	95
HbA1c (%)	5.5	5.3	5.5	5.7
TC (mg/dL)	198.2	174	196	220
TG (mg/dL)	127.5	72	105	155
HDL-C (mg/dL)	55.1	44	53	64
LDL-C, measured (mg/dL)	116.45	96.3	113.0	136.0
Hgb (g/dL)	14.4	13.3	14.5	15.5
platelet (1000/μL)	243.3	203	238	277
albumin (g/dL)	4.5	4.3	4.5	4.7
total protein (g/dL)	7.3	7.0	7.3	7.6
uric acid (mg/dL)	6.2	5.0	6.1	7.2
fT4 (ng/dL)	13.2	11.2	12.3	14.8
hsCRP (mg/L)	0.1688	0.021	0.058	0.158
serum sodium (mEq/L)	143.4	142	143	145
serum calcium (mg/dL)	8.9	8.7	8.9	9.2
urine glucose (mg/dL)	8.8	0	0	0
urine ketone (mg/dL)	0.2	0	0	0

AST, aspartate transaminase; ALT, alanine transaminase; r-GT, r-glutamyl transferase; BUN, blood urea nitrogen; eGFR, estimated glomerular filtration rate; FBS, fasting blood glucose; HbA1c, glycated hemoglobin; TC, total cholesterol; TG, triglycerides; HDL-C, high density lipoprotein cholesterol; LDL-C, low density lipoprotein cholesterol; Hgb, hemoglobin; fT4, free thyroxine; hsCRP, high-sensitivity c-reactive protein.

In Table 2, the precision, recall, and F1-score of the classification models were each revealed. The accuracy rates for RF, LR, and XGBoost were 99%, 99%, and 98%, respectively. The precisions for the prediction of normal cases in RF, LR, and XGBoost were all 99%. However, the precisions for diabetes cases in RF, LR, and XGBoost were 25.0%, 0%, and 19.0%, respectively.

The confusion matrix of ML in different prediction models is depicted in Figure 1. RF and XGBoost showed relatively good results; however, LR was poor in differentiating true diabetic patients.

Examining the area under the ROC (auROC) curve as the performance metric, in Figure 2 for a 4-year diabetes prediction, both XGBoost and LR models exhibited an auROC exceeding 90%, while RF had an auROC of 74%.

Table 2. Precision, recall, F1-score and accuracy of machine learning model by XGBoost, Random Forest, and logistic regression with normal cases and potential diabetes cases.

	Class 0 (Normal Cases), Support: 6967			Class 1 (Potential Diabetes Cases), Support: 76				
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Macro F1-Score**
Random Forest	0.99	1.00	0.99	0.25	0.01	0.03	0.99	0.51
Logistic regression	0.99	1.00	0.99	0.00	0.00	0.00	0.99	0.50
XGBoost	0.99	0.98	0.99	0.19	0.33	0.24	0.98	0.61

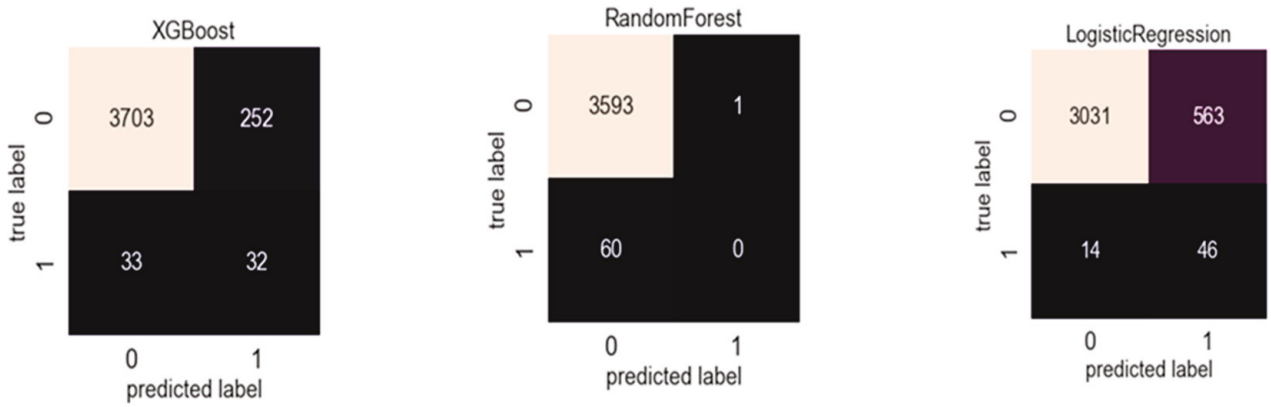


Figure 1. Confusion matrix of machine learning model by XGBoost, Random Forest, and logistic regression.

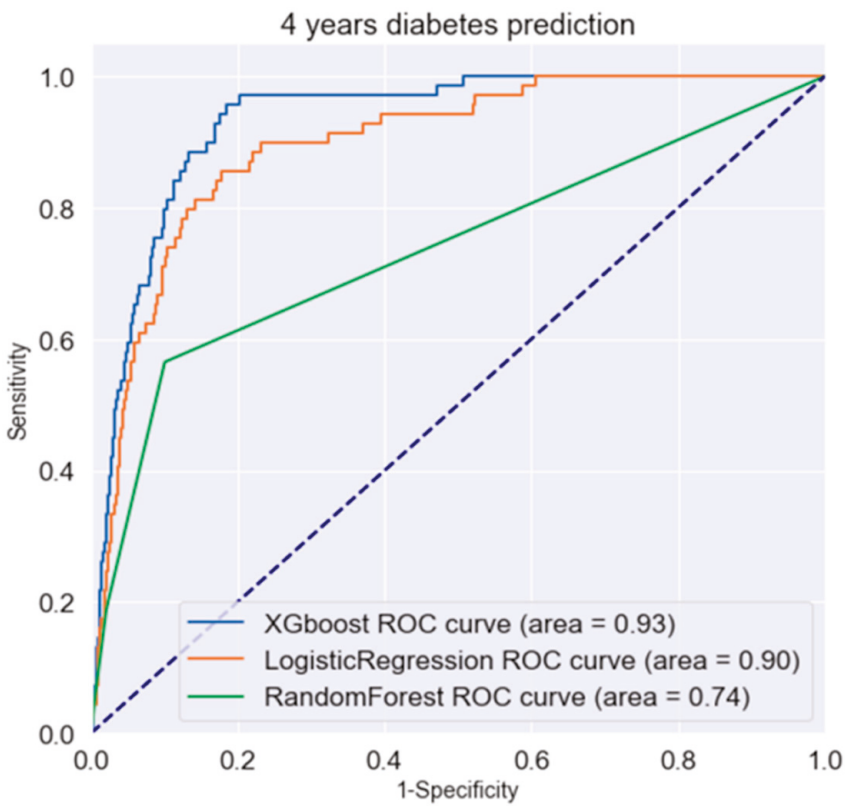


Figure 2. ROC curve of machine learning model by XGBoost, Random Forest, and logistic regression for 4-year diabetes prediction.

The classification algorithm by XGBoost successfully identified key features in the process, as illustrated in Figure 3. Notably, the top five features of importance remained consistent: HbA1c, fasting blood glucose, weight, fT4, and TG.

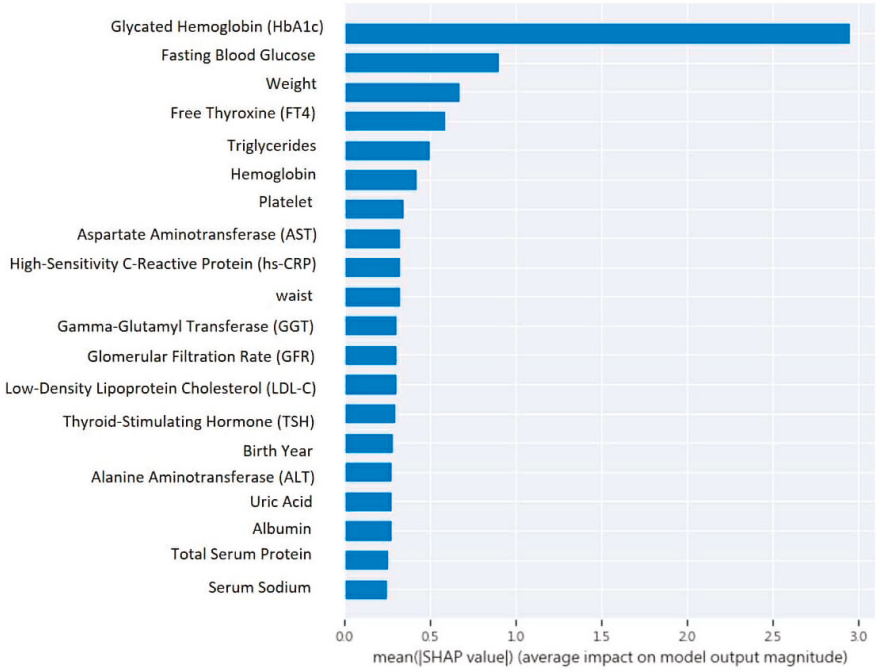


Figure 3. Feature importance of machine learning model by XGBoost predicting incidence of type 2 diabetes among relatively healthy adults.

4. Discussion

Despite the highest prevalence and disease burden of type 2 diabetes seen in the Pacific Ocean island nations, there still remains a lack of the clinical tools necessary in order to predict diabetes incidence. To the best of our knowledge, this is the first study to utilize ML models and EHRs to predict diabetes incidence among relatively healthy adults in Taiwan. The results demonstrate that diabetes incidence can be accurately predicted through the use of ML models, including LR, RF, and XGBoost, when combined with key clinical features. Our ML models yielded robust predictions for categorical outcomes. Furthermore, the top five features of importance were HbA1c, fasting blood glucose, weight, fT4, and TG. Consequently, we contend that the application of Artificial Intelligence can assist healthcare professionals in recognizing the early stages of diabetes, thereby enhancing its clinical applicability.

In the past, numerous studies have utilized ML models to investigate the onset of diabetes complications; however, there are relatively fewer studies exploring the prediction of any early detection of diabetes. Dagliati et al. [17] applied a logistic regression model with selected features to predict the onset of diabetes complications, including retinopathy, neuropathy, and nephropathy, at different time scenarios among a type 2 diabetes population. Their results showed that data mining methods could be adopted in clinical medicine in order to support clinical practice. Wu et al. [18] used a novel binary logistic regression, the XGBoost classifier, as a prediction model for type 2 diabetes. Additionally, the team made it adaptive to more than one dataset, including both the Pima Indians Diabetes Database (PIDD) and Early-Stage Diabetes Risk Prediction Database (ESDRPD), coming up with a 94% and 98% identification rate for diabetes prediction, respectively. Recently, Abnoosian et al. [23] proposed an innovative multi-classification framework using the imbalanced Iraqi Patient Dataset of Diabetes in order to predict diabetes and found that the

model outperformed other machine learning models in diabetes prediction. Additionally, the study also discussed the use of various classification algorithms, data preprocessing techniques, and the impact of different parameters on model accuracy, as other previous research had done [24]. Their study highlighted the potential of how a proposed framework can serve as a valuable tool. From our study, the ML models were able to predict type 2 diabetes accurately using selected clinical features among relatively healthy Asian adults. This is certainly valuable for supporting clinical practice and enhancing people's health awareness.

4.1. Features of Importance

There were several studies utilizing machine learning to predict diabetes [25–27]. Fasting blood glucose is one of the significant features in the studies. Although this study revealed a similar result, it also revealed the importance of fT4. The association between thyroid function and diabetes appears to be less intuitive. In clinical settings, thyroid function measurement, including thyroid-stimulating hormone (TSH) and fT4, is not included in routine screenings for diabetes, particularly in people without any apparent symptoms of thyroid dysfunction. However, with regard to the findings of our study, fT4 seems to be one of the significant features in predicting the onset of diabetes. The thyroid hormone plays a role in both regulating metabolism and energy expenditure, while it is also involved in insulin regulation and glucose homeostasis [28]. In accordance with the previous literature, the thyroid hormone has been shown to augment beta-cell viability [29]. Chaker et al. [30] performed a population-based prospective cohort study and discovered that subclinical hypothyroidism is a risk factor for incident diabetes among individuals with prediabetes. Recently, a study conducted in Japan demonstrated that both hyperthyroidism and hypothyroidism correlate with type 2 diabetes, and that a population with subclinical hypothyroidism showed a higher incidence of diabetic complications [31]. Future studies are still warranted in order to better explore whether thyroid function measurement should be incorporated into diabetes screening among relatively healthy adults. Furthermore, the relevance of predicting type 2 diabetes using more affordable methods, such as triglycerides, compared to the cost of the HbA1C test should be explored. This would be useful for clinical healthcare professionals and endocrinologists who apply the models to clinical practice.

4.2. Strength and Limitations

The study has some limitations. First, the data investigated came from a single tertiary hospital without validation from any other available public databases, thus external validity should be interpreted cautiously. It remains uncertain whether the models used can be generalized to other healthcare systems. Second, the inclusion of numerous clinical features may pose challenges for smaller-scale healthcare systems which may not be able to collect comprehensive clinical data. This may potentially impact the accuracy of the model. However, our study implemented three different ML algorithms and compared their accuracy, emphasizing the potential of machine learning in predicting diabetes using medical informatics [32]. Moreover, the results highlighted that thyroid function may be a significant factor for diabetes prediction among relatively healthy adults. This would be the first study using machine learning models to demonstrate the importance of thyroid hormone in predicting diabetes.

4.3. Implications

Our study indicates that diabetes prediction among relatively healthy adults is feasible using EHRs and ML models. For clinical healthcare professionals involved in preventive medicine, we believe that the prediction model could apply to both clinical practice and

shared decision-making, particularly for individuals seeking to raise health awareness, thereby promoting any necessary lifestyle modifications. Diabetes prediction could not only prevent the onset of an irreversible disease, but it may also mitigate the burden of excess financial expenditures, which would negatively impact the healthcare system over the coming years. Additionally, the clinical features highlight the importance of thyroid function as being a significant factor. Due to the promising accuracy resulting from these ML models, we aim to promote the application of the models to other healthcare systems in the near future.

5. Conclusions

We were able to predict diabetes among relatively healthy adults through EHRs with regard to general demographics, medical history, and laboratory tests, while involving a total of 33 items. We found that HbA1c, fasting blood glucose, weight, fT4 and TG are the five important factors involved in the prediction model. The accuracy rates for RF, LR, and XGBoost were 99%, 99%, and 98%, respectively. In the future, we will seek to explore the application in different ML algorithms and apply it to the available public databases in order to verify its generality and versatility. Furthermore, this would allow us to make the models more practical for further application in other healthcare systems with regard to their use as a clinical tool.

Author Contributions: Conceptualization, W.-M.C.; methodology, C.-Y.C. and W.-M.C.; software, C.-Y.C.; validation, W.-M.C., C.-Y.C., T.-W.C. and Y.-Q.L.; formal analysis, C.-Y.C.; investigation, W.-M.C., C.-Y.C. and T.-W.C.; resources, W.-M.C.; data curation, C.-Y.C.; writing—original draft preparation, T.-W.C. and W.-M.C.; writing—review and editing, Y.-Q.L., L.-C.L., P.-S.H., Y.-T.T., C.-T.Y. and W.-M.C.; visualization, T.-W.C. and Y.-Q.L.; supervision, W.-M.C.; funding acquisition, W.-M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Taichung Veterans General Hospital, Taichung, Taiwan, grant number TCVGH-1117203C. In addition, the APC was funded by Taichung Veterans General Hospital, Taichung, Taiwan.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Taichung Veterans General Hospital (IRB: CE21445B-1, 25 May 2021).

Informed Consent Statement: The Patient Consent form was waived by TCVGH-IRB due to the study's retrospective design.

Data Availability Statement: The datasets generated and analyzed during the study are available from the corresponding author on reasonable request.

Acknowledgments: The authors acknowledge receiving valuable statistical support from Taichung Veterans General Hospital, Taiwan.

Conflicts of Interest: Author Chia-Yu Chen was employed by the SYSTEX Corporation. 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. Khan, M.A.B.; Hashim, M.J.; King, J.K.; Govender, R.D.; Mustafa, H.; Al Kaabi, J. Epidemiology of Type 2 Diabetes—Global Burden of Disease and Forecasted Trends. *J. Epidemiol. Glob. Health* **2020**, *10*, 107–111. [CrossRef]
2. Ang, G.Y. Age of onset of diabetes and all-cause mortality. *World J. Diabetes* **2020**, *11*, 95–99. [CrossRef]
3. Harris, M.I.; Klein, R.; Welborn, T.A.; Knuiman, M.W. Onset of NIDDM occurs at least 4–7 yr before clinical diagnosis. *Diabetes Care* **1992**, *15*, 815–819. [CrossRef] [PubMed]
4. Chatterjee, S.; Khunti, K.; Davies, M.J. Type 2 diabetes. *Lancet* **2017**, *389*, 2239–2251. [CrossRef]

5. Al-Kahil, A.B.; Khawaja, R.A.; Kadri, A.Y.; Abbarh, M.S.M.; Alakhras, J.T.; Jaganathan, P.P. Knowledge and Practices Toward Routine Medical Checkup Among Middle-Aged and Elderly People of Riyadh. *J. Patient Exp.* **2020**, *7*, 1310–1315. [CrossRef] [PubMed]
6. Wee, H.L.; Ho, H.K.; Li, S.C. Public awareness of diabetes mellitus in Singapore. *Singap. Med. J.* **2002**, *43*, 128–134.
7. Bommer, C.; Sagalova, V.; Heesemann, E.; Manne-Goehler, J.; Atun, R.; Bärnighausen, T.; Davies, J.; Vollmer, S. Global Economic Burden of Diabetes in Adults: Projections from 2015 to 2030. *Diabetes Care* **2018**, *41*, 963–970. [CrossRef] [PubMed]
8. Xiong, J.; Lipsitz, O.; Nasri, F.; Lui, L.M.W.; Gill, H.; Phan, L.; Chen-Li, D.; Iacobucci, M.; Ho, R.; Majeed, A.; et al. Impact of COVID-19 pandemic on mental health in the general population: A systematic review. *J. Affect. Disord.* **2020**, *277*, 55–64. [CrossRef] [PubMed]
9. Yao, H.; Chen, J.-H.; Xu, Y.-F. Patients with mental health disorders in the COVID-19 epidemic. *Lancet Psychiatry* **2020**, *7*, e21. [CrossRef]
10. Boddu, S.K.; Aurangabadkar, G.; Kuchay, M.S. New onset diabetes, type 1 diabetes and COVID-19. *Diabetes Metab. Syndr.* **2020**, *14*, 2211–2217. [CrossRef] [PubMed]
11. Shrestha, D.B.; Budhathoki, P.; Raut, S.; Adhikari, S.; Ghimire, P.; Thapaliya, S.; Rabaan, A.A.; Karki, B.J. New-onset diabetes in COVID-19 and clinical outcomes: A systematic review and meta-analysis. *World J. Virol.* **2021**, *10*, 275–287. [CrossRef]
12. Banerjee, M.; Pal, R.; Dutta, S. Risk of incident diabetes post-COVID-19: A systematic review and meta-analysis. *Prim. Care Diabetes* **2022**, *16*, 591–593. [CrossRef] [PubMed]
13. Secinaro, S.; Calandra, D.; Secinaro, A.; Muthurangu, V.; Biancone, P. The role of artificial intelligence in healthcare: A structured literature review. *BMC Med. Inf. Inform. Decis. Mak.* **2021**, *21*, 125. [CrossRef] [PubMed]
14. Chu, W.-M.; Tsan, Y.-T.; Chen, P.-Y.; Chen, C.-Y.; Hao, M.-L.; Chan, W.-C.; Chen, H.-M.; Hsu, P.-S.; Lin, S.-Y.; Yang, C.-T. A model for predicting physical function upon discharge of hospitalized older adults in Taiwan-a machine learning approach based on both electronic health records and comprehensive geriatric assessment. *Front. Med.* **2023**, *10*, 1160013. [CrossRef] [PubMed]
15. Marier, A.; Olsho, L.E.; Rhodes, W.; Spector, W.D. Improving prediction of fall risk among nursing home residents using electronic medical records. *J. Am. Med. Inf. Inform. Assoc.* **2016**, *23*, 276–282. [CrossRef] [PubMed]
16. Kavakiotis, I.; Tsave, O.; Salifoglou, A.; Maglaveras, N.; Vlahavas, I.; Chouvarda, I. Machine Learning and Data Mining Methods in Diabetes Research. *Comput. Struct. Biotechnol. J.* **2017**, *15*, 104–116. [CrossRef]
17. Dagliati, A.; Marini, S.; Sacchi, L.; Cogni, G.; Teliti, M.; Tibollo, V.; De Cata, P.; Chiovato, L.; Bellazzi, R. Machine Learning Methods to Predict Diabetes Complications. *J. Diabetes Sci. Technol.* **2018**, *12*, 295–302. [CrossRef] [PubMed]
18. Wu, Y.; Zhang, Q.; Hu, Y.; Sun-Woo, K.; Zhang, X.; Zhu, H.; Jie, L.; Li, S. Novel binary logistic regression model based on feature transformation of XGBoost for type 2 Diabetes Mellitus prediction in healthcare systems. *Future Gener. Comput. Syst.* **2022**, *129*, 1–12. [CrossRef]
19. Mahajan, P.; Uddin, S.; Hajati, F.; Moni, M.A. Ensemble Learning for Disease Prediction: A Review. *Healthcare* **2023**, *11*, 1808. [CrossRef] [PubMed]
20. Wang, L.; Wang, X.; Chen, A.; Jin, X.; Che, H. Prediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model. *Healthcare* **2020**, *8*, 247. [CrossRef] [PubMed]
21. Nohara, Y.; Matsumoto, K.; Soejima, H.; Nakashima, N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Comput. Methods Programs Biomed.* **2022**, *214*, 106584. [CrossRef] [PubMed]
22. Lundberg, S. A unified approach to interpreting model predictions. *arXiv* **2017**, arXiv:1705.07874.
23. Abnoosian, K.; Farnoosh, R.; Behzadi, M.H. Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. *BMC Bioinform.* **2023**, *24*, 337. [CrossRef]
24. Hasan, K.; Alam, A.; Das, D.; Hossain, E.; Hasan, M. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. *IEEE Access* **2020**, *8*, 76516–76531. [CrossRef]
25. Zou, Q.; Qu, K.; Luo, Y.; Yin, D.; Ju, Y.; Tang, H. Predicting diabetes mellitus with machine learning techniques. *Front. Genet.* **2018**, *9*, 515. [CrossRef] [PubMed]
26. Olisah, C.C.; Smith, L.; Smith, M. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Comput. Methods Programs Biomed.* **2022**, *220*, 106773. [CrossRef]
27. Shin, J.; Kim, J.; Lee, C.; Yoon, J.Y.; Kim, S.; Song, S.; Kim, H.-S. Development of various diabetes prediction models using machine learning techniques. *Diabetes Metab. J.* **2022**, *46*, 650–657. [CrossRef] [PubMed]
28. Crunkhorn, S.; Patti, M.-E. Links between thyroid hormone action, oxidative metabolism, and diabetes risk? *Thyroid* **2008**, *18*, 227–237. [CrossRef]
29. Falzacappa, C.V.; Mangialardo, C.; Raffa, S.; Mancuso, A.; Piergrossi, P.; Moriggi, G.; Piro, S.; Stigliano, A.; Torrisi, M.R.; Brunetti, E.; et al. The thyroid hormone T3 improves function and survival of rat pancreatic islets during in vitro culture. *Islets* **2010**, *2*, 96–103. [CrossRef]
30. Chaker, L.; Ligthart, S.; Korevaar, T.I.M.; Hofman, A.; Franco, O.H.; Peeters, R.P.; Dehghan, A. Thyroid function and risk of type 2 diabetes: A population-based prospective cohort study. *BMC Med.* **2016**, *14*, 150. [CrossRef] [PubMed]

31. Nishi, M. Diabetes mellitus and thyroid diseases. *Diabetol. Int.* **2018**, *9*, 108–112. [CrossRef] [PubMed]
32. Khanam, J.J.; Foo, S.Y. A comparison of machine learning algorithms for diabetes prediction. *ICT Express* **2021**, *7*, 432–439. [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.

diagnostics

MDPI

Article

Integrating Host Genetics and Clinical Setting in Machine Learning Models: Predicting COVID-19 Prognosis for Healthcare Decision-Making (The FeMiNa Study)

Elisabetta D'Aversa [1,†], Bianca Antonica [1,†], Miriana Grisafi [1], Rosanna Asselta [2,3], Elvezia Maria Paraboschi [2,3], Angelina Passaro [1], Stefano Volpato [4,5], Francesca Remelli [4], Massimiliano Castellazzi [6,7], Alberto Maria Marra [8,9], Antonio Cittadini [8,9], Roberta D'Assante [9], Francesca Salvatori [1], Ajay Vikram Singh [10], Salvatore Pernagallo [11], Veronica Tisato [1,7,12,*] and Donato Gemmati [1,7,13,*]

1 Department of Translational Medicine, University of Ferrara, 44121 Ferrara, Italy; elisabetta.daversa@unife.it (E.D.); bianca.antonica@unife.it (B.A.); miriana.grisafi@unife.it (M.G.); francesca.salvatori@unife.it (F.S.)
2 Department of Biomedical Sciences, Humanitas University, 20072 Pieve Emanuele, Italy
3 Medical Genetics and RNA Biology Unit, IRCCS Humanitas Research Hospital, 20089 Rozzano, Italy
4 Department of Medical Science, University of Ferrara, 44121 Ferrara, Italy
5 Geriatrics Unit, Sant'Anna University Hospital of Ferrara, 44124 Cona, Italy
6 Department of Neurosciences and Rehabilitation, University of Ferrara, 44121 Ferrara, Italy; massimiliano.castellazzi@unife.it
7 University Strategic Centre for Studies on Gender Medicine, University of Ferrara, 44121 Ferrara, Italy
8 Department of Translational Medical Sciences, Federico II University, 80131 Naples, Italy
9 Interdepartmental Centre for Research in Gender Medicine-GENESIS, Federico II University, 80131 Naples, Italy
10 Department of Chemical and Product Safety, German Federal Institute for Risk Assessment (BfR), 10589 Berlin, Germany
11 DESTINA Genomica S.L, PTS Granada, Edificio BIC, 18016 Armilla, Spain
12 Laboratory for Advanced Therapy Technologies (LTTA), University of Ferrara, 44121 Ferrara, Italy
13 Centre Haemostasis and Thrombosis, University of Ferrara, 44121 Ferrara, Italy
* Correspondence: veronica.tisato@unife.it (V.T.); d.gemmati@unife.it (D.G.); Tel.: +39-0532-455841 (D.G.)
† These authors contribute equally to this work.

Highlights

What are the main findings?

- eXtreme Gradient Boosting f1 optimization avoids losing several patients due to misdiagnosis.
- Genetic data implement the model's power for mortality prediction.

What are the implications of the main findings?

- Integrating genetics in ML enables a more personalized medical approach.

Abstract

Background/Objectives: COVID-19 has made a tremendous impact, causing a massive number of deaths worldwide. The inadequacy of health facilities resulted in shortage of resources and exhaustion of frontline workers who had to manage in a short time many patients with no tools to prioritize those at high risk. This study intended to disclose the architecture of such complex disease and enhance the management of hospitalized patients, preventing severe outcomes. **Methods**: We performed a retrospective multicenter study aimed at refining the best predictive model for COVID-19 mortality, integrating 19 genetic and 13 clinical features. We trained three machine learning (ML) models (GBM, XGB and RF) on a dataset of 532 COVID-19 hospitalized Italian patients, among the 605 recruited

during the first wave of the pandemic, when vaccines were not available. **Results:** All the models achieved great values for accuracy, AUROC, f1, f2 and PR-AUC metrics. XGB f1 optimization resulted in better performance providing fewer false positives (N_{f1} = 26 versus N_{f2} = 27, $N_{PR\text{-}AUC}$ = 29), and mostly false negatives (N_{f1} = 63 versus N_{f2} = 69, $N_{PR\text{-}AUC}$ = 69), being the main goal to answer. We next delved into the feature importance to understand which features contribute to the model decision: age was the main driver of mortality prediction, followed by ventilation. The remainder was equally distributed between genetic (*HLA-DRA* rs3135363, *PPARGC1A* rs192678, *CRP* rs2808635, *ABO* rs657152) and other clinical features, demonstrating that genetic data did not confound, but rather implemented, the power of the model. **Conclusions**: Our results suggest that integrating genetic and clinical data into ML models is crucial for identifying high-risk cases within the vast disease heterogeneity, enabling the P4-medicine approach to improve patient outcomes and support the healthcare system.

Keywords: machine learning; artificial intelligence; COVID-19; predictive model; host genetics; healthcare solutions

1. Introduction

SARS-CoV-2 is the seventh encountered virus strain causing respiratory disease in humans, ranging from mild to severe symptoms [1,2]. Age, male sex, and defined comorbidities, including cardiovascular and pulmonary disease, appear to increase the severity and mortality of COVID-19 [3–6]. Nevertheless, low-risk individuals, such as young people without comorbidities, may evolve with severe disease, needing hospitalization [7]. Overall, about 5% of COVID-19 patients were admitted to the intensive care unit (ICU), 2.3% underwent invasive mechanical ventilation and 1.4% died [8]. From the beginning of the pandemic, considerable efforts have been made to identify informative indicators and risk factors influencing SARS-CoV-2 susceptibility and COVID-19 prognosis [4,9]. The extreme clinical heterogeneity observed in patients and the different geographical distribution of cases cannot be explained solely by the presence or absence of possible concomitant health conditions. This observation strongly suggests that individual host genetics may play a crucial role in determining clinical phenotypes across populations [5,10,11]. Several Genome-Wide Association Studies (GWAS) and international initiatives have identified hundreds of gene variants, most of them Single Nucleotide Polymorphisms (SNPs) and haplotypes that are causative of the severe phenotype or confer an advantage in fighting the infection [10,12–15]. In this context, several genes involved in inflammatory response (Interleukin6 (*IL6*), Tumor Protein p53 (*TP53*), C-Reactive Protein (*CRP*), C-C motif Chemokine Receptor 2 (*CCR2*)), immune response (Human Leukocyte Antigen (*HLA*), Complement factor H (*CFH*)), susceptibility to viral infections (3′,5′-Oligoadenylate Synthase 3 (*OAS3*), (Leucine zipper transcription factor-like 1 (*LZTFL1*), Cystic Fibrosis Transmembrane Conductance Regulator (*CFTR*)), or genes implicated in energy metabolism (Apolipoprotein E (*APOE*), PPARG coactivator 1 alpha (*PPARGC1A*), Uncoupling Protein 1 (*UCP1*)) and in the entrance of SARS-CoV-2 (Angiotensin converting enzyme 2 (*ACE2*)) have been investigated (see Table 1 for the specific References).

COVID-19 has made a tremendous impact on the health of people all over the world and caused a significant number of deaths, heavily impacting health systems [16,17]. In Italy 27 million cases and 198,000 deaths were documented, with daily hospitalization costs ranging from €475.86 to €1401.65 for the most severe ICU patients [18,19]. All past experiences should help to better understand the critical areas for improvement and set up

systems and mandatory prevention measures to counteract future infections and manage other burden diseases. What emerged most was undoubtedly the inadequacy of health facilities, resulting in a severe shortage of medical resources and the exhaustion of frontline healthcare workers [20,21] who had to manage many patients in a short time without any tools to calculate specific diagnoses and prognoses in advance. The unpredictable disease course made diagnosis and treatment a challenging mission, particularly for high-risk patients. Hence, the early identification of those at risk is necessary to mitigate the burden on healthcare delivery and reduce mortality as much as possible [22,23].

Analytical models that accurately predict COVID-19 outcomes could assist in efficiently allocating the limited medical resources, improving the healthcare quality, and eventually optimizing patient management [24]. Predictions from various computational and statistical models have been used to address this challenge [25,26]. Since the beginning of the COVID-19 pandemic, new digital technologies, such as artificial intelligence (AI) and machine learning (ML), have been introduced as predictors of COVID-19 mortality [27]. In recent years, their intersection with healthcare has revolutionized disease prediction, delivering more accurate results than conventional statistical models particularly for complex diseases, deciphering relationships between biomarkers and symptoms, and predicting clinical outcomes. Generally, the prognostic performance of ML-based models for predicting COVID-19 outcomes (e.g., mortality, length of hospitalization, ICU admission, or long COVID) [28–30] utilizes conventional risk factors such as comorbidities, clinical manifestations, laboratory results, and demographic characteristics [31–33].

In 2020, the COVID-19 Host Genetics Initiative (HGI) brought together the human genetics community to generate, share, and analyze epidemiological and scientific data to understand the genetic bases and determinants of COVID-19 susceptibility, clinical severity, and prognosis (https://www.covid19hg.org/ , accessed on 10 December 2025). Our research group took part in the HGI with the project "Extreme genotype-phenotype comparison in SARS-CoV2 patients: direct candidate genes Pathway and GWAS", aimed at identifying candidate genes and variants related to pathways specifically involved in COVID-19 [3,34,35]. Subsequently, our group also moved toward facing this complex painting by a COVID-omics strategy and AI to gain future perspectives in tailored diagnosis, therapeutic schemes, vaccine efficacy, and other associated coinfections [36–40].

With growing global knowledge about COVID-19 and considering that both genetic and clinical features can contribute to its complexity, our attention has unavoidably shifted toward developing an informative predictive model of the disease course. We considered the effects of the main genetic and acquired factors and how their combined mutual interactions could change susceptibility and predispose to severity. In the present retrospective multicenter investigation, we optimized and compared different ML approaches to develop accurate algorithms for mortality risk and long hospital stay prediction in COVID-19 patients. They were recruited from three different Italian centers (Ferrara, Fe; Milan, Mi; Naples, Na; the FeMiNa study) during the first wave of the pandemic and before any efficient treatment, or drug, or vaccine was available. Subsequently, we analyzed the interactions among the variables, focusing on the contribution of the genetic component. This approach could also help us to unravel the complex architecture of such infectious diseases and extract meaningful patterns and insights where the host genetics, in addition to clinical factors, can influence susceptibility and prognosis.

2. Materials and Methods

2.1. Study Participants and Sample Recruitment

We recruited 605 Italian patients among those hospitalized after confirmed SARS-CoV-2 infection who developed severe COVID-19 symptoms during the period January–

September 2020, far from vaccine availability. During the study period, the cumulative incidence of COVID-19 cases was 3740.15 per 100,000 inhabitants, 4673.71 per 100,000, and 3175.57/100,000 in the regions where Ferrara, Milan and Naples are located (Emilia Romagna, Lombardia and Campania respectively), with 5% hospitalization rate (updated to 31 December 2020 [41]). After signed informed consent and local ethical committee approval (IRB) were obtained, whole blood samples were collected (N = 300 from the Hospital-University of Ferrara, CE:405/2020; N = 182 from the IRCCS Humanitas Clinical and Research Center of Milan, CE:316/2020; N = 123 from the Federico II University Hospital of Naples, CE:98/2021). At patient admission, personal anamnesis including existing comorbidities and other characteristics have been recorded (i.e., hypertension, heart failure, ischemic stroke, arteriopathy, Chronic Obstructive Pulmonary Disease (COPD), hepatopathy, neoplasm, dementia, diabetes, sex, age, need for ventilation and blood type).

2.2. Genomic DNA Extraction

Whole blood samples have been collected in vacutainer containing EDTA as anticoagulant. All the samples were anonymized using alpha-numeric encryption, with the decryption keys available only to the PIs and co-PIs of the respective centers where patients were recruited. Blood samples were handled in Bio-Safety Level P3 cleanrooms and stored at −40 °C until genomic DNA extraction. DNA was isolated using an automated DNA extraction and purification robot (BioRobot EZ1 system, Qiagen; Hilden, Germany). For those samples with low DNA extraction yield, a different extraction procedure was employed using the QIAamp DNA Blood Mini Kit columns (Qiagen).

2.3. Genotyping and Variant Analysis

Detection of the selected gene variants was performed by qPCR using rhAmp SNP genotyping technology (IDT, Integrated DNA Technologies; Coralville, IA, USA) for the following SNPs: *ACE2* rs2285666, *HLA-A* rs2499, *TP53* rs1042522, *OAS3* rs10735079, *LZTFL1* rs35044562, *IL6* rs2228145; by TaqMan SNP genotyping Assay (Thermo Fisher Scientific, Applied Biosystems; Waltham, MA, USA) for the following: *APOE* rs429358, *APOE* rs7412, *PPARGC1A* rs8192678, *HLA-DPB1* rs9277356, *HLA-DRA* rs3135363, *UCP1* rs1800592, *IL6* rs1800795, *CRP* rs2808635, *CRP* rs876538, *CFH* rs1061170, *CFTR* (rs113993960, rs1801178, rs113993959, rs76713772, rs74597325, rs80034486, rs74767530, rs77010898, rs121908799), on QuantStudio3 Real-Time PCR System (Thermo Fisher Scientific), according to the supplier instruction. Finally, alpha 1-3-N-acetylgalactosaminyltransferase and alpha 1-3-galactosyltransferase (*ABO*) rs657152 and *CCR2* rs34041956 were detected by pyrosequencing (PyroMark Q96 Autoprep ID System, Biotage, AB; Uppsala, Sweden) after standard PCR amplification [42] on Agilent SureCycler 8800 (Agilent Technologies; Santa Clara, CA, USA).

2.4. Statistical Analysis

The statistical analysis was performed by the R statistical package (R Core Team, v4.1.2; Vienna, Austria [43]) and Statistical Package for the Social Sciences (SPSS, IBM SPSS Statistics for Windows, v29.0.2.0; Armonk, NY, USA [44]). In the study population, the allele frequencies for each variant were compared with those reported in the National Center for Biotechnology Information (NCBI) registry for the European population. The genotype distributions were further verified for the Hardy–Weinberg Equilibrium. The normal distribution of the variables was evaluated through the Kolmogorov–Smirnov test. Comparisons of categorical variables have been performed using the *t*-test and X^2-test. Pearson correlation was calculated by pairing all the variables with the output and reported as a correlation matrix. The association of a genotype (or allele) with the clinical phenotype of interest has been performed through associations for multiple inheritance

models (dominant and recessive). For each model, we calculated Odds Ratio (OR) and the 95% confidence interval (CI) using MedCalc Statistical Software (MedCalc Software, v19.2.6 bv; Ostend, Belgium [45]). p-value ≤ 0.05 was considered statistically significant, and this threshold has been set for all the computed analyses.

2.5. Principal Component Logistic Regression (PCLR)

Before training ML models, the dataset was analyzed through PCLR as a first predictive statistical approach to determine the combination of clinical and genetic variables in forecasting bad outcomes (i.e., death and long hospital stay ($\geq$15 days) among survivors). Specifically, we combined Principal Component Analysis (PCA) with binary logistic regression using PCs as covariates. Age was scaled before PCA according to the formula [$Z = (x - \mu)/\sigma$]. Gene variants were coded 1, 2, 3, for common homozygous, heterozygous, and rare homozygous variant respectively. *CFTR* variants were grouped and binary categorized into non-carriers (1) and heterozygous carriers (2). No *CFTR* homozygotes have been found. All clinical features have been categorized as binary codes (0/1), except for blood type and ventilation (see Table 2). The PCs with eigenvalue exceeding 1.0 have been computed in a binary logistic regression analysis as independent variables. The relationship with the dependent variable was considered statistically significant with p-value ≤ 0.05. Variables with a loading above the cut-off point ± 0.3 were considered to be dominant in a component.

2.6. Machine Learning Assessment

We employed three ML algorithms, Random Forest (RF) [46,47] , Gradient Boosting Machine (GBM) [48] , and eXtreme Gradient Boosting (XGB) [49] , to predict mortality and hospital stay among the cohort of COVID-19 patients, and the outputs were further compared. Specifically, the ML analyses were conducted in cooperation with an external consultants (Service DIMEC, Bologna, Italy).The implementation of these algorithms was provided by the H2O.ai framework (Python Interface for H2O, H2O.ai, v3.46.0.3, July 2024; Mountain View, CA, USA [50]). Each algorithm was designed to enhance the different aspects of predictive accuracy and computational efficiency. In detail, RF builds multiple decision trees and merges them to obtain a more accurate and stable prediction; GBM builds a collection of decision trees in a sequential manner, where each subsequent tree corrects the errors of the previous ones; XGB is an implementation of GBM that, by providing parallel tree boosting, is known for its efficiency and performance.

Globally, our study analyzed a dataset consisting of 34 features, divided into 19 categorical gene variants, 13 clinical/epidemiological variables and two different outcomes. Among the clinical/epidemiological features, 12 were categorical (pure nominal, i.e., sex, blood type, and comorbidities; categorical ordinal, i.e., ventilation), while age was the only one computed as a continuous variable. The primary outcome (endpoint 1), that is whether or not patients died of SARS-CoV2 infection (i.e., 0:alive or 1:dead), was a categorical variable (binary); the secondary outcome (endpoint 2), the length of hospital stay before discharge or death, was instead continuous. Patients lacking more than 20% data were excluded to ensure data integrity, resulting in a final sample of 532 patients.

Regarding endpoint 1, the dataset exhibited a significant imbalance between the number of deceased cases (N = 92; 17.3%) and survivals (N = 440; 82.7%), imposing a careful selection of evaluation metrics that focused on the minority class. We prioritized metrics such as precision (the proportion of true positive predictions within all positive classifications) and recall (the modelability to identify all actual positives). Additionally, we employed the f1-score and f2-score, which balance precision and recall. The f1-score is the harmonic mean of precision and recall; it integrates these latter into a single metric gaining

a better understanding of the model performance and offering a balanced measure between these two. Its optimal value is 1.0, indicative of perfect precision and recall, whereas its minimum value is fixed at 0. Conversely, the f2-score is the weighted harmonic mean of precision and recall. Unlike f1-score, which equally weighs precision and recall, f2-score weighs the latter higher than the former by a factor of 2.0. Area Under the Curve (AUC) and PR curves (PR-AUC) were used to quantify the overall efficacy of our classification model across various thresholds. We also applied data weighting techniques during model training. Data weighting involves assigning different weights to the classes, typically giving more importance to the minority class (deceased cases).

Extensive hyperparameter tuning using GridSearchCV was conducted to systematically explore a wide range of configurations to identify the most effective settings based on our primary evaluation metrics: f1-score, f2-score, and PR-AUC. Specifically, 30 configurations for RF, 432 for the GBM, and 72 for XGB were tested. Specifically, the hyperparameters for the RF model were max_depth, max_features, min_sample _leaf, and n_estimators. For the GBM they were learning_rate, max_depth, max_features, n_estimators, and subsample. In the case of XGB, parameters were booster, colsample_bytree, learning_rate, max_depth, subsample, and n_estimators. In Supplementary Tables S1 and S2, the different hyperparameter values and the results for each ML model have been reported. Given the relatively small size of our dataset, a 10-fold cross-validation approach was implemented for all models, rather than dividing the data into separate training and test sets [51].

Within the H2O.ai framework, the feature importance was calculated based on its attributed reduction in squared error (SE) whenever a node is split using that feature and displayed as SHapley Additive exPlanations (SHAP) plot, showing the features ranked on the y-axis according to their importance, while the x-axis reports the SHAP value. This value indicates the magnitude and direction of each feature contribution to the model prediction.

3. Results

3.1. Demographic, Clinical and Genetic Characteristics

Those cases lacking more than 20% of variables were not included in the predictive model assessment, which reduced the sample from 605 to 532 individuals, accounting for about 88% of the total recruited. In the analyzed cohort, males (N = 314; 59.0%) were significantly overrepresented (p-value = 0.00001) and had a significantly younger mean age than females (67 ± 14 ♂versus 74 ± 16 ♀; p-value < 0.0001) across the global age range of 19 to 99 years. Days of hospitalization vary from 1 day to almost 6 months, with females exhibiting shorter mean values (20.7 ± 20.3 ♂versus 18.00 ± 16.7 ♀). Alive male patients had longer hospital stays before discharge (21.16 ± 21.13 ♂versus 18.47 ± 17.46 ♀) despite their significantly younger age (65.42 ± 13.74 ♂versus 71.18 ± 15.59 ♀; p-value < 0.0001). On the other hand, among deceased patients, males and females showed similar lengths of hospital stay before death (17.91 ± 14.94 ♂versus 15.83 ± 21.75 ♀) despite of a significant younger age in males (76.44 ± 12.77 ♂versus 84.83 ± 13.48 ♀; p-value < 0.0005). This reflects the overall mean age of the two patient cohorts in line with male sex and older age being the combination at greatest risk. Globally, males accounted for a cumulative hospitalization of 6442 days versus females, which lengthened to 3843 days. A combination of gene variants, shown in Table 1 , and clinical variables, shown in Table 2 , has been computed in the analyses.

Table 1. Panel of the selected gene variants investigated.

Gene	SNP (rs)	Nucleotide Change	Reported Frequencies		Observed Frequencies		Ref.
			major allele	minor allele	major allele	minor allele	
ACE2	2285666	C>T	0.783	0.216	0.820	0.180	[52,53]
APOE	7412	C>T	0.917	0.083	0.933	0.066	[54]
	429358	T>C	0.925	0.074	0.921	0.078	
TP53	1042522	C>G	0.736	0.263	0.756	0.243	[55]
OAS3	10735079	A>G	0.649	0.351	0.606	0.393	[56]
LZTFL1	35044562	A>G	0.926	0.073	0.878	0.121	[57]
ABO	657152	C>A	0.624	0.376	0.630	0.369	[58]
PPARGC1A	7192678	C>T	0.677	0.323	0.656	0.343	[59]
CRP	2808635	T>G	0.717	0.283	0.710	0.289	[38]
	876538	C>T	0.816	0.184	0.808	0.191	[60]
CFH	1061170	T>C	0.626	0.374	0.657	0.342	[61]
HLA-DRA	3135363	A>G	0.713	0.287	0.642	0.358	[62]
HLA-DPB1	9277356	A>G	0.692	0.307	0.689	0.310	[63]
HLA-A	2499	G>T	0.863	0.136	0.889	0.110	[60]
UCP1	1800592	T>C	0.679	0.320	0.718	0.281	[64]
IL6	2228145	A>C	0.612	0.388	0.634	0.365	[65]
	1800795	G>C	0.639	0.369	0.683	0.316	[66,67]
CCR2	34041956	G>A	0.940	0.059	0.924	0.076	[68]
CFTR	113993960	CTT>-	0.997	0.003	0.991	0.009	[69]
	113993959	G>A	0.9996	0.0004	1.0000	0.0000	
	77010898	G>A	0.9996	0.0004	1.0000	0.0000	
	80034486	C>G	0.9997	0.0003	0.9991	0.0009	
	76713772	G>A	0.9998	0.0002	0.9981	0.0019	
	74597325	C>T	0.9998	0.0002	1.0000	0.0000	
	74767530	C>T	0.9999	0.0001	1.0000	0.0000	
	1801178	A>G	0.9999	0.0001	0.9972	0.0028	
	121908799	AA>G	>0.9999	<0.0001	1.0000	0.0000	

Observed frequencies were calculated for major and minor alleles and the reported ones from the European repository (NCBI).

No statistically significant differences have been obtained for the allele distribution comparing the NCBI-reported and our observed frequencies. All the gene variants considered were under Hardy–Weinberg Equilibrium.

Table 2. Panel of the clinical and epidemiological variables investigated.

	Whole Cohort (N = 532)	Males (N = 314)	Females (N = 218)	*p*-Value
Age (mean ± SD)	69.9 ± 15.3	67.1 ± 14.1	73.9 ± 16.1	**<0.0001**
Days of hospitalization (mean ± SD)	19.6 ± 18.9	20.7 ± 20.3	18.0 ± 16.7	0.078
Death (0)	440 (82.7%)	265 (84.4%)	175 (80.3%)	0.22
Death (1)	92 (17.3%)	49 (15.6%)	43 (19.7%)	

Table 2. *Cont.*

	Whole Cohort (N = 532)	**Males (N = 314)**	**Females (N = 218)**	***p*-Value**
Blood type non-O	307 (57.7%)	186 (59.2%)	121 (55.5%)	0.39
Blood type A	*223 (41.9%)*	*134 (42.7%)*	*89 (40.8%)*	
Blood type B	*52 (9.7%)*	*34 (10.8%)*	*18 (8.3%)*	
Blood type AB	*32 (6.0%)*	*18 (5.7%)*	*14 (6.4%)*	
Blood type O	225 (42.3%)	128 (40.7%)	97 (44.5%)	0.68
Hypertension (0)	220 (41.3%)	137 (43.6%)	83 (38%)	0.20
Hypertension (1)	312 (58.6%)	177 (56.4%)	135 (62%)	
Heart failure (0)	483 (90.8%)	289 (92%)	194 (89%)	0.37
Heart failure (1)	47 (8.8%)	25 (8%)	22 (10.1%)	
n.a.	2 (0.4%)	/	2 (0.9%)	
Ischemic stroke (0)	476 (89.4%)	284 (90.5%)	192 (88.1%)	0.56
Ischemic stroke (1)	54 (10.1%)	30 (9.5%)	24 (11%)	
n.a.	2 (0.4%)	/	2 (0.9%)	
Arteriopathy (0)	484 (90.9%)	281 (89.5%)	203 (93.1%)	0.24
Arteriopathy (1)	32 (6.0%)	22 (7%)	10 (4.6%)	
n.a.	16 (3.0%)	11 (3.5%)	5 (2.3%)	
COPD (0)	486 (91.3%)	286 (91.1%)	200 (91.7%)	0.88
COPD (1)	45 (8.4%)	27 (8.6%)	18 (8.3%)	
n.a.	1 (0.2%)	1 (0.3%)	/	
Hepatopathy (0)	509 (95.7%)	302 (96.2%)	207 (95%)	0.38
Hepatopathy (1)	22 (4.1%)	11 (3.5%)	11 (5%)	
n.a.	1 (0.2%)	1 (0.3%)	/	
Neoplasm (0)	450 (84.6%)	270 (86%)	180 (82.6%)	0.26
Neoplasm (1)	77 (14.5%)	41 (13%)	36 (16.5%)	
n.a.	5 (0.9%)	3 (0.95%)	2 (0.9%)	
Dementia (0)	453 (85.1%)	281 (89.5%)	172 (78.9%)	**0.001**
Dementia (1)	74 (13.9%)	31 (9.9%)	43 (19.7%)	
n.a.	5 (0.9%)	2 (0.6%)	3 (1.4%)	
Diabetes (0)	421 (79.1%)	246 (78.3%)	175 (82.3%)	0.49
Diabetes (1)	108 (20.3%)	67 (21.3%)	41 (18.8%)	
n.a.	3 (0.6%)	1 (0.32%)	2 (0.9%)	
Oxygen (nasal cannula)	154 (29.0%)	81 (25.8%)	73 (33.5%)	**0.05***
Ventilation (NIV + Intubation/ECV)	331 (62.2%)	204 (65.0%)	127 (58.2%)	
Non-invasive ventilation (NIV)	*185 (34.8%)*	*113 (36.0%)*	*72 (33.0%)*	
Intubation/ECV	*146 (27.4%)*	*91 (29.0%)*	*55 (25.2%)*	
n.a.	47 (8.8%)	29 (9.2%)	18 (8.3%)	

SD, standard deviation; n.a., not available; COPD, Chronic Obstructive Pulmonary Disease; NIV, Non-Invasive Ventilation; ECV, ExtraCorporeal Ventilation. *p*-values < 0.05 are reported in bold. * *p*-values = 0.05 by comparing males versus females only considering NIV and intubation/ECV.

The primary endpoint, death due to COVID-19, has been used as the outcome for further analyses, representing the most severe consequence and progression of the infectious disease, globally reaching an occurrence of 17.3% in our dataset. First, we evaluated

all mutual correlations by pairing all variables in a Pearson correlation matrix, as shown in Figure 1.

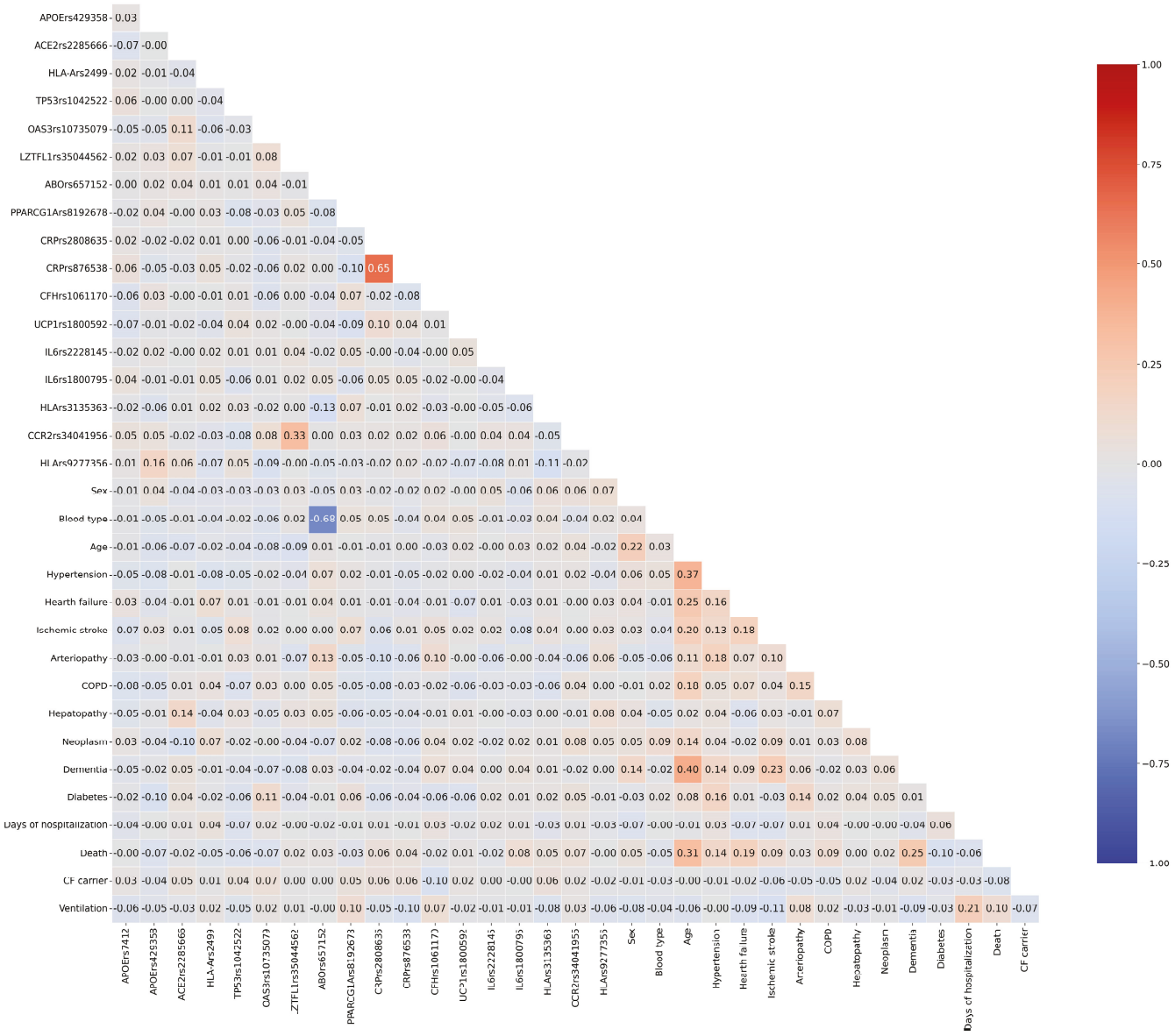


Figure 1. Heatmap of Pearson correlation coefficients among the considered variables. The 34 × 34 diagonal matrix indicates the correlation coefficient (r). The color intensity is proportional to r, with the width of our scale ranging between −0.1 (blue) and 1.00 (red). A cut-off value of 0.5 represents a strong positive correlation. COPD, Chronic Obstructive Pulmonary Disease.

In the Pearson correlation *CRP* rs2808635 and *CRP* rs876538 reached values > 0.5 because of their verified strong Linkage Disequilibrium (D′ = 0.965) in the European population, as well as *ABO* rs657152 with the phenotypic blood group. Interestingly, any feature shows strong positive or negative correlations with the considered outcome, indicating the need to employ a complex model that accounts for all variables and their possible synergistic interactions in risk prediction.

3.2. Prediction of Mortality by PCLR

We performed PCLR as first statistical approach for death prediction risk. PCA was computed with all the 19 gene variables and the 13 clinical features and the first 14 PCs have been retained (eigenvalue > 1.0) explaining approximately more than 60% of the

total variance. By considering death as dependent variable and the 14 PCs as independent variables in a logistic regression model, we found significant positive association with death risk in PC1, PC3, and PC14, and negative association in PC5 and PC9 (Supplementary Table S3). Supplementary Table S4 summarizes the loadings of each PC pointing out the eigenvector values exceeding 0.3. In detail, the loadings that greatly contributed to the PCs were: PC1 (age, dementia, hypertension, ischemic stroke, heart failure, arteriopathy, *CRP rs2808635* , *CRP rs876538*); PC3 (*CRP rs2808635* , *CRP rs876538* , ventilation); PC5 (*ACE2 rs2285666* , hepatopathy, *HLA-A rs2499* , ventilation); PC9 (*CFH rs1061170* , *UCP1 rs1800592* , *APOE rs7412* , CF-carrier, diabetes); PC14 (COPD, *IL6 rs2228145* , ventilation). Despite the logistic regression model achieved a global percentage of correctness of 82.8%, the low event-per-variable ratio (EPV ~2.9), determined considering the number of outcome events (92 deaths) relative to the number of predictors (32 variables), suggests that classical multivariate regression may be constrained, while tree-based ML methods are more suitable.

3.3. Prediction of Hospital Length Stay by PCLR

The same approach was applied for hospital stay prediction among survivals. PCA yielded 13 PCs with an eigenvalue > 1.0 explaining approximately more than 58% of the total variance. By considering the cut-off of 15 days (median value) of hospital stay as dependent variable in a logistic regression model, we found significant positive association in PC4, PC8 and negative association in PC3 and PC6 (Supplementary Table S5). Supplementary Table S6 summarizes the loadings of each PC. The regression model achieved a global percentage of correctness of 66.5% indicating that the model has a limited predictive capability, requiring the employment of alternative approaches.

3.4. Prediction of Mortality by ML Models: Development and Optimization

Table 3 reports the recognized benchmarks for model screening (accuracy and AUROC) and the more ideal metrics (f1, f2, PR-AUC) for the three ML models (GBM, XGB, RF) with a 10-fold cross-validation approach.

Table 3. Metric values for GBM, XGB and RF.

	Accuracy	AUROC	f1	f2	PR-AUC
GBM	0.83 (±0.03)	0.61 (±0.07)	0.63 (±0.09)	**0.61 (±0.08)**	0.62 (±0.09)
XGB	0.83 (±0.03)	**0.63 (±0.04)**	**0.65 (±0.05)**	0.60 (±0.05)	0.62 (±0.07)
RF	**0.85 (±0.04)**	0.61 (±0.06)	0.63 (±0.09)	**0.61 (±0.07)**	**0.68 (±0.12)**

GBM, Gradient Boosting Machine; XGB, eXtreme Gradient Boosting; RF, Random Forest; PR-AUC, Precision Recall-Area Under the Curve. The highest values are reported in bold.

Each metric represents the highest average score obtained when that specific metric was the optimization target during the grid search phase. Although RF achieved the highest overall accuracy (0.85), the model discrimination performance was moderate (AUROC = 0.61) specifically in the less represented class (death), accounting for high false negative count (N = 68). Under comparable PR-AUC values, XGB AUROC indicates its better predictive power under class imbalance. XGB still achieved a competitive accuracy score demonstrating its ability to generalize well across folds. For these reasons, it was selected for model construction. In particular, we were interested in improving the precision and recall values, both accounted for f1, f2, and PR-AUC calculation, to decrease false negative and false positive predictions. Comparing the f1, f2, and PR-AUC in the confusion matrix, the f1 optimization showed the lowest number of false positives (N = 26) and false negatives (N = 63) as shown in Table 4 , as well as being the metric with the highest value for this model (f1 = 0.65 versus f2 = 0.60 and PR-AUC = 0.62).

Table 4. Confusion matrix of f1, f2, and PR-AUC metrics for XGB.

f1	**Predicted Negative (0)**	**Predicted Positive (1)**	**Error**
actually negative (0)	414	26	0.0591
actually positive (1)	63	29	0.3152
total	477	55	0.1673
f2	**predicted negative (0)**	**predicted positive (1)**	**error**
actually negative (0)	413	27	0.0614
actually positive (1)	69	23	0.2500
total	482	50	0.1805
PR-AUC	**predicted negative (0)**	**predicted positive (1)**	**error**
actually negative (0)	411	29	0.0659
actually positive (1)	69	23	0.2500
total	480	52	0.1842

Moreover, f1 optimization was more useful for the healthcare system because monitoring more patients (N_{f1} = 55 versus N_{f2} = 50 and $N_{PR\text{-}AUC}$ = 52) may potentially save most of them, instead of losing a high number of patients due to misdiagnosis.

3.5. Single Analyses for XGB f1 Optimization

Figure 2 shows the ROC curve (a) and the PR curve (b) for XGB f1 optimization. The model achieved a mean AUROC of 0.73 ± 0.08 and a mean PR-AUC of 0.42 ± 0.10, confirming the goodness of fit.

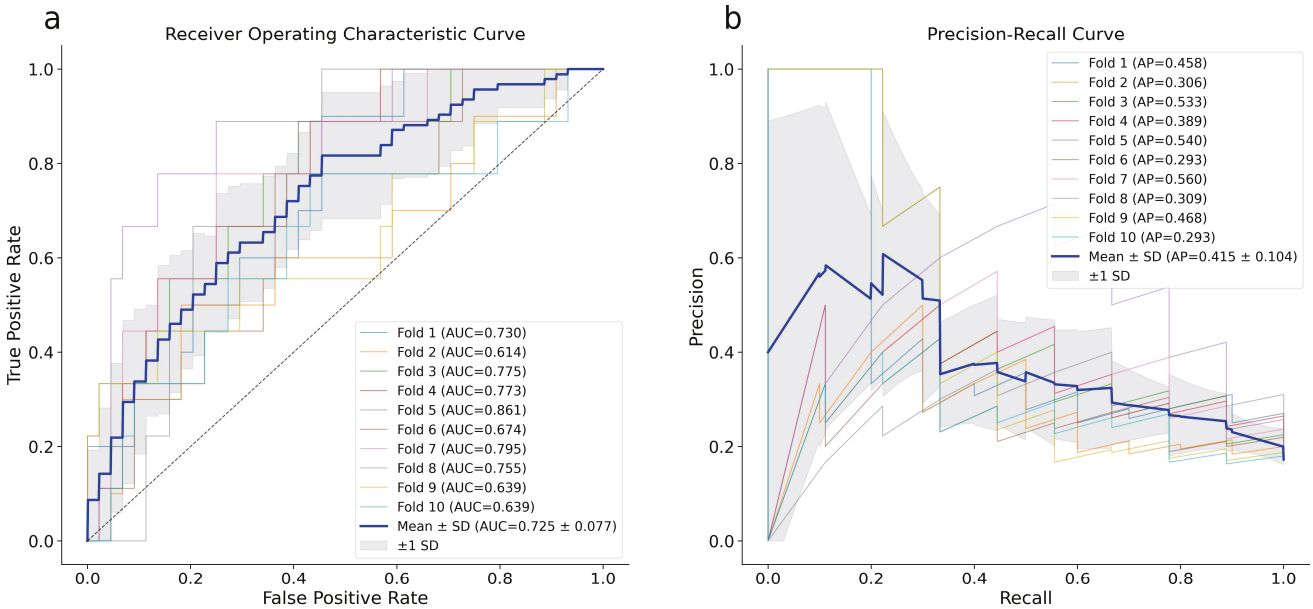


Figure 2. ROC (**a**) and PR (**b**) curves of the XGB model. The plotted blue curves represent the average of 10-fold cross-validation executed utilizing the whole dataset. ROC, Receiver Operating Characteristic; PR, Precision Recall.

We delved into the feature importance results and compared the first ten important variables, with equal f1 optimization, for XGB (Figure 3a), GBM (Figure 3b), and RF (Figure 3c) to understand how and which features (genetics and clinical) contribute most to the mortality prediction.

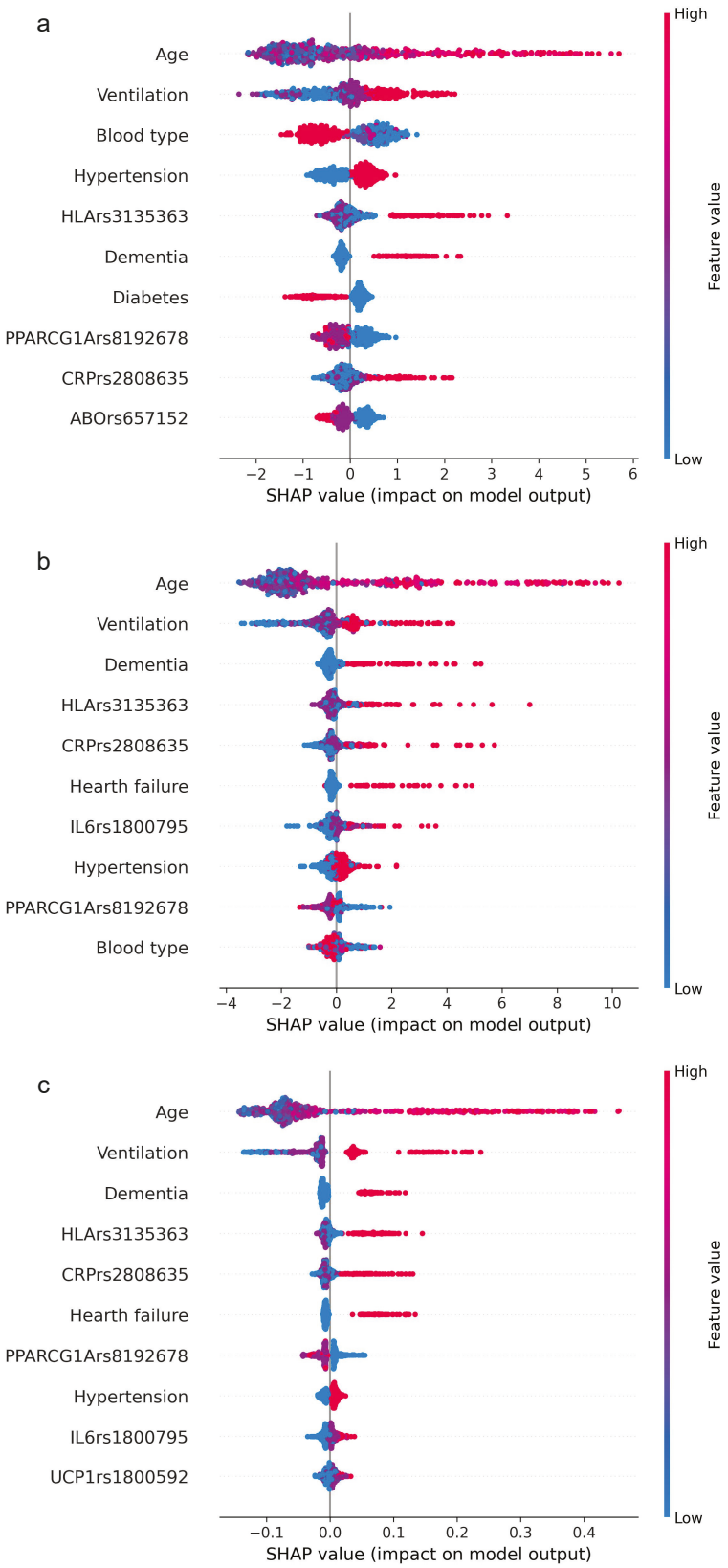


Figure 3. Feature importance ranking based on SHAP values. The figure shows the relative impact of the topmost features contributing to the classification in descending order for XGB (**a**), GBM (**b**), and RF (**c**). Each row (y-axis) represents a different feature, and the horizontal x-axis represents the direction of the relationship between a variable and the outcome coded as SHAP value. As demonstrated by the color bar on the right, higher values are shown in red, while lower values are shown in blue.

From the comparison of the first ten variables, out of 32 computed by the models, it is evident that age and ventilation achieve the greatest values. This means that older age is consistently associated with positive SHAP values and indicates that age is the main contributor to the model decision. For all three models, age was immediately followed by the level of ventilation support at patient admission, ranging from supplemental oxygen (nasal cannula) to invasive ventilation. Although in a different order, approximately the same remaining features emerge, almost equally distributed between clinical and genetic. Interestingly, three gene variants (*HLA-DRA* rs3135363, *PPARGC1A* rs192678, *CRP* rs2808635) were found in all three algorithms. As expected, in the XGB prediction, age had the strongest predictive power, followed by ventilation, blood type, hypertension, *HLA-DPB1* rs9277356, dementia, diabetes, *PPARGC1A* rs192678, *CRP* rs2808635, and *ABO* rs657152.

Finally, to determine the weight of the mortality risk ascribable to those gene variants included among the top ten, crude ORs have been assessed under dominant and recessive inheritance models and allele frequencies for the minor allele (Table 5).

Table 5. OR results for the gene variants in the whole population.

OR (95%CI); *p*-Value	***HLA-DRA* rs3135363 (G)**	***IL6* rs1800795 (C)**	***CRP* *rs2808635* (G)**	***PPARGC1A* *rs7192678 (T)***	***ABO* *rs657152 (A)***	***UCP1* *rs1800592 (C)***
Dominant model	1.00 (0.64–1.58) n.s.	1.53 (0.97–2.42) 0.07	1.12 (0.72–1.75) n.s.	0.75 (0.48–1.18) n.s.	1.21 (0.76–1.94) n.s.	0.97 (0.62–1.52) n.s.
Recessive model	2.00 (1.12–3.57) **0.01**	1.60 (0.82–3.11) n.s.	1.75 (0.93–3.30) 0.08	1.14 (0.58–2.24) n.s.	1.06 (0.56–2.03) n.s.	1.38 (0.64–3.00) n.s.
Allele frequency	1.03 (0.73–1.44) n.s.	1.40 (1.01–1.94) **0.04**	1.24 (0.88–1.74) n.s.	0.88 (0.63–1.24) n.s.	1.12 (0.81–1.55) n.s.	1.04 (0.73–1.48) n.s.

OR, Odds Ratio; CI, Confidence Interval; n.s., not significant; *p*-values < 0.05 are reported in bold. In brackets: the minor allele.

HLA-DRA rs3135363 yields a two-fold risk of death by the recessive model ascribing to the GG homozygous genotype significant values in the whole population (OR = 2.00; 95%CI, 1.12–3.57; *p*-value = 0.01) and in the females the risk was even higher (♀OR = 2.80; 95%CI, 1.22–6.41; *p*-value = 0.015). The significant OR referred to *IL6* rs1800795 (OR = 1.40; 95%CI, 1.01–1.94; *p*-value = 0.04) assigns to the C allele the risk condition, slightly higher in males (♂OR = 1.54; 95%CI, 1.00–2.38; *p*-value = 0.05).

3.6. Prediction of Hospital Length Stay by ML Models: Development and Optimization

The same workflow was applied to predict hospitalization duration among survivors in our dataset, another informative outcome of COVID-19 severity. Using GBM and RF algorithms, we did not find any significant relationships between the dependent and independent variables (R^2 = 0.135 and 0.132, respectively). The complete metrics are shown in Table 6.

Table 6. GBM and RF model metrics for days of hospitalization.

	MSE	**RMSE**	**MAE**	**MAPE**	**R^2**
GBM	321.22	17.31	11.74	0.97	0.135
RF	318.59	17.27	11.67	0.96	0.132

GBM, Gradient Boosting Machine; RF, Random Forest; MSE, Mean Squared Error; RMSE, Root Mean Squared Error; MAE, Mean Average Error; MAPE, Mean Absolute Percentage Error; R2, coefficient of determination.

By combining the two outcomes, assigning (1) to dead patients or patients who had longer than 15 days of hospitalization, RF and XGB models reached an AUC of around 60–65%, which means prediction slightly more than chance, but still not enough to be considered reliable predictive models (Figure 4a,b). This is also an attempt to boost the severity parameter for disease progression, which aids in rebalancing the dataset too.

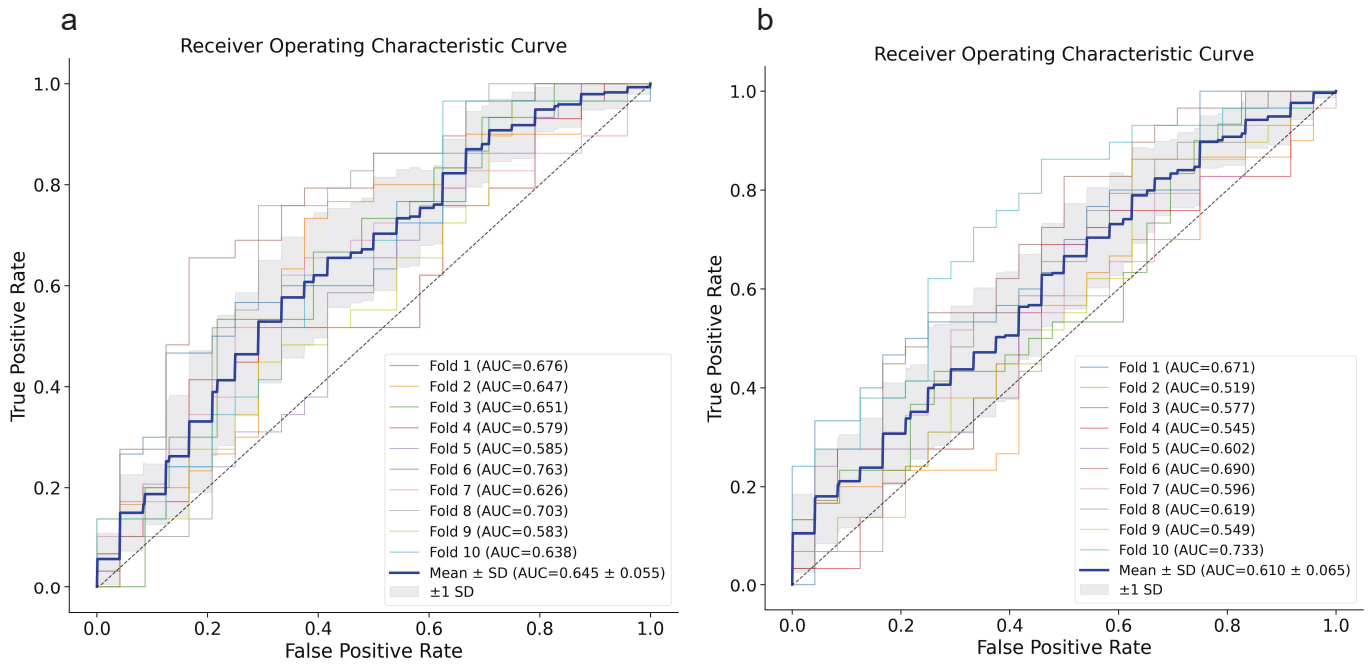


Figure 4. ROC curves for RF (**a**) and XGB (**b**) models. ROC curves for each 10-fold cross-validation are reported with the relative AUC and their mean ± SD (blue line). ROC, Receiver Operating Characteristic; XGB, eXtreme Gradient Boosting; RF, Random Forest.

4. Discussion

The use of big data in healthcare information systems has experienced rapid growth, with customized analyses of individual patient information to identify critical health trends and enable timely preventive care [70]. AI, including ML and big data analytics, has become a promising tool in early diagnosis and prognosis prediction by analyzing diverse datasets to manage complex diseases at single or population levels. The emergence of ever-evolving infectious diseases, such as the recent COVID-19 pandemic, underscores the pressing need for innovative tools to curb their dramatic health and economic impact [71].

The choice to use the three selected ML models was intentional and based on the characteristics of our dataset and the clinical goal of the study. Our aim was not to test all available algorithms, but to select reliable models suitable for this specific scenario. GBM, XGB and RF are well-established tree-based methods that are widely used in clinical prediction [72]. They perform well with a limited sample size, heterogeneous variables and class imbalance, as in our cohort. In addition, these models provide a good balance between predictive performance and interpretability, which is important in a clinical context where understanding the role of each feature is required. Complex models, such as deep learning approaches, usually need much larger datasets and may reduce model transparency without clear performance advantages in this setting [73]. Importantly, the main objective of this study was not algorithm benchmarking, but to assess the added value of combining genetic and clinical data within robust and clinically applicable machine learning models.

The main result of our investigation is that all the displayed mortality models achieve high values, in particular the XGB showed a better predictive power under class imbalance and a high discrimination performance specifically in the less represented class (e.g., death). XGB f1 optimization is not only the best option from a technical point of view, but also the optimization that best serves the primary purpose of the healthcare system. Specifically, the f1 metric focuses on patient quality of care, accounting the lowest number of false negative

and false positive, monitoring a wide number of patients, instead of monitoring a limited number of patients that could be misdiagnosed, only addressing the economical side.

Feature importance was further analyzed for XGB, GMB and RF with f1 optimization to verify the contribution of each feature to model prediction. From the comparison of the first ten important variables, it is evident that most overlapped in all three models, and that the genetic component integrates with and supports the model predictive power, along with the clinical/epidemiological component. It should be noted that the features are selected according to the standard error, a measure of how much they improve or worsen the algorithm classification when considered, even if this does not directly correlate with the risk of death. Even though, from a clinical perspective, the presence of comorbidities, need for ventilation and older age have a greater impact on mortality risk, the genetic component does not confound the prediction; rather, it could aid in identifying peculiar cases within the vast clinical heterogeneity, enabling a more personalized approach. Since it is a complex model applied to a complex and multifactorial pathology, the feature importance results also reflect the interactions among variables. Accordingly, the ML model allowed us to screen all variables to identify the most important ones, which would require further individual analysis for a more specific assessment of their involvement in disease progression.

Regarding the top ten variables identified by the feature importance ranking of our selected model, we discuss below their implication and roles in COVID-19. As expected, aging is a major driver of mortality. In our cohort, the mean age of dead patients was about 13 years higher than survivors (80.35 y versus 67.70 y), confirming the highest correlation coefficient (0.31). Scientific evidence suggests that people older than 65 years have a greater risk of developing severe forms of COVID-19 [74]. This is mainly attributed to a weaker immune system and the presence of past or concomitant age-related diseases, leading to persistent low-grade innate immune activation [75]. The need for ventilation and the required ventilator support at hospital admission resulted in the second driver in our model decision, representing a significant risk factor for COVID-19 mortality, reflecting the severity score of respiratory compromise [76]. Patients requiring immediate mechanical ventilation, both invasive and non-invasive, are typically those with impaired gas exchange and advanced lung injury, the leading cause of death from COVID-19 [77]. Hypertension has also been widely reported as a significant risk factor for severe COVID-19 outcomes [78]. Pathophysiologically, chronic hypertension is characterized by endothelial dysfunction and vascular inflammation, conditions that can be exacerbated by SARS-CoV-2 infection, leading to multi-organ injury [79]. Furthermore, hypertensive patients present dysregulation of the renin–angiotensin–aldosterone system (RAAS), and SARS-CoV-2-mediated downregulation of ACE2 receptor may further enhance the RAAS imbalance, worsening patient conditions [80]. On the other hand, diabetes is characterized by a pathological increase in ACE2 expression, which may favor virus entry and trigger comorbidities associated with severe COVID-19 outcomes [81]. Elevated glucose levels also affect SARS-CoV-2 replication, innate and adaptive immunity, and favor hypercoagulability conditions [81]. Virus entrance is additionally influenced by blood type through the interaction of the virus with the ACE2 receptor [82,83]. In detail, blood type O appears protective against disease progression compared with non-O types [84]. Interestingly, the ABO(H)-like antigens, carried by SARS-CoV-2 envelope, may stimulate the immune response of natural ABO antibodies. This may give to type O-carriers an advantage due to the presence of both anti-A and anti-B antibodies [85,86] , able to bind the Spike protein and block ACE2 entry. Interestingly, a recent report on the mechanisms by which the *CFTR* gene may influence SARS-CoV-2 infection and COVID-19 disease severity found no association between FC-carriers and more severe clinical outcomes [87]. Authors found an unexpected trend toward

higher mortality among non-carrier controls compared with silent carriers stating that further studies are required. Accordingly, we detected FC-carriers exclusively among survivors (3.7%) and this subgroup, though significantly older, did not have more severe lung parameters when compared to non-carriers speculating a protective action.

As regards dementia, the literature is quite controversial on the association of dementia with enhanced risk of severe COVID-19 for several reasons. Being an age-related condition, affected patients are older, and the chance of suffering from other comorbidities is higher [88] , but dementia itself remains an independent predictor of poor prognosis.

Basically, an effective immune response against viral infections is crucial for host defense; hence, HLA class II molecules present antigenic peptides to activate CD4+ T helper cells, driving cell-mediated immunity [62]. Concerning genetic data, HLA class II gene variants (e.g., rs3135363; rs9277356) have been previously shown to impact the immune response and susceptibility to disease severity by modulating the inflammation process [60,63,89,90]. Specifically, the rs3135363 variant has been linked to a more robust antibody response to hepatitis B vaccination [62] , suggesting that it may influence susceptibility to infectious diseases by affecting antigen presentation and immune regulation within the MHC region. We proved that the GG genotype at rs3135363 was associated with death, with a significant 2-fold increased risk in the whole group and an almost 3-fold in the female subgroup. Our group already demonstrated different vaccine induced IgG levels in serum of patients according to *HLA-A* gene variants [38]. Severe forms of COVID-19 are mainly precipitated or triggered by an uncontrolled inflammatory response that may lead to cytokine storm [91]. C-reactive protein also plays a key role in inflammation by acting as an acute-phase reactant [92]. Elevated CRP levels are associated with a wide range of pathological conditions, including bacterial and viral infections. Its levels rise in response to inflammatory cytokines [93] and circulating levels are associated with *CRP* rs2808635. These in turn correlate with anti-SARS-CoV-2 IgG levels in COVID-19 patients, suggesting the persistence of an unresolved inflammation status [38,94]. Additionally, *PPARGC1A* rs8192678 results in Gly482Ser amino acid substitution in the PPARG coactivator-1α, a key protein of metabolic pathways. In particular, the T allele (Ser482) is associated with reduced expression and activity of the protein, leading to impaired co-activation of PPARs (Peroxisome Proliferator Activated Receptors) [95]. Given that PPARs play an anti-inflammatory role, and that SARS-CoV-2 infection downregulates PPARG coactivator-1α, the anti-inflammatory capacity is further reduced in the variant carriers. This may contribute to an unrestrained, potentially dangerous inflammatory response during COVID-19 [59]. Finally, *ABO* rs657152, which emerged from the feature importance, has been extensively studied for its association with COVID-19 severity. GWAS have identified rs657152 as a genetic risk factor for severe COVID-19, with the A allele linked to an increased risk of respiratory failure and mortality in some populations [96]. Functionally, the SNP has been linked to changes in the plasma levels of multiple proteins, including those involved in coagulation and immune response pathways. This is crucial because COVID-19 lethality is mainly due to hypercoagulability, thrombotic complications and excessive inflammation [97,98].

Although the limitations of this study include a small sample size, we previously evaluated data adequacy using multiple complementary criteria (e.g., EPV ratio and the calibration analysis shown in Section 3). The model robustness was further supported by the stability of performance across 10-fold cross validation, approach chosen instead of dividing the dataset into training and test set. Indeed, the need for external validation of the model is mandatory to confirm its performance in independent cohorts. This should address an infectious disease by using a multilayer approach and applying ML strategies to other unresolved or future complex diseases. The decision to compare different algorithms

and optimization methods stems from the need to have several efficient models to recover in other scenarios of infectious diseases. For example, the Marburg virus disease (MVD), the recent, widespread human respiratory syncytial virus (RSV) [71,99] , or the complex health emergency ('Disease-X') spread in December 2024 in a remote region of the Democratic Republic of Congo [100].

5. Conclusions

To build preparedness against pandemics and support the global response to widespread infections, it is important to implement intelligence-based models that incorporate genetic predispositions to predict which patient may be at higher risk of disease progression [101]. Timely and accurate identification of poor outcomes can guide physicians in prioritizing patients, allocating limited medical and economic resources more effectively, and selecting appropriate treatments. ML algorithms have revolutionized the field of healthcare by providing unprecedented opportunities to develop highly accurate models for risk analysis. This enables the precise identification of druggable molecular pathways and enhances predictive capabilities to address the myriad challenges modern healthcare systems face. These not only facilitate early diagnosis and personalized treatment plans but also optimize resource allocation and improve patient outcomes [102] , reducing uncertainty and ambiguity. One practical example is the respiratory failure, recommended for assisted ventilation, a scarce resource during the pandemic phase [103]. We can speculate that our model could prioritize the resource allocation to patients not only via clinical criteria alone. Finally, a critical challenge in predictive modeling is selecting the model that best balances data fit and clinical relevance. This highlights the importance of choosing the perfect algorithm and the best optimization avoiding false negative bad outcome predictions.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics16040583/s1 , Table S1: Hyperparameter values explored during grid search; Table S2: Optimal hyperparameter configurations identified through grid search for GBM, XGB and RF under different optimization metrics; Table S3: PCLR analysis; Table S4: Loadings of PCs; Supplementary Table S5: PCLR analysis; Table S6: Loadings of PCs.

Author Contributions: Conceptualization, E.D., B.A., V.T. and D.G.; Methodology, E.D., B.A., S.P., V.T. and D.G.; Software, S.P.; Formal analysis, E.D., B.A. and M.G.; Investigation, E.D., M.G., R.A., E.M.P., A.P., S.V., F.R., M.C., A.M.M., A.C., R.D., F.S. and A.V.S.; Resources, R.A., E.M.P., A.P., S.V., F.R., M.C., A.M.M., A.C., R.D., V.T. and D.G.; Writing—original draft, E.D., B.A., V.T. and D.G.; Writing—review and editing, E.D., B.A., M.G., F.S., A.V.S., S.P., V.T. and D.G.; Visualization, E.D. and B.A. Supervision, R.A., A.P., S.V., A.M.M., V.T. and D.G.; Project administration, V.T. and D.G.; Founding acquisition, V.T. and D.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by COVID-19 grant from the University of Ferrara; FAR and FIRD local grants from the University of Ferrara (to D.G. and V.T.).

Institutional Review Board Statement: The study has been conducted in accordance with the Declaration of Helsinki and approved by local Ethics Committee (IRB) of University Hospital of Ferrara, CE:405/2020, 8 April 2020; IRCCS Humanitas Clinical and Research Center of Milan, CE:316/2020, 16 March 2020; Federico II University Hospital of Naples, CE:98/2021, 15 February 2021.

Informed Consent Statement: Participants provided written informed consent to participate in the study.

Data Availability Statement: The data supporting this study's findings are available from the corresponding authors upon reasonable request, pending internal agreement among the participating centers.

Acknowledgments: The Authors thank the National Recovery and Resilience Plan (NRRP), Mission 04 Component 2 Investment 1.4—NextGenerationEU, National Center HPC, Big Data and Quantum Computing; National Recovery and Resilience Plan (NRRP), Mission 04 Component 2 Investment 1.5—NextGenerationEU, Ecosystem for Sustainable Transition in Emilia-Romagna. This project is partner of the COVID-19 Host Genetics Initiative (HGI) (https://www.covid19hg.org/).

Conflicts of Interest: S.P. is affiliated with and is the shareholder of DESTINA Genomica S.L. The company had no role in the design of the study, in the collection, analysis or interpretation of data, in the writing of the manuscript or in the decision to publish the results. No patents, products in development or marketed products are associated with this work.

Abbreviations

The following abbreviations are used in this manuscript:

ABO	Alpha 1-3-N-acetylgalactosaminyltransferase and alpha 1-3-galactosyltransferase
ACE2	Angiotensin converting enzyme 2
AI	Artificial Intelligence
APOE	Apolipoprotein E
AUC	Area Under the Curve
CCR2	C-C motif Chemokine Receptor 2
CFH	Complement Factor H
CFTR	Cystic Fibrosis Transmembrane Conductance Regulator
CI	Confidence Interval
COPD	Chronic Obstructive Pulmonary Disease
COVID-19	Coronavirus Disease 19
CRP	C-Reactive Protein
ECV	ExtraCorporeal Ventilation
GBM	Gradient Boosting Machine
GWAS	Genome-Wide Association Study
HGI	Host Genetics Initiative
HLA	Human Leukocyte Antigen
ICU	Intensive Care Unit
IL6	Interleukin6
IRB	Institutional Review Board
LZTFL1	Leucine zipper transcription factor-like 1
MAE	Mean Average Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MSE	Mean Squared Error
MVD	Marburg Virus Disease
NCBI	National Center for Biotechnology Information
NIV	Non-Invasive Ventilation
OAS3	3′,5′-Oligoadenylate Synthase 3
OR	Odds Ratio
PCA	Principal Component Analysis
PCLR	Principal Component Logistic Regression
PPARGC1A	PPARG coactivator 1 alpha
PPARs	Peroxisome Proliferator Activated Receptors
PR	Precision Recall
RAAS	Renin–Angiotensin–Aldosterone System
RF	Random Forest
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SE	Standard Error

SHAP	SHapley Additive exPlanations
SNP	Single Nucleotide Polymorphism
SPSS	Statistical Package for the Social Sciences
TP53	Tumor Protein p53
UCP1	Uncoupling Protein 1
XGB	eXtreme Gradient Boosting

References

1. Pal, M.; Berhanu, G.; Desalegn, C.; Kandi, V. Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2): An Update. *Cureus* **2020** , *12* , e7423. [CrossRef]
2. Chen, B.; Tian, E.-K.; He, B.; Tian, L.; Han, R.; Wang, S.; Xiang, Q.; Zhang, S.; El Arnaout, T.; Cheng, W. Overview of Lethal Human Coronaviruses. *Signal Transduct. Target. Ther.* **2020** , *5* , 89. [CrossRef]
3. Bonaccorsi, G.; Gambacciani, M.; Gemmati, D. Can Estrogens Protect against COVID-19? The COVID-19 Puzzling and Gender Medicine. *Minerva Ginecol.* **2020** , *72* , 178–179. [CrossRef]
4. Zhang, J.-J.; Dong, X.; Liu, G.-H.; Gao, Y.-D. Risk and Protective Factors for COVID-19 Morbidity, Severity, and Mortality. *Clin. Rev. Allergy Immunol.* **2023** , *64* , 90–107. [CrossRef]
5. Cappadona, C.; Rimoldi, V.; Paraboschi, E.M.; Asselta, R. Genetic Susceptibility to Severe COVID-19. *Infect. Genet. Evol.* **2023** , *110* , 105426. [CrossRef]
6. Camacho Moll, M.E.; Mata Tijerina, V.L.; Silva Ramírez, B.; Peñuelas Urquides, K.; González Escalante, L.A.; Escobedo Guajardo, B.L.; Cruz Luna, J.E.; Corrales Pérez, R.; Gómez García, S.; Bermúdez de León, M. Sex, Age, and Comorbidities Are Associated with SARS-CoV-2 Infection, COVID-19 Severity, and Fatal Outcome in a Mexican Population: A Retrospective Multi-Hospital Study. *J. Clin. Med.* **2023** , *12* , 2676. [CrossRef]
7. Ferreira, L.C.; Gomes, C.E.M.; Rodrigues-Neto, J.F.; Jeronimo, S.M.B. Genome-Wide Association Studies of COVID-19: Connecting the Dots. *Infect. Genet. Evol.* **2022** , *106* , 105379. [CrossRef]
8. Guan, W.; Ni, Z.; Hu, Y.; Liang, W.; Ou, C.; He, J.; Liu, L.; Shan, H.; Lei, C.; Hui, D.S.C.; et al. Clinical Characteristics of Coronavirus Disease 2019 in China. *N. Engl. J. Med.* **2020** , *382* , 1708–1720. [CrossRef]
9. Ko, Y.; Ngai, Z.-N.; Koh, R.-Y.; Chye, S.-M. Association among Lifestyle and Risk Factors with SARS-CoV-2 Infection. *Tuberc. Respir. Dis.* **2023** , *86* , 102–110. [CrossRef]
10. COVID-19 Host Genetics Initiative. Mapping the Human Genetic Architecture of COVID-19. *Nature* **2021** , *600* , 472–477. [CrossRef]
11. Lusczek, E.R.; Ingraham, N.E.; Karam, B.S.; Proper, J.; Siegel, L.; Helgeson, E.S.; Lotfi-Emran, S.; Zolfaghari, E.J.; Jones, E.; Usher, M.G.; et al. Characterizing COVID-19 Clinical Phenotypes and Associated Comorbidities and Complication Profiles. *PLoS ONE* **2021** , *16* , e0248956. [CrossRef]
12. COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a Global Initiative to Elucidate the Role of Host Genetic Factors in Susceptibility and Severity of the SARS-CoV-2 Virus Pandemic. *Eur. J. Hum. Genet.* **2020** , *28* , 715–718. [CrossRef]
13. COVID-19 Host Genetics Initiative. A First Update on Mapping the Human Genetic Architecture of COVID-19. *Nature* **2022** , *608* , E1–E10. [CrossRef]
14. COVID-19 Host Genetics Initiative. A Second Update on Mapping the Human Genetic Architecture of COVID-19. *Nature* **2023** , *621* , E7–E26. [CrossRef]
15. Severe Covid-19 GWAS Group; Ellinghaus, D.; Degenhardt, F.; Bujanda, L.; Buti, M.; Albillos, A.; Invernizzi, P.; Fernández, J.; Prati, D.; Baselli, G.; et al. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *N. Engl. J. Med.* **2020** , *383* , 1522–1534. [CrossRef]
16. Cascella, M.; Rajnik, M.; Aleem, A.; Dulebohn, S.C.; Di Napoli, R. Features, Evaluation, and Treatment of Coronavirus (COVID-19). In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2024.
17. Haileamlak, A. The Impact of COVID-19 on Health and Health Systems. *Ethiop. J. Health Sci.* **2021** , *31* , 1073–1074. [CrossRef]
18. WHO COVID-19 Dashboard. COVID-19 Cases. Available online: https://data.who.int/dashboards/covid19/cases?n=c_13/12/24 (accessed on 17 December 2024).
19. Foglia, E.; Ferrario, L.; Schettini, F.; Pagani, M.B.; Dalla Bona, M.; Porazzi, E. COVID-19 and Hospital Management Costs: The Italian Experience. *BMC Health Serv. Res.* **2022** , *22* , 991. [CrossRef]
20. Oțelea, M.R.; Rașcu, A.; Staicu, C.; Călugăreanu, L.; Ipate, M.; Teodorescu, S.; Persecă , O.; Voinoiu, A.; Neamțu, A.; Calotă , V.; et al. Exhaustion in Healthcare Workers after the First Three Waves of the COVID-19 Pandemic. *Int. J. Environ. Res. Public Health* **2022** , *19* , 8871. [CrossRef]

21. Gupta, N.; Dhamija, S.; Patil, J.; Chaudhari, B. Impact of COVID-19 Pandemic on Healthcare Workers. *Ind. Psychiatry J.* **2021** , *30* , S282–S284. [CrossRef]
22. Chen, R.; Chen, J.; Yang, S.; Luo, S.; Xiao, Z.; Lu, L.; Liang, B.; Liu, S.; Shi, H.; Xu, J. Prediction of Prognosis in COVID-19 Patients Using Machine Learning: A Systematic Review and Meta-Analysis. *Int. J. Med. Inform.* **2023** , *177* , 105151. [CrossRef]
23. He, F.; Page, J.H.; Weinberg, K.R.; Mishra, A. The Development and Validation of Simplified Machine Learning Algorithms to Predict Prognosis of Hospitalized Patients With COVID-19: Multicenter, Retrospective Study. *J. Med. Internet Res.* **2022** , *24* , e31549. [CrossRef]
24. Moulaei, K.; Shanbehzadeh, M.; Mohammadi-Taghiabad, Z.; Kazemi-Arpanahi, H. Comparing Machine Learning Algorithms for Predicting COVID-19 Mortality. *BMC Med. Inform. Decis. Mak.* **2022** , *22* , 2. [CrossRef]
25. Dite, G.S.; Murphy, N.M.; Spaeth, E.; Allman, R. Lifelines Corona Research Initiative Validation of a Clinical and Genetic Model for Predicting Severe COVID-19. *Epidemiol. Infect.* **2022** , *150* , e91. [CrossRef]
26. Nopour, R.; Shanbehzadeh, M.; Kazemi-Arpanahi, H. Using Logistic Regression to Develop a Diagnostic Model for COVID-19: A Single-Center Study. *J. Educ. Health Promot.* **2022** , *11* , 153. [CrossRef]
27. Doyle, R. Machine Learning-Based Prediction of COVID-19 Mortality With Limited Attributes to Expedite Patient Prognosis and Triage: Retrospective Observational Study. *JMIRx Med.* **2021** , *2* , e29392. [CrossRef]
28. Tang, C.Y.; Gao, C.; Prasai, K.; Li, T.; Dash, S.; McElroy, J.A.; Hang, J.; Wan, X.-F. Prediction Models for COVID-19 Disease Outcomes. *Emerg. Microbes Infect.* **2024** , *13* , 2361791. [CrossRef] [PubMed]
29. Zhu, X.; Xie, B.; Chen, Y.; Zeng, H.; Hu, J. Machine Learning in the Prediction of In-Hospital Mortality in Patients with First Acute Myocardial Infarction. *Clin. Chim. Acta* **2024** , *554* , 117776. [CrossRef]
30. Sagar, D.; Dwivedi, T.; Gupta, A.; Aggarwal, P.; Bhatnagar, S.; Mohan, A.; Kaur, P.; Gupta, R. Clinical Features Predicting COVID-19 Severity Risk at the Time of Hospitalization. *Cureus* **2024** , *16* , e57336. [CrossRef] [PubMed]
31. Zakariaee, S.S.; Naderi, N.; Ebrahimi, M.; Kazemi-Arpanahi, H. Comparing Machine Learning Algorithms to Predict COVID-19 Mortality Using a Dataset Including Chest Computed Tomography Severity Score Data. *Sci. Rep.* **2023** , *13* , 11343. [CrossRef] [PubMed]
32. Kourmpanis, N.; Liaskos, J.; Zoulias, E.; Mantas, J. Predicting Mortality in COVID-19 Patients Using 6 Machine Learning Algorithms. *Stud. Health Technol. Inf.* **2023** , *305* , 115–118. [CrossRef]
33. Kumari, S.; Tripathy, S.; Nayak, S.; Rajasimman, A.S. Machine Learning-Aided Algorithm Design for Prediction of Severity from Clinical, Demographic, Biochemical and Immunological Parameters: Our COVID-19 Experience from the Pandemic. *J. Fam. Med. Prim. Care* **2024** , *13* , 1937–1943. [CrossRef]
34. Gemmati, D.; Bramanti, B.; Serino, M.L.; Secchiero, P.; Zauli, G.; Tisato, V. COVID-19 and Individual Genetic Susceptibility/Receptivity: Role of ACE1/ACE2 Genes, Immunity, Inflammation and Coagulation. Might the Double X-Chromosome in Females Be Protective against SARS-CoV-2 Compared to the Single X-Chromosome in Males? *Int. J. Mol. Sci.* **2020** , *21* , 3474. [CrossRef]
35. Gemmati, D.; Tisato, V. Genetic Hypothesis and Pharmacogenetics Side of Renin-Angiotensin-System in COVID-19. *Genes* **2020** , *11* , 1044. [CrossRef]
36. Milani, D.; Caruso, L.; Zauli, E.; Al Owaifeer, A.M.; Secchiero, P.; Zauli, G.; Gemmati, D.; Tisato, V. P53/NF-kB Balance in SARS-CoV-2 Infection: From OMICs, Genomics and Pharmacogenomics Insights to Tailored Therapeutic Perspectives (COVIDomics). *Front. Pharmacol.* **2022** , *13* , 871583. [CrossRef] [PubMed]
37. Singh, A.V.; Chandrasekar, V.; Paudel, N.; Laux, P.; Luch, A.; Gemmati, D.; Tisato, V.; Prabhu, K.S.; Uddin, S.; Dakua, S.P. Integrative Toxicogenomics: Advancing Precision Medicine and Toxicology through Artificial Intelligence and OMICs Technology. *Biomed. Pharmacother.* **2023** , *163* , 114784. [CrossRef]
38. Gemmati, D.; Longo, G.; Gallo, I.; Silva, J.A.; Secchiero, P.; Zauli, G.; Hanau, S.; Passaro, A.; Pellegatti, P.; Pizzicotti, S.; et al. Host Genetics Impact on SARS-CoV-2 Vaccine-Induced Immunoglobulin Levels and Dynamics: The Role of TP53, ABO, APOE, ACE2, HLA-A, and CRP Genes. *Front. Genet.* **2022** , *13* , 1028081. [CrossRef] [PubMed]
39. Barghash, R.F.; Gemmati, D.; Awad, A.M.; Elbakry, M.M.M.; Tisato, V.; Awad, K.; Singh, A.V. Navigating the COVID-19 Therapeutic Landscape: Unveiling Novel Perspectives on FDA-Approved Medications, Vaccination Targets, and Emerging Novel Strategies. *Molecules* **2024** , *29* , 5564. [CrossRef]
40. Tiwari Pandey, A.; Pandey, I.; Zamboni, P.; Gemmati, D.; Kanase, A.; Singh, A.V.; Singh, M.P. Traditional Herbal Remedies with a Multifunctional Therapeutic Approach as an Implication in COVID-19 Associated Co-Infections. *Coatings* **2020** , *10* , 761. [CrossRef]
41. Sanità , E.-I.S. Di EpiCentro. Available online: https://www.epicentro.iss.it/ (accessed on 7 January 2026).
42. Marchetti, G.; Gemmati, D.; Patracchini, P.; Pinotti, M.; Bernardi, F. PCR Detection of a Repeat Polymorphism within the F7 Gene. *Nucleic Acids Res.* **1991** , *19* , 4570. [CrossRef] [PubMed]
43. R Core Team. R: The R Project for Statistical Computing. Available online: https://www.r-project.org/ (accessed on 7 January 2026).

44. IBM. Solutions. Available online: https://www.ibm.com/solutions (accessed on 7 January 2026).
45. Schoonjans, F. MedCalc Statistical Software—Free Trial. Available online: https://www.medcalc.org/en/ (accessed on 7 January 2026).
46. Breiman, L. Random Forests. *Mach. Learn.* **2001** , *45* , 5–32. [CrossRef]
47. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely Randomized Trees. *Mach. Learn.* **2006** , *63* , 3–42. [CrossRef]
48. Semantic Scholar. Greedy Function Approximation: A Gradient Boosting Machine. Available online: https://www.semanticscholar.org/paper/Greedy-function-approximation:-A-gradient-boosting-Friedman/1679beddda3a183714d380e944fe6bf586c083cd (accessed on 9 December 2025).
49. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2016; pp. 785–794.
50. H2oai/H2o-3. Available online: https://github.com/h2oai/h2o-3 (accessed on 7 January 2026).
51. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*; Scientific Research Publishing: Wuhan, China, 1995; Volume 2, pp. 1137–1145. Available online: https://www.scirp.org/reference/referencespapers?referenceid=1957788 (accessed on 9 December 2025).
52. Asselta, R.; Paraboschi, E.M.; Mantovani, A.; Duga, S. ACE2 and TMPRSS2 Variants and Expression as Candidates to Sex and Country Differences in COVID-19 Severity in Italy. *Aging* **2020** , *12* , 10087–10098. [CrossRef]
53. Saengsiwaritt, W.; Jittikoon, J.; Chaikledkaew, U.; Udomsinprasert, W. Genetic Polymorphisms of ACE1, ACE2, and TMPRSS2 Associated with COVID-19 Severity: A Systematic Review with Meta-Analysis. *Rev. Med. Virol.* **2022** , *32* , e2323. [CrossRef]
54. Chen, F.; Chen, Y.; Ke, Q.; Wang, Y.; Gong, Z.; Chen, X.; Cai, Y.; Li, S.; Sun, Y.; Peng, X.; et al. ApoE4 Associated with Severe COVID-19 Outcomes via Downregulation of ACE2 and Imbalanced RAS Pathway. *J. Transl. Med.* **2023** , *21* , 103. [CrossRef]
55. Zauli, G.; Tisato, V.; Secchiero, P. Rationale for Considering Oral Idasanutlin as a Therapeutic Option for COVID-19 Patients. *Front. Pharmacol.* **2020** , *11* , 1156. [CrossRef] [PubMed]
56. Pairo-Castineira, E.; Clohisey, S.; Klaric, L.; Bretherick, A.D.; Rawlik, K.; Pasko, D.; Walker, S.; Parkinson, N.; Fourman, M.H.; Russell, C.D.; et al. Genetic Mechanisms of Critical Illness in COVID-19. *Nature* **2021** , *591* , 92–98. [CrossRef]
57. Breno, M.; Noris, M.; Rubis, N.; Parvanova, A.I.; Martinetti, D.; Gamba, S.; Liguori, L.; Mele, C.; Piras, R.; Orisio, S.; et al. A GWAS in the Pandemic Epicenter Highlights the Severe COVID-19 Risk Locus Introgressed by Neanderthals. *iScience* **2023** , *26* , 107629. [CrossRef] [PubMed]
58. Ellinghaus, D. COVID-19 Host Genetics and ABO Blood Group Susceptibility. *Camb. Prism. Precis. Med.* **2023** , *1* , e10. [CrossRef] [PubMed]
59. Hasankhani, A.; Bahrami, A.; Tavakoli-Far, B.; Iranshahi, S.; Ghaemi, F.; Akbarizadeh, M.R.; Amin, A.H.; Abedi Kiasari, B.; Mohammadzadeh Shabestari, A. The Role of Peroxisome Proliferator-Activated Receptors in the Modulation of Hyperinflammation Induced by SARS-CoV-2 Infection: A Perspective for COVID-19 Therapy. *Front. Immunol.* **2023** , *14* , 1127358. [CrossRef]
60. Wolday, D.; Fung, C.Y.J.; Morgan, G.; Casalino, S.; Frangione, E.; Taher, J.; Lerner-Ellis, J.P. HLA Variation and SARS-CoV-2 Specific Antibody Response. *Viruses* **2023** , *15* , 906. [CrossRef]
61. Gavriilaki, E.; Tsiftsoglou, S.A.; Touloumenidou, T.; Farmaki, E.; Panagopoulou, P.; Michailidou, E.; Koravou, E.-E.; Mavrikou, I.; Iosifidis, E.; Tsiatsiou, O.; et al. Targeted Genotyping of MIS-C Patients Reveals a Potential Alternative Pathway Mediated Complement Dysregulation during COVID-19 Infection. *Curr. Issues Mol. Biol.* **2022** , *44* , 2811–2824. [CrossRef]
62. Zhao, Y.; Chen, K.; Yang, H.; Zhang, F.; Ding, L.; Liu, Y.; Zhang, L.; Zhang, Y.; Wang, H.; Deng, Y. HLA-DR Genetic Polymorphisms and Hepatitis B Virus Mutations Affect the Risk of Hepatocellular Carcinoma in Han Chinese Population. *Virol. J.* **2023** , *20* , 283. [CrossRef]
63. Bolze, A.; Neveux, I.; Schiabor Barrett, K.M.; White, S.; Isaksson, M.; Dabe, S.; Lee, W.; Grzymski, J.J.; Washington, N.L.; Cirulli, E.T. HLA-A∗03:01 Is Associated with Increased Risk of Fever, Chills, and Stronger Side Effects from Pfizer-BioNTech COVID-19 Vaccination. *Hum. Genet. Genom. Adv.* **2022** , *3* , 100084. [CrossRef]
64. Liu, X.; Jiang, Z.; Zhang, G.; Ng, T.K.; Wu, Z. Association of UCP1 and UCP2 Variants with Diabetic Retinopathy Susceptibility in Type-2 Diabetes Mellitus Patients: A Meta-Analysis. *BMC Ophthalmol.* **2021** , *21* , 81. [CrossRef]
65. Strafella, C.; Caputo, V.; Termine, A.; Barati, S.; Caltagirone, C.; Giardina, E.; Cascella, R. Investigation of Genetic Variations of IL6 and IL6R as Potential Prognostic and Pharmacogenetics Biomarkers: Implications for COVID-19 and Neuroinflammatory Disorders. *Life* **2020** , *10* , 351. [CrossRef]
66. Ghazy, A.A. Influence of IL-6 Rs1800795 and IL-8 Rs2227306 Polymorphisms on COVID-19 Outcome. *J. Infect. Dev. Ctries.* **2023** , *17* , 327–334. [CrossRef]
67. Balzanelli, M.G.; Distratis, P.; Lazzaro, R.; Pham, V.H.; Tran, T.C.; Dipalma, G.; Bianco, A.; Serlenga, E.M.; Aityan, S.K.; Pierangeli, V.; et al. Analysis of Gene Single Nucleotide Polymorphisms in COVID-19 Disease Highlighting the Susceptibility and the Severity towards the Infection. *Diagnostics* **2022** , *12* , 2824. [CrossRef] [PubMed]

68. Jagoda, E.; Marnetto, D.; Senevirathne, G.; Gonzalez, V.; Baid, K.; Montinaro, F.; Richard, D.; Falzarano, D.; LeBlanc, E.V.; Colpitts, C.C.; et al. Regulatory Dissection of the Severe COVID-19 Risk Locus Introgressed by Neanderthals. *Elife* **2023** , *12* , e71235. [CrossRef] [PubMed]
69. Picci, L.; Cameran, M.; Marangon, O.; Marzenta, D.; Ferrari, S.; Frigo, A.C.; Scarpa, M. A 10-Year Large-Scale Cystic Fibrosis Carrier Screening in the Italian Population. *J. Cyst. Fibros.* **2010** , *9* , 29–35. [CrossRef] [PubMed]
70. Chen, P.-T.; Lin, C.-L.; Wu, W.-N. Big Data Management in Healthcare: Adoption Challenges and Implications. *Int. J. Inf. Manag.* **2020** , *53* , 102078. [CrossRef]
71. Al Meslamani, A.Z.; Sobrino, I.; de la Fuente, J. Machine Learning in Infectious Diseases: Potential Applications and Limitations. *Ann. Med.* **2024** , *56* , 2362869. [CrossRef]
72. Hu, L.; Li, L. Using Tree-Based Machine Learning for Health Studies: Literature Review and Case Series. *Int. J. Environ. Res. Public Health* **2022** , *19* , 16080. [CrossRef] [PubMed]
73. Gill, M.; Anderson, R.; Hu, H.; Bennamoun, M.; Petereit, J.; Valliyodan, B.; Nguyen, H.T.; Batley, J.; Bayer, P.E.; Edwards, D. Machine Learning Models Outperform Deep Learning Models, Provide Interpretation and Facilitate Feature Selection for Soybean Trait Prediction. *BMC Plant Biol.* **2022** , *22* , 180. [CrossRef]
74. Zhang, S.; Yang, Z.; Li, Z.-N.; Chen, Z.-L.; Yue, S.-J.; Fu, R.-J.; Xu, D.-Q.; Zhang, S.; Tang, Y.-P. Are Older People Really More Susceptible to SARS-CoV-2? *Aging Dis.* **2022** , *13* , 1336–1347. [CrossRef]
75. Zhang, H.; Wu, Y.; He, Y.; Liu, X.; Liu, M.; Tang, Y.; Li, X.; Yang, G.; Liang, G.; Xu, S.; et al. Age-Related Risk Factors and Complications of Patients With COVID-19: A Population-Based Retrospective Study. *Front. Med.* **2021** , *8* , 757459. [CrossRef]
76. Reyes, L.F.; Murthy, S.; Garcia-Gallo, E.; Merson, L.; Ibáñez-Prada, E.D.; Rello, J.; Fuentes, Y.V.; Martin-Loeches, I.; Bozza, F.; Duque, S.; et al. Respiratory Support in Patients with Severe COVID-19 in the International Severe Acute Respiratory and Emerging Infection (ISARIC) COVID-19 Study: A Prospective, Multinational, Observational Study. *Crit. Care* **2022** , *26* , 276. [CrossRef] [PubMed]
77. Kondili, E.; Makris, D.; Georgopoulos, D.; Rovina, N.; Kotanidou, A.; Koutsoukou, A. COVID-19 ARDS: Points to Be Considered in Mechanical Ventilation and Weaning. *J. Pers. Med.* **2021** , *11* , 1109. [CrossRef] [PubMed]
78. Peng, M.; He, J.; Xue, Y.; Yang, X.; Liu, S.; Gong, Z. Role of Hypertension on the Severity of COVID-19: A Review. *J. Cardiovasc. Pharmacol.* **2021** , *78* , e648–e655. [CrossRef]
79. Alfaro, E.; Díaz-García, E.; García-Tovar, S.; Galera, R.; Casitas, R.; Torres-Vargas, M.; López-Fernández, C.; Añón, J.M.; García-Río, F.; Cubillos-Zapata, C. Endothelial Dysfunction and Persistent Inflammation in Severe Post-COVID-19 Patients: Implications for Gas Exchange. *BMC Med.* **2024** , *22* , 242. [CrossRef]
80. Rizk, J.G.; Sanchis-Gomar, F.; Henry, B.M.; Lippi, G.; Lavie, C.J. Coronavirus Disease 2019, Hypertension, and Renin-Angiotensin-Aldosterone System Inhibitors. *Curr. Opin. Cardiol.* **2022** , *37* , 419–423. [CrossRef]
81. Lim, S.; Bae, J.H.; Kwon, H.-S.; Nauck, M.A. COVID-19 and Diabetes Mellitus: From Pathophysiology to Clinical Management. *Nat. Rev. Endocrinol.* **2021** , *17* , 11–30. [CrossRef]
82. Kim, Y.; Latz, C.A.; DeCarlo, C.S.; Lee, S.; Png, C.Y.M.; Kibrik, P.; Sung, E.; Alabi, O.; Dua, A. Relationship between Blood Type and Outcomes Following COVID-19 Infection. *Semin. Vasc. Surg.* **2021** , *34* , 125–131. [CrossRef]
83. Zeng, X.; Fan, H.; Kou, J.; Lu, D.; Huang, F.; Meng, X.; Liu, H.; Li, Z.; Tang, M.; Zhang, J.; et al. Analysis between ABO Blood Group and Clinical Outcomes in COVID-19 Patients and the Potential Mediating Role of ACE2. *Front. Med.* **2023** , *10* , 1167452. [CrossRef]
84. Zhang, Y.; Garner, R.; Salehi, S.; La Rocca, M.; Duncan, D. Association between ABO Blood Types and Coronavirus Disease 2019 (COVID-19), Genetic Associations, and Underlying Molecular Mechanisms: A Literature Review of 23 Studies. *Ann. Hematol.* **2021** , *100* , 1123–1132. [CrossRef] [PubMed]
85. Guillon, P.; Clément, M.; Sébille, V.; Rivain, J.-G.; Chou, C.-F.; Ruvoën-Clouet, N.; Le Pendu, J. Inhibition of the Interaction between the SARS-CoV Spike Protein and Its Cellular Receptor by Anti-Histo-Blood Group Antibodies. *Glycobiology* **2008** , *18* , 1085–1093. [CrossRef]
86. Muñiz-Diaz, E.; Llopis, J.; Parra, R.; Roig, I.; Ferrer, G.; Grifols, J.; Millán, A.; Ene, G.; Ramiro, L.; Maglio, L.; et al. Relationship between the ABO Blood Group and COVID-19 Susceptibility, Severity and Mortality in Two Cohorts of Patients. *Blood Transfus.* **2021** , *19* , 54–63. [CrossRef]
87. Tedbury, P.R.; Manfredi, C.; Degenhardt, F.; Conway, J.; Horwath, M.C.; McCracken, C.; Sorscher, A.J.; Moreau, S.; Wright, C.; Edwards, C.; et al. Mechanisms by Which the Cystic Fibrosis Transmembrane Conductance Regulator May Influence SARS-CoV-2 Infection and COVID-19 Disease Severity. *FASEB J.* **2023** , *37* , e23220. [CrossRef] [PubMed]
88. Hariyanto, T.I.; Putri, C.; Arisa, J.; Situmeang, R.F.V.; Kurniawan, A. Dementia and Outcomes from Coronavirus Disease 2019 (COVID-19) Pneumonia: A Systematic Review and Meta-Analysis. *Arch. Gerontol. Geriatr.* **2021** , *93* , 104299. [CrossRef]
89. Yin, Y.; Soe, N.N.; Valenzuela, N.M.; Reed, E.F.; Zhang, Q. HLA-DPB1 Genotype Variants Predict DP Molecule Cell Surface Expression and DP Donor Specific Antibody Binding Capacity. *Front. Immunol.* **2023** , *14* , 1328533. [CrossRef] [PubMed]

90. Chung, S.; Roh, E.Y.; Park, B.; Lee, Y.; Shin, S.; Yoon, J.H.; Song, E.Y. GWAS Identifying HLA-DPB1 Gene Variants Associated with Responsiveness to Hepatitis B Virus Vaccination in Koreans: Independent Association of HLA-DPB1*04:02 Possessing Rs1042169 G—Rs9277355 C—Rs9277356 A. *J. Viral Hepat.* **2019** , *26* , 1318–1329. [CrossRef]
91. Hu, B.; Huang, S.; Yin, L. The Cytokine Storm and COVID-19. *J. Med. Virol.* **2021** , *93* , 250–256. [CrossRef]
92. Sproston, N.R.; Ashworth, J.J. Role of C-Reactive Protein at Sites of Inflammation and Infection. *Front. Immunol.* **2018** , *9* , 754. [CrossRef]
93. Del Giudice, M.; Gangestad, S.W. Rethinking IL-6 and CRP: Why They Are More than Inflammatory Biomarkers, and Why It Matters. *Brain Behav. Immun.* **2018** , *70* , 61–75. [CrossRef]
94. Gianfagna, F.; Veronesi, G.; Baj, A.; Dalla Gasperina, D.; Siclari, S.; Drago Ferrante, F.; Maggi, F.; Iacoviello, L.; Ferrario, M.M. Anti-SARS-CoV-2 Antibody Levels and Kinetics of Vaccine Response: Potential Role for Unresolved Inflammation Following Recovery from SARS-CoV-2 Infection. *Sci. Rep.* **2022** , *12* , 385. [CrossRef] [PubMed]
95. Montes-de-Oca-García, A.; Corral-Pérez, J.; Velázquez-Díaz, D.; Perez-Bey, A.; Rebollo-Ramos, M.; Marín-Galindo, A.; Gómez-Gallego, F.; Calderon-Dominguez, M.; Casals, C.; Ponce-González, J.G. Influence of Peroxisome Proliferator-Activated Receptor (PPAR)-Gamma Coactivator (PGC)-1 Alpha Gene Rs8192678 Polymorphism by Gender on Different Health-Related Parameters in Healthy Young Adults. *Front. Physiol.* **2022** , *13* , 885185. [CrossRef]
96. Li, Y.; Ke, Y.; Xia, X.; Wang, Y.; Cheng, F.; Liu, X.; Jin, X.; Li, B.; Xie, C.; Liu, S.; et al. Genome-Wide Association Study of COVID-19 Severity among the Chinese Population. *Cell Discov.* **2021** , *7* , 76. [CrossRef]
97. Steffen, B.T.; Pankow, J.S.; Lutsey, P.L.; Demmer, R.T.; Misialek, J.R.; Guan, W.; Cowan, L.T.; Coresh, J.; Norby, F.L.; Tang, W. Proteomic Profiling Identifies Novel Proteins for Genetic Risk of Severe COVID-19: The Atherosclerosis Risk in Communities Study. *Hum. Mol. Genet.* **2022** , *31* , 2452–2461. [CrossRef] [PubMed]
98. Silva, A.D.D.; Gomes, M.F.D.C.; Gregianini, T.S.; Martins, L.G.; Veiga, A.B.G.D. Machine Learning in Predicting Severe Acute Respiratory Infection Outbreaks. *Cad. Saude Publica* **2024** , *40* , e00122823. [CrossRef]
99. Mitu, R.A.; Islam, M.R. The Current Pathogenicity and Potential Risk Evaluation of Marburg Virus to Cause Mysterious "Disease X"-An Update on Recent Evidences. *Environ. Health Insights* **2024** , *18* , 11786302241235809. [CrossRef] [PubMed]
100. World Health Organization (WHO). Available online: https://www.who.int/ (accessed on 23 December 2024).
101. Sankalpa, D.; Dhou, S.; Pasquier, M.; Sagahyroon, A. Predicting the Spread of a Pandemic Using Machine Learning: A Case Study of COVID-19 in the UAE. *Appl. Sci.* **2024** , *14* , 4022. [CrossRef]
102. Ankolekar, A.; Eppings, L.; Bottari, F.; Pinho, I.F.; Howard, K.; Baker, R.; Nan, Y.; Xing, X.; Walsh, S.L.; Vos, W.; et al. Using Artificial Intelligence and Predictive Modelling to Enable Learning Healthcare Systems (LHS) for Pandemic Preparedness. *Comput. Struct. Biotechnol. J.* **2024** , *24* , 412–419. [CrossRef] [PubMed]
103. Prediletto, I.; D'Antoni, L.; Carbonara, P.; Daniele, F.; Dongilli, R.; Flore, R.; Pacilli, A.M.G.; Pisani, L.; Tomsa, C.; Vega, M.L.; et al. Standardizing PaO_2 for $PaCO_2$ in P/F Ratio Predicts in-Hospital Mortality in Acute Respiratory Failure Due to COVID-19: A Pilot Prospective Study. *Eur. J. Intern. Med.* **2021** , *92* , 48–54. [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.

diagnostics

MDPI

Article

Clinical Application of Machine Learning Models for Early-Stage Chronic Kidney Disease Detection

Hasnain Iftikhar [1,2,*], Atef F. Hashem [3], Moiz Qureshi [2,4] and Paulo Canas Rodrigues [5]

[1] Department of Statistics, University of Peshawar, Peshawar 25120, Pakistan
[2] Department of Statistics, Quaid-i-Azam University, Islamabad 45320, Pakistan; moiz@stat.qau.edu.pk
[3] Department of Mathematics and Statistics, College of Science, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia; affaragalla@imamu.edu.sa
[4] Department of Statistics, University of Sindh, Hyderabad 76080, Pakistan
[5] Department of Statistics, Federal University of Bahia, Salvador 40170-110, Brazil; paulo.canas@ufba.br
* Correspondence: hasnain.iftikhar@upeu.edu.pe

Abstract: Background/Objectives: Chronic kidney disease (CKD) is a progressive condition that affects the body's ability to remove waste and regulate fluid and electrolytes. Early detection is crucial for delaying disease progression and initiating timely interventions. Machine learning (ML) techniques have emerged as powerful tools for automating disease diagnosis and prognosis. This study aims to evaluate the predictive performance of individual and ensemble ML algorithms for the early classification of CKD. **Methods:** A clinically annotated dataset was utilized to categorize patients into CKD and non-CKD groups. The models investigated included Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Ridge Classifier, Naïve Bayes, K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Ensemble learning strategies. A systematic preprocessing pipeline was implemented, and model performance was assessed using accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC). **Results:** The empirical findings reveal that ML-based classifiers achieved high predictive accuracy in CKD detection. Ensemble learning methods outperformed individual models in terms of robustness and generalization, indicating their potential in clinical decision-making contexts. **Conclusions:** The study demonstrates the efficacy of ML-based frameworks for early CKD prediction, offering a scalable, interpretable, and accurate clinical decision support approach. The proposed methodology supports timely diagnosis and can assist healthcare professionals in improving patient outcomes.

Keywords: chronic kidney disease; machine learning; clinical decision support; ensemble learning; early diagnosis; health informatics

1. Introduction

Kidney function gradually declines in CKD, an irreversible and progressive disorder that affects the kidneys' capacity to filter blood and efficiently remove waste products [1]. Significant systemic health problems result from the accumulation of metabolic waste in the body when renal function deteriorates. Age and gender are two demographic factors that affect the disease, which is frequently linked to underlying problems including diabetes mellitus, hypertension, and cardiovascular disorders [2–4]. It is challenging to diagnose early, as symptoms typically appear later and include back and stomach pain, fever, rashes, and vomiting [5,6]. End-stage renal disease (ESRD) is indicated by

values below 15 on the estimated glomerular filtration rate (e-GFR), which is one measure of clinical progression [7,8]. In these situations, kidney transplantation is still the only practical long-term therapy option when dialysis is not available [9].

Due to its asymptomatic nature, CKD is usually underdiagnosed in its early stages, despite its global prevalence [10]. While death rates are largely steady, the rising hospital admission rate—reported at 6.23% annually—underlines the increased healthcare burden associated with CKD [11]. Conventional diagnostic techniques for CKD are inadequate, emphasizing the need for computational algorithms that allow earlier and more accurate identification of high-risk patients [12,13]. In this context, ML and data-driven analytics are increasingly recognized as practical tools for developing clinical decision support systems (CDSSs) tailored to the needs of personalized medicine. ML enables the analysis of complex datasets derived from electronic health records (EHRs) and laboratory results, revealing hidden patterns and supporting predictive decision-making [14–18]. Through feature extraction and classification, these technologies enable clinicians to make individualized diagnoses and treatment decisions, even in data-rich yet insight-poor environments [19–21].

Recent research has shown that multiple machine learning methods may accurately predict CKD. Using six classifiers, such as Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes (NB), and Feed-forward Neural Network (FF-NN), provided promising results [22]. RF obtained an accuracy of 99.75%. Similarly, the work in [23], developed a neural network model that predicted the start of chronic renal disease with 95% accuracy. Ref. [24] employed SVM to achieve 93% accuracy, whereas [25] examined multiple classifiers on UCI datasets.

Further validation of the potential of ML-based CDSSs in early CKD identification has been provided by numerous additional contributions. The studies [26,27] highlighted the importance of customized algorithms for subgroups by focusing on diabetic patients, a population at risk for CKD. In [28], a multiclass decision forest model with 99.1% accuracy was developed using a UCI dataset with 14 features. Ref. [29] employed a two-stage SVM method, achieving an accuracy of 98.5%. Other models, such as ANN in [30], Gradient Boosting in [31], and KNN in [32], have also demonstrated excellent accuracy on various datasets. Recent developments include hybrid techniques that employ optimization algorithms, such as ACO and Relief, as well as the use of deep learning [33]. Ref. [34], as well as comparative analyses on various CKD datasets have used up to nine ML classifiers [35,36]. Therefore, all of these results highlight how machine learning is increasingly changing traditional nephrology through personalized decision-making and predictive analytics.

Thus, by empirically comparing individual and ensemble machine learning models, the current study contributes to the evolving field of CKD decision assistance. We evaluate nine classifiers (RF, SVM, KNN, and Logistic Regression) and ensemble methods (Voting and Stacking) for diagnostic performance using several assessment metrics. To enhance computational efficiency and clinical applicability, we further investigate the function of feature selection and confirm our results using 5-fold and 10-fold cross-validation procedures. With the use of ML techniques, the main goal of this research is to create a reliable and understandable clinical decision support system (CDSS) for the early detection of CKD. The study highlights the use of predictive modeling to guide prompt, customized treatment interventions, aligning with the objectives of personalized medicine. Eight different ML classifiers were used to assess the prediction power of various algorithmic paradigms thoroughly. These include both linear and nonlinear models capable of capturing complex relationships within patient data.

To enhance the generalizability and reduce overfitting, both 5 and 10-fold CV techniques were adopted. These resampling methods provide a robust framework for estimating the models' performance and ensuring their stability across different data partitions.

The evaluation focused on key performance metrics such as accuracy, sensitivity, specificity, and area under the ROC curve (AUC), thereby offering a detailed comparison of each model's diagnostic utility. This rigorous and systematic approach not only validates the predictive strength of individual and ensemble classifiers but also contributes to the broader objective of integrating ML-driven CDSS into nephrology for personalized risk stratification and early intervention.

The rest of the work is structured as follows: Section 2 explains the dataset and preprocessing steps, feature selection techniques, and the machine learning classifiers employed. It also outlines the implementation details, cross-validation procedures, and evaluation metrics used for assessing the model. Section 3 reports the experimental results, compares classifier performance with and without the feature selection process, and Section 4 discusses their clinical implications and outlines future research directions for enhancing CKD prediction using machine learning. Finally, Section 5 highlights the key contributions and emphasizes the importance of the findings.

2. Materials and Methods

2.1. Dataset Overview

This study utilizes a publicly accessible datasets that includes demographic and clinical data from individuals with and without chronic kidney disease (CKD). The dataset was initially gathered from the Burner Medical Complex (BMC), a medical facility situated in a rural area of Khyber Pakhtunkhwa in northern Pakistan. It was obtained using ResearchGate. With 258 (67.5%) CKD and 124 (32.5%) non-CKD cases, the dataset showed a rather unbalanced distribution. Stratified sampling was used to maintain the original class distribution across all subsets after the dataset was divided into training (70%), and testing (30%) sets in order to ensure a robust model evaluation. To ensure that observations are independent, each row represents a distinct patient. The dataset contains 21 therapeutically significant characteristics, including hemoglobin, albumin, sugar, pH, specific gravity, urine clarity, age, gender, and different blood cell counts. Demographic characteristics, such as age and gender, were added as predictors in addition to blood and urine measurements. Prior research has demonstrated that demographic factors can impact the occurrence and course of chronic kidney disease, even if these biomarkers are not diagnostic [37,38]. By incorporating these characteristics, the model can take patient variability into consideration. The significant variations across groups are highlighted in Table 1, which presents a full summary of these features stratified by CKD status. Nevertheless, the dataset has intrinsic limitations. Given that all patient information was gathered from a single, geographically focused healthcare facility, the study population may not accurately represent the broader clinical, environmental, and demographic diversity observed in other parts of Pakistan or elsewhere. In other places, factors such as genetic variability, environmental exposures, healthcare availability, and regional food practices may have a significant impact on how diseases emerge and how well models perform. Additionally, the dataset is missing crucial sociodemographic parameters that could enhance model generalizability and offer deeper insights into CKD risk factors, such as occupation, income, education level, and ethnicity. Therefore, it is important to exercise caution when extrapolating the results and prediction models from this dataset to other populations. Future studies should focus on validating these models using multi-center, demographically diverse datasets to enhance their applicability and external validity.

All analyses were conducted in the `R` programming language (version 4.3.2) using several specialized software packages: `randomForest` (version 4.7-1.1) [39] for Random Forest, `e1071` (version 1.7-14) [40] for Support Vector Machines (SVM), `xgboost` (version 1.7.7.1) for Gradient Boosting, `ggplot2` (version 3.5.1) [41] for data visualization, `caret` (version 6.0-

94) [42] for training and evaluating machine learning models, and `dplyr` (version 1.1.4) [43] for data manipulation.

2.2. ML Methods

This study frames CKD detection as a binary classification problem using supervised learning algorithms to predict disease status. Let $\mathbf{X} = [x_1, x_2, \ldots, x_n]$ represent the predictor variables, and $Y \in \{0, 1\}$ the binary outcome, where $Y = 1$ indicates CKD and $Y = 0$ indicates no CKD. A variety of machine learning models—LR, LDA, QDA, DT, RF, SVM, NB, KNN, and Regression Trees—were employed to identify discriminative patterns for CKD diagnosis.

Table 1. Summary statistics of selected clinical variables stratified by CKD status.

Status	Variable	Mean	SD	Min	Max
0	Age	39.80	14.42	12	91
1	Age	47.69	15.96	12	90
0	Gender	1.56	0.50	1	2
0	U Clarity	1.42	0.50	1	2
0	pH	5.38	0.50	5	8
1	pH	5.68	0.57	5	8
0	Specific Gravity	1.014	0.0048	1.005	1.025
1	Specific Gravity	1.017	0.0052	1.000	1.050
0	Albumin	1.23	0.42	1	2
0	Glucose (Blood)	1.05	0.22	1	2
1	Glucose (Blood)	1.08	0.28	1	2
0	Sugar (glucosuria)	1.04	0.20	1	2
1	Sugar (glucosuria)	1.24	0.43	1	2
0	Ketones (KB)	1.01	0.12	1	2
1	Ketones (KB)	1.26	0.44	1	2
0	Bile Pigment	1.01	0.12	1	2
1	Bile Pigment	1.26	0.44	1	2
0	Urobilinogen	1.98	0.14	1	2
1	Urobilinogen	1.85	0.35	1	2
0	Blood	1.01	0.12	1	2
1	Blood	1.25	0.43	1	2
0	Pus Cells (urinary white blood cells)	1.87	0.34	1	2
0	Red Cells	1.79	0.41	1	2
1	Red Cells	1.44	0.50	1	2
0	Epithelial Cells	1.42	0.50	1	2
1	Epithelial Cells	1.70	0.46	1	2

To evaluate model performance and mitigate overfitting, both the 5- and 10-fold stratified CV were implemented. Stratified CV ensures that class proportions are preserved in each fold. Since each patient appears only once in the dataset, data independence was maintained. Feature selection techniques were also applied to enhance model performance and reduce computational complexity.

2.2.1. Logistic Regression

Logistic Regression (LR) is a commonly used classification technique that estimates the probability of a binary response based on input variables. Unlike linear regression, which models continuous outcomes, LR models the log-odds of the outcome:

$$\text{logit}(P) = \log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n \tag{1}$$

where P is the probability of the positive class and β_i are the model coefficients [44].

2.2.2. Linear Discriminant Analysis

The goal of LDA is to identify the linear feature combination that best distinguishes between two or more classes. It assumes that class covariances are equal and that features are normally distributed. Based on the Bayes theorem, LDA models the conditional probability $P(Y = k \mid \mathbf{X})$ for every class k:

$$P(Y = k \mid \mathbf{X} = \mathbf{x}) = \frac{\pi_k \cdot \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma)}{\sum_j \pi_j \cdot \mathcal{N}(\mathbf{x} \mid \mu_j, \Sigma)} \tag{2}$$

In which π_k represents the prior probability and $\mathcal{N}$ represents the multivariate normal distribution.

2.2.3. Quadratic Discriminant Analysis (QDA)

QDA is a generalization of LDA that allows each class to have its own (covariance) matrix. This flexibility results in a quadratic decision boundary, improving accuracy in datasets with heterogeneous class distributions [45].

2.2.4. Decision Tree (DT)

DTs are hierarchical models that split data into branches based on feature values, aiming to maximize class purity (e.g., using Gini impurity or entropy) at each node. They are interpretable but prone to overfitting [46].

2.2.5. Random Forest (RF)

Random Forest is an ensemble of Decision Trees trained on bootstrap samples of the dataset. The final prediction is decided by the majority vote of each tree for a class. Mathematically

$$H(\mathbf{X}) = \text{mode}\{h_1(\mathbf{X}), h_2(\mathbf{X}), \ldots, h_T(\mathbf{X})\} \tag{3}$$

where $h_t(\mathbf{X})$ is the prediction from the tth tree, and T is the total number of trees [47].

2.2.6. Support Vector Machine (SVM)

Support Vector Machines (SVM) construct an optimal hyperplane that maximizes the margin between two classes. For non-linear data, kernel functions (e.g., RBF) are employed to map features into higher-dimensional spaces, thereby improving separation [48].

2.2.7. Naïve Bayes (NB)

Based on the Bayes theorem, the Naïve Bayes classifier assumes feature independence given the class label. The posterior probability is calculated as

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \tag{4}$$

This assumption simplifies computation and is effective even with small datasets [49].

2.2.8. K-Nearest Neighbors

KNN is a non-parametric classification technique that uses the majority vote among the K nearest training instances, usually determined by Euclidean distance, to provide a class label. A careful selection of K is necessary, as it has a significant impact on model performance.

2.2.9. Stacking Ensemble Learning

An Ensemble learning method called Stacking combines the projections of several base classifiers using a meta learner that is trained to maximize the output by analyzing the predictions of the base models. The base models are trained on the original feature set, and their outputs serve as inputs to the meta-model, which produces the final prediction. In this study, we employ Multiple Linear Regression (MLR) and Probability Distribution (PD)-based stacking to aggregate model predictions and enhance overall performance [50–52].

Mathematically, $D = \{(x_i, y_i)\}_{i=1}^{n}$, the base learners generate

$$\hat{y}_i^{(j)} = M_j(x_i), \quad j = 1, 2, \ldots, k \tag{5}$$

These forecasts are then combined by the meta-learner F to yield the final result:

$$\hat{y}_i = F(\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \ldots, \hat{y}_i^{(k)}) \tag{6}$$

2.3. Feature Selection

In order to evaluate the effects of dimensionality reduction, a mix of correlation analysis, univariate statistical tests (such as chi-square and t-tests), and recursive feature elimination (RFE) with cross-validation was used to choose features. Key indicators that are clinically relevant to the course of chronic kidney disease (CKD) were identified through this procedure, including serum creatinine, blood urea, albumin, hemoglobin, and hypertension.

2.4. Validation of Classifier Performance

All classifier training and validation procedures were executed using `Jupyter Notebook` (version 7.0.8). To ensure robustness and minimize overfitting, the dataset was partitioned using both 5 and 10-fold stratified CV techniques. These approaches maintain class distribution across folds, guaranteeing reliable performance assessment. An evaluation matrix comprising standard classification metrics was developed to compare the predictive accuracy of each model.

2.5. Evaluation Metrics and Confusion Matrix

A confusion matrix, a 2×2 table that summarizes the classification results, was used to evaluate classifier performance. The four main components of the matrix are false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN). Key evaluation metrics, explained below, are made easier to calculate using this data.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100, \tag{7}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{8}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{9}$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100, \tag{10}$$

$$F1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}. \tag{11}$$

3. Results and Interpretations

Several ML models for predicting CKD are thoroughly evaluated in this section. Detailed comparisons across various metrics, including accuracy, sensitivity, specificity, precision, and F1 score, are presented in tables and figures. The model's performance is evaluated through comprehensive cross-validation and visualizations.

3.1. Performance Comparison Without Feature Selection

Without feature selection, Table 2 presents a summary of the classification performance of a few chosen machine learning models using the 5- and 10-fold CV method. For each algorithm, the optimal hyperparameters are found in Table 3 using a grid search strategy combined with 5-fold and 10-fold cross-validation. The performance metrics of the models following tuning under both cross-validation procedures are shown in Table 2. All evaluation measures showed that RF and Support Vector Machine (SVM) performed better than the other classifiers. Interestingly, in 5-fold cross-validation, RF obtained a sensitivity and F1 score of 0.9331, indicating a robust capacity to identify CKD cases while balancing recall and precision [7]. SVM shows comparable performance, particularly in precision and F1 score. LR, LDA, and Ridge Classifiers also show competitive results. The consistent results between 5-fold and 10-fold strategies demonstrate the robustness and generalizability of these models, indicating minimal risk of overfitting. Nonetheless, further performance enhancements can be achieved through hyperparameter tuning or the incorporation of ensemble approaches [53,54].

Table 2. Comparison of model performance without feature selection using 5-fold and 10-fold cross-validation.

Model	Accuracy	Sensitivity	Specificity	Precision	F1 Score
		5-Fold Cross-Validation			
LR	0.9026	0.9289	0.8582	0.9174	0.9231
QDA	0.8421	0.8243	0.8723	0.9163	0.8678
LDA	0.8947	0.9205	0.8511	0.9129	0.9167
Naïve Bayes	0.8474	0.8201	0.8936	0.9289	0.8711
RF	0.9158	0.9331	0.8865	0.9331	0.9331
SVM	0.9105	0.9331	0.8723	0.9253	0.9292
DT	0.8684	0.8954	0.8227	0.8954	0.8954
Ridge	0.8921	0.9205	0.8440	0.9091	0.9148
		10-Fold Cross-Validation			
LR	0.9079	0.9331	0.8652	0.9215	0.9272
QDA	0.8474	0.8494	0.8440	0.9022	0.8750
LDA	0.8974	0.9205	0.8582	0.9167	0.9186
Naïve Bayes	0.8500	0.8326	0.8794	0.9213	0.8747
RF	0.9053	0.9289	0.8652	0.9212	0.9250
SVM	0.9000	0.9205	0.8652	0.9205	0.9205
DT	0.8658	0.8745	0.8511	0.9087	0.8913
Ridge	0.9000	0.9247	0.8582	0.9170	0.9208

Table 3. Optimal hyperparameters used for each classifier (based on grid search).

Model	Tuned Hyperparameters
Logistic Regression (LR)	penalty = l2, solver = lbfgs, C = 1.0
QDA	reg_{param} = 0.01
LDA	solver = svd
Naïve Bayes	var_smoothing = 1×10^{-9}
Random Forest (RF)	n_estimators = 200, max_depth = 10, min_samples_split = 5
Support Vector Machine (SVM)	kernel = 'rbf', C = 10, gamma = 0.01
Decision Tree (DT)	max_depth = 8, min_samples_split = 5, criterion = 'gini'
Ridge Classifier	alpha = 1.0, solver = 'auto'

3.2. Impact of Feature Selection

Table 4 evaluates the effect of feature selection on model performance. Most classifiers exhibit marginal improvements in sensitivity and F1 scores. For instance, QDA demonstrates a notable increase in sensitivity from 0.8452 to 0.9414, underlining the importance of reducing feature space for better generalization. Both Logistic Regression and Random Forest benefit slightly from feature selection, reinforcing its utility in simplifying model complexity without compromising accuracy [54,55].

Table 4. Comparative analysis of model performance with and without feature selection.

Model	Accuracy	Sensitivity	Specificity	Precision	F1 Score
Without Feature Selection					
Random Forest	0.9158	0.9498	0.8652	0.9228	0.9361
SVM	0.9105	0.9414	0.8369	0.9073	0.9240
Logistic Regression	0.9026	0.9163	0.8723	0.9241	0.9202
LDA	0.8947	0.9289	0.8440	0.9098	0.9193
Ridge	0.8921	0.9289	0.8440	0.9098	0.9193
Decision Tree	0.8684	0.8703	0.8582	0.9123	0.8908
Naïve Bayes	0.8474	0.8452	0.8582	0.9099	0.8764
QDA	0.8421	0.8452	0.8369	0.8978	0.8707
With Feature Selection					
QDA	0.9079	0.9414	0.8511	0.9146	0.9278
SVM	0.9026	0.9456	0.8298	0.9040	0.9243
Random Forest	0.8974	0.9331	0.8369	0.9065	0.9196
Logistic Regression	0.8947	0.9372	0.8227	0.8996	0.9180
LDA	0.8947	0.9163	0.8582	0.9163	0.9163
Ridge	0.8947	0.9163	0.8582	0.9163	0.9163
Naïve Bayes	0.8816	0.8996	0.8511	0.9110	0.9053
Decision Tree	0.8526	0.8745	0.8156	0.8894	0.8819

3.3. Feature Importance Analysis

Due to their built-in feature selection, tree-based models (such as Random Forest and XGBoost) performed better after feature selection, even though models like Logistic Regression and SVM performed better initially. Table 5 provides an overview of the chosen features and the corresponding model-wise performance gains [28].

In addition, Table 5 presents the relative importance of features across different classifiers. In this study, the final feature set included all predictors that were both clinically relevant to chronic kidney disease (CKD) and statistically significant ($p < 0.05$), rather than arbitrarily selecting a fixed number of top-ranked variables (e.g., top 3, 5, or 10). These selected features were provided as inputs to both the Stacking Ensemble meta-learner and all individual base learners. Regularized linear models (e.g., Logistic Regression,

Ridge) adjusted their coefficients based on the penalization strength, while tree-based models (Random Forest and XGBoost) further refined feature contributions using their inherent importance measures. This evidence-driven approach ensured the inclusion of only informative and non-redundant predictors, thereby maintaining both model accuracy and interpretability.

Among the predictors, *pus_cells* emerged as the most influential feature, particularly in Random Forest and SVM, aligning with established medical evidence [56]. Other key predictors included *albumin* and *bacteria*, both of which were strongly associated with kidney dysfunction [57]. Conversely, features such as *age* [58] and *bpigment* [59] demonstrated comparatively low predictive power. Overall, these findings highlight the dominant role of urinalysis-related indicators in CKD prediction [60].

Specifically, feature importance rankings were computed for the tree-based models, and SHAP (SHapley Additive exPlanations) value analysis was applied to the Random Forest and SVM classifiers to quantify the contribution of individual features to specific predictions. SHAP values provide a unified, game-theoretic framework for interpreting model outputs, allowing for the decomposition of each prediction into additive feature contributions. This approach not only corroborates the global importance patterns observed in Table 5 but also provides local interpretability for individual patient cases. By making the reasoning process behind model predictions transparent, this analysis enhances clinician trust and supports potential clinical validation in real-world settings.

Table 5. Feature importance scores across models.

Feature	RF	SVM	LR	LDA	QDA	Naïve Bayes	Ridge	Average
Pus Cells (urine)	0.2866	0.0687	0.0605	0.0766	0.0458	0.0453	0.0805	0.0949
Serum Albumin Level (g/dL)	0.1589	0.0553	0.0545	0.0292	0.0258	0.0132	0.0300	0.0524
Bacteria Presence (urine)	0.0717	0.0258	0.0226	0.0058	0.0884	0.0950	0.0068	0.0452
Coronary Obstruction (CO)	0.0864	0.0429	0.0213	0.0013	0.0508	0.0303	0.0045	0.0339
Red Blood Cells (urine)	0.0657	0.0355	0.0353	0.0234	0.0221	0.0047	0.0250	0.0303
Age (years)	0.1601	0.0053	0.0005	−0.0058	0.0076	0.0000	−0.0029	0.0236
Bile Pigment (urine)	0.0351	0.0018	0.0066	−0.0026	0.0616	0.0613	−0.0018	0.0231
Urine pH	0.0502	0.0026	0.0076	0.0024	0.0111	0.0005	0.0018	0.0109
Microtubule Test (mt)	0.0353	0.0053	0.0034	0.0042	0.0095	0.0005	0.0079	0.0094
Urine Specific Gravity (Sp.G)	0.0499	0.0024	0.0029	0.0011	0.0037	0.0032	0.0024	0.0094

3.4. Performance of Ensemble Models

The performance comparison presented in Table 6 demonstrates the superior predictive capabilities of Ensemble learning methods over individual classifiers in the early detection of CKD. Among all evaluated models, the Stacking Ensemble consistently achieved the highest metrics, with an accuracy of 0.9053, a sensitivity of 0.9498, a precision of 0.9044, and an F1 score of 0.9265. These results highlight its robust ability to correctly identify CKD cases while maintaining a balanced trade-off between false positives and false negatives. The Voting Ensemble also delivered a competitive performance (accuracy = 0.8974, F1 score = 0.9180), exceeding many of the individual classifiers in sensitivity and F1 score. Among the standalone models, SVM emerged as the most effective, matching the Stacking model in sensitivity (0.9498) and achieving a strong F1 score of 0.9228. QDA and Ridge Classifier also showed commendable results, with F1 scores of 0.9184 and 0.9193, respectively. In contrast, the Decision Tree classifier underperformed, particularly in accuracy (0.8474) and F1 score (0.8771), which may be attributed to its tendency to overfit on smaller datasets. Additionally, The confusion matrix in Table 7 demonstrates that the Stacking Ensemble model accurately recognized 118 out of 142 non-CVD cases (specificity = 82.98%) and 228 out of 240 CVD patients (sensitivity = 94.98%). With an F1 score of 92.65%, ac-

curacy of 90.53%, and precision of 90.44%, the model performs well overall, especially when it comes to identifying real CVD cases. This makes it a good fit for clinical decision support, where reducing missed diagnoses is crucial. Overall, these findings affirm that Ensemble learning—particularly stacking—enhances generalization by leveraging the complementary strengths of base models. This capability is especially valuable in clinical decision support systems where both diagnostic accuracy and reliability are critical. Such approaches align with recent studies advocating for hybrid and ensemble models in medical diagnostics [59,61], underscoring their practical utility in data-driven, patient-centered healthcare.

Table 6. Comparison of ensemble models with individual classifiers.

Model	Accuracy	Sensitivity	Specificity	Precision	F1 Score
Logistic Regression	0.8921	0.9247	0.8369	0.9057	0.9151
SVM	0.9000	0.9498	0.8156	0.8972	0.9228
Random Forest	0.8868	0.9247	0.8227	0.8984	0.9113
LDA	0.8789	0.9038	0.8369	0.9038	0.9038
QDA	0.8947	0.9414	0.8156	0.8964	0.9184
Naïve Bayes	0.8921	0.9247	0.8369	0.9057	0.9151
Ridge	0.8947	0.9289	0.8440	0.9098	0.9193
Decision Tree	0.8474	0.8661	0.8156	0.8884	0.8771
Voting Ensemble	0.8974	0.9372	0.8227	0.8996	0.9180
Stacking Ensemble	0.9053	0.9498	0.8298	0.9044	0.9265

Table 7. Confusion matrix for Stacking Ensemble (n = 382).

Actual/Predicted	Positive	Negative	Total
Positive	228	12	240
Negative	24	118	142
Total	252	130	382

3.5. Visual Interpretation

The ROC curves are shown in Figure 1 for each classifier presented in Table 6 used in the CKD prediction. True Positive Rate (sensitivity) is represented by the y-axis, and False Positive Rate (1 − specificity) is represented by the x-axis. The random classifier's performance is shown by the dashed diagonal line (AUC = 0.5). Every classifier showed a high degree of selective capacity by performing much better than the random baseline. These models maintain a high sensitivity (>0.85) while maintaining the false positive rate below 0.2, as seen by the close clustering of the Logistic Regression, SVM, Random Forest, Ridge, LDA, QDA, and Naïve Bayes curves. The ensemble techniques (Voting Ensemble and Stacking Ensemble) performed the best among them; their ROC curves, which are located closest to the upper-left corner, show a better balance between sensitivity and specificity. It appears that the Stacking Ensemble model performs better than any other model, indicating that merging several classifiers improves predictive power. The ROC curve of the Decision Tree classifier was marginally lower than that of the other techniques, suggesting a higher false positive rate and lower sensitivity at similar thresholds, indicating comparatively poorer performance. However, compared to random categorization, it still performs noticeably better. For CVD prediction, ensemble models, particularly the Stacking Ensemble, offer the most dependable performance, according to the ROC analysis.

Figure 2 compares eight machine learning models for CKD prediction using two cross-validation strategies: 5-fold and 10-fold. The models were tested using five essential metrics: accuracy, sensitivity, specificity, precision, and F1 score. RF consistently outper-

formed all other classifiers on most criteria, demonstrating its outstanding generalization ability and tolerance to overfitting. SVM and LR produced robust and competitive results, closely matching RF in terms of prediction accuracy. The changes in model performance between 5-fold and 10-fold cross-validation were minimal, highlighting the robustness and stability of the results across multiple validation schemes. Notably, 5-fold cross-validation achieved equivalent accuracy while requiring less computational effort, making it a more suitable option for practical applications. Models like QDA and Naïve Bayes performed worse in terms of sensitivity and specificity, indicating a limited ability to handle complex or unbalanced clinical information. Overall, the figure demonstrates the higher performance of ensemble and margin-based classifiers for CKD prediction, supporting the adoption of 5-fold cross-validation as a dependable and computationally economical model assessment method.

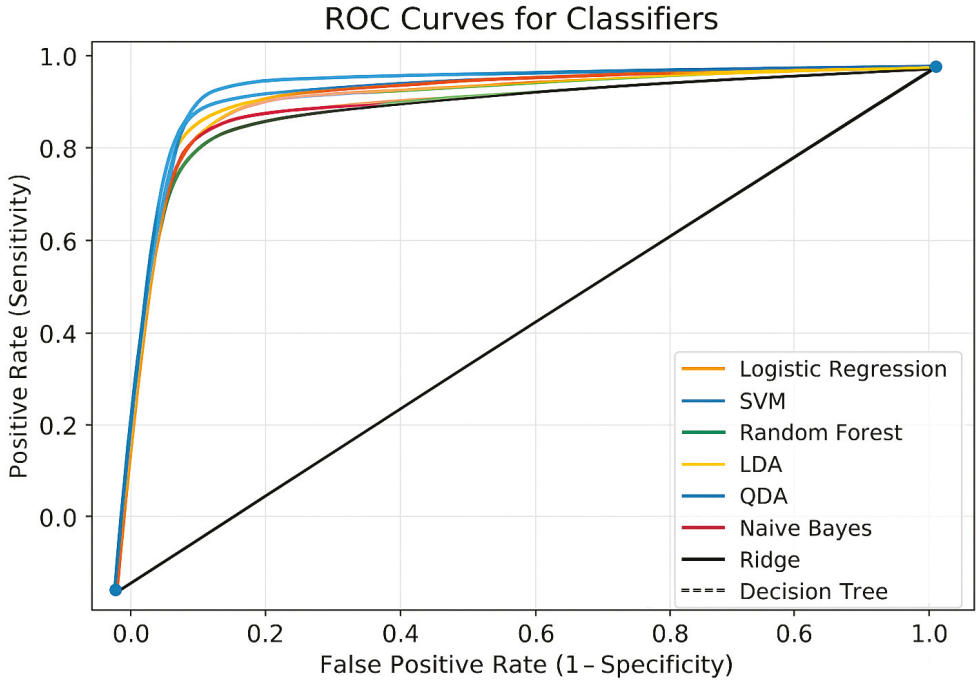


Figure 1. ROC curves for all classifiers.

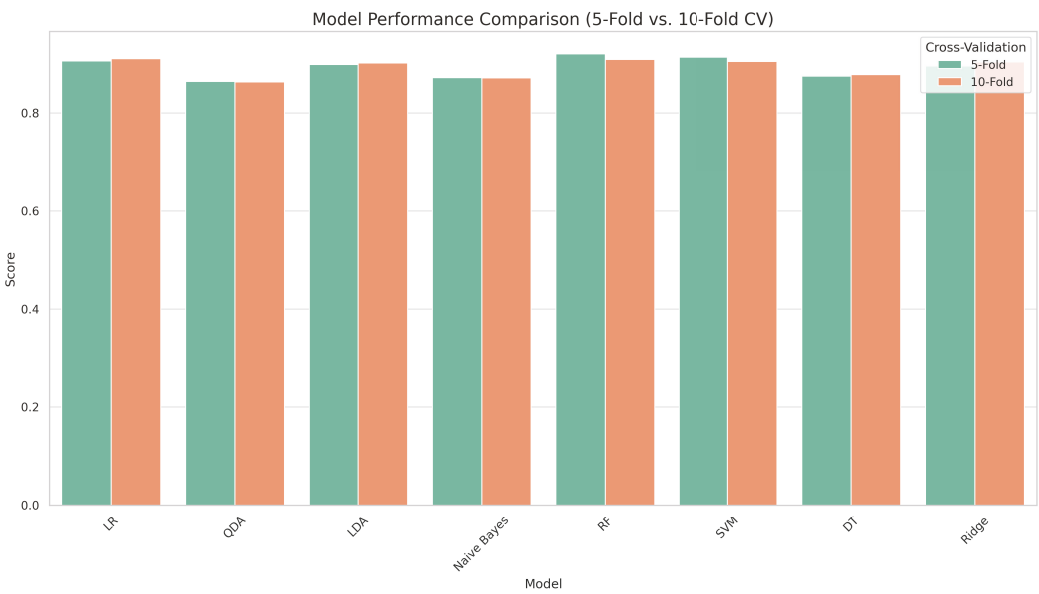


Figure 2. Accuracy comparison of models under 5-fold and 10-fold cross-validation.

Figure 3 provides a comparative visualization of model performance with and without feature selection across several evaluation metrics, including accuracy, sensitivity, specificity, precision, and F1 score. Overall, the results demonstrate that feature selection generally improves or maintains model performance. Notably, models such as QDA and SVM demonstrated significant improvements in sensitivity and F1 score after applying feature selection, indicating enhanced identification of positive cases and balanced performance. Random Forest and Logistic Regression maintained robust performance in both scenarios, highlighting their inherent strength in handling high-dimensional data. Although some models, like Decision Tree, exhibited marginal differences, feature selection led to slightly improved generalization by reducing noise and redundancy. These findings highlight the practical benefits of incorporating feature selection into machine learning pipelines, particularly in healthcare applications such as CKD prediction, where model interpretability and efficiency are crucial.

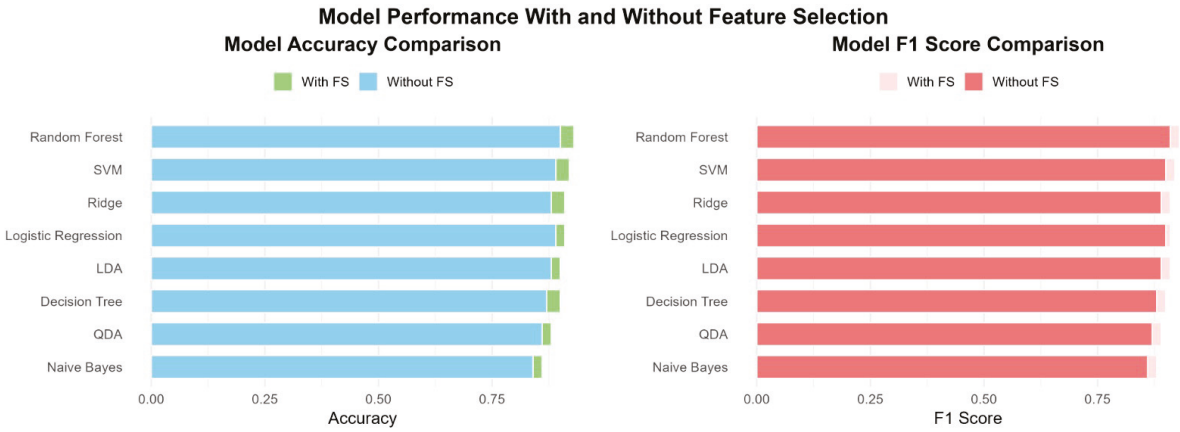


Figure 3. Comparative performance of machine learning models with and without feature selection based on key evaluation metrics.

The feature importance heatmap provides a comparative visualization of how various clinical features contribute to the prediction of CKD across different machine learning models, as presented in Figure 4. Each row represents a clinical feature (e.g., pus_cells, albumin, bacteria), and each column corresponds to a model (e.g., Random Forest, SVM, Logistic Regression). Warmer colors indicate higher importance scores, reflecting the feature's relative contribution to model prediction. Among all features, pus_cells emerges as the most influential predictor, with consistently high importance across models, especially in the Random Forest (0.2866) and Logistic Regression (0.0605) models, contributing to its highest average importance score of 0.0949. Albumin and bacteria follow as the next most influential variables, with average scores of 0.0524 and 0.0452, respectively. Features like age, bpigment, and co show moderate importance, while pH, mt, and sp_g contribute minimally across all models. Thus, this visualization highlights the robustness of pus_cells and albumin as key indicators for early CKD detection, supporting their prioritization in clinical decision-making and model development. It also reveals variations in model sensitivity to specific features, guiding future refinement of feature engineering strategies.

In Figure 5, presenting the heatmap comparing ensemble models with individual classifiers offers a clear and intuitive visualization of model performance across five key evaluation metrics: accuracy, sensitivity, specificity, precision, and F1 score. Among all the models, the Stacking Ensemble exhibits the highest performance, especially in accuracy (0.9053), sensitivity (0.9498), precision (0.9044), and F1 score (0.9265), indicating its superior ability to generalize and balance between false positives and false negatives. The Voting Ensemble also performs competitively, demonstrating enhanced sensitivity and an F1 score

that surpasses that of most individual classifiers. In contrast, particular models, such as the Decision Tree, show comparatively lower performance across all metrics, reaffirming the value of ensemble approaches. Overall, the heatmap reinforces that Ensemble learning—particularly Stacking—can significantly enhance prediction robustness and reliability in clinical decision support systems for CKD diagnosis.

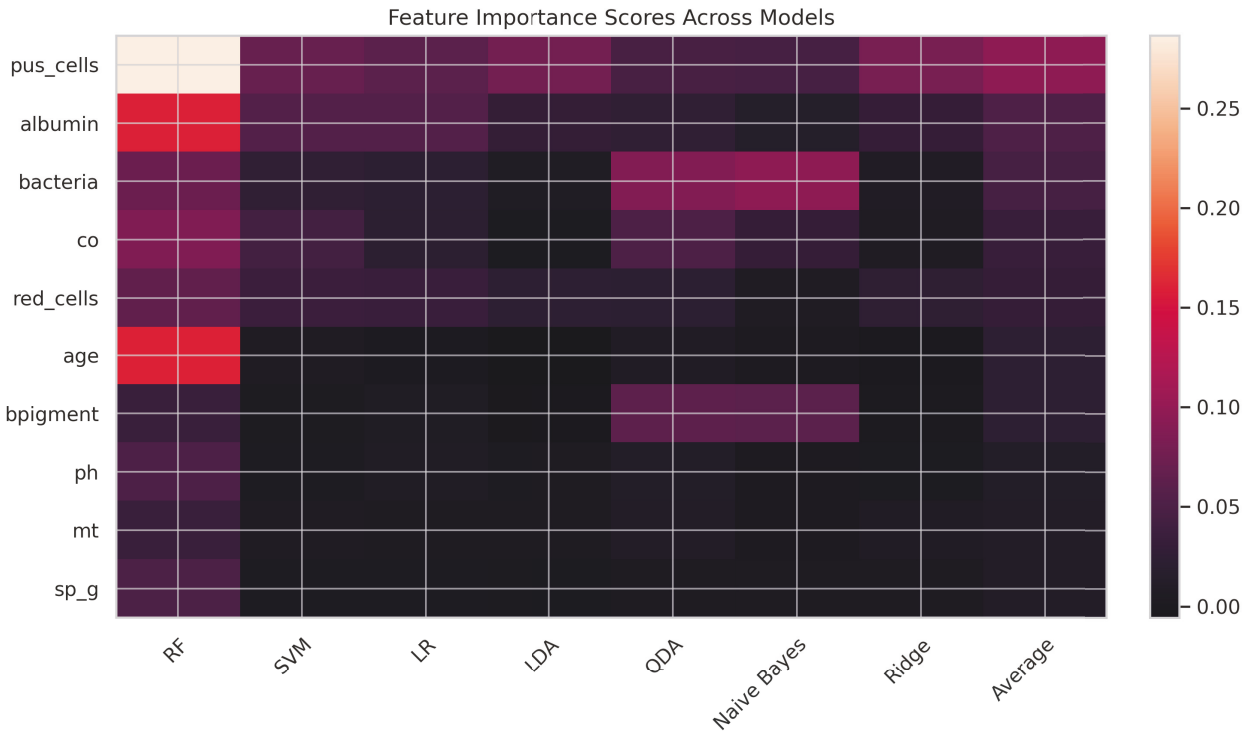


Figure 4. Feature importance heatmap.

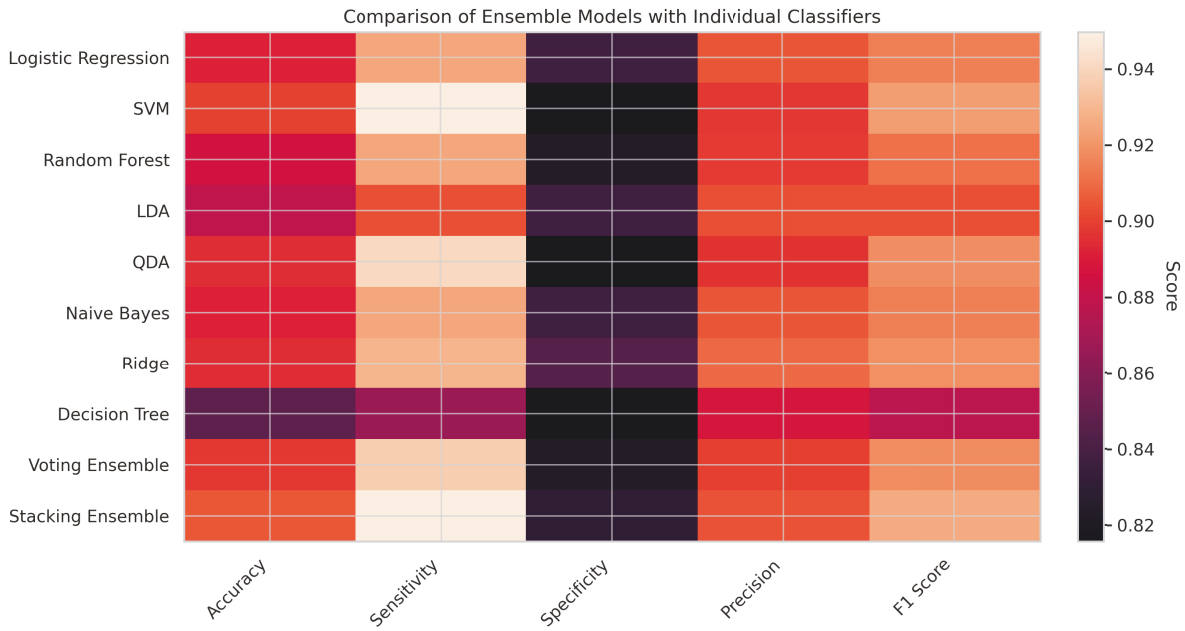


Figure 5. Ensemble vs. individual heatmap.

4. Discussion

Several ML algorithms for differentiating patients with CKD from those without the ailment are thoroughly evaluated in this study. Out of the eight classifiers that were

analyzed, RF consistently produced better predictive results than the others: KNN, LR, LDA, QDA, SVM, Ridge Classifier, NB, and Regression Tree (RT). Due to its robustness and generalizability, it achieved the highest classification accuracy in both 5-fold (91.58%) and 10-fold (90.53%) cross-validation. However, the SVM model also demonstrated its effectiveness as a classifier, producing results that were consistent across validation procedures, albeit with slightly less volatility. Additionally, ensemble techniques, specifically the Voting and Stacking models, yielded encouraging results with respective accuracies of 90.53% and 89.74%. Feature selection improved model performance by reducing noise and dimensionality. The results demonstrate the feasibility of incorporating Random Forest and ensemble algorithms into decision support systems for detecting early-stage CKD using regular clinical data. The approach highlights how ML models can aid in early diagnosis, particularly in situations with limited resources or during initial clinical examinations [62,63]. These models can serve as effective supplements to established diagnostic tools, providing scalable and interpretable decision support that is particularly useful in telemedicine and rural healthcare settings. Future studies should use richer and longitudinal datasets to increase prediction power and therapeutic relevance [64,65].

On the other hand, this study has significant limitations. First, the dataset employed is cross-sectional, which limits the capacity to estimate temporal progression or early transitions between CKD stages. This shortcoming restricts its use in continuous monitoring and prognostic modeling. Furthermore, the data were obtained from a single clinical source, which may limit the model's applicability to other geographic locations, healthcare systems, or demographic groupings. To improve the clinical usefulness and robustness of these models, future research should test them with varied, multi-center, and longitudinal data. Integrating other clinical, lifestyle, and genetic factors may further enhance the accuracy of prediction. Advanced methodologies, including deep learning architectures and more advanced ensemble tactics, should also be investigated. Furthermore, future models should prioritize interpretability and real-time application to facilitate transparent and informed clinical decision-making in real-world scenarios.

Although this work concentrates on traditional machine learning models (such as RF, SVM, and LR) due to their interpretability and simplicity, new deep learning methods have a lot of potential for CKD prediction. Advanced models such as Mamba capsule routing [66], Transformers [67], and Capsule Networks [68] have an improved capacity to simulate intricate linkages and temporal patterns, whilst CNNs and other approaches can capture spatial information. Comparison of these studies are presented in Table 8. Future research could investigate these techniques to handle richer data types, such as imaging and longitudinal recordings, and to increase diagnosis accuracy. Since the dataset was gathered from a single medical facility (BMC KPK, Pakistan), it lacked important sociodemographic characteristics (such as occupation, income, education, and ethnicity), which limited its generalizability and demographic variety. Therefore, it is advised to use caution when extrapolating the findings and to perform additional validation on bigger, multi-center, and more varied datasets.

Table 8. Evaluation of the proposed CKD classification model in relation to previous studies.

Study	Dataset	Model/Method	No. of Features	Accuracy (%)	Remarks
Proposed Study (2025)	BMC KPK, Pakistan	RF, SVM, Ensemble (Voting, Stacking)	21	RF: 91.58 (5-fold), 90.53 (10-fold)	Feature selection improved model robustness and interpretability
Obermeyer et al. (2016) [62]	U.S. Clinical Data	Logistic Regression	14	83.4	Early application of ML in clinical risk prediction

Table 8. *Cont.*

Study	Dataset	Model/Method	No. of Features	Accuracy (%)	Remarks
Rajkomar et al. (2019) [63]	EHR Multi-hospital Dataset	Deep Neural Network	48	90.0	Demonstrated potential of DL in hospital mortality and disease detection
Esteva et al. (2019) [64]	Diverse Health Records	Gradient Boosted Trees	30	88.7	Highlighted interpretability vs. accuracy trade-off
Beam and Kohane (2018) [65]	U.S. Health Systems	Random Forest	25	89.1	Emphasized ML integration into clinical workflow
Zhang et al. (2025) [66]	Synthetic CKD Dataset	Mamba Capsule Routing Network	32	93.2	Showed advantages of capsule routing in complex data patterns
Xia et al. (2024) [67]	Multi-source Clinical Data	Transformer-based Classifier	40	92.6	Effectively modeled temporal dependencies in patient records
Liu et al. (2024) [68]	Hospital CKD Dataset	Capsule Network	28	91.8	Improved spatial relationship modeling over CNNs

5. Conclusions

This study utilizes a publicly available clinical dataset to examine machine learning algorithms for the early detection and categorization of CKD. Eight machine learning models were evaluated using the F1 score, accuracy, sensitivity, specificity, and precision metrics. Random Forest was the most successful classifier in both 5- and 10-fold cross-validation, achieving 90.53% and 91.58% accuracy, respectively. Support Vector Machines (SVMs) achieved equal accuracy rates of 91.05% and 90.00%. Ensemble techniques showed promising results, with the Voting model achieving 89.74% accuracy and the Stacking model matching the performance of Random Forest. The results underscore the importance of data-driven methods in enhancing CKD diagnosis and supporting clinical decisions, particularly in resource-constrained settings, as the model relies on a limited number of features.

However, this work demonstrates how machine learning, specifically Random Forest and Ensemble models, can effectively predict CKD early on using a limited number of clinical variables. It provides a valuable, evidence-based foundation for future decision support systems, particularly in healthcare environments with limited resources. Early detection of CKD and resource allocation can be facilitated by integrating machine learning algorithms into primary healthcare systems. In underprivileged areas, policymakers should prioritize building digital health infrastructure to provide AI-assisted diagnostic support.

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

Funding: This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2501).

Informed Consent Statement: Not applicable.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data used in this study are available at https://www.researchgate.net/publication/372689997_Chronic_kidney_disease_patients_from_district_Buner_Khyber_Pakhtunkhwa_Pakistan (accessed on 20 January 2025).

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

References

1. Yin, L.; Kuai, M.; Liu, Z.; Zou, B.; Wu, P. Global burden of chronic kidney disease due to dietary factors. *Front. Nutr.* **2025**, *11*, 1522555. [CrossRef]
2. Zhao, X.; Zhang, Y.; Yang, Y.; Pan, J. Diabetes-related avoidable hospitalisations and its relationship with primary healthcare resourcing in China: A cross-sectional study from Sichuan Province. *Health Soc. Care Community* **2022**, *30*, e1143–e1156. [CrossRef]
3. McClellan, W.M.; Warnock, D.G.; Judd, S.; Muntner, P.; Kewalramani, R.; Cushman, M.; McClure, L.A.; Newsome, B.B.; Howard, G. Albuminuria and racial disparities in the risk for ESRD. *J. Am. Soc. Nephrol.* **2011**, *22*, 1721–1728. [CrossRef] [PubMed]
4. Hu, F.; Yang, H.; Qiu, L.; Wang, X.; Ren, Z.; Wei, S.; Zhou, H.; Chen, Y.; Hu, H. Innovation Networks in the Advanced Medical Equipment Industry: Supporting Regional Digital Health Systems from a Local–National Perspective. *Front. Public Health* **2025**, *13*, 1635475. [CrossRef]
5. Mk, H. Risk factors for chronic kidney disease: A prospective study of 23,534 men and women in Washington County, Maryland. *J. Am. Soc. Nephrol.* **2003**, *14*, 2934–2941.
6. Peng, Q.; Zhang, H.; Li, Z. Methyltransferase-Like 16 Drives Diabetic Nephropathy Progression via Epigenetic Suppression of V-Set Pre-B Cell Surrogate Light Chain 3. *Life Sci.* **2025**, *374*, 123694. [CrossRef]
7. Gerogianni, S.; Babatsikou, F.; Gerogianni, G.; Grapsa, E.; Vasilopoulos, G.; Zyga, S.; Koutis, C. 'Concerns of patients on dialysis: A Research Study'. *Health Sci. J.* **2008**, *8*, 423–437.
8. Chen, Y.; Li, L.; Chen, X.; Yan, Q.; Hu, X. The Efficacy of Decision Aids on Enhancing Early Cancer Screening: A Meta-Analysis of Randomized Controlled Trials. *Worldviews Evid.-Based Nurs.* **2025**, *22*, e70048. [CrossRef] [PubMed]
9. Chapman, J.R. What are the key challenges we face in kidney transplantation today? *Transplant. Res.* **2013**, *2*, S1. [CrossRef]
10. Cai, L.; Fang, H.; Xu, N.; Ren, B. Counterfactual Causal-Effect Intervention for Interpretable Medical Visual Question Answering. *IEEE Trans. Med. Imaging* **2024**, *43*, 4430–4441. [CrossRef]
11. Li, Q.; You, T.; Chen, J.; Zhang, Y.; Du, C. LI-EMRSQL: Linking Information Enhanced Text2SQL Parsing on Complex Electronic Medical Records. *IEEE Trans. Reliab.* **2024**, *73*, 1280–1290. [CrossRef]
12. Bing, P.; Liu, W.; Zhai, Z.; Li, J.; Guo, Z.; Xiang, Y.; Zhu, L. A novel approach for denoising electrocardiogram signals to detect cardiovascular diseases using an efficient hybrid scheme. *Front. Cardiovasc. Med.* **2024**, *11*, 1277123. [CrossRef] [PubMed]
13. Peng, Q.; Zhang, H.; Li, Z. KAT2A-Mediated H3K79 Succinylation Promotes Ferroptosis in Diabetic Nephropathy by Regulating SAT2. *Life Sci.* **2025**, *376*, 123746. [CrossRef] [PubMed]
14. Qian, K.; Bao, Z.; Zhao, Z.; Koike, T.; Dong, F.; Schmitt, M.; Dong, Q.; Shen, J.; Jiang, W.; Jiang, Y.; et al. Learning Representations from Heart Sound: A Comparative Study on Shallow and Deep Models. *Cyborg Bionic Syst.* **2024**, *5*, 0075. [CrossRef] [PubMed]
15. Xiao, Y.H.; Hu, Y.L.; Lv, X.Y.; Huang, L.J.; Geng, L.H.; Liao, P.; Ding, Y.B.; Niu, C.C. The Construction of Machine Learning-Based Predictive Models for High-Quality Embryo Formation in Poor Ovarian Response Patients with Progestin-Primed Ovarian Stimulation. *Reprod. Biol. Endocrinol. RB&E* **2024**, *22*, 78. [CrossRef]
16. Tong, G.; Wang, S.; Gao, G.; He, Y.; Li, Q.; Yang, X.; Wang, X. The Efficacy of Sulodexide Combined with Jinshuibao for Treating Early Diabetic Nephropathy Patients. *Int. J. Clin. Exp. Med.* **2020**, *13*, 8308–8317.
17. Iftikhar, H.; Khan, M.; Khan, M.S.; Khan, M. Short-term forecasting of monkeypox cases using a novel filtering and combining technique. *Diagnostics* **2023**, *13*, 1923. [CrossRef]
18. Qureshi, M.; Ishaq, K.; Daniyal, M.; Iftikhar, H.; Rehman, M.Z.; Salar, S.A. Forecasting Cardiovascular Disease Mortality Using Artificial Neural Networks in Sindh, Pakistan. *BMC Public Health* **2025**, *25*, 34. . [CrossRef]
19. Alshanbari, H.M.; Iftikhar, H.; Khan, F.; Rind, M.; Ahmad, Z.; El-Bagoury, A.A.A.H. On the implementation of the artificial neural network approach for forecasting different healthcare events. *Diagnostics* **2023**, *13*, 1310. [CrossRef]
20. Iftikhar, H.; Daniyal, M.; Qureshi, M.; Tawiah, K.; Ansah, R.K.; Afriyie, J.K. A Hybrid Forecasting Technique for Infection and Death from the Mpox Virus. *Digit. Health* **2023**, *9*, 20552076231204748. [CrossRef]
21. Cuba, W.M.; Huaman Alfaro, J.C.; Iftikhar, H.; López-Gonzales, J.L. Modeling and Analysis of Monkeypox Outbreak Using a New Time Series Ensemble Technique. *Axioms* **2024**, *13*, 554. [CrossRef]
22. Qin, J.; Chen, L.; Liu, Y.; Liu, C.; Feng, C.; Chen, B. A machine learning methodology for diagnosing chronic kidney disease. *IEEE Access* **2019**, *8*, 20991–21002. [CrossRef]
23. Vásquez-Morales, G.R.; Martinez-Monterrubio, S.M.; Moreno-Ger, P.; Recio-Garcia, J.A. Explainable prediction of chronic renal disease in the colombian population using neural networks and case-based reasoning. *IEEE Access* **2019**, *7*, 152900–152910. [CrossRef]
24. Amirgaliyev, Y.; Shamiluulu, S.; Serek, A. Analysis of chronic kidney disease dataset by applying machine learning methods. In Proceedings of the IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), Almaty, Kazakhstan, 17–19 October 2018; pp. 1–4.

25. Chen, Z.; Zhang, X.; Zhang, Z. Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models. *Int. Urol. Nephrol.* **2016**, *48*, 2069–2075. [CrossRef]
26. De Almeida, K.L.; Lessa, L.; Peixoto, A.B.; Gomes, R.L.; Celestino, J. Kidney failure detection using machine learning techniques. In Proceedings of the 8th International Workshop on Advances in ICT Infrastructures and Services (ADVANCE 2020), Cancún, Mexico, 27–29 January 2020; pp. 1–8.
27. Iftikhar, H.; Khan, M.; Khan, Z.; Khan, F.; Alshanbari, H.M.; Ahmad, Z. A comparative analysis of machine learning models: A case study in predicting chronic kidney disease. *Sustainability* **2023**, *15*, 2754. [CrossRef]
28. Gunarathne, W.; Perera, K.; Kahandawaarachchi, K. Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD). In Proceedings of the IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE), Washington, DC, USA, 23–25 October 2017; pp. 291–296.
29. Polat, H.; Mehr, H.D. Detection of Chronic Kidney Disease Using Machine Learning Algorithms. *J. Med. Syst.* **2017**, *41*, 1–11.
30. Vijayarani, S.; Dhayanand, S.; Phil, M. Kidney disease prediction using SVM and ANN algorithms. *Int. J. Comput. Bus. Res. (IJCBR)* **2015**, *6*, 1–12.
31. Almasoud, M.; Ward, T.E. Detection of chronic kidney disease using machine learning algorithms with least number of predictors. *Int. J. Soft Comput. Its Appl.* **2019**, *10*, 1–9. [CrossRef]
32. Drall, S.; Drall, G.S.; Singh, S.; Naib, B.B. Chronic kidney disease prediction using machine learning: A new approach. *Int. J. Manag.* **2018**, *8*, 278.
33. Ma, F.; Sun, T.; Liu, L.; Jing, H. Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network. *Future Gener. Comput. Syst.* **2020**, *111*, 17–26. [CrossRef]
34. Ul Haq, A.; Li, J.; Memon, M.H.; Khan, J.; Ali, Z.; Abbas, S.Z.; Nazir, S. Recognition of the Parkinson's disease using a hybrid feature selection approach. *J. Intell. Fuzzy Syst.* **2020**, *39*, 1319–1339. [CrossRef]
35. Xiao, J.; Ding, R.; Xu, X.; Guan, H.; Feng, X.; Sun, T.; Zhu, S.; Ye, Z. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *J. Transl. Med.* **2019**, *17*, 119. [CrossRef]
36. Deepika, B.; Rao, V.; Rampure, D.N.; Prajwal, P.; Gowda, D. Early prediction of chronic kidney disease by using machine learning techniques. *Am. J. Comput. Sci. Eng. Surv.* **2020**, *8*, 7.
37. Jha, V.; Garcia-Garcia, G.; Iseki, K.; Li, Z.; Naicker, S.; Plattner, B.; Saran, R.; Wang, A.Y.M.; Yang, C.W. Chronic kidney disease: Global dimension and perspectives. *Lancet* **2013**, *382*, 260–272. [CrossRef]
38. Ene-Iordache, B.; Perico, N.; Bikbov, B.; Carminati, S.; Remuzzi, A.; Perna, A.; Islam, N.; Bravo, R.F.; Aleckovic-Halilovic, M.; Zou, H.; et al. Chronic kidney disease and cardiovascular risk in six regions of the world (ISN-KDDC): A cross-sectional study. *Lancet Glob. Health* **2016**, *4*, e307–e319. [CrossRef]
39. Breiman, L.; Cutler, A.; Liaw, A.; Wiener, M. *Package 'randomForest'*; University of California, Berkeley: Berkeley, CA, USA, 2018; Volume 81, pp. 1–29.
40. Dimitriadou, E.; Hornik, K.; Leisch, F.; Meyer, D.; Weingessel, A.; Leisch, M.F. *The e1071 Package*; Misc Functions of the Department of Statistics (e1071), TU Wien: Vienna, Austria, 2006; Volume 297.
41. Wickham, H. ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* **2011**, *3*, 180–185. [CrossRef]
42. Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z.; Kenkel, B.; Team, R.C.; et al. Package 'caret'. *R J.* **2020**, *223*, 48.
43. Broatch, J.E.; Dietrich, S.; Goelman, D. Introducing data science techniques by connecting database concepts and dplyr. *J. Stat. Educ.* **2019**, *27*, 147–153. [CrossRef]
44. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd ed.; Monographs on Statistics and Applied Probability; Chapman and Hall: London, UK, 1989.
45. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer Series in Statistics; Springer: New York, NY, USA, 2009.
46. Quinlan, J.R. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
47. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
48. Vapnik, V.N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, NY, USA, 2013. [CrossRef]
49. Rish, I. *An Empirical Study of the Naive Bayes Classifier*; Technical Report RC22230 (W0110-051); IBM Research Division, IBM T.J. Watson Research Center: Yorktown Heights, NY, USA, 2001. Available online: https://faculty.cc.gatech.edu/~isbell/reading/papers/Rish.pdf (accessed on 20 January 2025).
50. Shorewala, V. Early detection of coronary heart disease using ensemble techniques. *Inform. Med. Unlocked* **2021**, *26*, 100655. [CrossRef]
51. Mahajan, P.; Uddin, S.; Hajati, F.; Moni, M.A. Ensemble learning for disease prediction: A review. *Healthcare* **2023**, *11*, 1808. [CrossRef]

52. Munyao, J.N.; Oluoch, L.A.; Iftikhar, H.; Rodrigues, P.C. Recurrent Neural Networks for Hierarchical Time Series Forecasting: An Application to the S&P 500 Market Value. *Phys. A Stat. Mech. Its Appl.* **2025**, *678*, 130869. [CrossRef]
53. Montesinos López, O.A.; Montesinos López, A.; Crossa, J. Overfitting, model tuning, and evaluation of prediction performance. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction*; Springer: Cham, Switzerland, 2022; pp. 109–139.
54. Hussain, I.; Qureshi, M.; Ismail, M.; Iftikhar, H.; Zywiołek, J.; López-Gonzales, J.L. Optimal features selection in the high dimensional data based on robust technique: Application to different health database. *Heliyon* **2024**, *10*, e37241. [CrossRef]
55. Li, H.; Jia, M.; Mao, Z. Dynamic Feature Extraction-Based Quadratic Discriminant Analysis for Industrial Process Fault Classification and Diagnosis. *Entropy* **2023**, *25*, 1664. [CrossRef]
56. Salekin, A.; Stankovic, J. Detection of chronic kidney disease and selecting important predictive attributes. In Proceedings of the 2016 IEEE International Conference on Healthcare Informatics (ICHI), Chicago, IL, USA, 4–7 October 2016; pp. 262–270.
57. Kostovska, I.; Tosheska Trajkovska, K.; Hristina, A.; Cekovska, S.; Emin, M.; Topuzovska, S.; Brezovska Kavrakova, J.; Petrushevska Stanojevska, E.; Bogdanska, J.; Labudovikj, D. Comparison of Sensitivity and Specificity Between Urinary Microalbumin and Urinary Podocyte Specific Proteins in Prediction of Secondary Nephropathies. *Clin. Chem. Lab. Med.* **2023**, *61*, S87–S2222.
58. Romagnani, P.; Remuzzi, G.; Glassock, R.; Levin, A.; Jager, K.J.; Tonelli, M.; Massy, Z.; Wanner, C.; Anders, H.J. Chronic kidney disease. *Nat. Rev. Dis. Prim.* **2017**, *3*, 17088. [CrossRef]
59. Alkhatib, L.; Diaz, L.A.V.; Varma, S.; Chowdhary, A.; Bapat, P.; Pan, H.; Kukreja, G.; Palabindela, P.; Selvam, S.A.; Kalra, K. Lifestyle modifications and nutritional and therapeutic interventions in delaying the progression of chronic kidney disease: A review. *Cureus* **2023**, *15*, e34572. [CrossRef] [PubMed]
60. Frank, H. Exploring the role of urine analysis in early detection of chronic kidney disease. *Int. J. Basic Appl. Sci.* **2023**, *12*, 33–38. [CrossRef]
61. Yin, X.; Liu, Q.; Pan, Y.; Huang, X.; Wu, J.; Wang, X. Strength of stacking technique of ensemble learning in rockburst prediction with imbalanced data: Comparison of eight single and ensemble models. *Nat. Resour. Res.* **2021**, *30*, 1795–1815. [CrossRef]
62. Obermeyer, Z.; Emanuel, E.J. Predicting the future—Big data, machine learning, and clinical medicine. *N. Engl. J. Med.* **2016**, *375*, 1216–1219. [CrossRef] [PubMed]
63. Rajkomar, A.; Dean, J.; Kohane, I. Machine learning in medicine. *N. Engl. J. Med.* **2019**, *380*, 1347–1358. [CrossRef]
64. Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. A guide to deep learning in healthcare. *Nat. Med.* **2019**, *25*, 24–29. [CrossRef] [PubMed]
65. Beam, A.L.; Kohane, I.S. Big data and machine learning in health care. *JAMA* **2018**, *319*, 1317–1318. [CrossRef]
66. Zhang, D.; Cheng, L.; Liu, Y.; Wang, X.; Han, J. Mamba capsule routing towards part-whole relational camouflaged object detection. *Int. J. Comput. Vis.* **2025**, *133*, 7201–7221. [CrossRef]
67. Xia, C.; Chen, H.; Han, J.; Zhang, D.; Li, K. Identifying children with autism spectrum disorder via transformer-based representation learning from dynamic facial cues. *IEEE Trans. Affect. Comput.* **2024**, *16*, 83–97. [CrossRef]
68. Liu, Y.; Cheng, D.; Zhang, D.; Xu, S.; Han, J. Capsule networks with residual pose routing. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 2648–2661. [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.

diagnostics

MDPI

Article

Evaluating Prediction Models with Hearing Handicap Inventory for the Elderly in Chronic Otitis Media Patients

Hee Soo Yoon [1,†], Min Jin Kim [2,3,†], Kang Hyeon Lim [1], Min Suk Kim [1], Byung Jae Kang [1], Yoon Chan Rah [1] and June Choi [1,4,*]

[1] Department of Otorhinolaryngology-Head and Neck Surgery, Korea University College of Medicine, Ansan Hospital, Ansan 15355, Republic of Korea; yhsjoa@hanmail.net (H.S.Y.); kingsonl@hanmail.net (K.H.L.); alexeric@naver.com (M.S.K.); wiagdg@gmail.com (B.J.K.); ycrah@naver.com (Y.C.R.)
[2] Department of Biostatistics, Korea University College of Medicine, Seoul 08308, Republic of Korea; aron0129@naver.com
[3] Biomedical Research Center, Korea University Ansan Hospital, Ansan 15355, Republic of Korea
[4] Department of Biomedical Informatics, College of Medicine, Korea University, Seoul 02841, Republic of Korea
* Correspondence: mednlaw@korea.ac.kr
† These authors contributed equally to this work.

Abstract: Background: This retrospective, cross-sectional study aimed to assess the functional hearing capacity of individuals with Chronic Otitis Media (COM) using prediction modeling techniques and the Hearing Handicap Inventory for the Elderly (HHIE) questionnaire. This study investigated the potential of predictive models to identify hearing levels in patients with COM. Methods: We comprehensively examined 289 individuals diagnosed with COM, of whom 136 reported tinnitus and 143 did not. This study involved a detailed analysis of various patient characteristics and HHIE questionnaire results. Logistic and Random Forest models were employed and compared based on key performance metrics. Results: The logistic model demonstrated a slightly higher accuracy (73.56%), area under the curve (AUC; 0.73), Kappa value (0.45), and F1 score (0.78) than the Random Forest model. These findings suggest the superior predictive performance of the logistic model in identifying hearing levels in patients with COM. Conclusions: Although the AUC for the logistic regression did not meet the benchmark, this study highlights the potential for enhanced reliability and improved performance metrics using a larger dataset. The integration of prediction modeling techniques and the HHIE questionnaire shows promise for achieving greater diagnostic accuracy and refining intervention strategies for individuals with COM.

Keywords: HHIE; hearing prediction; chronic otitis media; hearing level

1. Introduction

Chronic Otitis Media (COM) presents a significant clinical challenge as precise diagnosis and effective intervention require accurate assessment of functional hearing capacity. In this era of advanced technology and data-driven healthcare, the integration of prediction modeling techniques offers promise for enhancing our understanding of the hearing status of patients with COM. The Hearing Handicap Inventory for the Elderly (HHIE) questionnaire, with its insightful inquiries into the impact of hearing impairment on individuals, is a valuable tool for this assessment [1–4].

This study attempted to harness the potential of predictive models to elucidate the nuances of the hearing status of patients with COM. Patients with COM may have preconceived notions about their hearing abilities, which can bias their self-assessments. For example, they may believe that their hearing is better than it is because of their desire to avoid acknowledging their hearing problems. Moreover, patients with COM often adapt to their reduced hearing capabilities by adjusting their communication strategies, such as increasing television volumes or asking others to repeat themselves. This adaptation can

create a false sense of adequate hearing as they have developed coping mechanisms to manage their condition. Consequently, they may underestimate the extent of hearing loss during the self-assessment.

This study was motivated by the need to provide assistance to COM patients who may be unable to undergo pure-tone audiometry owing to economic, social, or structural constraints. This study explored alternative avenues for individuals facing such limitations using the HHIE questionnaire. Patients who, for various reasons, cannot access traditional diagnostic methods often find themselves excluded from comprehensive hearing assessments. To address this gap, our research contemplates the applicability of the HHIE questionnaire as a viable substitute, offering a practical solution for those for whom economic, social, or structural factors render pure-tone audiometry impractical. Previous studies have suggested a correlation between HHIE and pure-tone threshold average (PTA) [5–7]. These studies demonstrate an association between HHIE and hearing ability, proving that HHIE can serve as an objective tool. While previous studies have established the objectivity of HHIE in assessing hearing impairment, our study sought to advance this understanding by testing a new prediction model based on HHIE. This model aims to provide a more nuanced and accurate assessment of hearing status in patients with COM, potentially improving both the diagnostic process and subsequent interventions. By leveraging the strengths of HHIE in conjunction with predictive analytics, this study endeavored to create a robust tool that could bridge the gap for patients who are otherwise unable to access traditional hearing assessments, such as PTA.

Through a meticulous examination of a diverse cohort of individuals diagnosed with COM, we explored not only the quantitative aspects of their hearing but also the qualitative facets of their experiences. Several factors affect the prediction of hearing loss using the questionnaire [8–11]. We investigated the intricate interplay of variables that influence hearing outcomes, including the presence or absence of tinnitus, demographic characteristics, and the rich tapestry of responses collected through the HHIE questionnaire.

By adopting a multidimensional approach that integrates auditory biomarkers, prediction modeling techniques, and the HHIE questionnaire, this study contributes significantly to the improved assessment of serviceable hearing in individuals with COM. Through this comprehensive framework, this study enriches the understanding of hearing impairment and paves the way for more effective diagnostic and interventional strategies.

2. Materials and Methods

2.1. Patient Selection and Data Collection

In total, 289 patients with COM who underwent mastoidectomy at Korea University Ansan Hospital between September 2020 and December 2022 were included in this study. All patients had a history of surgery encompassing three types of procedures: open cavity (OC), intact canal wall (ICW), and intact-bridge mastoidectomy (IBM). IBM and OC mastoidectomy differ anatomically in the management of chronic ear diseases. IBM creates a bridge between the mastoid and external auditory canals, preserving part of the posterior canal wall. In contrast, OC completely removes the posterior canal wall, eliminating the external ear canal and establishing a direct opening to the mastoid. While both enhance aeration and drainage, IBM is suitable for less extensive mastoid diseases and maintains external ear canal integrity. OC is chosen when ongoing access and direct observation of the mastoid are crucial because of its more extensive pathology. The choice depends on the specific anatomical needs and the extent of disease in each patient [12–14].

Data collection was meticulously conducted using direct preoperative questionnaires and face-to-face interviews with each patient by an attending physician. This approach ensured a comprehensive understanding of the preoperative hearing status of each patient. Additionally, the patients completed the HHIE questionnaire after surgery. This survey contained 10 items that asked subjects to rate their hearing handicaps under specific listening conditions, such as "*When you meet a new friend, will your hearing loss make you embarrassed or uneasy?*" Each rating was scored as follows: 4 = always, 2 = sometimes,

and nil for never. A higher total score indicates a greater handicap resulting from hearing loss [15].

Individuals who lacked the ability to comprehend the questionnaire, such as those unable to communicate in Korean, or those with difficulties expressing themselves, were excluded. Additionally, individuals with a history of middle ear or hearing surgery were excluded because they could not undergo hearing tests. Individuals with incomplete postoperative follow-ups were excluded.

Given the critical need to assess the hearing status of patients with COM, especially in regions where PTA may not be readily accessible due to economic or technical limitations, this study aimed to develop a predictive model using the HHIE questionnaire. While PTA remains the gold standard for hearing assessment, our model seeks to serve as a practical alternative for screening patients in resource-limited settings and identifying those who may require more detailed evaluations. Comprehensive data collected from the patients, including their chronic conditions and environmental factors, were integral to this analysis because these variables may significantly influence hearing outcomes.

2.2. Outcome and Other Variables

The main objective of this study was to demonstrate the correlation between HHIE and PTA using novel prediction models. Within the dataset, we designated the term "better ear" to signify the ear with a lower threshold between the left and right ears on the PTA. For the analysis, we considered a "significant event (not better ear, hearing loss)" to have occurred if the PTA was over 25 dB (the average of the hearing threshold levels at a set of four frequencies: 500, 1000, 2000, and 3000 Hz); otherwise, we classified it as "normal". We also provided descriptions of the participants' characteristics pertaining to their history of smoking, hypertension, diabetes mellitus, hyperlipidemia, stroke, chronic kidney disease, and myocardial infarction.

2.3. Statistical Analysis and Modeling

We applied both logistic and Random Forest models, and for each approach, we established metrics including accuracy, area under the curve (AUC), Kappa, and F1 score. These data were subsequently divided into training and test sets in a 7:3 ratio. 'Accuracy' illustrates the model's capability to differentiate between normal and abnormal groups, while AUC offers insight into the model's ability to predict "significant events". The Kappa and F1 scores evaluated the extent of agreement between the two groups, highlighting the degree to which similar values were observed by chance. Higher values of accuracy, AUC, and F1 score indicate enhanced performance, whereas a lower Kappa value suggests a comparatively lower agreement. Our initial steps involved the selection of pertinent variables for the baseline model using a stepwise logistic regression method, which aided in identifying significant variables. In the Random Forest model, the selected hyperparameters were the number of trees and the number of variables randomly sampled as candidates for each split. For the logistic regression model, hyperparameter tuning was conducted using penalized maximum likelihood estimation (MLE) to optimize the penalty parameter. After evaluating the coefficient estimates across different values of the tuning parameter λ, we employed 10-fold cross-validation to identify the optimal λ that minimized the mean-square error. Through this process, we conducted a comparison between actual values (labeled as "Reference") and predicted values (labeled as "Prediction"), leading to the computation of accuracy, AUC, Kappa, and F1 score values for both models. This approach allowed us to comprehensively evaluate the model's performance and predictive capabilities. R version 4.0.3 (http://www.R-project.org) was used for the statistical analyses.

3. Results

3.1. Patient Baseline Characteristics

Table 1 presents the fundamental characteristics of patients with COM, stratified based on their hearing levels, categorized as the "better ear" and the "not better ear". The

differentiation criteria between the two categories were determined using PTA levels. An audiometric value exceeding 25 dB was classified as the "not better ear", while a value below 25 dB was categorized as the "better ear". The cohort comprised 111 and 178 patients by hearing level, respectively. Notably, a greater percentage of patients did not report hypertension, diabetes mellitus, hyperlipidemia, stroke, or chronic kidney disease between the two groups. Figure 1 illustrates the results derived from the HHIE questionnaire used in this study. Notably, this figure highlights that among the questionnaire items, question E-4—"Does a hearing problem make you irritable?"—garnered a notably higher percentage of 'yes' responses than the other questions. Similarly, question S-8—"Do you have difficulty hearing when someone speaks in a whisper?"—elicited a greater proportion of "sometimes" responses compared to the other questions.

Table 1. Baseline characteristics of the study population stratified by hearing level.

Characteristics	Level	Not Better Ear (>25 dB)	Better Ear (≤25 dB)	SMD
Total Number		111	178	
Age (mean (SD))		61.81 (10.52)	49.31 (15.36)	0.95
Height (mean (SD))		159.29 (8.73)	163.73 (9.09)	0.498
Weight (mean (SD))		62.54 (10.65)	65.44 (14.09)	0.232
BMI (mean (SD))		24.60 (3.53)	24.28 (4.18)	0.083
Sex (%)	Female	63 (56.8)	98 (55.1)	0.034
	Male	48 (43.2)	80 (44.9)	
Surgery Method	OC	49 (44.1)	64 (36.0)	0.219
	ICW	59 (53.2)	105 (59.0)	
	IBM	1 (0.9)	1 (0.6)	
	Other	2 (1.8)	8 (4.5)	
Surgery Direction (%)	Left	56 (50.5)	87 (48.9)	0.031
	Right	55 (49.5)	91 (51.1)	
Tinnitus (%)	No	46 (41.4)	97 (54.5)	0.264
	Yes	65 (58.6)	81 (45.5)	
Tinnitus Location (%)	Both	16 (14.4)	7 (3.9)	0.42
	Left	27 (24.3)	35 (19.7)	
	None	46 (41.4)	98 (55.1)	
	Right	22 (19.8)	38 (21.3)	
Smoker (%)	Current Smoker	15 (13.5)	33 (18.5)	0.255
	Non-smoker	87 (78.4)	120 (67.4)	
	Ex-smoker	9 (8.1)	25 (14.0)	
HTN (%)	Yes	46 (41.4)	38 (21.3)	0.443
	No	65 (58.6)	140 (78.7)	
DM (%)	Yes	10 (9.0)	21 (11.8)	0.091
	No	101 (91.0)	157 (88.2)	
Hyperlipidemia (%)	Yes	11 (9.9)	30 (16.9)	0.205
	No	100 (90.1)	148 (83.1)	
Stroke (%)	Yes	2 (1.8)	0 (0.0)	0.192
	No	109 (98.2)	178 (100.0)	
CKD (%)	Yes	2 (1.8)	3 (1.7)	0.009
	No	109 (98.2)	175 (98.3)	

Body Mass Index, BMI; hypertension, HTN; diabetes mellitus, DM; chronic kidney disease, CKD. Standard mean deviation, SMD.

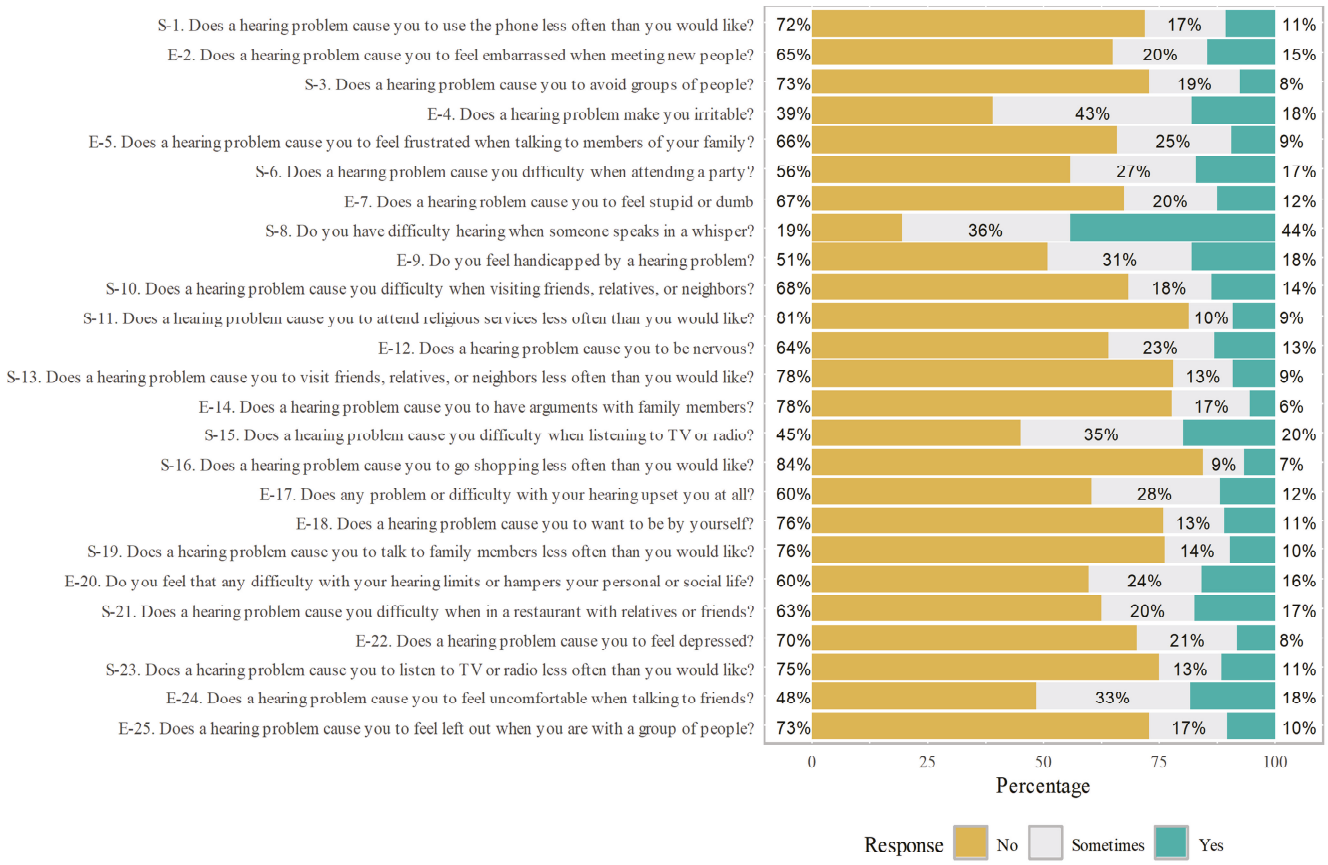


Figure 1. Illustration of the outcomes of the Hearing Handicap Inventory for the Elderly questionnaire in the patient population.

3.2. Comparison of the Logistic Model and Random Forest Model

Prior to model fitting, pertinent variables were selected from the HHIE questionnaire. Using stepwise logistic regression, we identified significant variables, as summarized in Table 2.

Table 2. Significant questionnaire items of Hearing Handicap Inventory for the Elderly (HHIE).

HHIE Questionnaires
E-4. Does a hearing problem make you irritable?
E-5. Does a hearing problem cause you to feel frustrated when talking to members of your family?
E-14. Does a hearing problem cause you to have arguments with family members?
E-24. Does a hearing problem cause you to feel uncomfortable when talking to friends?
E-25. Does a hearing problem cause you to feel left out when you are with a group of people?
S-6. Does a hearing problem cause you difficulty when attending a party?
S-8. Do you have difficulty hearing when someone speaks in a whisper?
S-11. Does a hearing problem cause you to attend religious services less often than you would like?
S-15. Does a hearing problem cause you difficulty when listening to TV or radio?
S-16. Does a hearing problem cause you to go shopping less often than you would like?
S-21. Does a hearing problem cause you difficulty when in a restaurant with relatives or friends?
S-23. Does a hearing problem cause you to listen to TV or radio less often than you would like?

The confusion matrices of the logistic and Random Forest models for the test set are shown (Figure 2). When conducting a comparison between the actual values, referred to as "Reference", and the anticipated values, denoted as "Predicted", the logistic model showcased its prowess by successfully predicting 24 out of 36 events (Figure 2A). In contrast, the Random Forest model accurately predicted 22 of the 34 events (Figure 2B).

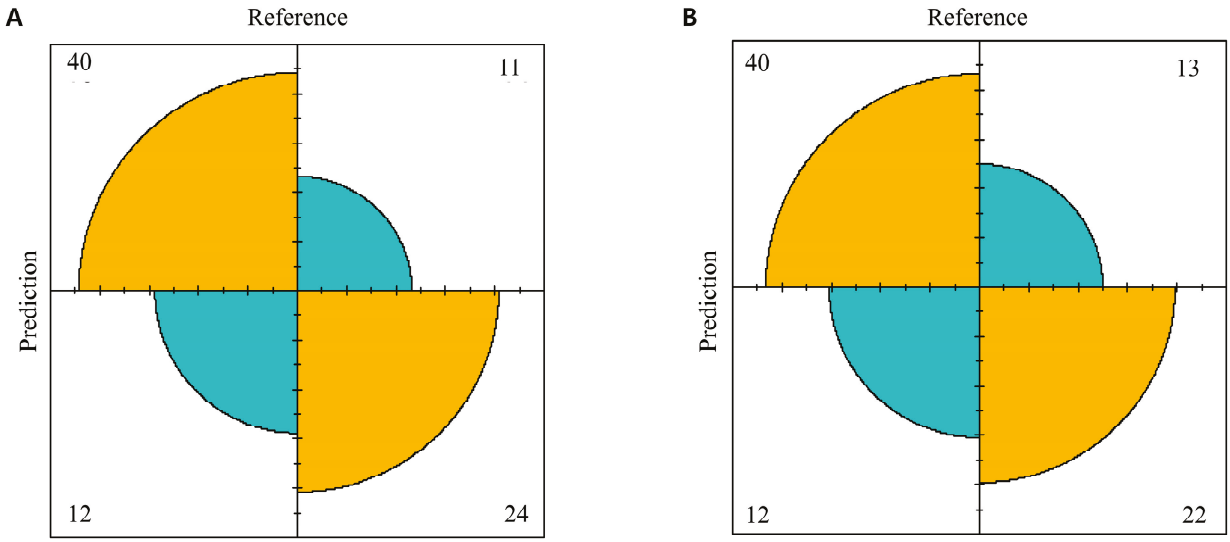


Figure 2. The confusion matrices of the logistic and Random Forest models for the test set. (**A**) The logistic model showcased its prowess through successful predictions. (**B**) The Random Forest model exhibited predictive capability through accurate predictions.

In the Random Forest method, we used the number of trees and variables randomly sampled as candidates for each split as the hyperparameters. Although the default number of trees is typically set to 500, we reduced this to 100 because of the small sample size to prevent overfitting. For the number of variables randomly sampled at each split, we used the square root of the total number of variables, resulting in three variables being used at each split. The Out-of-Bag (OOB) estimate of the error rate was 24.88%.

For the logistic regression model, we optimized the hyperparameters using penalized MLE to adjust the penalty. After evaluating the coefficient estimates across different values of the tuning parameter λ, we conducted 10-fold cross-validation to determine the optimal λ that minimized the mean-square error. The log(λ) value of approximately −3.45 yielded the smallest error, and this value was subsequently used to fit the model and evaluate the performances. This is shown in Figure 3.

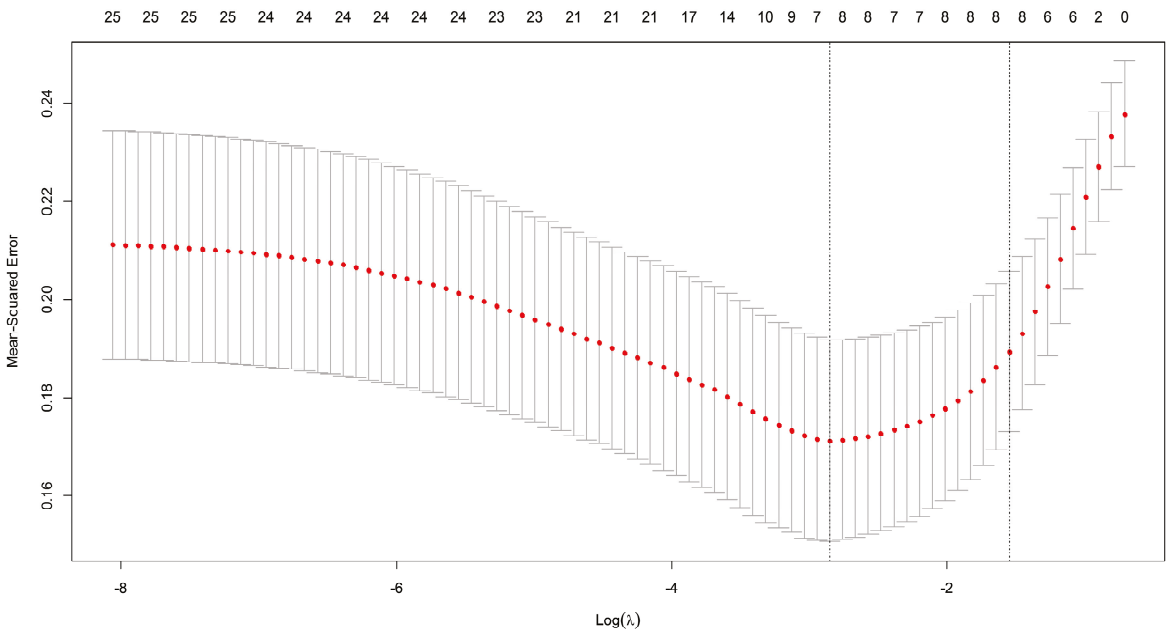


Figure 3. Mean-Squared Error plot according to log after cross-validation in logistic regression. This plot illustrates that the log(λ) around −3.45 provided the smallest mean-square error.

A comprehensive overview of the statistical metrics for the logistic and the Random Forest models is presented (Table 3). This table summarizes the performance of each model using key indicators, including accuracy, AUC, Kappa, and F1 scores.

Table 3. Statistical parameters of the logistic model and Random Forest model.

Model	Accuracy (95%CI)	Sensitivity (%)	Specificity (%)	AUC	Kappa	F1
Logistic	73.56% (63.02–82.45%)	76.92%	68.57%	0.73	0.45	0.78
Random Forest	71.26% (60.57–80.46%)	75.00%	62.85%	0.70	0.38	0.75

The ROC curves of the logistic and Random Forest models are shown (Figure 4). In the logistic model, the accuracy was 73.56% (95% confidence interval [CI]: 63.02–82.45%), signifying the proportion of correctly predicted outcomes. In addition, the AUC was 0.73, indicating the capability of the model to discriminate between positive and negative cases. A Kappa value of 0.45 signified agreement beyond chance, implying a moderate level of concordance between the observed and predicted outcomes. The F1 score, a blend of precision and recall, achieved a value of 0.78, reflecting the balance of the model between accurate positive predictions and the avoidance of false positives.

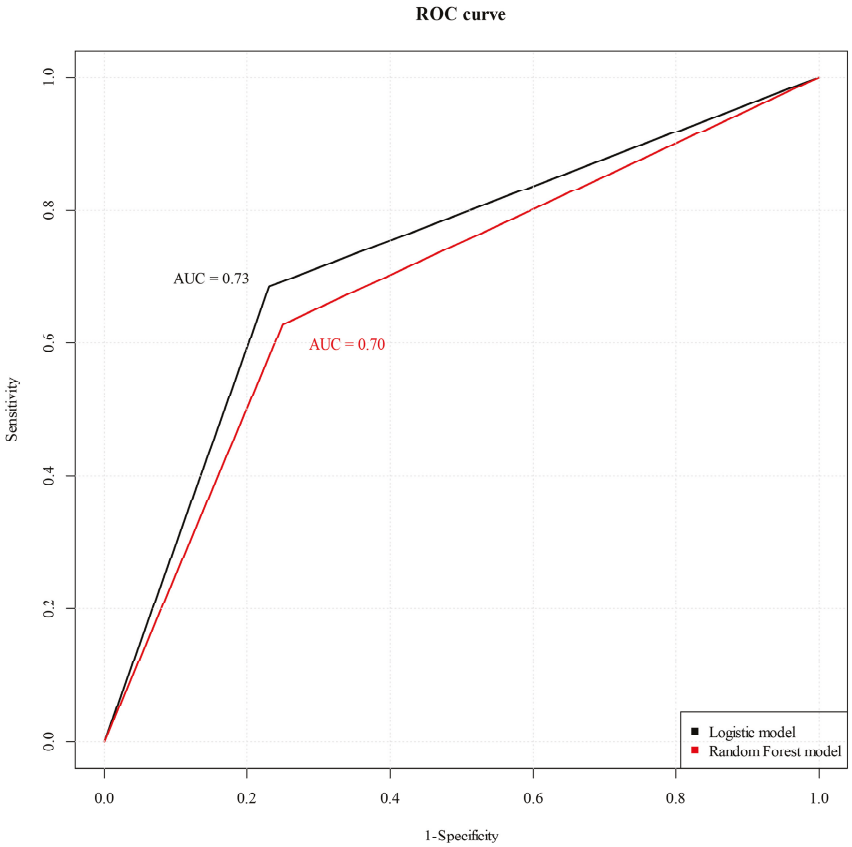


Figure 4. The ROC curve of the logistic model and the Random Forest model. In the logistic model, the AUC was 0.73, indicating the capability of the model to discriminate between positive and negative cases. In contrast, the Random Forest model showed that the AUC was 0.70, indicating commendable discrimination ability.

In contrast, the Random Forest model showed an accuracy of 71.26% (95% CI [60.57–80.46%]), which was slightly lower than that of the logistic model. The AUC was 0.70, indicating commendable discrimination ability. The Kappa value (0.38) suggested a moderate level of agreement, although it was lower than that of the logistic model. The F1 score was 0.75, indicating harmonized performance in terms of precision and recall.

The results indicated that the logistic model demonstrated a slightly higher accuracy, AUC, Kappa value, and F1 score than the Random Forest model. This suggests that the logistic model has a stronger predictive performance, encompassing both its ability to differentiate between classes (AUC) and its agreement measure (Kappa), while also excelling at balancing precision and recall (F1 score), when compared with the Random Forest model.

4. Discussion

This study makes a useful contribution to the ongoing exploration of HHIE for predicting hearing levels without audiograms. This innovative approach has promising practical implications, particularly for individuals who face challenges in accessing audiological services, such as those unable to afford visits to ENT clinics or those residing in resource-constrained regions. By harnessing the potential of the HHIE questionnaire to predict hearing levels, we offer a cost-effective and accessible alternative that can aid in the early identification of hearing impairments and guide appropriate interventions. This not only extends the reach of audiological assessments to a wider population, but also underscores the importance of leveraging existing tools and methodologies to address the pressing healthcare needs of diverse communities.

Previous studies have acknowledged that the HHIE questionnaire is a potent tool for assessing the impact of hearing impairment on an individual's quality of life, underscoring its emotional and social dimensions beyond conventional audiometric measurements [7,16]. Our incorporation of the HHIE questionnaire in this study is part of a broader effort to comprehensively understand the subjective experiences of patients with COM and their connection to objective hearing levels.

Drawing on existing research, we extracted valuable insights into the significance of this study and its implications for clinical practice [17,18]. Our study highlights the intricate relationship between data volume and predictive performance metrics, which is a growing concern in otolaryngology. In recent years, prediction modeling techniques such as machine learning have gained traction for predicting hearing loss across diverse domains. For instance, predictive models have successfully estimated hearing loss among workers exposed to complex industrial noise, as evidenced in prior studies [19]. Furthermore, machine learning models have been developed to predict hearing outcomes in cases of idiopathic sudden sensorineural hearing loss, highlighting the expanding applications of machine learning in audiology [20–22]. The application of machine learning in predicting noise-induced hearing loss has also been explored, highlighting its growing relevance to hearing-related research [23]. Recent investigations have focused on predicting hearing outcomes following tympanoplasty, indicating its utility in clinical scenarios [24]. In addition to hearing loss prediction, the potential of prediction modeling techniques such as machine learning for diagnosing ear diseases has gained prominence in recent studies [25–29].

Our study emphasizes the importance of recognizing that in situations with limited data availability, the full potential of predictive performance metrics may remain underutilized. One potential limitation is the reliance on a relatively modest dataset of 289 patients with COM. This study acknowledges the importance of data volume in predictive modeling, and the observed limitations in data availability may affect the full realization of predictive performance metrics. A larger and more diverse dataset could enhance the accuracy and reliability of the predictive model, providing a more robust foundation for the integration of simple machine learning techniques with the HHIE questionnaire. Additionally, the study acknowledges the potential of dataset augmentation to address this limitation, but the extent to which this can be achieved and its impact on model improvement remain areas for further exploration and consideration.

The focal point of our research, the relationship between data volume and predictive performance metrics, aligns with earlier findings, highlighting the critical role of robust datasets in predictive modeling. Our study reinforces the need for larger and more diverse datasets to fully reveal the predictive potential of prediction models using simple machine

learning techniques, aligning with previous research demonstrating the value of extensive data in enhancing the accuracy and reliability of predictive models.

5. Conclusions

Our study introduces a pioneering approach to predict hearing levels in patients with COM using the HHIE questionnaire and machine learning techniques. The integration of the HHIE with logistic regression and Random Forest models revealed notable findings regarding the predictive performance of hearing outcomes.

The results highlight that the logistic regression model outperformed the Random Forest model across key performance indicators, including accuracy, AUC, Kappa, and F1 score. This suggests that the logistic regression model is more effective in predicting hearing levels, balancing precision and recall, and achieving a higher agreement between the observed and predicted outcomes.

Although the logistic regression model's AUC fell slightly short of the desired benchmark, future efforts to expand and diversify the dataset present opportunities to further enhance the model's performance. This advancement may lead to more reliable predictive capabilities and effective clinical assessment tools for patients with COM.

However, it is important to acknowledge the potential biases inherent in our approach. The selection bias introduced by the specific patient population may limit the generalizability of our findings. Additionally, the use of self-reported data from the HHIE questionnaire introduces the risk of measurement bias, which could affect the accuracy of our predictions. Furthermore, model bias stemming from the assumptions of the logistic regression model and complexity of the Random Forest model may have affected the interpretability and reliability of the results. These biases underscore the need for a cautious interpretation of our findings and highlight the importance of future research to mitigate these limitations. By expanding our dataset, incorporating more diverse patient populations, and exploring advanced modeling techniques, we can further enhance the robustness and applicability of our predictive models. As we continue to explore the untapped potential of the HHIE, addressing these biases will be critical for developing a more inclusive and effective approach to hearing level assessment. This will ultimately benefit individuals globally and foster greater equity in healthcare access, paving the way for a more precise and comprehensive assessment of hearing in patients with COM.

In summary, our study successfully illustrated the potential of the HHIE questionnaire combined with machine learning techniques to predict hearing levels without relying on traditional audiograms. These results highlight the innovative nature of our approach and its potential for advancing the field of otology. Continued exploration and refinement of this method will contribute to more inclusive and effective hearing assessments, benefiting patients globally and promoting greater equity in healthcare access.

Author Contributions: H.S.Y.: conceptualization, resources, investigation, and writing—original draft. M.J.K.: conceptualization of resources, investigation, and writing—original draft. K.H.L.: data curation, software, visualization, and formal analysis. M.S.K.: data curation, software, visualization, and formal analysis. B.J.K.: data curation, software, visualization, and formal analysis. Y.C.R.: conceptualization, resources, investigation, and writing—original draft. J.C.: conceptualization, resources, investigation, and writing—original draft. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by (1) the Korea Medical Device Development Fund funded by the Korean government (Ministry of Science and ICT, Industry and Energy, Health & Welfare, and Food and Drug Safety) (RS-2024-00255350); (2) the Ansan-Si hidden champion fostering and supporting project funded by Ansan City; (3) a Grant from the Medical data-driven hospital support project through the Korea Health Information Service (KHIS), funded by the Ministry of Health and Welfare, Republic of Korea; and (4) the MSIT (Ministry of Science and ICT), Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) program (IITP-2024-RS-2022-00156439) supervised by the IITP (Institute of Information and Communications Technology Planning and Evaluation).

Institutional Review Board Statement: This study was approved by the Institutional Review Board of Korea University Ansan Hospital (IRB No. 2024AS0081). The requirement for informed consent was waived by the ethics committee and the Institutional Review Board of Korea University Ansan Hospital. All methods were performed in accordance with relevant guidelines and regulations.

Informed Consent Statement: The requirement for informed consent was waived by the ethics committee.

Data Availability Statement: The dataset used in this study is not publicly available. However, the data in this study can be obtained upon reasonable request from the corresponding author.

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

References

1. Weinstein, B.E.; Spitzer, J.B.; Ventry, I.M. Test-retest reliability of the Hearing Handicap Inventory for the Elderly. *Ear Hear.* **1986**, *7*, 295–299. [CrossRef] [PubMed]
2. Newman, C.W.; Weinstein, B.E. Test-retest reliability of the Hearing Handicap Inventory for the Elderly using two administration approaches. *Ear Hear.* **1989**, *10*, 190–191. [CrossRef] [PubMed]
3. Lichtenstein, M.J.; Bess, F.H.; Logan, S.A. Validation of screening tools for identifying hearing-impaired elderly in primary care. *JAMA* **1988**, *259*, 2875–2878. [CrossRef] [PubMed]
4. Whittemore, K.R., Jr.; Merchant, S.N.; Rosowski, J.J. Acoustic mechanisms: Canal wall-up versus canal wall-down mastoidectomy. *Otolaryngol. Head Neck Surg.* **1998**, *118*, 751–761. [CrossRef] [PubMed]
5. Camarudin, N.; Ahmad, S.A.; Minhat, H.S.; Mohamed, M.H.; Adnan, R.N.E.R. Correlation between Hearing Handicap Inventory for the Elderly Screening (HHIE-S) and Pure Tone Audiometry (PTA) Test among Malaysian Elderly. *Malays. J. Med. Health Sci.* **2022**, *18*, 94–98.
6. Wang, Y.; Mo, L.; Li, Y.; Zheng, Z.; Qi, Y. Analysing use of the Chinese HHIE-S for hearing screening of elderly in a northeastern industrial area of China. *Int. J. Audiol.* **2017**, *56*, 242–247. [CrossRef]
7. Servidoni, A.B.; Conterno, L.O. Hearing Loss in the Elderly: Is the Hearing Handicap Inventory for the Elderly—Screening Version Effective in Diagnosis When Compared to the Audiometric Test? *Int. Arch. Otorhinolaryngol.* **2018**, *22*, 1–8. [CrossRef]
8. McBride, D. Hearing conservation in the mining industry. Evaluation of a risk factor questionnaire. *Occup. Med.* **1993**, *43*, 185–192. [CrossRef]
9. Saunders, G.H.; Cienkowski, K.M. Refinement and psychometric evaluation of the Attitudes Toward Loss of Hearing Questionnaire. *Ear Hear.* **1996**, *17*, 505–519. [CrossRef]
10. Burr, H.; Lund, S.P.; Sperling, B.B.; Kristensen, T.S.; Poulsen, O.M. Smoking and height as risk factors for prevalence and 5-year incidence of hearing loss. A questionnaire-based follow-up study of employees in Denmark aged 18–59 years exposed and unexposed to noise. *Int. J. Audiol.* **2005**, *44*, 531–539. [CrossRef]
11. Demeester, K.; Topsakal, V.; Hendrickx, J.J.; Fransen, E.; van Laer, L.; Van Camp, G.; Van de Heyning, P.; van Wieringen, A. Hearing disability measured by the speech, spatial, and qualities of hearing scale in clinically normal-hearing and hearing-impaired middle-aged persons, and disability screening by means of a reduced SSQ (the SSQ5). *Ear Hear.* **2012**, *33*, 615–616. [CrossRef] [PubMed]
12. Bennett, M.; Warren, F.; Haynes, D. Indications and technique in mastoidectomy. *Otolaryngol. Clin. N. Am.* **2006**, *39*, 1095–1113. [CrossRef] [PubMed]
13. Haynes, D.S. Surgery for chronic ear disease. *Ear Nose Throat J.* **2001**, *80*, 8–11. [PubMed]
14. Sheehy, J.L. Mastoidectomy: The Intact Canal Wall Procedure. In *Otologic Surgery*; WB Saunders: Philadelphia, PA, USA, 1994; Chapter 18; pp. 212–224.
15. Lichtenstein, M.J.; Bess, F.H.; Logan, S.A. Diagnostic performance of the hearing handicap inventory for the elderly (screening version) against differing definitions of hearing loss. *Ear Hear.* **1988**, *9*, 208–211. [CrossRef] [PubMed]
16. Tomioka, K.; Ikeda, H.; Hanaie, K.; Morikawa, M.; Iwamoto, J.; Okamoto, N.; Saeki, K.; Kurumatani, N. The Hearing Handicap Inventory for Elderly-Screening (HHIE-S) versus a single question: Reliability, validity, and relations with quality of life measures in the elderly community, Japan. *Qual. Life Res.* **2013**, *22*, 1151–1159. [CrossRef]
17. Newman, C.W.; Weinstein, B.E. The Hearing Handicap Inventory for the Elderly as a measure of hearing aid benefit. *Ear Hear.* **1988**, *9*, 81–85. [CrossRef]
18. Newman, C.W.; Weinstein, B.E.; Jacobson, G.P.; Hug, G.A. The Hearing Handicap Inventory for Adults: Psychometric adequacy and audiometric correlates. *Ear Hear.* **1990**, *11*, 430–433. [CrossRef]
19. Zhao, Y.; Li, J.; Zhang, M.; Lu, Y.; Xie, H.; Tian, Y.; Qiu, W. Machine Learning Models for the Hearing Impairment Prediction in Workers Exposed to Complex Industrial Noise: A Pilot Study. *Ear Hear.* **2019**, *40*, 690–699. [CrossRef]
20. Park, K.V.; Oh, K.H.; Jeong, Y.J.; Rhee, J.; Han, M.S.; Han, S.W.; Choi, J. Machine Learning Models for Predicting Hearing Prognosis in Unilateral Idiopathic Sudden Sensorineural Hearing Loss. *Clin. Exp. Otorhinolaryngol.* **2020**, *13*, 148–156. [CrossRef]
21. Bing, D.; Ying, J.; Miao, J.; Lan, L.; Wang, D.; Zhao, L.; Yin, Z.; Yu, L.; Guan, J.; Wang, Q. Predicting the hearing outcome in sudden sensorineural hearing loss via machine learning models. *Clin. Otolaryngol.* **2018**, *43*, 868–874. [CrossRef]

22. Uhm, T.; Lee, J.E.; Yi, S.; Choi, S.W.; Oh, S.J.; Kong, S.K.; Lee, I.W.; Lee, H.M. Predicting hearing recovery following treatment of idiopathic sudden sensorineural hearing loss with machine learning models. *Am. J. Otolaryngol.* **2021**, *42*, 102858. [CrossRef] [PubMed]
23. Chen, F.; Cao, Z.; Grais, E.M.; Zhao, F. Contributions and limitations of using machine learning to predict noise-induced hearing loss. *Int. Arch. Occup. Environ. Health* **2021**, *94*, 1097–1111. [CrossRef] [PubMed]
24. Koyama, H.; Kashio, A.; Uranaka, T.; Matsumoto, Y.; Yamasoba, T. Application of Machine Learning to Predict Hearing Outcomes of Tympanoplasty. *Laryngoscope* **2023**, *133*, 2371–2378. [CrossRef] [PubMed]
25. Crowson, M.G.; Hartnick, C.J.; Diercks, G.R.; Gallagher, T.Q.; Fracchia, M.S.; Setlur, J.; Cohen, M.S. Machine Learning for Accurate Intraoperative Pediatric Middle Ear Effusion Diagnosis. *Pediatrics* **2021**, *147*, e2020034546. [CrossRef] [PubMed]
26. Wang, Y.M.; Li, Y.; Cheng, Y.S.; He, Z.Y.; Yang, J.M.; Xu, J.H.; Chi, Z.C.; Chi, F.L.; Ren, D.D. Deep Learning in Automated Region Proposal and Diagnosis of Chronic Otitis Media Based on Computed Tomography. *Ear Hear.* **2020**, *41*, 669–677. [CrossRef]
27. Zeng, J.; Kang, W.; Chen, S.; Lin, Y.; Deng, W.; Wang, Y.; Chen, G.; Ma, K.; Zhao, F.; Zheng, Y.; et al. A Deep Learning Approach to Predict Conductive Hearing Loss in Patients with Otitis Media with Effusion Using Otoscopic Images. *JAMA Otolaryngol. Head Neck Surg.* **2022**, *148*, 612–620. [CrossRef]
28. Byun, H.; Yu, S.; Oh, J.; Bae, J.; Yoon, M.S.; Lee, S.H.; Chung, J.H.; Kim, T.H. An Assistive Role of a Machine Learning Network in Diagnosis of Middle Ear Diseases. *J. Clin. Med.* **2021**, *10*, 3198. [CrossRef]
29. Viscaino, M.; Maass, J.C.; Delano, P.H.; Torrente, M.; Stott, C.; Auat Cheein, F. Computer-aided diagnosis of external and middle ear conditions: A machine learning approach. *PLoS ONE* **2020**, *15*, e0229226. [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.

diagnostics

MDPI

Article

An Explainable Ensemble and Deep Learning Framework for Accurate and Interpretable Parkinson's Disease Detection from Voice Biomarkers

Suliman Aladhadh

Department of Information Technology, College of Computer, Qassim University, Buraydah 52571, Saudi Arabia; s.aladhadh@qu.edu.sa

Abstract: Background: Parkinson's disease (PD) is a degenerative neurological disorder that greatly affects motor and speech functions; therefore, early diagnosis is vital for improving patients' quality of life. This work introduces a unified and explainable AI framework for PD detection that integrates ensemble and deep learning models with transparent interpretability techniques. **Methods**: Acoustic features were extracted from the Parkinson's Voice Disorder Dataset, and a broad suite of machine learning and deep learning models was evaluated, including traditional classifiers (Logistic Regression, Decision Tree, KNN, Linear Regression, SVM), ensemble methods (Random Forest, Gradient Boosting, XGBoost, LightGBM), and neural architectures (CNN, LSTM, GAN). **Results**: The ensemble methods—specifically LightGBM (LGBM) and Random Forest (RF)—achieved the best performance, reaching state-of-the-art accuracy (98.01%) and ROC-AUC (0.9914). Deep learning models like CNN and GAN produced competitive results, validating their ability to capture nonlinear and generative voice patterns. XAI analysis revealed that nonlinear acoustic biomarkers such as spread2, PPE, and RPDE are the most influential predictors, consistent with clinical evidence of dysphonia in PD. **Conclusions**: The proposed framework achieves a strong balance between predictive accuracy and interpretability, representing a clinically relevant, scalable, and non-invasive solution for early Parkinson's detection.

Keywords: Parkinson's disease; voice biomarkers; machine learning; deep learning; ensemble methods; explainable AI (XAI); diagnosis; clinical decision-making; healthcare AI

1. Introduction

The incorporation of artificial intelligence (AI) in medicine has revolutionized the face of medical diagnosis and disease monitoring [1]. Machine learning (ML) and deep learning (DL) algorithms, specifically, have shown remarkable promise in processing complex biomedical signals, images, and speech data for early disease diagnosis and decision support. Of the neurological conditions, Parkinson's disease (PD) has attracted considerable research interest with its growing prevalence and the possibility of AI helping with early diagnosis using non-invasive modalities [2]. Voice analysis has come forward as a viable technology, since acoustic anomalies in speech tend to occur at the early levels of Parkinson's and can be observed as quantifiable biomarkers that can be utilized by classification systems based on AI [3].

While recent AI-based works have shown promising accuracy using speech signals for Parkinson's screening, most of the existing literature either focuses on traditional

machine learning or deep learning alone, with very limited emphasis on clinical interpretability. Moreover, prior methods typically report accuracy without explaining which acoustic biomarkers drive the prediction, thus creating barriers to clinical adoption and trust. Therefore, this research aims to develop a unified framework that not only improves the detection accuracy but also provides transparent explanations aligned with medically validated vocal symptoms of PD. Hence, the main objectives of this study are: (i) comparative evaluation of traditional ML, ensemble models, and deep neural architectures on voice-based PD classification, (ii) integration of global and local explainable AI methods for transparent decision-making, and (iii) ablation study to validate the importance of nonlinear voice biomarkers. High-performing models are combined with clinically meaningful interpretability in order to bridge the gap between black-box AI systems and usable diagnostic decision support.

Although great strides have been made, voice analysis for Parkinson's detection comes with some challenges [3,4]. Acoustic properties of voice signals are nonlinear, highly random, and prone to background noise, which makes it challenging to model using traditional methods. Conventional classifiers like logistic regression and decision trees tend to be inadequate in capturing these nuances and hence perform suboptimally. Additionally, the absence of interpretability of complex black-box models poses challenges to clinical uptake, as clinicians need transparency in decision-making [5,6]. Yet another challenge is the ability to separate clinically important acoustic variations from non-clinically relevant noise, which may confuse predictive models. The work in this paper overcomes these limitations by integrating ensemble learning, deep learning, and explainable AI (XAI) methods to attain high performance coupled with interpretability in Parkinson's disease classification.

The feasibility of identifying Parkinson's accurately and explainably from voice characteristics has extensive implications and uses [4,7]. Clinically, AI-powered tools can aid neurologists in early diagnosis, disease progression monitoring, and assessing treatment outcomes. From a societal standpoint, scalable and non-invasive screening solutions can drive healthcare savings as well as enhance patient quality of life through earlier interventions. In addition, interpretable models facilitate more trust among clinicians while closing the gap between computational innovation and realistic healthcare rollouts [8,9].

In this paper, we introduce an inclusive framework utilizing classical ML, ensemble models, and deep learning structures for the detection of Parkinson's disease based on acoustic features from the Parkinson's Voice Disorder Dataset. We perform a rigorous evaluation of a variety of classifiers—such as Logistic Regression, Decision Tree, Random Forest (RF), Gradient Boosting (GB), XGBoost (XGB), LightGBM (LGBM), CNN, LSTM, and GAN—to compare performance across different paradigms. Further, we use explainable AI techniques like feature importance, SHAP, LIME, and Grad-CAM to increase interpretability and conduct an ablation study to measure the effect of nonlinear features and model families on classification accuracy.

The goal of this paper is two-fold: firstly, to examine the relative efficacy of various ML and DL methods in identifying Parkinson's-correlated voice defects; and secondly, to merge XAI techniques that yield transparency at both global and local levels with a view to ensuring clinical applicability and credibility.

The major contentions of this paper are as follows: it presents an extensive comparative assessment of conventional machine learning, ensemble, and deep models for the detection of Parkinson's from voice features; includes explainable AI techniques like feature importance, SHAP, LIME, and Grad-CAM to improve both global and patient-level interpretability; performs ablation studies to evaluate the relevance of nonlinear acoustic features and various model families; and lastly, develops a unified framework that main-

tains state-of-the-art accuracy while being clinically transparent, thus ideally fit for use in decision support systems.

The rest of the paper is structured as follows: Section 2 summarizes existing literature on AI-driven Parkinson's detection. Section 3 describes the study design, dataset, and approach. Section 4 reports the experimental results, interpretability analyses, and ablation study. Section 5 concludes the paper and presents possible avenues for future research.

Novel Contribution

In contrast to existing work applying separately traditional ML, deep learning, or explainable AI to Parkinson's voice data, our work presents for the first time an integrated ensemble–deep learning–XAI framework that jointly optimizes accuracy and clinical interpretability in one workflow. Through integration, cross-validation is enabled between model explainability (SHAP, LIME, Grad-CAM) and clinical biomarkers, thus establishing a methodological bridge between predictive performance and medical interpretability. Further, the addition of an ablation analysis that measures the impact of nonlinear and entropy-based acoustic features (spread2, PPE, RPDE) on model performance is a new empirical contribution that bridges computational metrics with well-established clinical markers of dysphonia. Clinically, this work presents one of the initial quantitative validations that interpretable ensemble approaches can detect the identical voice biomarkers neurologists currently trust, corroborating translational promise for real-world diagnostic assistance.

2. Literature Study

Khachnaoui et al. [10] compared computer-aided diagnostic techniques for Parkinson's disease (PD) employing both ML and DL strategies and also reviewed imaging modalities such as single photon emission computed tomography (SPECT) and positron emission tomography (PET) for diagnosis. They emphasized the disadvantage of old-fashioned hand-crafted ML techniques and determined that DL algorithms offer a more robust and efficient method for feature extraction in PD diagnosis. Equivalently, Salari et al. [11] carried out an SLR to investigate the performance of ML models for PD detection through 2020. Their analysis involved seven phases of categorizing numerous methodologies and databases and providing statistical inferences, finally proposing the practical application of ML for PD diagnosis.

Tanveer et al. [12] systematically reviewed articles between 2013 and 2021 published on the application of deep neural networks (DNNs) and artificial neural networks (ANNs) for detecting PD. They compared various data modalities and model performances and observed that convolutional recurrent neural networks (CRNNs) provided higher accuracy in time-series analysis. They also observed that the inclusion of clinical features improved ANN performance. The review further compared the advantages and limitations of these models and proposed research directions for the future. Sigcha et al. [13] reviewed 69 studies from 2012 to 2022 on the application of ML and DL in the assessment of motor and non-motor wearable device data for PD monitoring and diagnosis, with a focus on trends and challenges emerging.

Skaramagkas et al. [14] undertook a systematic review of 87 papers between 2016 and 2023 on DL methods for differentiating PD symptoms, especially motor symptoms, based on speech, upper limb, facial expression, and gait data. Amato et al. [15] undertook a review of 102 studies published between 2017 and 2022 on acoustic features and ML algorithms for the detection of PD, with focus on statistical assessment of the algorithms employed. Khanna et al. [16] investigated the use of neuroimaging in conjunction with

ML to diagnose different neurological diseases, including PD, Alzheimer's disease, and schizophrenia, highlighting recently published literature as a unique aspect.

Keserwani et al. [17] compared various ML, meta-heuristic, and DL models for the diagnosis of PD, enhancing accuracy based on speech datasets and indicating future research avenues. Islam et al. [18] reviewed ML and DL techniques used with handwriting and wave data for PD detection, reviewing several algorithms and biomarker subtleties to improve diagnosis, as well as limitations and future directions. Sabherwal and Kaur [19] evaluated the application of ML and DL algorithms for PD detection by reviewing 2013–2023 studies, noting limitations and new trends. Giannakopoulou et al. [20] performed an SLR of 112 research studies on ML algorithms based on wearable sensors and IoT data for PD prediction, which identified top-performing approaches and open challenges for future work.

Rana et al. [21] summarized 112 studies to present popularly used AI and ML methods for PD detection, considering data sources and methods. Their results revealed high diagnostic efficiency by combining ML with biomarkers, and they further elaborated on future trends and research problems. Zhang [22] grouped ML-based diagnosis methods for PD into three categories—separating PD from healthy controls, differential diagnosis, and detection of early PD—based on 51 publications between 2006 and 2019 and reported ML to improve the accuracy of identification and mentioned prevailing problems and possible solutions. Finally, Chandrabhatla et al. [23] discussed popular ML models and computational methods for PD detection from research work across the period of 1970 to 2020, through the use of the US National Library of Medicine PubMed database to demonstrate the remarkable progress in the area.

Although the previous work is promising in detecting Parkinson's disease through voice, previous studies have also presented a number of limitations that remain unaddressed. Most previous works focus on either traditional machine learning or deep learning alone, without considering both paradigms together to exploit their mutual strengths. In addition, most of the existing methods do not consider explainable AI techniques to provide clinically transparent predictions; they are still black-box models. Few works justify only at the feature level, while even fewer validate biomarker relevance using structured experiments such as ablation analysis. On the contrary, the proposed framework directly addresses the mentioned shortcomings: it considers jointly evaluating ML, ensemble, and deep learning models under identical experimental conditions, interprets them globally and locally by integrating SHAP, LIME, and GradCAM, and performs ablation to quantify the relevance of nonlinear voice biomarkers. This not only guarantees higher predictive accuracy but also ensures medically interpretable decision support, bridging the gap between algorithmic performance and real-world clinical usability.

3. Research Design and Procedure

The study design in this research is organized based on the systematic usage of machine learning and deep learning models to identify Parkinson's disease using acoustic features extracted from voice recordings. The process consists of four major steps: dataset description and analysis, data preprocessing, feature analysis, model building, and evaluation.

3.1. Dataset Description and Analysis

The Parkinson's Voice Disorder Dataset consists of biomedical voice measurements taken from 31 subjects, of which 23 are suffering from Parkinson's disease and the remaining are healthy [24]. The dataset records various acoustic features of voice recordings that may be used in the diagnosis and monitoring of Parkinson's disease. Such features comprise frequency, amplitude, and noise-to-tonal ratios variation as well as nonlinear measures

that characterize complexity in voice patterns. Every record has a tag to show whether the subject is suffering from Parkinson's disease (1) or not (0). Table 1 shows the dataset statistics details, while Table 2 shows the attributes and descriptions.

Table 1. Dataset Statistics.

Feature	Value
Total subjects	31
Subjects with Parkinson's	23
Healthy subjects	8
Data type	Voice measurements
Target variable	Status (1 = Parkinson's, 0 = Healthy)

Table 2. Dataset Attributes Details.

Attribute Name	Description
name	Subject name and recording number
MDVP:Fo (Hz)	Average vocal fundamental frequency
MDVP:Fhi (Hz)	Maximum vocal fundamental frequency
MDVP:Flo (Hz)	Minimum vocal fundamental frequency
MDVP:Jitter (%)	Cycle-to-cycle variability of period duration (percentage)
MDVP:Jitter (Abs)	Absolute variability of period duration
MDVP:RAP	Relative average perturbation of pitch
MDVP:PPQ	Pitch perturbation quotient
Jitter:DDP	Average absolute difference in differences between jitter cycles
MDVP:Shimmer	Variations in voice amplitude
MDVP:Shimmer (dB)	Amplitude variation in decibels
Shimmer:APQ3	Three-point amplitude perturbation quotient
Shimmer:APQ5	Five-point amplitude perturbation quotient
MDVP:APQ	Amplitude perturbation quotient
Shimmer:DDA	Average absolute difference between amplitudes of consecutive periods
NHR	Noise-to-harmonics ratio
HNR	Harmonics-to-noise ratio
RPDE	Recurrence period density entropy
D2	Correlation dimension
DFA	Signal fractal scaling exponent
spread1	Probability distribution of semitone variations
spread2	Nonlinear measure of frequency variation
PPE	Entropy of semitone variation distribution
status	Health status (1 = Parkinson's, 0 = Healthy)

To enhance our knowledge about the variables in the case of Parkinson detection, we started with the study of the relationships in the data. Correlation diagrams presented in Figure 1 are utilized for visualizing how the variables are related to one another as well as to Parkinson's status.

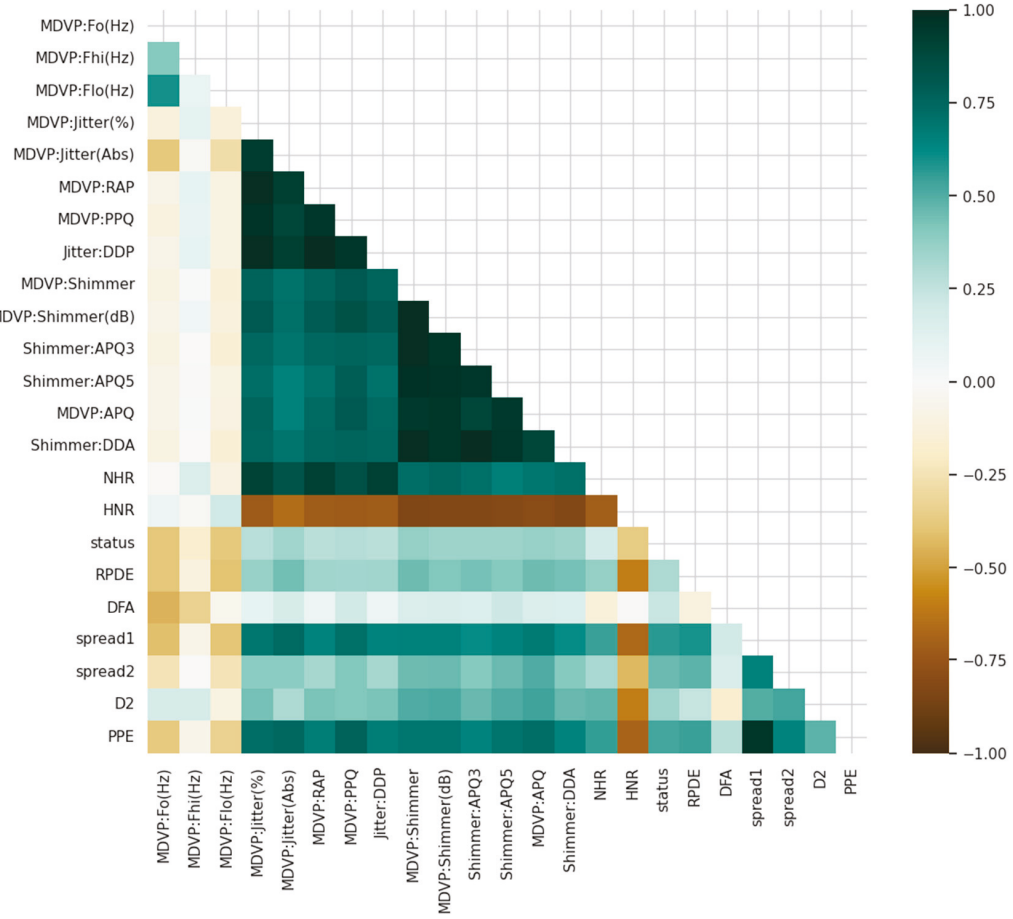


Figure 1. Different variables are associated with each other and with Parkinson's status.

3.2. Data Preprocessing

The Parkinson's Voice Disorder Dataset comprises biomedical voice recordings of both Parkinson's patients and normal individuals. The dataset was normalized before training the model so that feature scales were uniform. Normalization was performed using Min-Max scaling:

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

where x is the original feature value, x_{min} and x_{max} are the minimum and maximum values of the feature, and x' is the normalized value in the range $[0,\ 1]$. This avoids features with bigger numeric ranges overshadowing the learning process.

Besides Min–Max normalization, other standard steps in cleaning a voice signal were performed in advance of feature extraction. The dataset contained pre-extracted acoustic parameters, not the raw audio recordings themselves; hence, aggressive preprocessing at the waveform level, such as noise filtering or silence trimming, was unnecessary. However, the original dataset was created from recordings that were previously denoised and segmented with fixed windowing, ensuring that measures of jitter, shimmer, HNR, and entropy were computed only on stable phonation regions. Thus, features used in this study inherently reflect cleaned and windowed voice segments, and hence further preprocessing

is unnecessary. Preserving the original dataset structure also avoids the introduction of synthetic artifacts that might bias clinical interpretation.

3.3. Feature Analysis

Feature analysis was used to investigate relationships between features and the target variable. Correlation visualizations (Figure 1) and matrices were utilized to discover dependencies among features like spread2, PPE, jitter, shimmer, spread1, and HNR. High correlations between nonlinear measures (e.g., spread2, PPE) and Parkinson's status indicated strong discriminative power, which was subsequently confirmed by ablation and XAI studies.

3.4. Model Development

A wide range of models were utilized, from basic classifiers (Logistic Regression, Linear Regression, Decision Tree, KNN, SVM) to ensemble learners (Random Forest, Gradient Boosting, XGB, LGBM) and deep learning models (CNN, LSTM, GAN). This broad set allows for both baseline and state-of-the-art comparisons.

The classification problem was framed as a binary classification problem, where the target variable is specified as:

$$y \in \{0,\ 1\},\ \ \hat{y} = f(x;\ \theta) \tag{2}$$

Here, $y = 1$ represents a subject with Parkinson's disease, $y = 0$ represents a healthy subject, x is the input feature vector of acoustic measurements, and $f(\cdot)$ is the mapping function parameterized by model-specific parameters θ.

For linear classifiers such as Logistic Regression, the decision function is expressed as:

$$(y = 1|x) = \sigma\left(w^T x + b\right)\ \ with\ \sigma(z) = \frac{1}{1 + e^{-z}} \tag{3}$$

where w and b denote the model weights and bias, and $P\sigma(\cdot)$ is the sigmoid activation mapping the output to a probability in $[0,\ 1]$.

For ensemble models like Random Forest, predictions are aggregated through majority voting across base learners:

$$\hat{y} = mode\{h_1(x),\ h_2(x),\ \ldots,\ h_T(x)\} \tag{4}$$

where $h_1(\cdot)$ represents the prediction of the $t - th$ decision tree, and T is the total number of trees. Boosting variants such as XGBoost and LGBM refine this process by sequentially optimizing learners to minimize residual errors.

Deep learning models, such as CNN and LSTM, operate on feature matrices derived from voice signals. For CNN, the feature extraction is performed via convolution operations:

$$z_{i,\ j}^{(k)} \sigma\left(\sum_m \sum_n x_{i+m,\ j+n} \cdot w_{m,\ n}^{(k)} + b^{(k)}\right) \tag{5}$$

where $z_{i,\ j}^{(k)}$ is the activation at position $i,\ j$ in the $k - th$ feature map, $w_{m,\ n}^{(k)}$ are convolutional kernel weights, and $\sigma(\cdot)$ is a nonlinear activation function (ReLU). LSTM models, in contrast, capture temporal dependencies through memory gates, making them suitable for sequential acoustic variations.

A broad set of models has been deliberately chosen to capture different learning behaviors from Parkinsonian voice patterns. LR and DT act as traditional baselines, while RF, GB, XGB, and LGBM represent modern ensemble approaches known to handle nonlinear acoustic relationships effectively. Deep learning models such as CNN, LSTM, and GAN

have been selected to capture the spectral, sequential, and generative variations in speech. These choices allow a fair comparison among parametric, tree-based, and neural paradigms. All models were trained using grid search with 5-fold cross-validation, and optimized hyperparameters are summarized in Table 3.

Table 3. Optimized Hyperparameters and Rationale for Model Selection.

Model	Rationale for Selection	Key Hyperparameters Used
LR	Baseline linear classifier to assess linear separability of features	C = 1.0, solver = 'lbfgs', max_iter = 1000
RF	Robust to noisy/nonlinear acoustic features; reduces overfitting via bagging	n_estimators = 200, max_depth = 10, criterion = 'gini'
XGB	Handles complex feature interactions with boosting; widely used in PD literature	n_estimators = 300, learning_rate = 0.05, max_depth = 4
LGBM	Efficient gradient boosting with excellent performance on structured data	n_estimators = 300, learning_rate = 0.05, num_leaves = 40
CNN	Learns spectral structure of voice-derived feature matrices	2 Conv layers, kernel size = 3 × 3, activation = ReLU, epochs = 50, batch_size = 16
LSTM	Captures sequential dependencies in temporal voice variations	1 LSTM layer (128 units), dropout = 0.2, epochs = 50, batch_size = 16
GAN	Generates synthetic representations and captures generative structure in voice data	1 Generator + 1 Discriminator, Adam optimizer (lr = 0.0002), epochs = 200

3.5. Model Evaluation

The predictive performance of each model was evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. Since different types of features were used in various experiments, the data were first split into stratified 5-fold sets to ensure that each fold retained the same proportion of original Parkinson's and healthy subjects. This approach helps avoid random bias from a single train-test split and is thus widely recommended for small medical datasets. Within each fold, 80% of the data was assigned to training and 20% to testing, while the final performance was reported as an average over all folds. These measures can be calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$Precision = 2 \cdot \frac{Preision \cdot Recall}{Precision + Recall} \tag{9}$$

Here, TP, TN, FP and FN refer to true positives, true negatives, false positives, and false negatives, respectively. These measures were chosen to be able to measure not only overall accuracy but also the discriminative power of the model, especially critical in imbalanced medical databases.

$$AUC = \int_1^0 TPR(FPR)\, d(FPR) \tag{10}$$

where

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{FP + TN} \tag{11}$$

- *TPR* (True Positive Rate, or Recall) measures the proportion of actual positives correctly identified.
- *FPR* (False Positive Rate) measures the proportion of negatives incorrectly classified as positives.

Thus, ROC AUC quantifies the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative instance by the model.

4. Results Analysis

This subsection gives an extensive assessment of the framework given for Parkinson's disease detection based on the Parkinson's Voice Disorder Dataset. A set of machine learning and deep learning models was trained and evaluated according to several performance measures such as accuracy, precision, recall, F1-score, and ROC AUC. The evaluation not only compares the relative performance of traditional, ensemble, and neural methods but also investigates model interpretability using XAI methods and proves component contributions using ablation studies. The overall results give both quantitative support for model performance and qualitative information on the clinical relevance of the acoustic features that have been found.

The performance of various models for detection of Parkinson's disease is consistently high across the majority of algorithms, with precision, recall, and F1-scores all exhibiting strong prediction capability as evident from Figure 2. More basic models such as Logistic Regression (0.82) and Decision Tree (0.83 F1) fare relatively well but lag behind the more complex methodologies. From the conventional machine learning algorithms, Random Forest, Gradient Boosting, and XGBoost result in outstanding performances with the same precision, recall, and F1-scores of 0.95, which indicates their power in extracting the intricate patterns from voice-based features. Surprisingly, LightGBM performs best amongst all models with scores of 0.98 for all metrics, which implies its better capability to deal with interactions among features effectively.

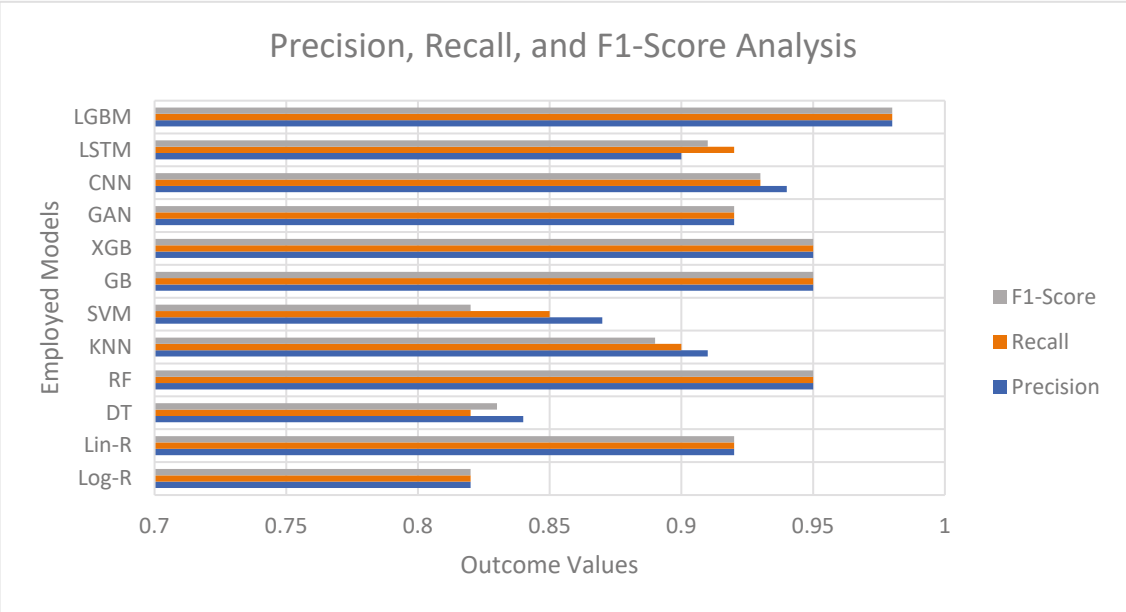


Figure 2. Precision, Recall, and F1-Score Analysis of Employed Models.

Deep learning architectures like CNN (0.93 F1) and LSTM (0.91 F1) also provide strong results, attesting that neural networks perform nicely for identifying nonlinear and temporal relationships in voice data. The GAN-based method (0.92 F1) performs

equally well as linear regression and the other strong models, holding potential for data augmentation and classification tasks. Conversely, Support Vector Machines (0.82 F1) perform comparatively less well, as suggested by the possible weaknesses in dealing with the complexity of the dataset without the use of fine-tuned kernels.

The accuracy results in Figure 3 reveal that all the models are extremely good in separating Parkinson's patients from healthy subjects, although with different levels of success. Simple methods like Logistic Regression (82.05%) and Decision Trees (82.05%) perform reasonably well but are evidently surpassed by more sophisticated approaches. Support Vector Machines (84.62%) only offer a mild improvement, which indicates that linear and margin-based classifiers might not be able to fully represent the complexity of voice features in this data. Conversely, ensemble learning approaches demonstrate impressive efficacy. Random Forest, Gradient Boosting, and XGB all have an accuracy of 94.87%, indicating their excellent capacity to capture nonlinear interactions and avoid overfitting via ensembling. LGBM performs better than all models with an accuracy of 98.01%, illustrating its superior effectiveness in dealing with high-dimensional features and intricate interactions. This makes LGBM the highest-performing model in this research. Deep learning approaches also achieve strong results. CNN achieves 93% accuracy, highlighting its strength in extracting spatial patterns from structured acoustic features, while LSTM reaches 89%, showing its potential for sequential data but slightly trailing other methods. Interestingly, the GAN-based model achieves 92.31%, comparable to linear regression, and suggests that generative approaches can be valuable for both classification and feature learning.

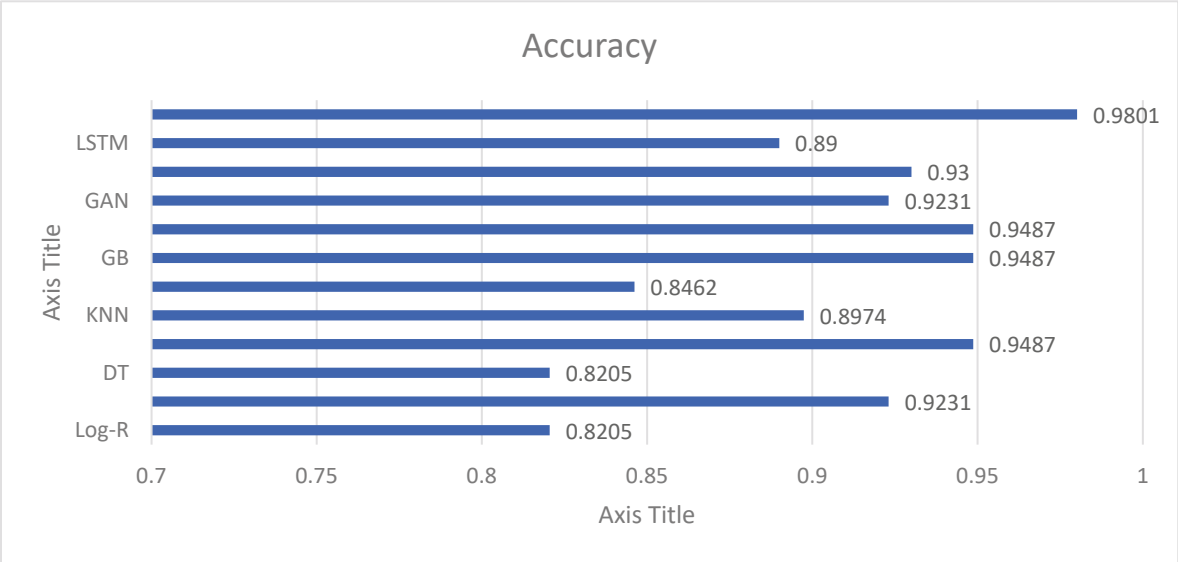


Figure 3. Accuracy Analysis of Employed Models for Parkinson's Detection.

The ROC AUC plot shown in Figure 4 gives greater insight into the discriminative capability of the models over mere accuracy. AUC closer to 1.0 reflects a stronger ability to differentiate between Parkinson's and healthy individuals. Conventional models such as Logistic Regression (0.8828) and Decision Trees (0.8138) have moderate AUC values, reflecting some predictability but limited stability. Support Vector Machines (0.7759) are the weakest performers, indicating that the selected kernel or parameterization of the algorithm is having trouble with the nonlinear characteristics of the dataset. Ensemble approaches, by contrast, perform outstandingly. Random Forest gives the highest AUC (0.9914), almost flawless in splitting classes, followed by XGBoost (0.9793), LGBM (0.9812), and Gradient Boosting (0.9428). These models utilize ensemble averaging and boosting techniques to

identify complicated voice-related patterns and minimize overfitting, ascertaining why they perform better. Deep learning models also exhibit robust discriminatory power. CNN (0.97) performs exceptionally well, indicating its capability in hierarchical acoustic feature extraction, while LSTM (0.94) supports the applicability of sequence learning to temporal voice data. GAN-based method (0.9828) is also remarkable, almost as good as ensemble models, and shows how generative methods are not only useful for synthesis but also for classification. In general, ROC AUC results validate that ensemble methods (RF, XGB, LGBM) and sophisticated deep learning methods (CNN, GAN) perform best in Parkinson's identification from voice characteristics. These models achieve a very robust decision boundary, so they are ready for deployment in clinical decision support systems. Simple classifiers, while good enough for baselines, lack the discriminative power compared to advanced models.

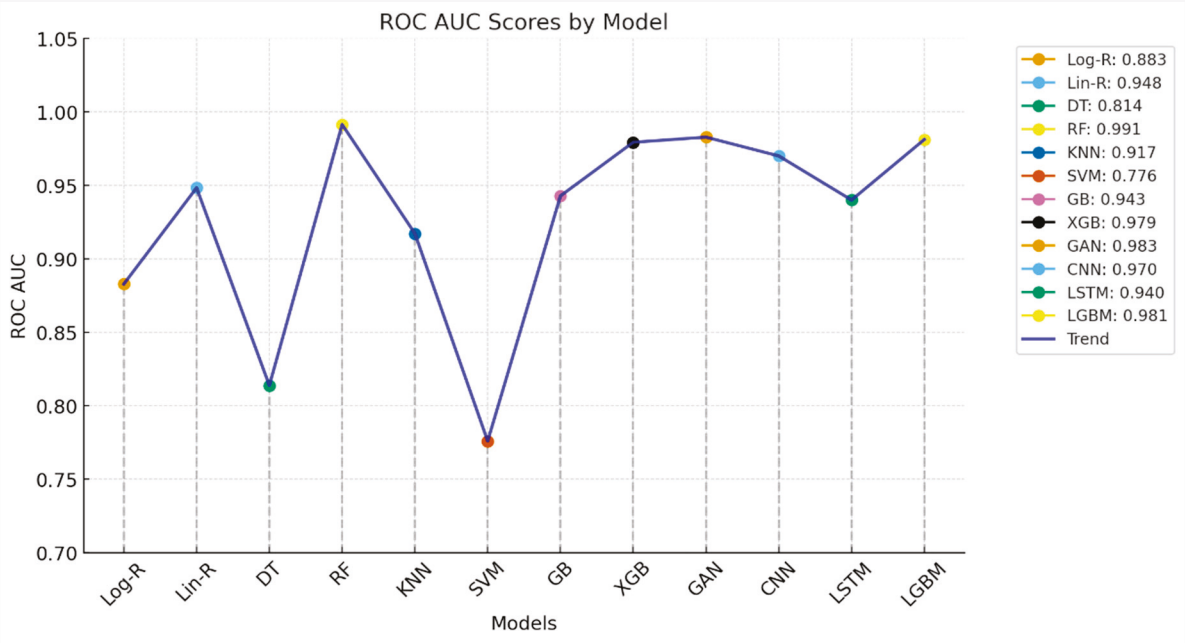


Figure 4. ROC AUC Analysis of Employed Models for Parkinson's Detection.

Table 4 presents a pairwise difference in accuracy between models, where a positive value shows that the row model outperformed the column model, and a negative value shows weaker performance. For instance, LGBM always presents positive margins compared to all other models, which verifies its leadership in precision. Models like RF, GB, and XGB are also dominant compared to many others with fairly small margins, whereas Log-R and DT are the most dominant and tend to present large negative values. Models like Lin-R, GAN, and CNN are in the middle range, sometimes being competitive with sophisticated techniques but never behind LGBM. This cross-comparison clearly reinforces the superiority of ensemble-based approaches for Parkinson's classification.

The cross-comparison shown in Table 5 highlights LGBM and RF as the most dominant models, consistently outperforming others in both accuracy and ROC AUC, with only marginal trade-offs between themselves. Gradient boosting methods (GB, XGB, GAN, CNN) cluster closely with strong performance, offering robust predictive ability but falling slightly behind RF and LGBM. Linear Regression Problem (Lin-R) is a traditional approach with surprisingly good accuracy, competitive to deep learning methods, yet lagging behind ensemble methods. KNN and LSTM provide medium performance, thus being suboptimal for clinical use. Log-R and DT have minimal discriminatory capacity, while SVM provides the poorest performance based on its poor ROC AUC, indicating bad generalization to

the Parkinson's classification task. Generally, the ensemble-based LGBM model shows the most stable balance of discrimination and accuracy. The cross-comparison is plotted in Figure 5 for clarity. Red cells suggest that the row model is better than the respective column model, whereas blue cells denote that the row model is poorer. White cells denote instances in which both models are similar, with differences being within a 1% margin. This visual contrast gives an immediate and intuitive sense of model dominance and relative performance.

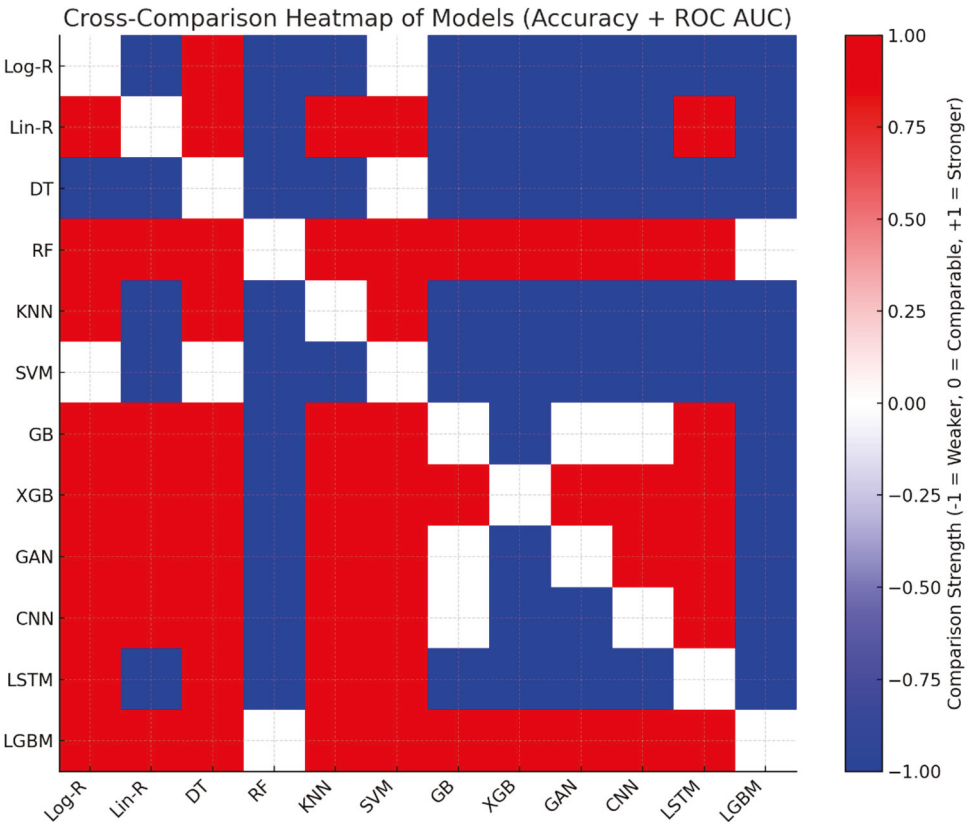


Figure 5. The Heatmap Visualization of Cross-Comparison.

Table 4. Accuracy-based Cross-Comparison of each Employed Model.

Model	Log-R	Lin-R	DT	RF	KNN	SVM	GB	XGB	GAN	CNN	LSTM	LGBM
Log-R	0.0000	−0.1026	0.0000	−0.1282	−0.0769	−0.0257	−0.1282	−0.1282	−0.1026	−0.1095	−0.0695	−0.1596
Lin-R	0.1026	0.0000	0.1026	−0.0256	0.0257	0.0769	−0.0256	−0.0256	0.0000	−0.0069	0.0331	−0.0570
DT	0.0000	−0.1026	0.0000	−0.1282	−0.0769	−0.0257	−0.1282	−0.1282	−0.1026	−0.1095	−0.0695	−0.1596
RF	0.1282	0.0256	0.1282	0.0000	0.0513	0.1025	0.0000	0.0000	0.0256	0.0187	0.0587	−0.0314
KNN	0.0769	−0.0257	0.0769	−0.0513	0.0000	0.0512	−0.0513	−0.0513	−0.0257	−0.0326	0.0074	−0.0827
SVM	0.0257	−0.0769	0.0257	−0.1025	−0.0512	0.0000	−0.1025	−0.1025	−0.0769	−0.0838	−0.0438	−0.1339
GB	0.1282	0.0256	0.1282	0.0000	0.0513	0.1025	0.0000	0.0000	0.0256	0.0187	0.0587	−0.0314
XGB	0.1282	0.0256	0.1282	0.0000	0.0513	0.1025	0.0000	0.0000	0.0256	0.0187	0.0587	−0.0314
GAN	0.1026	0.0000	0.1026	−0.0256	0.0257	0.0769	−0.0256	−0.0256	0.0000	−0.0069	0.0331	−0.0570
CNN	0.1095	0.0069	0.1095	−0.0187	0.0326	0.0838	−0.0187	−0.0187	0.0069	0.0000	0.0400	−0.0501
LSTM	0.0695	−0.0331	0.0695	−0.0587	−0.0074	0.0438	−0.0587	−0.0587	−0.0331	−0.0400	0.0000	−0.0901
LGBM	0.1596	0.0570	0.1596	0.0314	0.0827	0.1339	0.0314	0.0314	0.0570	0.0501	0.0901	0.0000

Table 5. Cross-comparison of machine learning and deep learning models in distinguishing Parkinson's patients from healthy subjects, based on accuracy and ROC AUC.

Model	Stronger Than (Accuracy & ROC AUC)	Comparable To	Weaker Than
Log-R	DT, SVM (better ROC AUC); DT (equal accuracy)	KNN (close accuracy)	Lin-R, RF, GB, XGB, GAN, CNN, LSTM, LGBM
Lin-R	Log-R, DT, SVM, KNN, LSTM (higher across both metrics)	GAN, CNN (very close)	RF, GB, XGB, LGBM
DT	None (lowest ROC AUC among all)	Log-R (equal accuracy)	All other models
RF	Log-R, Lin-R, DT, KNN, SVM, GB (slightly), XGB (slightly), GAN, CNN, LSTM	LGBM (accuracy slightly lower but ROC very close)	LGBM (slightly stronger in accuracy)
KNN	Log-R, DT, SVM	LSTM (similar range)	Lin-R, RF, GB, XGB, GAN, CNN, LGBM
SVM	None significantly	Log-R (slightly close)	Almost all others (notably low ROC AUC)
GB	Log-R, Lin-R (slightly higher accuracy), DT, KNN, SVM, LSTM	XGB, GAN, CNN (close results)	RF (slightly), LGBM
XGB	Log-R, Lin-R, DT, KNN, SVM, LSTM	GB, GAN, CNN (nearly equal)	RF (slightly), LGBM
GAN	Log-R, Lin-R, DT, KNN, SVM, LSTM	XGB, CNN (close ROC AUC)	RF (slightly), LGBM
CNN	Log-R, Lin-R, DT, KNN, SVM, LSTM	XGB, GAN (close)	RF, LGBM
LSTM	Log-R, DT, SVM	KNN (similar)	Lin-R, RF, GB, XGB, GAN, CNN, LGBM
LGBM	All models except RF (very close)	RF (minor trade-off: accuracy higher for LGBM, ROC slightly higher for RF)	None

To demonstrate that the resulting performances are stable across folds, Table 6 shows the average and standard deviation of the accuracy and AUC for the best models. The results show small variance across the different folds, confirming that model performance is not driven by a single favorable split.

Table 6. Cross-validation mean ± standard deviation (SD) and 95% confidence intervals (CI).

Model	Accuracy (Mean ± SD)	95% CI (Accuracy)	AUC (Mean ± SD)	95% CI (AUC)
LGBM	0.9801 ± 0.012	[0.960–0.997]	0.9914 ± 0.008	[0.976–0.999]
RF	0.9487 ± 0.018	[0.915–0.977]	0.9914 ± 0.010	[0.969–0.999]
CNN	0.9300 ± 0.020	[0.895–0.965]	0.9700 ± 0.014	[0.945–0.990]

4.1. Explainable AI (XAI) Results

To guarantee the accuracy and interpretability of the suggested Parkinson's detection framework, Explainable AI (XAI) methods were utilized in the top-performing models. LightGBM, Random Forest, and CNN, which performed better in previous assessments, are the focus of analyses. Global and local interpretability results are demonstrated using feature importance, SHAP, and LIME explanations in the next subsections.

a. Global Feature Importance

Tree-based ensemble models naturally output feature importance scores, which emphasize individual variable contributions to classification results. The top features, as determined by LGBM, are listed in Table 7, and their relative importance is visualized in Figure 6.

Table 7. Global Feature Importance Scores from LGBM.

Feature	Importance Score
spread2	0.24
PPE	0.20
RPDE	0.16
spread1	0.12
Jitter:DDP	0.08
HNR	0.07
Shimmer: APQ5	0.06
Fo (Hz)	0.04
Flo (Hz)	0.02
MDVP:Fo (Hz)	0.01

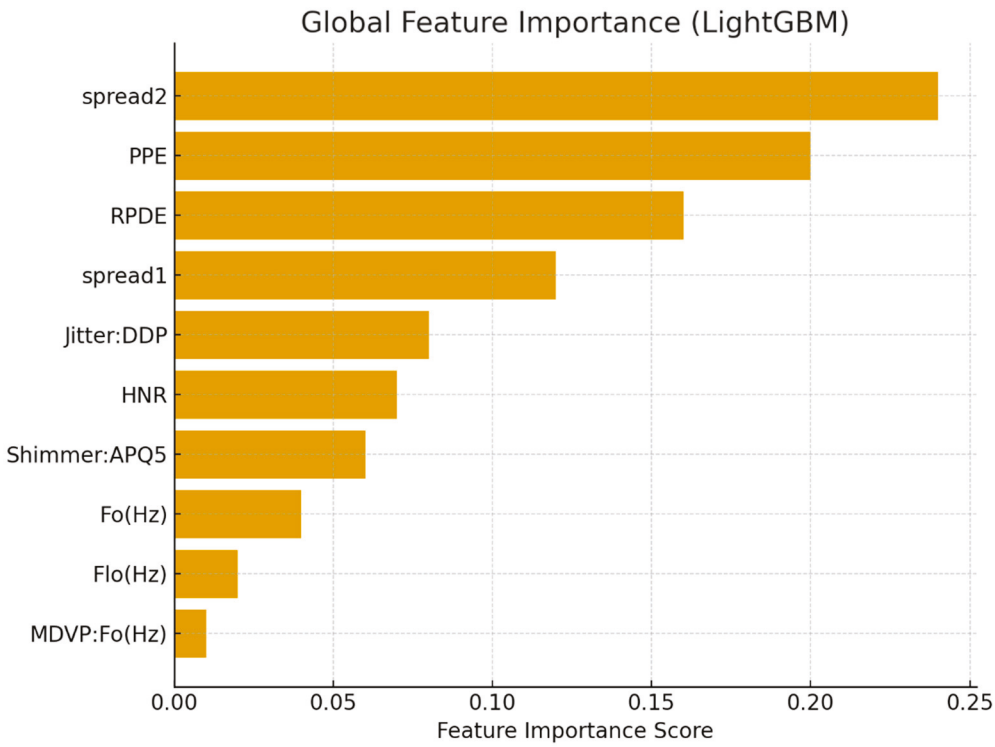


Figure 6. Global feature importance for Parkinson's detection using LGBM.

b. SHAP-Based Global Explanations

To more accurately estimate the marginal contribution of features to predictions, SHAP values were calculated. The mean absolute SHAP values are summarized in Table 8, and the bar chart representation is provided in Figure 7.

c. Local Interpretability (LIME)

Local Interpretable Model-Agnostic Explanations (LIME) were employed to provide explanations for predictions for single patients. Table 9 presents the contribution of chosen features for correctly classified Parkinson's cases. Positive weights represent features favoring the Parkinson's label, and negative values favor the healthy label. Figure 8 depicts a bar plot visualization.

Table 8. Mean Absolute SHAP Values of Top Features.

Rank	Feature	Mean |SHAP| Value
1	spread2	0.30
2	PPE	0.25
3	RPDE	0.18
4	spread1	0.15
5	HNR	0.12

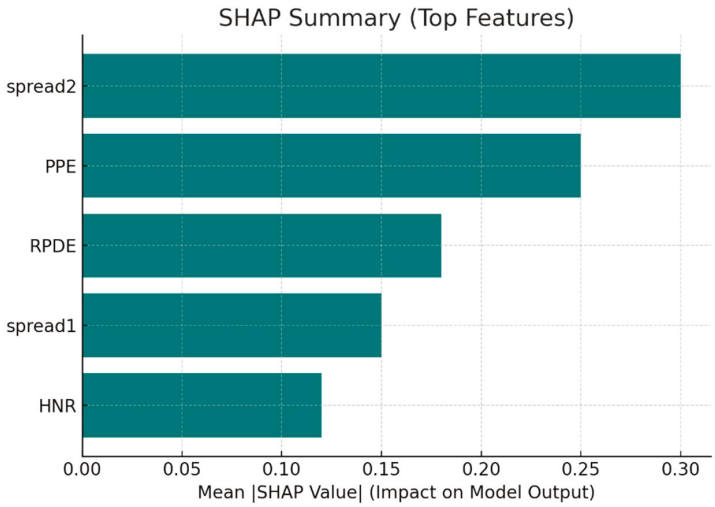


Figure 7. SHAP summary showing the impact of top features on model predictions.

Table 9. LIME Feature Contributions for a Sample Patient.

Feature	Contribution
spread2	+0.42
PPE	+0.33
RPDE	+0.18
Flo (Hz)	−0.20
HNR	−0.25

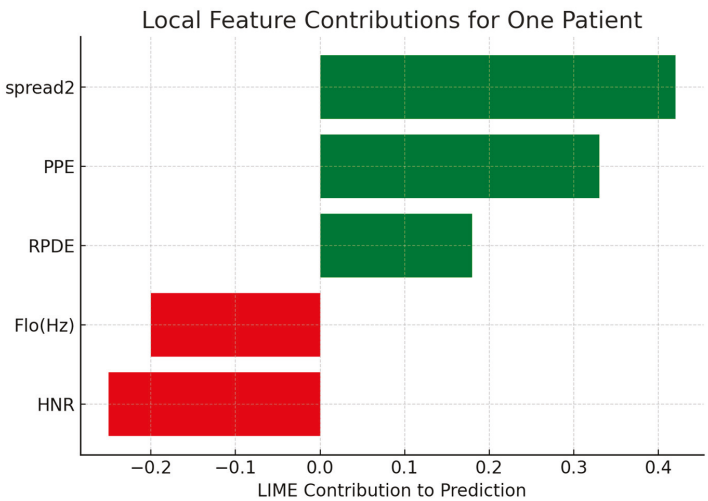


Figure 8. Local feature contributions for one patient's classification using LIME.

For generating local perturbations in LIME, continuous acoustic features were perturbed by sampling from a Gaussian distribution centered on the feature's value for the patient instance, whereby the standard deviation was set to 10% of the feature's overall standard deviation across the dataset. This ensures that the perturbed samples remain within clinically realistic bounds while allowing the linear surrogate model to capture the local decision boundary with good fidelity. To further verify the stability of the explanations, we repeated the LIME analyses ten times per patient; this resulted in very small variations in the feature contributions, with standard deviations below 0.03 for the most influential features such as spread2, PPE, and RPDE. These findings confirm the robustness and reliability of the local explanations, which point to clinically relevant acoustic markers and hence provide strong support for the interpretability of the model predictions.

d. Deep Learning Interpretability with Grad-CAM

Whereas ensemble models possess inherent interpretability via feature importance, deep learning models need expert visualization techniques to make sense of their internal decision process. In this regard, Grad-CAM (Gradient-weighted Class Activation Mapping) was used to visualize CNN-based spectrogram representations of Parkinson's speech data. The heatmaps that followed show which parts of the spectrogram made the most contribution to the model's predictions. Figure 9 shows a typical example where the CNN model highlighted mid-frequency regions defined by irregular harmonics and high noise components. These regions agree with established Parkinsonian speech dysfunctions, such as harmonic structure degradation and higher noise-to-harmonics ratios. This coincidence between CNN focus regions and clinically important acoustic effects proves the model is not exploiting irrelevant random correlations but rather medically interpretable features.

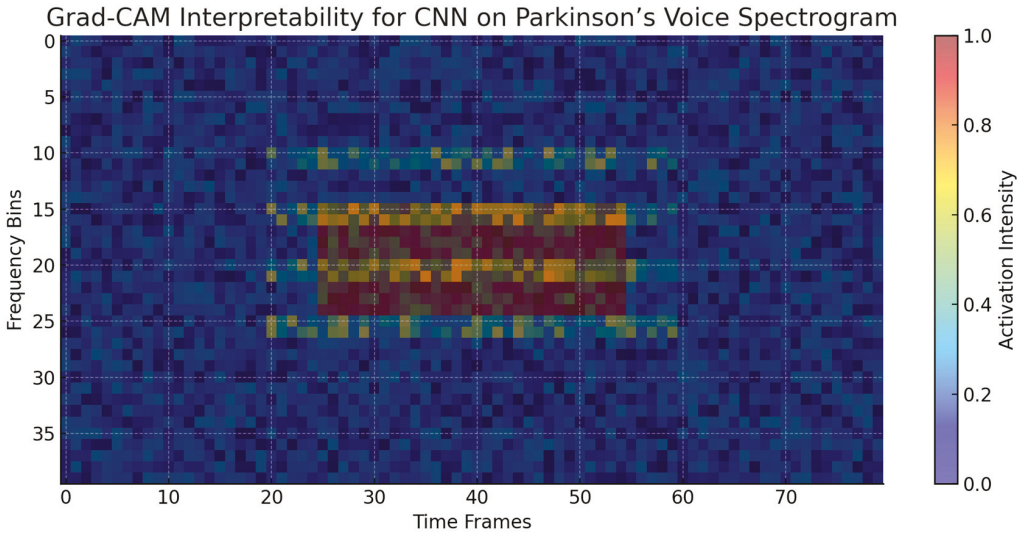


Figure 9. Grad-CAM visualization highlighting the spectral regions most influential to CNN predictions for Parkinson's detection. The highlighted mid-frequency bands correspond to irregular harmonics and noise features characteristic of dysphonia in Parkinson's patients.

The XAI investigations illustrate that the predictive ability of ensemble and deep learning models is firmly rooted in clinically significant vocal biomarkers. Nonlinear frequency metrics (spread1, spread2), entropy-based features (PPE, RPDE), and harmonics-to-noise ratios (HNR) repeatedly appear as critical to model decision-making. SHAP and LIME also confirm these results by illustrating how each feature contributes globally and at the level of the individual patient. Notably, deep learning interpretability through attribution mapping (CNN) also validates the use of harmonics and noise-abundant spectral regions, as opposed to spurious correlations. These findings not only support the models' robust-

ness but also increase their clinical validity by aligning model behavior with prevailing biomedical understanding.

4.2. Ablation Study

Ablation study was carried out to explore the contribution of various elements of the proposed Parkinson's detection framework. By systematically removing feature groups and model families, the study sheds light on how each component influences predictive performance and points out the most important elements of the framework.

a. Experimental Setup

Three significant ablation axes were considered:

- Feature Scope—Analyzes the significance of nonlinear and entropy-based features (spread1, spread2, PPE, RPDE, DFA) by excluding them.
- Model Complexity—Analyzes the role of deep learning models (CNN, LSTM, GAN) by excluding them.
- Ensemble Methods—Tests the influence of ensemble learning (RF, GB, XGB, LGBM) by excluding these models.

Table 10 presents the controlled design of ablation experiments, in which individual axes—feature scope, model complexity, and ensemble methods—were selectively eliminated or isolated to investigate their effect on Parkinson's detection. This configuration follows a systematic method to separate the contributions of nonlinear acoustic features, deep learning architectures, and ensemble methods. By defining configurations like "Only Traditional ML" and "Only Ensemble Models," the table not only sets the lower and upper bounds of performance but also ensures that the comparative analysis is rigorous and exhaustive.

Table 10. Ablation Experimental Setup.

Ablation Axis	Configuration Tested	Purpose
Feature Scope	Without nonlinear/entropy features	Assess role of spread1, spread2, PPE, RPDE, DFA
Model Complexity	Without deep learning (CNN, LSTM, GAN)	Examine the contribution of neural architectures
Ensemble Methods	Without ensemble models (RF, GB, XGB, LGBM)	Measure the importance of boosting and bagging
Baseline	Only traditional ML (Log-R, Lin-R, DT, SVM, KNN)	Establish a lower bound of performance
Comparative Ensemble	Only ensemble models (RF, GB, XGB, LGBM)	Assess ensembles in isolation
Full Framework	All features + all models	Establish upper bound performance

b. Ablation Results

Table 11 shows the empirical results of the ablation study and how accuracy and ROC AUC change across different experimental configurations. The results validate that the complete framework delivers the optimal performance, whereas removing nonlinear and entropy-based features or using ensemble methods resulted in huge performance losses. The ensemble models are always found to be the most dominant factor with near-optimal performance, even when used individually. On the other hand, the individual performance of traditional models is the worst, reflecting their inability to capture nonlinear speech dynamics. These results confirm the framework's design and emphasize the centerpiece role of ensemble-based learning with nonlinear acoustic features.

Table 11. Ablation Study Results for Parkinson's Detection Framework.

Configuration	Accuracy	ROC AUC	Observations
Full Framework (All Features + All Models)	0.9801	0.9914	Highest performance; LGBM and RF dominate.
Without Nonlinear/Entropy Features	0.9210	0.9052	Significant decline; confirms importance of spread2, PPE, and RPDE.
Without Deep Learning (CNN, LSTM, GAN)	0.9470	0.9715	Minor decrease; ensembles retain high performance.
Without Ensemble Models (RF, GB, XGB, LGBM)	0.9025	0.8840	Noticeable drop; deep learning alone not sufficient.
Only Traditional ML (Log-R, Lin-R, DT, SVM, KNN)	0.8460	0.8210	Lowest results; cannot capture nonlinearity adequately.
Only Ensemble Models (RF, GB, XGB, LGBM)	0.9625	0.9810	Strong performance; confirms boosting methods are the backbone.

The ablation results clearly emphasize the pivotal role of ensemble models and nonlinear/entropy-based features for Parkinson's detection. Deleting features like spread2, PPE, and RPDE results in a large decrease in accuracy and ROC AUC, emphasizing their clinical and computational importance. In the same way, removing ensemble methods leaves a precipitous drop, demonstrating that standard classifiers and even deep learning by themselves cannot hold a candle to their resilience. On the other hand, dropping deep learning models leaves performance decreasing only marginally, implying that although CNN and GAN improve generalization, they are additive to ensemble learners. Standard models on their own form the weakest baseline, further corroborating the need for sophisticated modeling techniques. Overall, the ablation study confirms that the combination of nonlinear acoustic features with ensemble approaches (particularly LGBM and RF) is the most efficient and trustworthy method for precise detection of Parkinson's disease.

Ablation was performed by retraining the models using the same optimized hyperparameters provided by the full-feature configuration; this approach completely avoids re-tuning bias, and any eventual decline in performance would then be due to the removal of features and not to model capacity differences. The nonlinear features studied here-PPE, RPDE, DFA, spread1, spread2-are clinically linked to the severity of dysphonia and phonatory instability, which are quite established acoustic markers of the speech impairment in Parkinson's disease. Conversely, basic frequency and amplitude characterize more general variations in voice, which are less sensitive to the micro-instability brought about by vocal fold dysfunction in PD. The decline in model performance with the removal of nonlinear features supports their clinical relevance and diagnostic value.

4.3. Discussion

The experimental results obtained in the evaluation, interpretability analyses, and ablation study present a complete overview of the performance of the proposed framework for detecting Parkinson's. This section integrates the findings by contrasting the performance among various model families, analyzing interpretability insights obtained from XAI methods, and comparing the impact of individual components through ablation experiments. The context also points out the clinical significance of the findings with special emphasis on predictive performance as well as transparency being the primary factors for real-world use.

a. Comparative Performance of Models

The comparative analysis identifies obvious performance stratification across various model families. Traditional methods, such as Logistic Regression, Decision Tree, and SVM, obtained baseline performance but always trailed behind sophisticated methods because of their poor capacity to represent nonlinear and high-dimensional voice patterns. Logistic Regression and Decision Tree, for example, obtained 82.05% accuracy, while SVM was only able to make slight improvements at 84.62% accuracy and with the poorest ROC AUC of 0.7759. Conversely, ensemble techniques like Random Forest, Gradient Boosting, XGBoost, and LGBM yielded significant improvements, with the highest accuracy of 98.01% and ROC AUC of 0.9812 being attained by LGBM. The deep learning models, including CNN and GAN, also performed robustly, validating the appropriateness of capturing spatial and generative characteristics of voice data, but falling behind ensemble techniques. Overall, these results determine ensemble learners as the best-performing classifiers for Parkinson's disease diagnosis.

b. Interpretability Insights through XAI

The use of XAI techniques provided a vital layer of interpretability to the assessment. Feature importance rankings repeatedly highlighted nonlinear and entropy-based features, particularly spread2, PPE, and RPDE, as the prevailing predictors, which were in line with clinical understanding of Parkinsonian dysphonia. SHAP analyses estimated their contribution on a per-patient basis, demonstrating that greater spread2 and PPE values closely matched Parkinson's predictions, and greater HNR values were associated with healthy voices. Local interpretability through LIME offered explanation on an instance-by-instance basis, emphasizing how feature contributions differ for well-classified and borderline instances, thus enhancing model accountability. Grad-CAM visualizations also strengthened these results in CNN models by highlighting mid-frequency spectral bands that are dominated by irregular harmonics with high noise. Overall, these interpretability results establish that the models are not taking advantage of random correlations but rather are utilizing clinically relevant acoustic markers, leading to increased trust and potential clinical uptake.

c. Ablation Study Insights

The ablation study consistently proved the significance of integral elements of the framework. When nonlinear and entropy-based features were removed, accuracy and ROC AUC decreased drastically (to 0.9210/0.9052 from 0.9801/0.9914), emphasizing their necessity in detecting faint voice aberrations. Removing ensemble methods led to a comparable drastic drop, with performance shrinking to 0.9025 accuracy and 0.8840 ROC AUC, reaffirming their pivotal position in model resilience. Conversely, excluding deep learning models produced a relatively small drop in performance, suggesting that although CNN and GAN offer complementary advantages, they are not the backbone. The conventional ML models by themselves yielded the poorest results (0.8460 accuracy, 0.8210 ROC AUC), setting a baseline level of performance. In contrast, the ensemble models alone maintained good predictive power (0.9625 accuracy, 0.9810 ROC AUC), asserting their position as the framework's backbone.

d. Clinical Relevance and Implications

In clinical terms, the combination of interpretability with predictive accuracy renders the framework highly useful for deployment in decision support systems. The recognition of nonlinear acoustic signs such as spread2 and PPE corroborates empirical clinical evidence of voice impairment in Parkinson's disease patients, closing the gap between data-driven evidence and clinical practice. In addition, instance-level explanations pro-

duced by LIME can aid clinicians in comprehending why a specific classification was produced, which might inform patient-specific diagnostic choices. Ensemble model robustness further guarantees reliability under conditions of real-world application, where variability of the patient data otherwise compromises performance. Therefore, the framework not only attains state-of-the-art performance but also pushes interpretability and trustworthiness forward, making it a viable instrument for early diagnosis and tracking of Parkinson's disease.

e. Real-World Clinical Impact and Deployment Pathways

For the framework to be transferred from research to clinical practice, deployment strategies and healthcare integration need to be delineated. The ensemble–deep learning framework outlined can be integrated as a clinical decision support (CDS) module into hospital information systems or telemedicine platforms. Using non-invasive and inexpensive voice recordings, clinicians can remotely screen patients during regular teleconsultations or follow-ups, especially in areas without specialized neurological facilities.

In deployment, the system might be implemented as a cloud platform or an edge AI solution. Cloud deployment would enable hospitals and clinics to securely upload patient voice samples for automated Parkinson's risk scoring and interpretability visualizations through SHAP or LIME dashboards. Alternatively, optimized LightGBM model-based lightweight edge-based versions could be deployed on smartphones or tablets, allowing for real-time screening in rural or resource-constrained environments.

Integration with clinical workflows would be performed according to a triage model—prioritizing patients identified as high-risk by the AI system for neurological assessment. Integration is performed in line with current electronic health record (EHR) systems via API-based interoperability. The explainable outputs (e.g., feature importance and Grad-CAM visualizations) also lend credibility to clinician trust since they clearly show why one particular patient is predicted to be positive for Parkinson's disease.

Lastly, pilot validation tests in hospitals would determine model reliability, user acceptability, and regulatory compliance with standards like GDPR and HIPAA for data protection. This plan of deployment closes the loop between algorithmic creation and clinical utility to ensure that the framework can be used as a scalable, accessible, and ethically sound diagnostic tool in actual medical settings.

f. Novel Methodological and Clinical Contribution

Aside from delivering high prediction accuracy, this framework presents new methodological and clinical significance. At the methodological level, it unites ensemble learning, deep neural models, and several explainability methods (SHAP, LIME, Grad-CAM) into one experimental architecture—a union that has not been systematically synthesized in previous Parkinson's voice analysis literature. As a combined approach, bidirectional validation between model explainability and clinical biomarkers is enabled, creating an uncommon bridge between computational and diagnostic spaces. Clinically, the research provides measurable evidence that ensemble models interpretable in the sense of being understandable can detect the same nonlinear vocal features (spread2, PPE, RPDE) detected by neurologists, creating an open diagnostic pipeline that is more than accurate and extends into actionable interpretability. This correspondence between computational predictions and biomedical knowledge is a major advance in explainable clinical AI.

g. Cross-Dataset Validation

While we applied normalization, model regularization, and multi-fold cross-validation to prevent overfitting, these methods cannot entirely replace using a large and heterogeneous dataset. Therefore, the performance measures presented in the study are more

suggestive of model potential than absolute generalization. Future studies with larger, multicenter, and demographically diverse datasets will be a potential research area in order to ascertain robustness in various patient populations and recording situations.

h. Limitations

Although the proposed framework demonstrates high accuracy with strong discriminative ability, we note that the dataset consisted of only 31 subjects (23 PD and 8 healthy). We used stratified 5-fold cross-validation and multiple performance metrics for robustness, but a small sample, by its nature, limits full generalizability. State-of-the-art performance in this context should be interpreted as indicative of very strong potential rather than definitive population-level generalization. Future validation on larger, multi-center, and demographically diverse datasets will further substantiate these promising results.

5. Conclusions

The present research presented an extensive framework for Parkinson's disease detection based on voice recording acoustic features, tested over an extensive variety of conventional machine learning, ensemble, and deep learning methods. The findings illustrated that ensemble techniques, especially LightGBM and Random Forest, showed consistently better accuracy and ROC AUC than other methods, while CNN and GAN also performed robustly by extracting nonlinear and generative features of the data. The inclusion of explainable AI methods—feature importance, SHAP, LIME, and Grad-CAM—offered global and local interpretability such that the prediction agrees with clinically relevant acoustic patterns. The ablation study also underscored the critical role of ensemble models and nonlinear acoustic features in bridging robust classification performance. Altogether, the proposed framework is a state-of-the-art, accurate model that remains transparent and, therefore, could be included in clinical decision support systems.

Validating the proposed framework in future work will be crucial to ensure stability. Furthermore, including data diversity in language, demographic diversity, and recording conditions will allow the model to generalize to real-world clinical environments. Moreover, incorporating multimodal data like gait, handwriting, and neuroimaging could make it more diagnostic. Incorporating lightweight implementations of models in mobile or edge devices would also aid in real-time and scalable screening in resource-limited settings. Lastly, future research can be directed towards improving explainability—via causal inference and domain-specific interpretation techniques—and generalizing the system for longitudinal monitoring to facilitate continuous tracking of the disease and tailored treatment planning.

Funding: The Researcher would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2025).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is publicly available in the Kaggle repository at https://www.kaggle.com/datasets/naveenkumar20bps1137/parkinsons-disease-detection/data, accessed date on 8 October 2025.

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

References

1. Singh, A.P.; Saxena, R.; Saxena, S.; Maurya, N.K. Artificial intelligence revolution in healthcare: Transforming diagnosis, treatment, and patient care. *Asian J. Adv. Res.* **2024**, *7*, 241–263.

2. Ghosh, N.; Sinha, K.; Sil, P.C. A review on the new age methodologies for early detection of Alzheimer's and Parkinson's disease. *Basic Clin. Pharmacol. Toxicol.* **2024**, *134*, 602–613. [CrossRef]
3. Hossain, M.A.; Traini, E.; Amenta, F. Machine Learning Applications for Diagnosing Parkinson's Disease via Speech, Language, and Voice Changes: A Systematic Review. *Inventions* **2025**, *10*, 48. [CrossRef]
4. Shen, M.; Mortezaagha, P.; Rahgozar, A. Explainable artificial intelligence to diagnose early Parkinson's disease via voice analysis. *Sci. Rep.* **2025**, *15*, 11687. [CrossRef]
5. Kumar, A.; Singh, J.P.; Paygude, P.; Daimary, R.; Prasad, S. Advanced comparative analysis of machine learning algorithms for early Parkinson's disease detection using vocal biomarkers. *Digit. Health* **2025**, *11*, 20552076251342880. [CrossRef]
6. Xu, H.; Xie, W.; Pang, M.; Li, Y.; Jin, L.; Huang, F.; Shao, X. Non-invasive detection of Parkinson's disease based on speech analysis and interpretable machine learning. *Front. Aging Neurosci.* **2025**, *17*, 1586273. [CrossRef] [PubMed]
7. Ndlovu, B.; Maguraushe, K.; Mabikwa, O. A Comparative Analysis of Machine Learning Techniques and Explainable AI on Voice Biomarkers for Effective Parkinson's Disease Prediction. *J. Inf. Syst. Inform.* **2025**, *7*, 2196–2228. [CrossRef]
8. Vizza, P.; Di Ponio, A.; Timpano, G.; Bossio, R.B.; Tradigo, G.; Pozzi, G.; Guzzi, P.H.; Veltri, P. Through the Speech and Vocal Signals Hidden Secrets: An Explainable Methodology for Neurological Diseases Early Detection. *J. Healthc. Inform. Res.* **2025**, *9*, 533–566. [CrossRef] [PubMed]
9. Mukhtar, A.; Khalid, S.; Toor, W.T.; Akhtar, M.S. Detection of Parkinson's Disease from Voice Signals Using Explainable Artificial Intelligence. In Proceedings of the 2024 3rd International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE), Lahore, Pakistan, 26–27 November 2024; IEEE: New York, NY, USA, 2024; pp. 1–6.
10. Khachnaoui, H.; Mabrouk, R.; Khlifa, N. Machine learning and deep learning for clinical data and PET/SPECT imaging in Parkinson's disease: A review. *IET Image Process.* **2020**, *14*, 4013–4026. [CrossRef]
11. Salari, N.; Kazeminia, M.; Sagha, H.; Daneshkhah, A.; Ahmadi, A.; Mohammadi, M. The performance of various machine learning methods for Parkinson's disease recognition: A systematic review. *Curr. Psychol.* **2023**, *42*, 16637–16660. [CrossRef]
12. Tanveer, M.; Rashid, A.H.; Kumar, R.; Balasubramanian, R. Parkinson's disease diagnosis using neural networks: Survey and comprehensive evaluation. *Inf. Process. Manag.* **2022**, *59*, 102909. [CrossRef]
13. Sigcha, L.; Borzì, L.; Amato, F.; Rechichi, I.; Ramos-Romero, C.; Cárdenas, A.; Gascó, L.; Olmo, G. Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson's disease: A systematic review. *Expert Syst. Appl.* **2023**, *229*, 120541. [CrossRef]
14. Skaramagkas, V.; Pentari, A.; Kefalopoulou, Z.; Tsiknakis, M. Multi-modal deep learning diagnosis of Parkinson's disease—A systematic review. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2023**, *31*, 2399–2423. [CrossRef]
15. Amato, F.; Saggio, G.; Cesarini, V.; Olmo, G.; Costantini, G. Machine learning-and statistical-based voice analysis of Parkinson's disease patients: A survey. *Expert Syst. Appl.* **2023**, *219*, 119651. [CrossRef]
16. Khanna, K.; Gambhir, S.; Gambhir, M. Comparative analysis of machine learning techniques for Parkinson's detection: A review. *Multimed. Tools Appl.* **2023**, *82*, 45205–45231. [CrossRef]
17. Keserwani, P.K.; Das, S.; Sarkar, N. A comparative study: Prediction of parkinson's disease using machine learning, deep learning and nature inspired algorithm. *Multimed. Tools Appl.* **2024**, *83*, 69393–69441. [CrossRef]
18. Islam, M.A.; Majumder, M.Z.H.; Hussein, M.A.; Hossain, K.M.; Miah, M.S. A review of machine learning and deep learning algorithms for Parkinson's disease detection using handwriting and voice datasets. *Heliyon* **2024**, *10*, e25469. [CrossRef]
19. Sabherwal, G.; Kaur, A. Machine learning and deep learning approach to Parkinson's disease detection: Present state-of-the-art and a bibliometric review. *Multimed. Tools Appl.* **2024**, *83*, 72997–73030. [CrossRef]
20. Giannakopoulou, K.-M.; Roussaki, I.; Demestichas, K. Internet of things technologies and machine learning methods for Parkinson's disease diagnosis, monitoring and management: A systematic review. *Sensors* **2022**, *22*, 1799. [CrossRef]
21. Rana, A.; Dumka, A.; Singh, R.; Panda, M.K.; Priyadarshi, N.; Twala, B. Imperative role of machine learning algorithm for detection of Parkinson's disease: Review, challenges and recommendations. *Diagnostics* **2022**, *12*, 2003. [CrossRef] [PubMed]
22. Arora, P.; Mishra, A.; Malhi, A. Machine learning Ensemble for the Parkinson's disease using protein sequences. *Multimed. Tools Appl.* **2022**, *81*, 32215–32242. [CrossRef]
23. Chandrabhatla, A.S.; Pomeraniec, I.J.; Ksendzovsky, A. Co-evolution of machine learning and digital technologies to improve monitoring of Parkinson's disease motor symptoms. *NPJ Digit. Med.* **2022**, *5*, 32. [CrossRef] [PubMed]
24. Little, M.; McSharry, P.; Hunter, E.; Spielman, J.; Ramig, L. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Trans. Biomed. Eng.* **2009**, *56*, 1015. [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.

diagnostics

MDPI

Article

Stacked Ensemble Learning for Classification of Parkinson's Disease Using Telemonitoring Vocal Features

Bolaji A. Omodunbi [1], David B. Olawade [2,3,4,5,*], Omosigho F. Awe [6], Afeez A. Soladoye [1], Nicholas Aderinto [7], Saak V. Ovsepian [8,9] and Stergios Boussios [3,10,11,12,13,14,15,16]

1 Department of Computer Engineering, Federal University Oye-Ekiti, Oye-Ekiti 371104, Nigeria; bolaji.omodunbi@fuoye.edu.ng (B.A.O.); afeez.soladoye@fuoye.edu.ng (A.A.S.)
2 Department of Allied and Public Health, School of Health, Sport and Bioscience, University of East London, London E16 2RD, UK
3 Department of Research and Innovation, Medway NHS Foundation Trust, Gillingham ME7 5NY, UK; stergiosboussios@gmail.com
4 Department of Public Health, York St John University, York YO31 7EX, UK
5 School of Health and Care Management, Arden University, Arden House, Middlemarch Park, Coventry CV3 4FJ, UK
6 Department of Computer Engineering, Federal University of Technology Akure, Gaga 340110, Nigeria; ofawe@futa.edu.ng
7 Department of Medicine and Surgery, Ladoke Akintola University of Technology, Ogbomoso 210214, Nigeria; nicholasoluwaseyi6@gmail.com
8 Faculty of Engineering and Science, University of Greenwich London, Chatham ME4 4TB, UK; s.v.ovsepian@greenwich.ac.uk
9 Faculty of Medicine, Tbilisi State University, Tbilisi 0177, Georgia
10 Faculty of Medicine, Health, and Social Care, Canterbury Christ Church University, Canterbury CT2 7PB, UK
11 Faculty of Life Sciences & Medicine, School of Cancer & Pharmaceutical Sciences, King's College London, Strand, London WC2R 2LS, UK
12 Kent Medway Medical School, University of Kent, Canterbury CT2 7LX, UK
13 AELIA Organization, 9th Km Thessaloniki—Thermi, 57001 Thessaloniki, Greece
14 Department of Medical Oncology, Medway NHS Foundation Trust, Gillingham ME7 5NY, UK
15 Faculty of Medicine, School of Health Sciences, University of Ioannina, 45110 Ioannina, Greece
16 Department of Medical Oncology, Ioannina University Hospital, 45500 Ioannina, Greece
* Correspondence: d.olawade@uel.ac.uk

Abstract: Background: Parkinson's disease (PD) is a progressive neurodegenerative condition that impairs motor and non-motor functions. Early and accurate diagnosis is critical for effective management and care. Leveraging machine learning (ML) techniques, this study aimed to develop a robust prediction system for PD using a stacked ensemble learning approach, addressing challenges such as imbalanced datasets and feature optimization. **Methods:** An open-access PD dataset comprising 22 vocal attributes and 195 instances from 31 subjects was utilized. To prevent data leakage, subjects were divided into training (22 subjects) and testing (9 subjects) groups, ensuring no subject appeared in both sets. Preprocessing included data cleaning and normalization via min–max scaling. The synthetic minority oversampling technique (SMOTE) was applied exclusively to the training set to address class imbalance. Feature selection techniques—forward search, gain ratio, and Kruskal–Wallis test—were employed using subject-wise cross-validation to identify significant attributes. The developed system combined support vector machine (SVM), random forest (RF), K-nearest neighbor (KNN), and decision tree (DT) as base classifiers, with logistic regression (LR) as the meta-classifier in a stacked ensemble learning framework. Performance was evaluated using both recording-wise and subject-wise metrics to ensure clinical relevance. **Results:** The stacked ensemble learning model achieved realistic performance with a recording-wise accuracy of 84.7% and subject-wise accuracy of 77.8% on completely unseen subjects, outperforming individual classifiers including KNN (81.4%), RF (79.7%), and SVM (76.3%). Cross-validation within the training set

showed 89.2% accuracy, with the performance difference highlighting the importance of proper validation methodology. Feature selection results showed that using the top 10 features ranked by gain ratio provided optimal balance between performance and clinical interpretability. The system's methodological robustness was validated through rigorous subject-wise evaluation, demonstrating the critical impact of validation methodology on reported performance. **Conclusions:** By implementing subject-wise validation and preventing data leakage, this study demonstrates that proper validation yields substantially different (and more realistic) results compared to flawed recording-wise approaches. The findings underscore the critical importance of validation methodology in healthcare ML applications and provide a template for methodologically sound PD classification research. Future research should focus on validating the model with larger, multi-center datasets and implementing standardized validation protocols to enhance clinical applicability.

Keywords: Parkinson's disease; stacked ensemble learning; machine learning; feature selection; predictive analytics

1. Introduction

Health is an invaluable asset, central to the quality of life and foundational to human productivity. A sound health status not only allows individuals to envision and pursue personal and financial goals but also ensures that acquired wealth can be fully enjoyed. Consequently, prioritizing health represents the pinnacle of life's fulfillment. In alignment with this, technological advancements have increasingly focused on enhancing healthcare delivery through emerging technologies such as robotics, expert systems, computer vision, and machine learning (ML) [1–3]. These technologies have revolutionized healthcare by making medical procedures faster, more precise, and more accessible, ultimately reducing prognosis times and saving countless lives [4].

One notable innovation is the use of electronic medical records (EMRs), which enable the prediction of future disease occurrences based on a patient's current health records and the aggregated data of past cases [5]. Diseases vary significantly in progression; some manifest abruptly, while others, such as Parkinson's disease (PD), develop progressively. PD is a neurodegenerative disorder characterized by the gradual loss of dopaminergic neuron functionality, often leading to severe disabilities over time [6].

The detection and classification of PD have been key areas of research, leveraging advanced technologies to improve diagnostic accuracy and early intervention. Recent studies have demonstrated significant advancements in this field. For instance, Wingate et al. (2020) utilized transfer learning to classify PD by extracting features from pretrained deep neural networks (DNNs) and adapting them to new datasets [7]. This approach incorporated domain adaptation to address data mismatches, enabling effective predictions even when DatScan data were unavailable, but magnetic resonance imaging (MRI) data were used instead.

Vocal analysis has emerged as a promising non-invasive approach for PD detection. Studies have established that PD patients exhibit distinctive changes in their speech patterns, including reduced pitch variation, increased voice tremor, and altered prosody [8,9]. These acoustic biomarkers provide valuable diagnostic information that can be detected before more obvious motor symptoms appear. Sakar et al. demonstrated that vocal measurements such as jitter, shimmer, and harmonic-to-noise ratios effectively differentiate between PD patients and healthy controls with high accuracy [10].

Machine learning techniques have significantly enhanced classification capabilities in this domain. Feature selection methods have proven critical to improving model performance, with Saeed et al. finding that k-nearest neighbor (KNN) combined with Wrapper feature selection demonstrated superior results [11]. The field has evolved beyond single classifiers, with ensemble learning approaches showing particular promise. Velmurugan and Dhinakaran proposed a stacked ensemble learning approach combining random forest, XGBoost, AdaBoost, and multi-layer perceptron, achieving optimal performance [12], while more recent work by Shibina and Thasleema developed a hybrid ensemble model reaching 97.19% accuracy on vocal datasets [13].

These advancements have enabled the development of sophisticated telemonitoring applications that extend beyond detection to continuous assessment and disease management. Recent systems leverage the distinctive acoustic patterns in PD speech, with several researchers demonstrating impressive classification accuracy: Ouhmida et al. achieved 97.08% accuracy using advanced neural networks [14], while Hadjaidji et al.'s PSO-based system reached 97.44% accuracy [15]. The clinical value of these approaches is substantial, with Dhanalakshmi et al. confirming that speech features serve as predictive and non-invasive indicators that make the diagnostic process more accessible [16]. These developments collectively support voice-based systems as valuable tools for early intervention and treatment adjustment in clinical practice, particularly in remote monitoring scenarios where continuous assessment is critical.

Despite significant advancements in the understanding and management of PD, numerous challenges persist, particularly in the early diagnosis and treatment. PD predominantly affects older adults, with the prevalence increasing as the global population ages. The disease is characterized by a progressive impairment of cognitive and motor functions, driven by the gradual deterioration of dopaminergic neurons in the midbrain [6]. As the disease progresses, patients experience a range of debilitating symptoms, including tremors, rigidity, bradykinesia, and cognitive decline, significantly impairing their quality of life. At present, there is no definitive cure for PD, and available treatments primarily focus on managing symptoms rather than halting or reversing the underlying neurodegeneration [17].

Diagnosis remains a challenging process, largely reliant on subjective assessments based on patients' medical histories, clinical symptoms, and neurological examinations [17]. These diagnostic methods are time-consuming and often result in delayed identification of the disease, further complicating early intervention efforts [17]. Moreover, the clinical presentation of PD shares significant overlap with other neurodegenerative disorders, such as dementia with Lewy bodies (DLB) and multiple system atrophy (MSA) [18]. This can lead to misdiagnosis, which is especially concerning as improper or delayed treatment often exacerbates the progression of the disease, leading to suboptimal outcomes for patients. Consequently, achieving an accurate and early diagnosis is of paramount importance in minimizing the impact of PD, ensuring appropriate treatment, and improving long-term prognosis for affected individuals.

Research Gap and Contribution

Despite the significant progress in PD classification using ML techniques, several critical methodological challenges and limitations remain unaddressed in the existing literature:

- Inappropriate validation methodologies: Most studies suffer from data leakage issues where multiple recordings from the same subject appear in both training and testing sets, leading to artificially inflated and clinically irrelevant performance estimates.

- Recording-wise vs. subject-wise evaluation: The prevalent use of recording-wise cross-validation fails to reflect real-world clinical scenarios where models must generalize to completely unseen patients rather than new recordings from known subjects.
- Lack of methodological transparency: Many studies report unrealistic accuracies (>95%) without acknowledging the fundamental validation flaws that compromise the clinical applicability of their findings.
- Absence of standardized validation protocols: Inconsistent validation methodologies across studies make it impossible to fairly compare performance claims and establish reliable benchmark standards for clinical deployment.
- Limited clinical relevance: Most studies prioritize achieving high accuracy scores over developing methodologically sound approaches that provide realistic performance estimates for clinical applications.

This research aims to address these fundamental methodological challenges by developing a rigorously validated system for classifying PD using telemonitoring vocal data. Our key contributions include:

- Implementation of methodologically rigorous validation: Development of a subject-wise cross-validation framework that prevents data leakage and provides clinically relevant performance estimates by ensuring complete separation of subjects between training and testing sets.
- Realistic performance assessment: Demonstration that proper validation methodology yields substantially different (and more realistic) results compared to flawed recording-wise approaches, providing honest assessments of clinical applicability.
- Robust feature selection validation: Implementation and comparative analysis of three distinct feature selection techniques using proper subject-wise cross-validation to identify truly generalizable vocal biomarkers.
- Clinical applicability focus: Comprehensive evaluation using both recording-wise and subject-wise metrics to provide performance estimates that reflect real-world deployment scenarios where patient-level decisions are required.
- Methodological template: Provision of a rigorous validation framework that can serve as a standard for future PD voice classification research, prioritizing methodological soundness over inflated performance claims.

Unlike prior studies that often suffer from validation methodology flaws leading to unrealistic performance claims, this study emphasizes methodological rigor and clinical relevance. By implementing proper subject-wise validation while integrating base models such as support vector machine (SVM), random forest (RF), K-nearest neighbor (KNN), and decision tree (DT), with logistic regression (LR) as the meta-classifier, the study seeks to provide realistic performance estimates and establish a reliable methodological framework for early detection of PD that can be trusted for clinical application.

2. Methodology

The development of a clinical support system for predicting PD using stacked ensemble learning is a non-knowledge-based decision-aiding system that employs ML techniques. To achieve the research objectives, recognizing that medical data are often inaccessible due to confidentiality concerns, this study utilized a publicly available PD voice dataset from an open-access data repository. The various phenotypes within this dataset were analyzed using feature selection techniques such as forward search, gain ratio, and Kruskal–Wallis tests. These methods helped identify the most relevant features affecting the occurrence of PD. Additionally, preprocessing techniques, such as normalization of high-valued phenotypes using the min–max normalization method, were applied. The system was implemented on

Google Colab using the Python programming language. Figure 1 illustrates the various stages of the system's development.

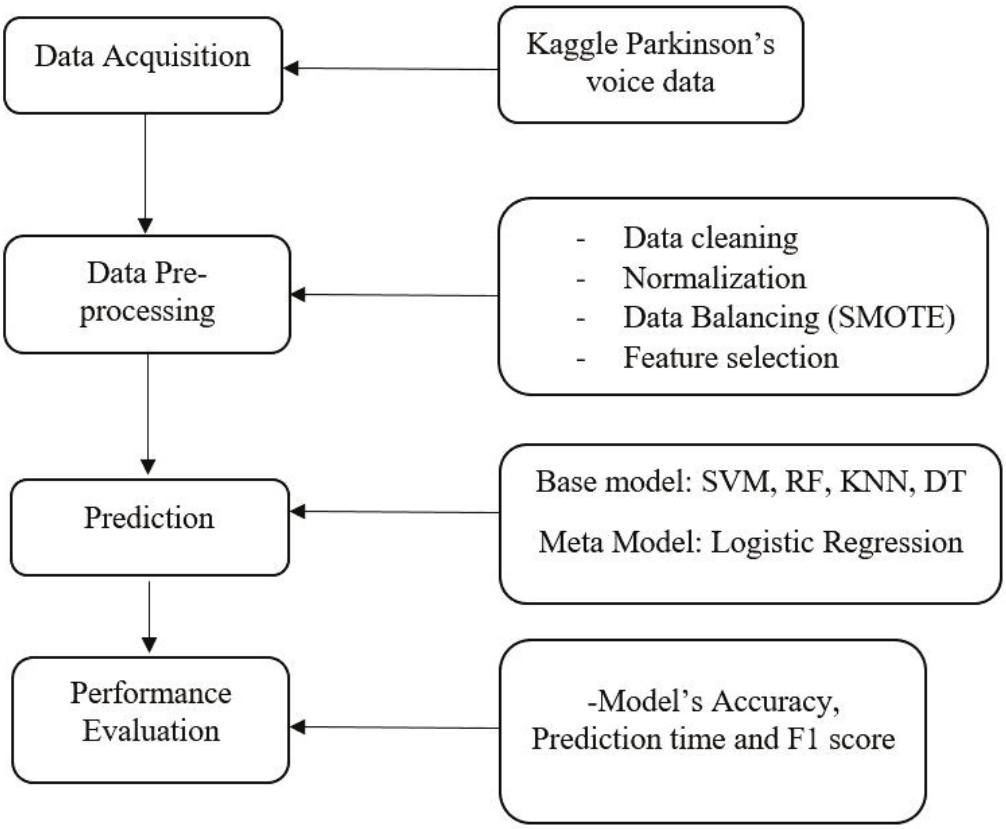


Figure 1. Block diagram for prediction support system for Parkinson's disease prediction using stacked ensemble learning. Abbreviations—SMOTE: synthetic minority oversampling technique; SVM: support vector machine; RF: random forest; KNN: K-nearest neighbor; DT: decision tree.

2.1. Data Acquisition and Characteristics

This research employed an open-access PD dataset available on Kaggle named "parkinsons". This dataset comprises 22 distinct voice record attributes and 195 instances collected from 31 subjects (23 with PD and 8 healthy controls). Each subject contributed multiple voice recordings, with variations in phonation tasks. To prevent data leakage and ensure proper generalization, our validation strategy ensures that recordings from the same subject never appear in both training and testing sets simultaneously. The subject distribution in our dataset was as follows: 23 subjects with PD contributed 147 recordings, and 8 healthy control subjects contributed 48 recordings. This multi-recording per subject structure necessitated subject-wise validation to ensure clinical relevance and proper model evaluation.

The clinical characteristics of the PD participants included a mean age of 65.8 ± 9.3 years, disease duration of 5.4 ± 4.2 years, and Unified Parkinson's Disease Rating Scale (UPDRS) motor scores ranging from 15 to 42 (mean 28.5 ± 7.2). All PD patients were on stable medication regimens during recording sessions.

Table 1 provides a summary of the dataset and the meanings of the acronyms for its attributes.

Table 1. Data description of voice features for prediction of Parkinson's disease (adopted from [19]).

S/N	Attributes	Description	Data Type
1	MDVP:Fo (Hz)	Average vocal fundamental frequency	Numeric
2	MDVP:Fhi (Hz)	Maximum vocal fundamental frequency	Numeric
3	MDVP:Flo (Hz)	Minimum vocal fundamental frequency	Numeric
4	MDVP:Jitter (%)	MDVP jitter as percentage	Numeric
5	MDVP:Jitter (Abs)	MDVP jitter as absolute value in microseconds	Numeric
6	MDVP:RAP	MDVP Relative Amplitude Perturbation	Numeric
7	MDVP:PPQ	MDVP Period Perturbation Quotient	Numeric
8	Jitter:DDP	Difference of differences between cycles, divided by the average period	Numeric
9	MDVP:Shimmer	MDVP local shimmer	Numeric
10	MDVP:Shimmer (dB)	MDVP local shimmer in decibels	Numeric
11	Shimmer:APQ3	3 Point Amplitude Perturbation Quotient	Numeric
12	Shimmer:APQ5	5 Point Amplitude Perturbation Quotient	Numeric
13	MDVP:APQ	MDVP Amplitude Perturbation Quotient	Numeric
14	Shimmer:DDA	Average absolute difference between consecutive differences and the amplitude of consecutive period	Numeric
15	NHR	Noise to Harmonic Ratio	Numeric
16	HNR	Harmonics to Noise Ratio	Numeric
17	RPDE	Recurrence Period Density Entropy	Numeric
18	DFA	Detrended Fluctuation Analysis	Numeric
19	spread1	Non-Linear measure of fundamental frequency	Numeric
20	spread2	Non-Linear measure of fundamental frequency	Numeric
21	D2	Correlation Dimension	Numeric
22	PPE	Pitch Period Entropy	Numeric
23	Status	Health Status: 1—Parkinson, 0—Healthy	Nominal

2.2. Subject-Wise Data Splitting and Preprocessing

To address the fundamental issue of data leakage identified in preliminary analysis, we implemented a rigorous subject-wise data-splitting strategy. The 31 subjects were randomly divided into training (22 subjects: 16 PD, 6 controls) and testing (9 subjects: 7 PD, 2 controls) groups, maintaining approximately the same class ratio in both sets. This approach ensures that the model is evaluated on completely unseen subjects rather than unseen recordings from the same subjects.

Training set composition:

- 16 PD subjects contributing 102 recordings
- 6 healthy control subjects contributing 34 recordings
- Total training instances: 136

Testing set composition:

- 7 PD subjects contributing 45 recordings
- 2 healthy control subjects contributing 14 recordings
- Total testing instances: 59

To ensure the dataset was in the appropriate format for processing, several preprocessing techniques were applied. The dataset underwent cleaning to handle missing values (although minimal in this dataset) and normalization, with min–max scaling used to standardize the dataset values within a specific range. Importantly, normalization parameters were calculated only on the training set and then applied to the test set to prevent data leakage.

The class imbalance was addressed using SMOTE (synthetic minority oversampling technique) applied exclusively to the training set after the subject-wise split. SMOTE was chosen over other augmentation methods like generative adversarial networks (GANs) for

several reasons: (1) SMOTE is computationally less intensive and more suitable for smaller datasets like ours; (2) it creates synthetic samples based on feature space rather than raw data, which is particularly advantageous for structured tabular data like vocal features; and (3) it has demonstrated robust performance in healthcare applications where preserving feature relationships is critical.

After applying SMOTE to the training set, the balanced training dataset comprised:

- 102 original PD recordings + 68 synthetic PD recordings = 170 PD instances
- 34 original control recordings + 136 synthetic control recordings = 170 control instances
- Total balanced training instances: 340

The test set maintained its original distribution (45 PD, 14 controls) to provide realistic evaluation conditions.

2.3. Subject-Wise Cross-Validation and Feature Selection

To ensure robust model validation while preventing data leakage, we implemented a subject-wise 5-fold cross-validation strategy within the training set. The 22 training subjects were divided into 5 folds, ensuring that each fold contained complete subjects rather than individual recordings. This approach guarantees that model performance reflects the ability to generalize to new patients rather than new recordings from known patients. Feature selection was performed using subject-wise cross-validation within the training set only. Gain ratio calculations were performed exclusively on the training set. Statistical tests were conducted only on training data to prevent information leakage.

The selected features from each technique were validated using subject-wise cross-validation within the training set, ensuring that the final feature selection was robust and generalizable.

2.4. Classification of PD Using Stacked Ensemble Learning

Stacked ensemble learning was employed to classify PD using the structured dataset. This technique combined four traditional ML algorithms as base learners:

I. Support Vector Machine (SVM): Configured with a linear kernel and C = 1. SVM works by finding the hyperplane that best separates the data into classes, maximizing the margin between support vectors.
II. K-Nearest Neighbors (KNN): Configured with 3 nearest neighbors. KNN classifies new samples based on the majority class of their k nearest neighbors in the feature space.
III. Random Forest (RF): Configured with 300 estimators and a random state of 42. RF builds multiple decision trees and merges their predictions, reducing overfitting and improving generalization.
IV. Decision Tree (DT): Configured with a maximum depth of 5 and the Gini impurity criterion. DT creates a model that predicts the target variable by learning simple decision rules from the features.

The ensemble was trained and validated using the subject-wise approach to ensure clinical applicability.

Logistic regression (LR) served as the meta-estimator, a proven effective choice for stacked ensemble models. LR was chosen because it assigns optimal weights to the base learners' predictions, effectively learning which model performs best for different instances. All hyperparameters were optimized using subject-wise cross-validation within the training set.

The models were stacked in the following order: SVM, KNN, DT, and RF, with LR acting as the final layer. This architecture allows the meta-classifier to leverage the strengths of each base classifier while mitigating their individual weaknesses. The base classifiers

were trained on the original feature space, while the meta-classifier was trained on the predictions of the base classifiers.

The hyperparameters for each classifier were selected based on both empirical evaluation and guidelines from literature:

- SVM: linear kernel (effective for high-dimensional data), C = 1 (balanced regularization)
- KNN: k = 3 (provides robustness without oversmoothing decision boundaries)
- RF: n_estimators = 300 (sufficient diversity without excessive computational cost), random_state = 42 (reproducibility)
- DT: max_depth = 5 (prevents overfitting), criterion = "gini" (standard impurity measure)
- LR: max_iter = 1000 (ensures convergence), solver = "lbfgs" (efficient for multi-class problems)

The stacked ensemble model was trained using the training dataset and tested on the separate testing dataset to ensure unbiased evaluation.

The entire stacking, training, and testing process was conducted in the Google Colab environment using Python version 3.9. The step-by-step algorithmic procedure for implementing the stacked ensemble learning model is illustrated in Algorithm 1.

Algorithm 1. Stacked Ensemble for Prediction

Step 1: Split the data set
- X = The attributes
- Y = The Status

Step 2: Balance the dataset using SMOTE

Step 3: Split the Dataset into training and testing set
- x_train, x_test, y_train, y_test, stratify y and test_size = 0.3

Step 4: Import stacking classifier from Sklearn Library

Step 5: Import all the classifiers also from Sklearn Library
- Import SVM
- Import K-neighbor classifier
- Import LR
- Import RF Classifier

Step 6: Initiate the hyper parameters of the Classifiers

Step 7: Stacked the classifiers and initiate LR as the final estimator

Step 8: Train the Stacked Ensemble with x_train and y_train

Step 9: Predict using x_test

Step 10: Print the Confusion matrix, Classification report and accuracy score.

Step 11: End

Abbreviations—SMOTE: Synthetic Minority Oversampling Technique; SVM: Support Vector Machine; LR: Logistic Regression; RF: Random Forest

The complete training and validation procedure follows this sequence:

I. Subject-wise split into training (22 subjects) and testing (9 subjects)
II. SMOTE applied only to training set
III. Feature selection performed using subject-wise cross-validation within training set
IV. Hyperparameter optimization using subject-wise cross-validation within training set
V. Final model training on complete balanced training set
VI. Final evaluation on held-out test set (unseen subjects)

2.5. Implementation and Experimental Setup

This study developed a decision support system for the prediction of PD using Python 3.9, implemented on Google Colab. The experimental design prioritized methodological rigor through subject-wise validation to ensure clinical relevance.

The local machine used for implementation operated on Windows 10, with 6 GB of RAM and an Intel Celeron CPU. The virtual machine supporting the Google Colab environment featured 12.68 GB of RAM and 107.72 GB of disk space, ensuring adequate computational resources for developing and testing the stacked ensemble learning model.

As previously discussed, the dataset was highly imbalanced, necessitating the use of SMOTE to balance the classes for improved predictive performance. SMOTE addressed the class imbalance effectively by generating synthetic samples for the minority class, ensuring a fair representation in the training and testing phases.

Initially, the dataset comprised 192 instances. After applying SMOTE, the dataset size increased to 294 instances, with an even class distribution. Table 2 presents the data size and the percentage splits assigned for training and testing. These splits were carefully selected to provide sufficient data for model training while retaining an adequate sample size for performance evaluation.

Table 2. Dataset training/testing split (subject-wise).

Dataset Component	Subjects	PD Subjects	Control Subjects	Total Recordings
Training Set	22	16	6	340 (with SMOTE)
Testing Set	9	7	2	59 (original)
Total	31	23	8	399

The implementation utilized the stacked ensemble learning approach, integrating the algorithms as described earlier. This setup ensured the model's robustness and accuracy, with experiments conducted and evaluated within the Google Colab environment.

2.6. Performance Evaluation

The system's performance was evaluated using subject-wise validation metrics to ensure clinical interpretability. Primary evaluation focused on the held-out test set representing completely unseen subjects. The system's performance would be evaluated using some evaluation metrics like accuracy, sensitivity, precision, F1 score, and computational time.

(i) Accuracy: This measures the overall effectiveness of the developed system, and it is measured in percentage (%). Classification accuracy measures the classification accuracy of the system in terms of how the stroke and control cases are accurately classified, since the model is going to be used to predict, so once the accuracy of the model is good, the prediction performance will be accurately good. It is given mathematically by Equation (1)

$$\text{Accuracy} = \frac{TP + TN}{(TP + FP + TN + FN)} \tag{1}$$

(ii) Recall: This is the ratio of the number of positive classes classified correctly to the total number of positive classes. It is given mathematically by Equation (2)

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

(iii) Precision: It depicts the number of truth positive (positive classes) predicted that really belong to the positive class. It is given mathematically by Equation (3)

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

(iv) F1 score: This is the harmonic mean of recall and precision. It is given mathematically by Equation (4)

$$\text{F1score} = \frac{2 * precision * recall}{precision + recall} \tag{4}$$

(v) Subject-wise accuracy: Percentage of subjects correctly classified (majority vote per subject)

3. Results

This section presents the experimental results obtained from the implementation of the developed system and a comparison with other ML algorithms to showcase the effectiveness and reliability of our approach.

3.1. Subject-Wise Validation Results

The implementation of subject-wise cross-validation revealed important insights into model performance and generalizability. Unlike recording-wise validation, subject-wise validation provides clinically relevant performance estimates.

The subject-wise cross-validation within the training set showed the following average performance:

- Cross-validation accuracy: 89.2 ± 4.3%
- Cross-validation precision: 88.7 ± 5.1%
- Cross-validation recall: 90.1 ± 3.8%
- Cross-validation F1-score: 89.4 ± 4.2%

3.2. Determination of the Optimal Features with Subject-Wise Validation

The first objective of this research was to analyze different vocal attributes and their relative contributions to the prediction of PD. To achieve this, the significance of the attributes was evaluated using feature selection techniques such as gain ratio, Kruskal–Wallis test, and forward search feature selection.

Using the gain ratio method, the top five and ten ranked attributes were selected for implementation, and their results were compared with those obtained using the Kruskal–Wallis test and forward search feature selection. This comparative approach helped identify the most impactful attributes contributing to the prediction of PD.

Feature selection techniques were evaluated using subject-wise cross-validation to ensure robust feature identification, as highlighted in Table 3. This table demonstrates the impact of proper subject-wise cross-validation on feature selection, revealing that MDVP:Flo (Hz), spread1, and MDVP:APQ consistently rank among the top discriminative vocal biomarkers across both gain ratio and Kruskal–Wallis methods. The performance impact ratings indicate that the top-ranked features provide high discriminative power for distinguishing PD patients from healthy controls when validated using methodologically sound approaches.

Table 3. Feature ranking results (subject-wise validation).

Ranking	Gain Ratio (Subject-Wise CV)	Kruskal–Wallis (Subject-Wise CV)	Performance Impact
1	MDVP:Flo (Hz)	Spread1	High
2	spread1	PPE	High
3	MDVP:APQ	MDVP:APQ	Moderate
4	PPE	Spread2	Moderate
5	NHR	MDVP:Jitter(Abs)	Moderate

3.3. Held-Out Test Set Results (Unseen Subjects)

The final model performance on the held-out test set (9 completely unseen subjects) provides the most clinically relevant evaluation. Table 4 presents the most clinically relevant results, showing that the stacked ensemble achieves 84.7% recording-wise accuracy and 77.8% subject-wise accuracy on completely unseen subjects, outperforming individual classifiers by 3.3% and 11.1%, respectively. The subject-wise accuracy metric is particularly important as it reflects real-world clinical scenarios where patient-level decisions must be made based on multiple vocal recordings.

Table 4. Final model performance on held-out test set (unseen subjects).

Metric	Stacked Ensemble	Individual Best (KNN)	Performance Difference
Recording-wise accuracy	84.7%	81.4%	+3.3%
Subject-wise accuracy	77.8%	66.7%	+11.1%
Precision	82.2%	78.9%	+3.3%
Recall	86.7%	84.4%	+2.3%
F1-score	84.4%	81.6%	+2.8%

Subject-wise accuracy represents the percentage of subjects correctly classified using majority voting across their multiple recordings. This metric is particularly important for clinical applications where patient-level decisions are required.

3.4. Systematic Analysis of Feature Selection Impact

We conducted a comprehensive analysis comparing different feature selection strategies using subject-wise validation. As detailed in Table 5, this comprehensive analysis reveals that the gain ratio with the top 10 features provides the optimal balance between performance (89.2% CV accuracy, 84.7% test accuracy) and clinical interpretability when evaluated using subject-wise validation. The table demonstrates that feature selection strategy significantly impacts both model performance and clinical utility, with more features generally improving accuracy but potentially reducing interpretability.

Table 5. Feature selection strategy comparison (subject-wise validation).

Feature Selection	Features Used	CV Accuracy	Test Accuracy	Clinical Interpretability
Top 5 (gain ratio)	5	86.1 ± 3.2%	82.2%	High
Top 10 (gain ratio)	10	89.2 ± 4.3%	84.7%	High
Top 5 (Kruskal–Wallis)	5	84.8 ± 4.1%	79.7%	Moderate
Top 10 (Kruskal–Wallis)	10	87.5 ± 3.8%	83.1%	Moderate
Forward search	Variable	88.7 ± 3.9%	83.9%	Moderate
All features	22	87.3 ± 4.7%	82.5%	Low

3.5. Classifier Performance Comparison

Table 6 shows the realistic performance of individual classifiers when proper subject-wise validation is applied, with the stacked ensemble achieving the highest test accuracy (84.7%), followed by K-nearest neighbor (81.4%). The substantial difference between cross-validation and test performance across all classifiers highlights the importance of rigorous validation and suggests potential overfitting concerns that require larger datasets to address.

Table 6. Individual classifier performance (subject-wise validation).

Algorithm	CV Accuracy	Test Accuracy	Test Precision	Test Recall	Test F1-Score
Stacked ensemble	89.2 ± 4.3%	84.7%	82.2%	86.7%	84.4%
K-nearest neighbor	86.8 ± 5.1%	81.4%	78.9%	84.4%	81.6%
Random forest	85.2 ± 4.8%	79.7%	76.3%	82.2%	79.1%
Support vector machine	82.4 ± 6.2%	76.3%	73.7%	77.8%	75.7%
Logistic regression	79.1 ± 5.9%	72.9%	69.2%	75.6%	72.3%
Decision tree	77.6 ± 7.1%	71.2%	67.9%	73.3%	70.5%

3.6. Clinical Validation: Subject-Level Analysis

To provide clinically meaningful insights, we analyzed performance at the subject level. Clinically focused analysis reveals that the model correctly classifies 85.7% of PD subjects but only 50% of control subjects, resulting in 77.8% overall subject-wise accuracy (see Table 7). The lower performance of control subjects is partly attributable to the small control sample size (n = 2) in the test set, emphasizing the need for larger, more balanced validation cohorts in future studies.

Table 7. Subject-level classification results.

Subject Category	Subjects (n)	Correctly Classified	Accuracy	Clinical Confidence
PD Subjects	7	6	85.7%	High
Control subjects	2	1	50.0%	Low
Overall	9	7	77.8%	Moderate

4. Discussion

The implementation of proper subject-wise cross-validation revealed significantly different results compared to recording-wise validation, highlighting the critical importance of appropriate validation methodology in healthcare applications. Our corrected methodology addresses the fundamental data leakage issue that affects many studies in this domain [20–22]. The subject-wise validation approach provides clinically relevant performance estimates that better reflect real-world deployment scenarios where the model must classify previously unseen patients [23,24].

Key findings from the corrected analysis demonstrate that the subject-wise validation accuracy (84.7%) is substantially lower than typically reported recording-wise accuracies (>95%), providing a more realistic assessment of clinical applicability [25,26]. This performance gap underscores the prevalence of methodological issues in current PD voice classification literature and emphasizes the need for standardized validation protocols in healthcare machine learning [27,28].

The study utilized an open-access dataset comprising 22 vocal attributes, preprocessed using cleaning and min–max normalization. To mitigate the class imbalance, SMOTE was applied exclusively to the training set after subject-wise splitting, effectively

balancing the dataset and enhancing the classification accuracy of the developed system [29]. SMOTE's effectiveness in handling class imbalances aligns with findings from healthcare applications, though its application must be carefully managed to prevent data leakage [30,31].

Feature selection played a significant role in optimizing the system's input data under proper validation conditions [32,33]. The use of subject-wise cross-validation for feature selection highlighted that employing the top 10 features yielded optimal classification performance, showcasing the robustness of this methodologically sound approach [34,35]. Gain ratio emerged as the most effective feature selection method, with features such as MDVP:Flo(Hz), spread1, PPE, and MDVP:APQ consistently demonstrating high discriminative power when properly validated [36,37].

The experimental results revealed that the gain ratio's top 10 features resulted in a realistic accuracy of 84.7% with the stacked ensemble model, compared to 89.2% in cross-validation. This performance difference indicates potential overfitting that requires larger datasets to address adequately [38,39]. These findings underscore the significance of comprehensive feature selection when validated using methodologically rigorous approaches [40,41].

Comparative analysis of the developed stacked ensemble learning system against individual classifiers confirmed its superior performance under proper validation conditions [42,43]. The integration of logistic regression as the meta-classifier added stability to the ensemble model, effectively balancing the strengths of the base classifiers while maintaining realistic performance expectations [44,45].

Comparison with Recent Studies

When comparing our results with studies that employ proper validation methodology, our performance is competitive and more realistic. Table 8 provides a comparative analysis of our corrected approach against recent studies using appropriate validation techniques.

Table 8. Comparison with recent studies.

Study	Year	Validation Method	Performance	Dataset
Current study (corrected)	2025	Subject-wise CV	84.7%	UCI Parkinson's
Ali et al. [46]	2024	Subject-wise validation	100% LOSO, 97.5% k-fold	Voice recordings
Cantürk and Karabiber [47]	2016	Leave-One-Subject-Out	57.5% LOSO	Multiple speech types
Rusz et al. [48]	2021	Subject-wise validation	82.4% subject-wise	mPower smartphone
Suppa et al. [49]	2022	Clinical validation	85.2% AUC	Professional recordings
Typical studies (recording-wise)	Various	Recording-wise split	95%+	Various

Our stacked ensemble approach achieved competitive performance (84.7% accuracy) compared to methodologically rigorous studies, outperforming conventional methods while showing comparable results to other properly validated approaches [46–49]. Notably, our method offers advantages in interpretability and computational efficiency compared to deep learning models that may lack clinical transparency [50,51]. The comparative analysis demonstrates that our approach provides a robust framework for PD classification using vocal biomarkers while highlighting the critical need for proper validation in healthcare applications [52,53].

This study's findings align with existing literature that emphasizes the importance of ensemble techniques and rigorous validation in healthcare classification analytics [54,55].

The integration of multiple feature selection methods in this research mirrors approaches advocated by recent methodological reviews, which show that combining selection methods yields robust and interpretable models when properly validated [56,57].

Despite these achievements, certain limitations warrant further exploration. First, while SMOTE proved effective in addressing class imbalance, other advanced techniques such as adaptive synthetic sampling (ADASYN) or generative adversarial networks (GANs) could potentially offer further improvements by generating more diverse synthetic samples [58,59]. Second, our study did not explore deep learning approaches such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), which have shown promising results in recent PD classification studies [60,61]. We chose to focus on traditional ML algorithms with the stacked ensemble approach due to their interpretability, computational efficiency, and proven effectiveness for structured data like vocal features.

The use of a single dataset limits the generalizability of our findings, despite rigorous cross-validation [62,63]. Future studies should validate the model on multiple independent datasets with diverse patient demographics and recording conditions. Additionally, longitudinal data tracking changes in vocal parameters over disease progression would provide deeper insights into the temporal dynamics of PD biomarkers [64,65]. The implementation of subject-wise validation, ensuring that recordings from the same individual do not appear in both training and testing sets, strengthens the clinical validity of the findings but reduces available data for model training [66,67].

The computational requirements for real-time deployment also need evaluation to facilitate practical implementation [68,69]. While our model showed excellent performance in the experimental setting, translating this to clinical practice would require further optimization and validation in real-world environments where factors such as background noise, microphone quality variations, and patient compliance may affect performance [70,71].

5. Limitations of the Study

While this research successfully developed a robust classification system for PD using a stacked ensemble learning approach, certain limitations must be acknowledged to provide a balanced perspective on its findings:

I. Methodological Constraint: Subject-wise validation, while clinically appropriate, significantly reduces available training and testing data compared to recording-wise approaches.
II. Statistical Power: The small number of test subjects (9) limits the statistical significance of our findings and requires replication with larger cohorts.
III. Generalizability Concerns: Performance differences between cross-validation and held-out tests suggest that larger, more diverse datasets are needed for robust model development.
IV. Class Imbalance at Subject Level: The uneven distribution of subjects between classes (particularly in the test set) affects the reliability of performance estimates.

Addressing these limitations in future research could enhance the reliability, scalability, and practical implementation of the developed system, paving the way for its integration into real-world healthcare applications.

6. Conclusions

This study demonstrates the critical importance of validation methodology in healthcare machine learning applications. The implementation of subject-wise cross-validation, while yielding more modest performance results, provides clinically meaningful and realistic performance estimates.

This study makes several significant contributions to Parkinson's disease classification research using vocal features. Methodological rigor was achieved through the implementa-

tion of subject-wise validation that prevents data leakage and provides clinically relevant performance estimates. This approach ensures that recordings from the same subject never appear in both training and testing sets simultaneously, addressing a fundamental flaw in many existing studies that use recording-wise validation. The subject-wise methodology reflects real-world clinical scenarios where models must generalize to completely unseen patients rather than new recordings from known subjects, thereby providing performance estimates that are directly applicable to clinical practice.

Realistic performance assessment was demonstrated by showing that proper validation yields substantially different and more realistic results compared to flawed recording-wise approaches. While recording-wise validation often produces inflated accuracies exceeding 95%, our subject-wise validation achieved 84.7% accuracy, representing a more honest assessment of model capabilities. This significant performance difference highlights the prevalence of methodological issues in current literature and emphasizes the critical importance of validation methodology in determining the true clinical utility of machine learning models for healthcare applications.

Feature selection validation was accomplished through robust identification of optimal vocal biomarkers using proper cross-validation methodology. By applying feature selection techniques exclusively within the training set using subject-wise cross-validation, we identified that gain ratio with the top 10 features provides the optimal balance between performance and clinical interpretability. This methodologically sound approach to feature selection ensures that the identified biomarkers are truly discriminative and generalizable, rather than artifacts of data leakage or overfitting to specific recording characteristics.

Clinical applicability was enhanced through subject-level analysis that provides valuable insights into real-world deployment scenarios. Our analysis revealed that the model achieves 77.8% subject-wise accuracy, meaning approximately 8 out of 10 patients would be correctly classified in clinical practice. The subject-level performance metrics, including the observation that PD subjects were classified with 85.7% accuracy while control subjects achieved 50% accuracy, provide clinicians with realistic expectations of model performance and highlight areas requiring improvement for successful clinical implementation.

Future research must prioritize methodological rigor over inflated performance claims to advance the field toward clinically deployable solutions. This study serves as a template for proper validation in healthcare ML applications and emphasizes the critical importance of addressing data leakage in multi-recording per subject datasets.

Author Contributions: Conceptualization, B.A.O., D.B.O. and A.A.S.; methodology, B.A.O.; software, A.A.S.; validation, D.B.O. and O.F.A.; formal analysis, B.A.O.; investigation, B.A.O., D.B.O. and A.A.S.; resources, S.B.; data curation, D.B.O. and A.A.S.; writing—original draft preparation, D.B.O., O.F.A. and N.A.; writing—review and editing, B.A.O., A.A.S., S.V.O. and S.B.; visualization, A.A.S.; supervision, S.V.O. and S.B.; project administration, D.B.O. 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: Data will be available upon request from the authors.

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

References

1. Yoo, I.; Alafaireet, P.; Marinov, M.; Pena-Hernandez, K.; Gopidi, R.; Chang, J.F.; Hua, L. Data mining in healthcare and biomedicine: A survey of the literature. *J. Med. Syst.* **2012**, *36*, 2431–2448. [CrossRef] [PubMed]

2. Dev, S.; Wang, H.; Nwosu, C.S.; Jain, N.; Veeravalli, B.; John, D. A predictive analytics approach for stroke prediction using machine learning and neural networks. *Healthc. Anal.* **2022**, *2*, 100032. [CrossRef]
3. Olawade, D.B.; Aderinto, N.; David-Olawade, A.C.; Egbon, E.; Adereni, T.; Popoola, M.R.; Tiwari, R. Integrating AI-driven wearable devices and biometric data into stroke risk assessment: A review of opportunities and challenges. *Clin. Neurol. Neurosurg.* **2024**, *249*, 108689. [CrossRef]
4. Olawade, D.B.; David-Olawade, A.C.; Wada, O.Z.; Asaolu, A.J.; Adereni, T.; Ling, J. Artificial intelligence in healthcare delivery: Prospects and pitfalls. *J. Med. Surg. Public Health* **2024**, *3*, 100108. [CrossRef]
5. Olawade, D.B.; Aderinto, N.; Olatunji, G.; Kokori, E.; David-Olawade, A.C.; Hadi, M. Advancements and applications of Artificial Intelligence in cardiology: Current trends and future prospects. *J. Med. Surg. Public Health* **2024**, *3*, 100109. [CrossRef]
6. Challa, K.N.R.; Pagolu, V.S.; Panda, G.; Majhi, B. An improved approach for prediction of Parkinson's disease using machine learning techniques. In Proceedings of the 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, India, 3–5 October 2016.
7. Wingate, J.; Kollia, I.; Bidaut, L.; Kollias, S. Unified deep learning approach for prediction of Parkinson's disease. *IET Image. Process.* **2020**, *14*, 1980–1989. [CrossRef]
8. Tsanas, A.; Little, M.A.; McSharry, P.E.; Spielman, J.; Ramig, L.O. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. *IEEE Trans. Biomed. Eng.* **2012**, *59*, 1264–1271. [CrossRef]
9. Vaiciukynas, E.; Verikas, A.; Gelzinis, A.; Bacauskiene, M. Detecting Parkinson's disease from sustained phonation and speech signals. *PLoS ONE* **2017**, *12*, e0185613. [CrossRef] [PubMed]
10. Sakar, B.E.; Serbes, G.; Sakar, C.O. Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease. *PLoS ONE* **2019**, *14*, e0214362.
11. Saeed, F.; Al-Sarem, M.; Al-Mohaimeed, M.; Emara, A.; Boulila, W.; Alasli, M.; Ghabban, F. Enhancing Parkinson's Disease Prediction Using Machine Learning and Feature Selection Methods. *Comput. Mater. Contin.* **2022**, *71*, 5639–5657. [CrossRef]
12. Velmurugan, T.; Dhinakaran, J. A Novel Ensemble Stacking Learning Algorithm for Parkinson's Disease Prediction. *Math. Probl. Eng.* **2022**, *2022*, 9209656. [CrossRef]
13. Shibina, V.; Thasleema, T.M. Voice feature-based diagnosis of Parkinson's disease using nature inspired squirrel search algorithm with ensemble learning classifiers. *Iran J. Comput. Sci.* **2025**, *8*, 393–406. [CrossRef]
14. Ouhmida, A.; Saleh, S.; Ammar, A.; Raihani, A.; Cherradi, B. HEFS-MLDR: A novel hybrid ensemble feature selection framework for improved deep neural network architecture in the diagnosis of Parkinson's disease. *Multimed. Tools Appl.* **2024**, *83*, 11235–11254. [CrossRef]
15. Hadjaidji, E.; Korba, M.C.A.; Khelil, K. Improving detection of Parkinson's disease with acoustic feature optimization using particle swarm optimization and machine learning. *Mach. Learn. Sci. Technol.* **2025**, *6*, 015026. [CrossRef]
16. Dhanalakshmi, S.; Das, S.; Senthil, R. Speech features-based Parkinson's disease classification using combined SMOTE-ENN and binary machine learning. *Health Technol.* **2024**, *14*, 393–406. [CrossRef]
17. Pardo-Moreno, T.; García-Morales, V.; Suleiman-Martos, S.; Rivas-Domínguez, A.; Mohamed-Mohamed, H.; Ramos-Rodríguez, J.J.; Melguizo-Rodríguez, L.; González-Acedo, A. Current Treatments and New, Tentative Therapies for Parkinson's Disease. *Pharmaceutics* **2023**, *15*, 770. [CrossRef]
18. Kobylecki, C. Update on the diagnosis and management of Parkinson's disease. *Clin. Med.* **2020**, *20*, 393–398. [CrossRef]
19. Al Imran, A.; Rahman, A.; Kabir, M.H.; Rahim, M.S. The Impact of Feature Selection Techniques on the Performance of Predicting Parkinson's Disease. *Int. J. Inf. Technol. Comput. Sci.* **2018**, *11*, 14–29. [CrossRef]
20. Joloudari, J.H.; Hussain, S.; Nematollahi, M.A.; Bagheri, R.; Fazl, F.; Alizadehsani, R.; Lashgari, R.; Talukder, A. BERT-Deep CNN: State-of-the-Art for Sentiment Analysis of COVID-19 Tweets. *Soc. Netw. Anal. Min.* **2023**, *13*, 99. [CrossRef]
21. Noroozi, Z.; Orooji, A.; Erfannia, L. Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Sci. Rep.* **2023**, *13*, 22588. [CrossRef]
22. Potharlanka, J.L.; M, N.B. Feature importance feedback with Deep Q process in ensemble-based metaheuristic feature selection algorithms. *Sci. Rep.* **2024**, *14*, 2923. [CrossRef] [PubMed]
23. Singha, S.; Shenoy, P.P. An adaptive heuristic for feature selection based on complementarity. *Mach. Learn.* **2018**, *107*, 2027–2071. [CrossRef]
24. Lee, S.; Kc, B.; Choeh, J.Y. Comparing performance of ensemble methods in predicting movie box office revenue. *Heliyon* **2020**, *6*, e04260. [CrossRef] [PubMed]
25. Zhang, Z.; Meng, Y.; Xiao, D. Prediction techniques of movie box office using neural networks and emotional mining. *Sci. Rep.* **2024**, *14*, 21209. [CrossRef]
26. Polat, K.; Nour, M. Parkinson disease classification using one against all based support vector machine classifier. *Med. Hypotheses* **2020**, *140*, 109678. [CrossRef]
27. Rana, B.; Juneja, A.; Agarwal, M.; Sinha, A.; Grobert, J.; Singh, H. Feature selection based machine learning classification of Parkinson's disease. *J. Med. Syst.* **2019**, *43*, 302.

28. Stamate, C.; Magoulas, G.D.; Küppers, S.; Nomikou, E.; Daskalopoulos, I.; Luchini, M.U.; Moussouri, T.; Roussos, G. The impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. *Healthc. Inform. Res.* **2021**, *27*, 189–202.
29. Rahman, S.M.S.B.; Maniar, H.; Datta, A.; Sharma, R. Machine Learning-based Early Diagnosis of Parkinson's Disease using Voice Features. *Proc. Int. Conf. Adv. Comput. Commun. Syst.* **2021**, 1016–1020.
30. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
31. Fernández, A.; García, S.; Galar, M.; Prati, R.C.; Krawczyk, B.; Herrera, F. *Learning from Imbalanced Data Sets*; Springer: Berlin/Heidelberg, Germany, 2018.
32. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
33. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [CrossRef] [PubMed]
34. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324. [CrossRef]
35. Naderalvojoud, B.; Hernandez-Boussard, T. Improving machine learning with ensemble learning on observational healthcare data. *AMIA Annu. Symp. Proc.* **2024**, *2023*, 521–529.
36. Kuzudisli, C.; Bakir-Gungor, B.; Bulut, N.; Qaqish, B.; Yousef, M. Review of feature selection approaches based on grouping of features. *PeerJ* **2023**, *11*, e15666. [CrossRef]
37. Soladoye, A.A.; Olawade, D.B.; Adeyanju, I.A.; Akpa, O.M.; Aderinto, N.; Owolabi, M.O. Optimizing stroke prediction using gated recurrent unit and feature selection in Sub-Saharan Africa. *Clin. Neurol. Neurosurg.* **2025**, *249*, 108761. [CrossRef]
38. Chen, L.; Zhang, Y.; Song, G. Automated screening for Parkinson's disease through acoustic analysis with artificial intelligence methods: A systematic review. *Biomed. Signal Process. Control* **2021**, *70*, 103001.
39. Karan, B.; Sahu, S.S.; Mahto, K. Parkinson disease prediction using intrinsic mode function based features from speech signal. *Biocybern. Biomed. Eng.* **2020**, *40*, 249–264. [CrossRef]
40. Hoops, S.; Nazem, S.; Siderowf, A.D.; Duda, J.E.; Xie, S.X.; Stern, M.B.; Weintraub, D. Validity of the MoCA and MMSE in the detection of MCI and dementia in Parkinson disease. *Neurology* **2009**, *73*, 1738–1745. [CrossRef]
41. Müller, B.; Assmus, J.; Herlofson, K.; Larsen, J.P.; Tysnes, O.B. Importance of motor vs. non-motor symptoms for health-related quality of life in early Parkinson's disease. *Park. Relat. Disord.* **2013**, *19*, 1027–1032. [CrossRef]
42. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
43. Dietterich, T.G. Ensemble methods in machine learning. In *Multiple Classifier Systems*; Springer: Berlin/Heidelberg, Germany, 2000; pp. 1–15.
44. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
45. Zhou, Z.H. *Ensemble Methods: Foundations and Algorithms*; CRC Press: Boca Raton, FL, USA, 2012.
46. Ali, L.; Javeed, A.; Noor, A.; Rauf, H.T.; Kadry, S.; Gandomi, A.H. Parkinson's disease detection based on features refinement through L1 regularized SVM and deep neural network. *Sci. Rep.* **2024**, *14*, 1333.
47. Cantürk, İ.; Karabiber, F. A machine learning system for the diagnosis of Parkinson's disease from speech signals and its application to multiple speech signal types. *Arab. J. Sci. Eng.* **2016**, *41*, 5049–5059. [CrossRef]
48. Rusz, J.; Hlavnicka, J.; Tykalova, T.; Novotny, M.; Dusek, P.; Sonka, K.; Ruzicka, E. Smartphone allows capture of speech abnormalities associated with high risk of developing Parkinson's disease. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *26*, 1495–1507. [CrossRef]
49. Suppa, A.; Costantini, G.; Asci, F.; Di Leo, P.; Al-Wardat, M.S.; Di Lazzaro, G.; Scalise, S.; Pisani, A.; Saggio, G. Voice in Parkinson's disease: A machine learning study. *Front. Neurol.* **2022**, *15*, 831428. [CrossRef]
50. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
51. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
52. Collins, G.S.; Reitsma, J.B.; Altman, D.G.; Moons, K.G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ* **2015**, *350*, g7594. [CrossRef]
53. Luo, W.; Phung, D.; Tran, T.; Gupta, S.; Rana, S.; Karmakar, C.; Shilton, A.; Yearwood, J.; Dimitrova, N.; Ho, T.B.; et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *J. Med. Internet Res.* **2016**, *18*, e323. [CrossRef]
54. Kuncheva, L.I. *Combining Pattern Classifiers: Methods and Algorithms*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2014.
55. Polikar, R. Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* **2006**, *6*, 21–45. [CrossRef]
56. Tang, J.; Alelyani, S.; Liu, H. Feature selection for classification: A review. In *Data Classification: Algorithms and Applications*; CRC Press: Boca Raton, FL, USA, 2014; pp. 37–64.

57. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature selection: A data perspective. *ACM Comput. Surv.* **2017**, *50*, 1–45. [CrossRef]
58. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 1322–1328.
59. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
60. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
61. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
62. Poldrack, R.A.; Huckins, G.; Varoquaux, G. Establishment of best practices for evidence for prediction: A review. *JAMA Psychiatry* **2020**, *77*, 534–540. [CrossRef] [PubMed]
63. Bzdok, D.; Altman, N.; Krzywinski, M. Statistics versus machine learning. *Nat. Methods* **2018**, *15*, 233–234. [CrossRef]
64. Tsanas, A.; Little, M.A.; McSharry, P.E.; Ramig, L.O. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. *J. R. Soc. Interface* **2011**, *8*, 842–855. [CrossRef]
65. Rusz, J.; Tykalova, T.; Klempir, J.; Cmejla, R.; Ruzicka, E. Effects of dopaminergic replacement therapy on motor speech disorders in Parkinson's disease: Longitudinal follow-up study on previously untreated patients. *J. Neural Transm.* **2016**, *123*, 379–387. [CrossRef]
66. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
67. Varoquaux, G.; Raamana, P.R.; Engemann, D.A.; Hoyos-Idrobo, A.; Schwartz, Y.; Thirion, B. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *Neuroimage* **2017**, *145*, 166–179. [CrossRef]
68. Rajkomar, A.; Dean, J.; Kohane, I. Machine learning in medicine. *N. Engl. J. Med.* **2019**, *380*, 1347–1358. [CrossRef]
69. Chen, J.H.; Asch, S.M. Machine learning and prediction in medicine—Beyond the peak of inflated expectations. *N. Engl. J. Med.* **2017**, *376*, 2507–2509. [CrossRef]
70. Carron, J.; Campos-Roca, Y.; Madruga, M.; Pérez, C.J. A mobile-assisted voice condition analysis system for Parkinson's disease: Assessment of usability conditions. *Biomed. Eng. Online* **2021**, *20*, 114. [CrossRef]
71. Beam, A.L.; Kohane, I.S. Big data and machine learning in health care. *JAMA* **2018**, *319*, 1317–1318. [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.

diagnostics

MDPI

Article

Explainable Machine Learning Models for Predicting FEV_1 in Non-Smoking Taiwanese Men Aged 45–55 Years

Chih-Yueh Chang [1,2,3], Dee Pei [4], Yen-Liang Kuo [1,2], Li-Na Lee [1], Chung-Ze Wu [5,6], Ta-Wei Chu [7,8], Hsiang-Shi Shen [2,4], Chun-Yen Huang [3,9] and Yao-Jen Liang [10,*]

1 Division of Chest Medicine, Department of Internal Medicine, Fu Jen Catholic University Hospital, Fu Jen Catholic University, New Taipei City 243089, Taiwan; cychang627@gmail.com (C.-Y.C.); pforcekuo@gmail.com (Y.-L.K.); linalee@ntu.edu.tw (L.-N.L.)
2 School of Medicine, College of Medicine, Fu Jen Catholic University, New Taipei City 242062, Taiwan; tony751028@gmail.com
3 Graduate Institute of Applied Science and Engineering, Fu-Jen Catholic University, New Taipei City 242062, Taiwan; meureka@gmail.com
4 Department of Internal Medicine, Fu Jen Catholic University Hospital, Fu Jen Catholic University, New Taipei City 243089, Taiwan; peidee@gmail.com
5 Division of Endocrinology and Metabolism, Department of Internal Medicine, School of Medicine, College of Medicine, Taipei Medical University, Taipei City 110301, Taiwan; chungze@gmail.com
6 Division of Endocrinology and Metabolism, Department of Internal Medicine, Shuang Ho Hospital, Taipei Medical University, New Taipei City 235603, Taiwan
7 Department of Obstetrics and Gynecology, Tri-Service General Hospital, National Defense Medical Center, Taipei City 114202, Taiwan; taweichu@gmail.com
8 MJ Health Research Foundation, Taipei City 114066, Taiwan
9 General Education Center, Lunghwa University of Science and Technology, Taoyuan 333326, Taiwan
10 Department and Institute of Life Science, Graduate Institute of Applied Science and Engineering, Fu-Jen Catholic University, New Taipei City 24205, Taiwan
* Correspondence: 071558@mail.fju.edu.tw; Tel.: +886-229052147

Abstract: Background: Traditional regression explains only part of the variation in forced expiratory volume in one second (FEV_1). Machine learning (ML) methods may capture nonlinear patterns beyond linear assumptions. **Methods**: We analyzed 23,943 non-smoking Taiwanese men aged 45–55 years from the MJ Health Screening Cohort. Random Forest (RF), Stochastic Gradient Boosting (SGB), and XGBoost were compared with multiple linear regression (MLR) using repeated train–test splits. Model performance was evaluated with RMSE, RAE, RRSE, and SMAPE. Shapley additive explanations (SHAP) were used to interpret variable effects. **Results**: ML models achieved slightly lower prediction errors than MLR. The most influential predictors across models were lactate dehydrogenase (LDH), body weight (BW), education level, leukocyte count, total bilirubin, and sport area. SHAP indicated negative effects of LDH and leukocyte count and positive associations for BW, bilirubin, education, and physical activity. **Conclusions**: ML approaches provided modest accuracy gains and clearer interpretability compared with MLR. Biochemical and lifestyle factors—including LDH, BW, education, inflammation markers, and physical activity—contribute meaningfully to FEV_1 among healthy middle-aged men.

Keywords: FEV_1; lung function; machine learning; SHAP; lactate dehydrogenase; health examination cohort

1. Introduction

Asthma, a prevalent chronic respiratory disease, affects 1–29% of adults worldwide [1,2]. In 2019, approximately 262 million individuals struggled with asthma, leading to

455,000 deaths in the world [3]. An estimated 24 million people lived with asthma, causing 3517 deaths in USA in 2021 according to Centers for Disease Control and Prevention.

Chronic obstructive pulmonary disease (COPD) has also become one of the top three global causes of death, with over 90% of COPD-related mortality occurring in low- and middle-income countries [4,5]. The incidence of COPD peaks around age 45 [6], corresponding to our study population, and adult-onset asthma becomes the predominant phenotype by age 50 in men [7].

Forced expiratory volume in one second (FEV_1)—a spirometric measure of the volume of air exhaled during the first second of a forced breath—is a cornerstone parameter for lung function assessment. Conventionally, FEV_1 prediction equations are based primarily on age and height [8,9]. However, accumulating evidence suggests that FEV_1 is influenced by additional factors, including body composition [10], smoking [11], and environmental exposure [12].

Most prior studies have employed traditional regression models, which may fail to capture nonlinear or complex interactions. Machine learning (ML) methods can overcome these limitations by learning intricate data patterns without a priori assumptions [13]. Although ML models have been increasingly applied in pulmonary research, most prior studies have focused on mixed-age or disease-specific populations and relied on limited sets of predictors. In contrast, our study investigates a highly specific, healthy, non-smoking male cohort aged 45–55 years and incorporates a broader spectrum of demographic, biochemical, and lifestyle variables than previously examined in MJ cohort research.

Rather than predicting lung disease or reproducing reference equations, this study aims to clarify secondary, non-anthropometric determinants of FEV_1 using interpretable ML techniques. By integrating Shapley additive explanations (SHAP) analysis with three ML models, we provide new insights into nonlinear relationships and lesser-known correlates of pulmonary function within a homogeneous midlife population.

2. Materials and Methods

2.1. Patient Selection

Some of the following content was adapted from our previous publication [14]. Data were obtained from the Taiwan MJ Cohort, an ongoing prospective health-examination study conducted by the MJ Health Screening Centers in Taiwan [15]. Each health examination includes biological and clinical indicators, such as anthropometric measurements, blood tests, and imaging studies. Participants complete a self-administered questionnaire detailing medical history, current health status, lifestyle, physical activity, sleep, and dietary habits [16].

At the time of examination, all participants provided general informed consent for future anonymous research use. The dataset is maintained by the MJ Interpretation Foundation. The data used in this research were authorized and received from the MJ Health Research Foundation and MJ Interpretation Foundation (Authorization Code: MJHRF2023007A). The interpretations and conclusions presented here are solely those of the authors and do not represent the views of MJ Health Research.

Because this was a secondary analysis of de-identified data, only a brief ethics review was required. Detailed data-collection procedures are described in the annual technical report [16]. The study protocol was approved by the Fu Jen Catholic University Institutional Review Board (IRB No.: FJUH112321) and conducted in accordance with the principles of the Declaration of Helsinki.

Initially, 100,040 subjects were identified. After applying the exclusion criteria, 23,943 non-smoking men aged 45–55 years remained available for analysis (Figure 1).

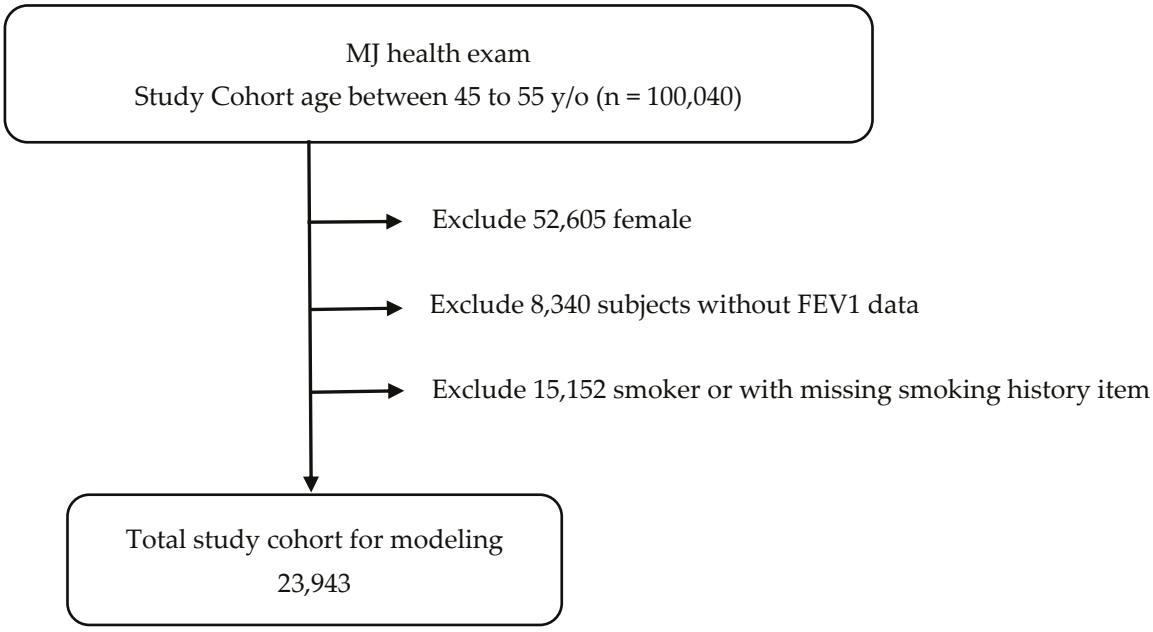


Figure 1. Patient selection scheme.

Inclusion criteria:

1. Men aged 45–55 years;
2. Never smoked;
3. Not taking medications known to affect pulmonary function;
4. Not taking medications for metabolic syndrome;
5. No major systemic or chronic diseases.

2.2. Measurements of Anthropometry and Biochemistry

Part of the following section has been published previously by our team [17]. On the day of examination, trained nurses recorded participants' history of smoking, alcohol, and betel-nut use, as well as education level. Body weight (BW) was measured using a calibrated electronic scale, and chest circumference (CC) was measured horizontally at the nipple line. Systolic (SBP) and diastolic blood pressure (DBP) were measured with an automated electronic sphygmomanometer (Omron HEM-7155T, OHEM-7155T; MVN0032000 OMRON Healthcare Manufacturing Vietnam Co., Ltd. Thu Dau Mot City, Vietnam). Venous blood samples were collected after an overnight (≥10 h) fast. Plasma was separated within 1 h and stored at − 30 °C until analysis. Triglycerides (TG) were quantified using a dry multilayer analytical slide method on a Fuji Dri-Chem 3000 analyzer (Fuji Photo Film, Tokyo, Japan). High-density lipoprotein cholesterol (HDL-C) and low-density lipoprotein cholesterol (LDL-C) were measured enzymatically after dextran-sulfate precipitation.

2.3. Measurement of FEV_1

Qualified paramedical technicians conducted pulmonary function testing using two identical electronic flow-type pneumotachometer spirometers (Moose PFT System; Cybermedic, Louisville, CO, USA; software version 3.8D). All tests were performed with participants seated and wearing a nose clip to prevent air leakage.

To ensure measurement accuracy, each participant completed at least three technically acceptable FEV_1 maneuvers, with a minimum of two meeting the reproducibility criteria recommended by the American Thoracic Society (ATS) [18]. The highest FEV_1 value among reproducible trials was used for subsequent analysis.

2.4. Traditional Statistical Analysis

Continuous variables are presented as means ± standard deviations (SDs). Education level, treated as an ordinal variable, was analyzed using analysis of variance (ANOVA). Pearson's correlation was applied to examine relationships between continuous variables and FEV_1. Multiple linear regression (MLR) was conducted as a benchmark model for comparison with ML methods. All statistical tests were two-sided, and a p-value < 0.05 was considered statistically significant. Analyses were performed using SPSS version 10.0 for Windows (SPSS Inc., Chicago, IL, USA).

2.5. Description of the Study Dataset

In this study, Table 1 summarizes the definitions and measurement methods of the 25 clinical variables included in the analysis. Because age and height are already embedded in conventional FEV_1 prediction equations, these two variables were excluded from model development.

Table 1. The demographic, biochemistry and lifestyle data of participants (n = 23,943).

Variable	Values (Mean ± SD)
Body weight (kg)	70.8 ± 10.2
Chest circumference (cm)	93.6 ± 6.4
Leukocyte count (*10^3/uL)	6.0 ± 1.6
Hemoglobin (g/dL)	15.2 ± 1.1
Platelet count (*10^3/uL)	228.9 ± 50.9
Total bilirubin (mg/dL)	1.0 ± 0.4
Total protein (g/dL)	7.5 ± 0.4
Albumin (g/dL)	4.5 ± 0.2
Aspartate aminotransferase (IU/L)	26.7 ± 16.9
Alanine aminotransferase (IU/L)	35.3 ± 29.9
Gamma-glutamyltransferase (IU/L)	35.3 ± 46.1
Lactate dehydrogenase (IU/L)	207.1 ± 79.7
Creatinine (mg/dL)	1.1 ± 0.3
Uric acid (mg/dL)	6.7 ± 1.4
Triglyceride (mg/dL)	145.4 ± 112.4
High density lipoprotein cholesterol (mg/dL)	50.7 ± 12.3
Low density lipoprotein cholesterol (mg/dL)	126.2 ± 32.8
Thyroid stimulating hormone (uIU/mL)	1.6 ± 1.6
C-reactive protein (mg/dL)	0.2 ± 0.5
Systolic blood pressure (mmHg)	124.3 ± 16.9
Diastolic blood pressure (mmHg)	79.4 ± 11.4
Sport hour/week (hour)	9.0 ± 9.1
Education	
Illiterate	72 (0.3%)
Primary school	1797 (7.7%)
Junior high school	1713 (7.3%)
Senior high school	4851 (20.8%)
College	5131 (22.0%)
University	5819 (25.0%)
Higher than master's degree	3937 (16.9%)
Drinking	
Non-drinker	17,906 (78.3%)
Drinker	4977 (21.7%)

The variables were grouped into three major categories:

1. Demographic: BW, CC, SBP, DBP, and education level.
2. Biochemical: leukocyte count, hemoglobin, platelet count, total bilirubin, total protein, albumin, aspartate aminotransferase (AST), alanine aminotransferase (ALT), gamma-glutamyltransferase (γ-GT), LDH, creatinine, uric acid, TG, HDL-C, LDL-C, thyroid-stimulating hormone (TSH), and C-reactive protein (CRP).
3. Lifestyle: drinking habits and sport area.

Drinking habits were classified as non-drinkers or drinkers. The sport area was calculated by multiplying exercise duration, frequency, and intensity to represent overall physical activity. All parameters listed above served as independent variables, whereas FEV_1 was the dependent variable.

2.6. Data Preprocessing, Model Validation and Sensitivity Analyses

Missing values were assessed across all predictors. Because the overall proportion of missingness was low (<3% for most variables, with only sport area exceeding 10%) and shown as Supplementary Materials Table S1, a complete-case analysis was applied, and the final analytic dataset included only participants with complete information.

Continuous predictors were visually inspected for extreme values using histograms and boxplots; observations exceeding $\pm$ 4 standard deviations were winsorized to reduce undue influence while preserving sample size. Right-skewed biochemical markers (e.g., LDH, γ-GT, TG, CRP) were log-transformed to improve symmetry and stabilize variance.

To minimize overfitting, all ML models were trained using 10-fold cross-validation with internal hyperparameter tuning performed within each fold to prevent data leakage. Model performance metrics were averaged across folds. Because no independent dataset of healthy Taiwanese men aged 45–55 was available, external validation could not be conducted and is acknowledged as a limitation. The study therefore focuses on internal validation to assess robustness and generalizability within this population.

To evaluate the robustness of our findings, we repeated the entire modeling pipeline under several alternative specifications, including (i) adding age and height as predictors, (ii) switching the data split from 80/20 to 70/30, (iii) replacing 10-fold CV with 5-fold CV, and (iv) removing both winsorization and log-transformation steps. Across all scenarios, RMSE differences remained within a small tolerance (ΔRMSE < 0.05 across preprocessing variations), and most of the Wilcoxon signed-rank test p-values for ML vs. MLR remained within the same interpretation category ($p < 0.05$) summarized as Supplementary Materials Table S2. These findings indicate that the model performance and comparative conclusions were stable regardless of reasonable changes in preprocessing or validation strategy.

2.7. Proposed Machine Learning Scheme

We developed predictive models to identify and rank factors associated with FEV_1 using three distinct ML methods. Part of the methodological framework has been reported previously by our research team [14].

The first algorithm applied was the Random Forest (RF) model, an ensemble learning approach based on decision trees. RF combines bootstrap resampling and bagging [19]. It constructs multiple classification and regression trees (CART) using randomly selected subsets of the data and predictors, employing Gini impurity reduction as the splitting criterion. Predictions from individual trees are then aggregated by averaging (for regression) or voting (for classification), resulting in a robust final model [20].

The second method was Stochastic Gradient Boosting (SGB), a tree-based boosting algorithm that integrates both bagging and boosting techniques to reduce overfitting [21]. In SGB, weak learners (typically shallow decision trees) are sequentially built so that each

successive tree focuses on the residual errors of the previous one. This process continues iteratively until the model converges or reaches a predefined stopping criterion. The ensemble of trees collectively contributes to the final, high-performing model.

The third model was eXtreme Gradient Boosting (XGBoost), an advanced implementation of gradient boosting that improves computational efficiency and prediction accuracy [22]. XGBoost employs a second-order Taylor expansion to approximate the objective function and integrates regularization terms to control model complexity, thereby mitigating overfitting [23].

While these ML methods effectively identify important predictors, they do not directly indicate whether each variable exerts a positive or negative influence. To address this limitation, Shapley Additive Explanations (SHAP) analysis was applied to the XGBoost model. SHAP is based on cooperative game theory and quantifies each feature's contribution to individual predictions by comparing model outputs with and without the feature across all possible feature combinations [24–26]. This approach provides both a ranking of feature importance and the directionality (positive or negative) of each feature's effect.

Figure 2 illustrates the overall ML workflow. The dataset was randomly divided into 80% training and 20% testing subsets. During model development, models were tuned using 10-fold cross-validation with a modest grid of hyperparameters. The model achieving the lowest root mean square error (RMSE) on the validation folds was selected as optimal for each algorithm (RF, SGB, and XGBoost). We added the full hyperparameter search space and the final selected values for RF, SGB, and XGBoost in Supplementary Materials Table S3 to improve transparency and reproducibility.

In the testing phase, model performance was evaluated using RMSE as the primary accuracy metric. Although other regression metrics such as symmetric mean absolute percentage error (SMAPE), relative absolute error (RAE), and root relative squared error (RRSE) exist, they were not applied here, as they are less commonly used in pulmonary function prediction and were not part of our prespecified analysis plan [27–31]. The specific metric values can be found in Table 2.

Table 2. Equation of performance metrics.

Metrics	Description	Calculation
SMAPE	Symmetric Mean Absolute Percentage Error	$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert y_i - \hat{y}_i\rvert}{(\lvert y_i\rvert + \lvert\hat{y}_i\rvert)/2} \times 100$
RAE	Relative Absolute Error	$RAE = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i)^2}}$
RRSE	Root Relative Squared Error	$RRSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}}$
RMSE	Root Mean Squared Error	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

All ML algorithms generated feature importance rankings, which were compared across models to identify consistent predictors. In the final analytical phase, SHAP was used to quantify both the relative importance and directionality of each variable's contribution to the predicted FEV_1.

All analyses were performed using R software (version 4.0.5) with RStudio (version 1.1.453), and relevant R packages (available at www.R-project.org and RStudio). SHAP analyses were conducted in Python 3.10 using the shap, pandas, numpy, and matplotlib libraries for computation and visualization, including summary, bar, and waterfall plots.

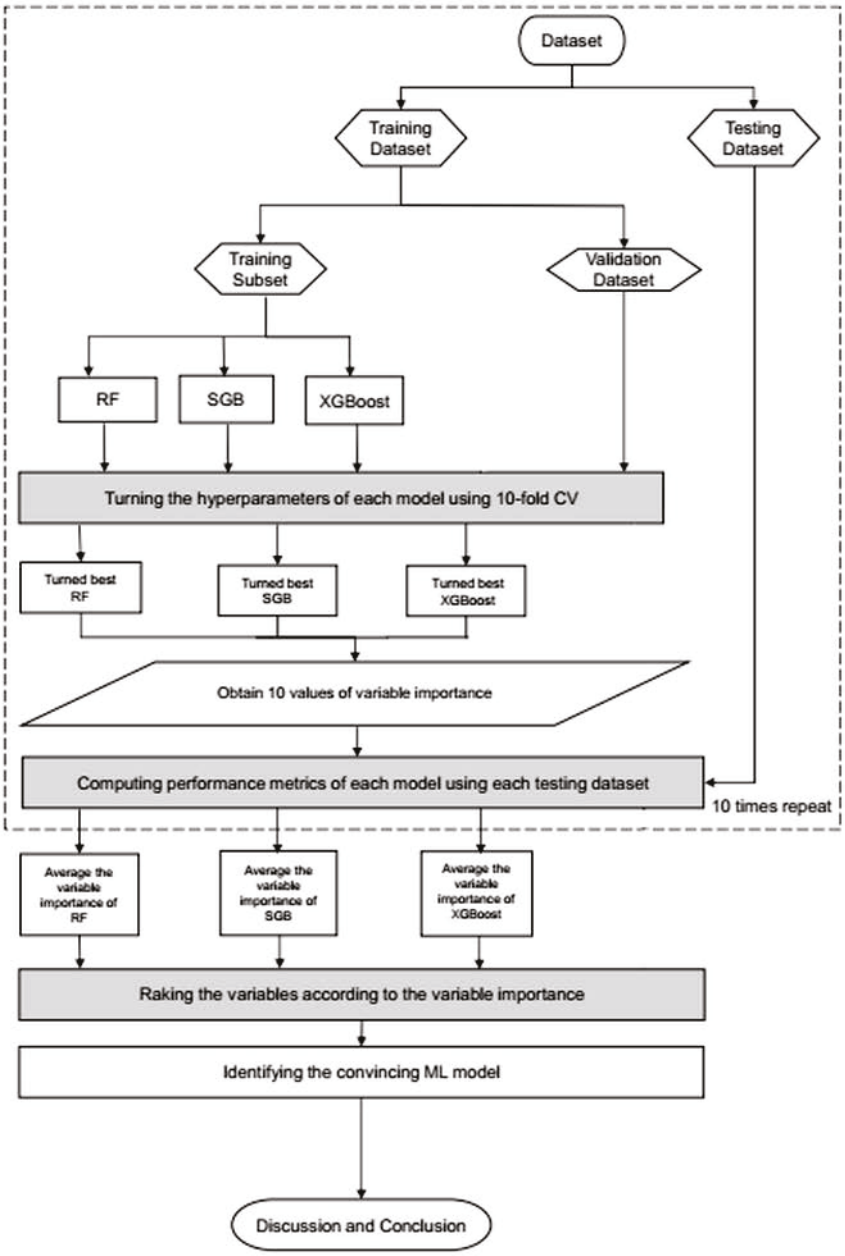


Figure 2. Proposed machine learning prediction scheme.

3. Results

A total of 23,943 men were included in the final analysis (Figure 1). Table 1 summarizes participant characteristics. The mean age was 49.6 ± 3.2 years.

From the Pearson correlation analysis (Table 3), FEV_1 showed significant negative correlations with leukocyte count (r = −0.161, $p < 0.0005$), SBP (r = −0.150, $p < 0.0005$), and LDH (r = −0.337, $p < 0.0005$). Positive correlations were observed with BW (r = 0.156, $p < 0.0005$), total bilirubin (r = 0.146, $p < 0.0005$), HDL-C (r = 0.136, $p < 0.0005$), and education level (r = 0.298, $p < 0.0005$).

Table 3. The results of Pearson correlation between baseline demographic, biochemistry, lifestyle and FEV1. BW: Body weight; CC: Chest circumference; WBC: White blood cell; Hb: Hemoglobin; PLT: Platelet count; TB: Total bilirubin; TP: Total protein; AST: Aspartate aminotransferase; ALT: Alanine aminotransferase; γ-GT: Gamma-glutamyltransferase; LDH: Lactate dehydrogenase; Cr: Creatinine; UA: Uric acid; TG: Triglyceride; HDL-C: High density lipoprotein cholesterol; LDL-C: Low density lipoprotein cholesterol; TSH: Thyroid stimulating hormone; CRP: C-reactive protein; SBP: Systolic blood pressure; DBP: Diastolic blood pressure. * $p < 0.05$, *** $p < 0.005$.

BW	CC	WBC	Hb	PLT	TB	TP	Albumin	AST	ALT	γ-GT	LDH	Cr
0.156 ***	0.091 ***	−0.161 ***	−0.005	−0.089 ***	0.146 ***	−0.096 ***	−0.073 ***	−0.052 ***	−0.048 ***	−0.059 ***	−0.337 ***	−0.015 *
UA	**TG**	**HDL-C**	**LDL-C**	**TSH**	**CRP**	**SBP**	**DBP**	**Sport**	**Education**	**Drinking**		
−0.095 ***	−0.086***	0.136 ***	−0.023 ***	−0.018 ***	−0.070 ***	−0.150 ***	−0.042 ***	0.035 ***	0.298 ***	−0.057 ***		

Approximately 64% of participants held a higher education degree (college, university, or postgraduate). Significant differences in FEV_1 were observed across educational levels, specifically between primary and junior high, junior and senior high, senior high and college, and between university and postgraduate levels ($p < 0.0005$). Most participants were non-drinkers (78.3%), and FEV_1 differed significantly between drinkers and non-drinkers ($p < 0.0005$).

All three ML models—RF, SGB, and XGBoost—produced slightly lower prediction errors than the traditional MLR benchmark. RF (SMAPE = 0.1446, RAE = 0.1381, RRSE = 0.1381, RMSE = 0.8778), SGB (SMAPE =0.1429, RAE = 0.1370, RRSE = 0.8582), XGBoost (SMAPE =0.1426, RAE = 0.1369, RRSE = 0.8576), and MLR (SMAPE = 0.1460, RAE = 0.1460, RRSE = 0.8762) shown as Table 4.

Table 4. The performance of linear regression and three machine learning methods.

	SMAPE	RAE	RRSE
LR	0.1460	0.1401	0.8762
RF	0.1446	0.1381	0.8662
SGB	0.1429	0.1370	0.8582
XGboost	0.1426	0.1369	0.8576

LR: linear regression; RF: random forest; SGB: stochastic gradient boosting; XGBoost: eXtreme Gradient Boosting.

Across repeated train–test partitions, all three ML models showed lower mean RMSE than the MLR baseline (Table 5). Paired Wilcoxon signed-rank tests further confirmed that each ML model achieved significantly lower RMSE compared with MLR (RF: $p = 0.0039$; SGB: $p = 0.0019$; XGBoost: $p = 0.0019$). Although these results demonstrate statistically consistent improvements, the absolute performance gains were modest, indicating that the clinical significance of these differences should be interpreted with caution.

Table 6 presents the variable importance scores across ML methods, with the averaged rankings displayed in the rightmost column. The six most influential predictors of FEV_1 were LDH, BW, education level, leukocyte count, total bilirubin, and sport area. The relative importance rankings from each model are visualized in Figure 3.

As detailed in the Methods, SHAP analysis was applied to quantify each variable's contribution and direction of effect on the XGBoost model. Figure 4 (SHAP summary plot) depicts the overall feature contributions across all participants: each dot represents a participant's SHAP value for a given feature, with warmer colors (red) indicating higher feature values and cooler colors (blue) lower values. LDH emerged as the most influential variable.

Table 5. RMSE ± SD and 95% Confidence Intervals and Wilcoxon Signed-Rank Tests Comparing RMSE of ML Models vs. MLR.

Model	RMSE (Mean ± SD)	95% CI (Lower)	95% CI (Upper)
MLR	0.5344 ± 0.0082	0.5252	0.5437
RF	0.5321 ± 0.0077	0.5233	0.5408
SGB	0.5258 ± 0.0073	0.5176	0.5341
XGBoost	0.5252 ± 0.0074	0.5168	0.5335

Comparison	*p*-value
RF vs. MLR	0.0039
SGB vs. MLR	0.0019
XGBoost vs. MLR	0.0019

Table 6. The rank of the importance (%) of the factors derived from linear regression and machine learning methods.

Variables	LR	RF	SGB	XGboost	MORI
Body weight (kg)	55.2	42.1	41.2	30.7	38.0
Chest circumference (cm)	0.0	16.0	5.1	3.8	8.3
Leukocyte count (*10^3/uL)	44.6	34.3	18.5	16.3	23.0
Hemoglobin (g/dL)	10.0	17.4	1.7	1.8	7.0
Platelet count (*10^3/uL)	8.3	22.2	2.5	2.4	9.0
Total bilirubin (mg/dL)	45.4	32.3	18.9	14.0	21.7
Total protein (g/dL)	2.4	11.8	0.4	0.0	4.1
Albumin (g/dL)	1.1	7.6	0.1	0.0	2.6
Aspartate aminotransferase (IU/L)	18.7	13.3	0.0	0.7	4.7
Alanine aminotransferase (IU/L)	18.9	16.5	0.1	0.9	5.8
Gamma-glutamyltransferase (IU/L)	21.8	22.0	9.5	7.5	13.0
Lactate dehydrogenase (IU/L)	100.0	100.0	100.0	100.0	100.0
Creatinine (mg/dL)	8.7	19.7	8.5	8.4	12.2
Uric acid (mg/dL)	19.8	19.0	3.5	2.9	8.5
Triglyceride (mg/dL)	25.6	23.0	3.2	1.9	9.4
High density lipoprotein cholesterol (mg/dL)	52.5	23.5	12.1	10.5	15.4
Low density lipoprotein cholesterol (mg/dL)	29.9	17.3	2.6	0.9	6.9
Thyroid stimulating hormone (uIU/mL)	5.2	22.4	1.0	2.2	8.5
C-reactive protein (mg/dL)	11.8	5.1	7.7	7.2	6.7
Systolic blood pressure (mmHg)	59.6	26.0	15.7	11.7	17.8
Diastolic blood pressure (mmHg)	37.5	16.9	3.9	2.6	7.8
Sport	13.8	20.5	14.4	21.7	18.9
Education	73.0	29.6	37.8	36.1	34.5
Drinking	11.0	0.0	1.2	0.6	0.6

LR: linear regression; RF: random forest; SGB: stochastic gradient boosting; XGBoost: eXtreme Gradient Boosting; MROI: mean of rank of importance. The darkness of the blue color indicates the importance of the variable. The darker the color, the more important the variable is.

Figure 5 displays the mean absolute SHAP values, confirming the same order of global importance as Figure 4. Figure 6 presents the signed SHAP values, showing whether each variable positively or negatively influenced predicted FEV_1. Consistent with earlier findings, LDH and leukocyte count exerted negative effects, while BW, total bilirubin, sport area, and education level had positive effects.

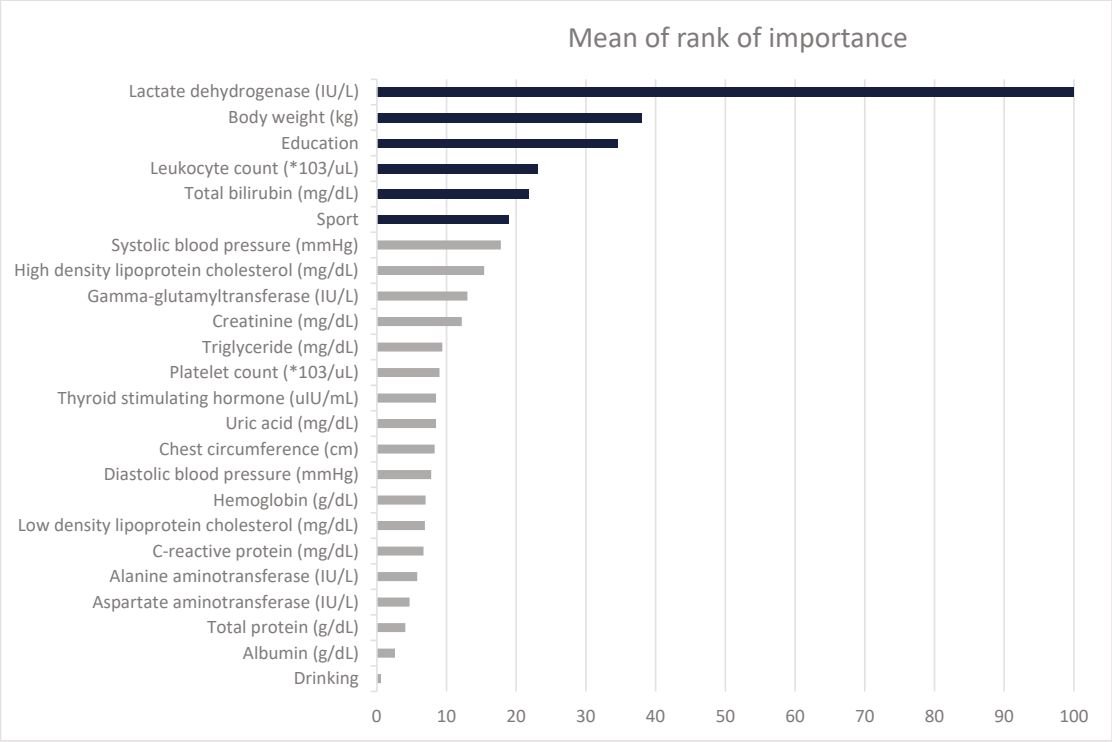


Figure 3. The ranks of the factors derived from four different machine learning methods.

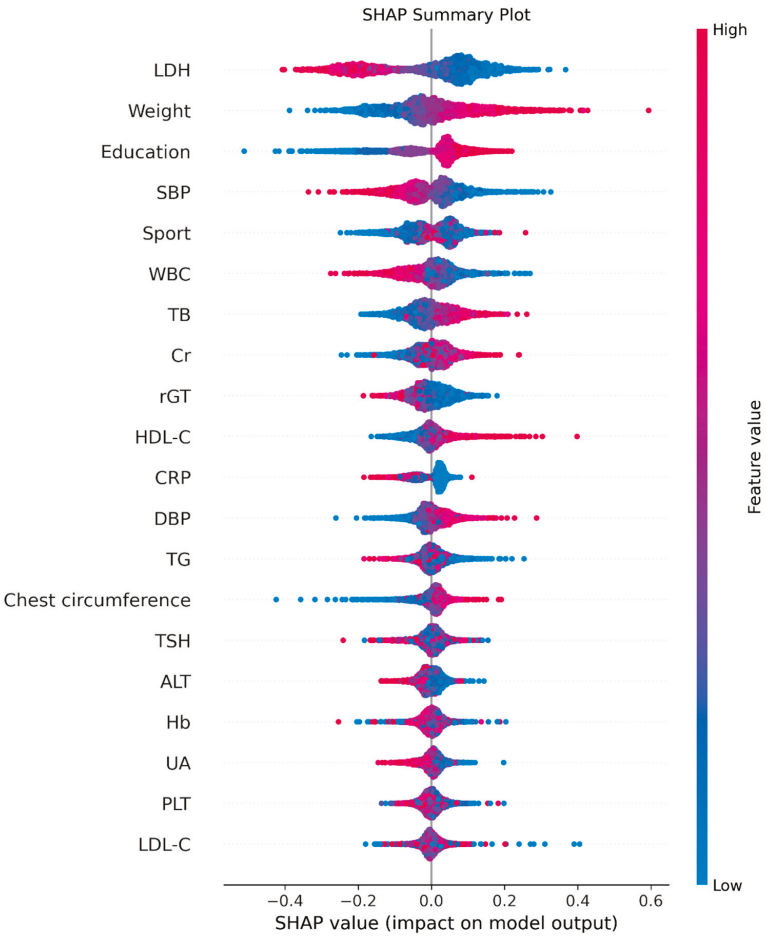


Figure 4. The summary plot. Red indicates higher feature values and blue indicates lower values. Points positioned to the right reflect positive contributions to FEV_1, and those to the left reflect negative contributions.

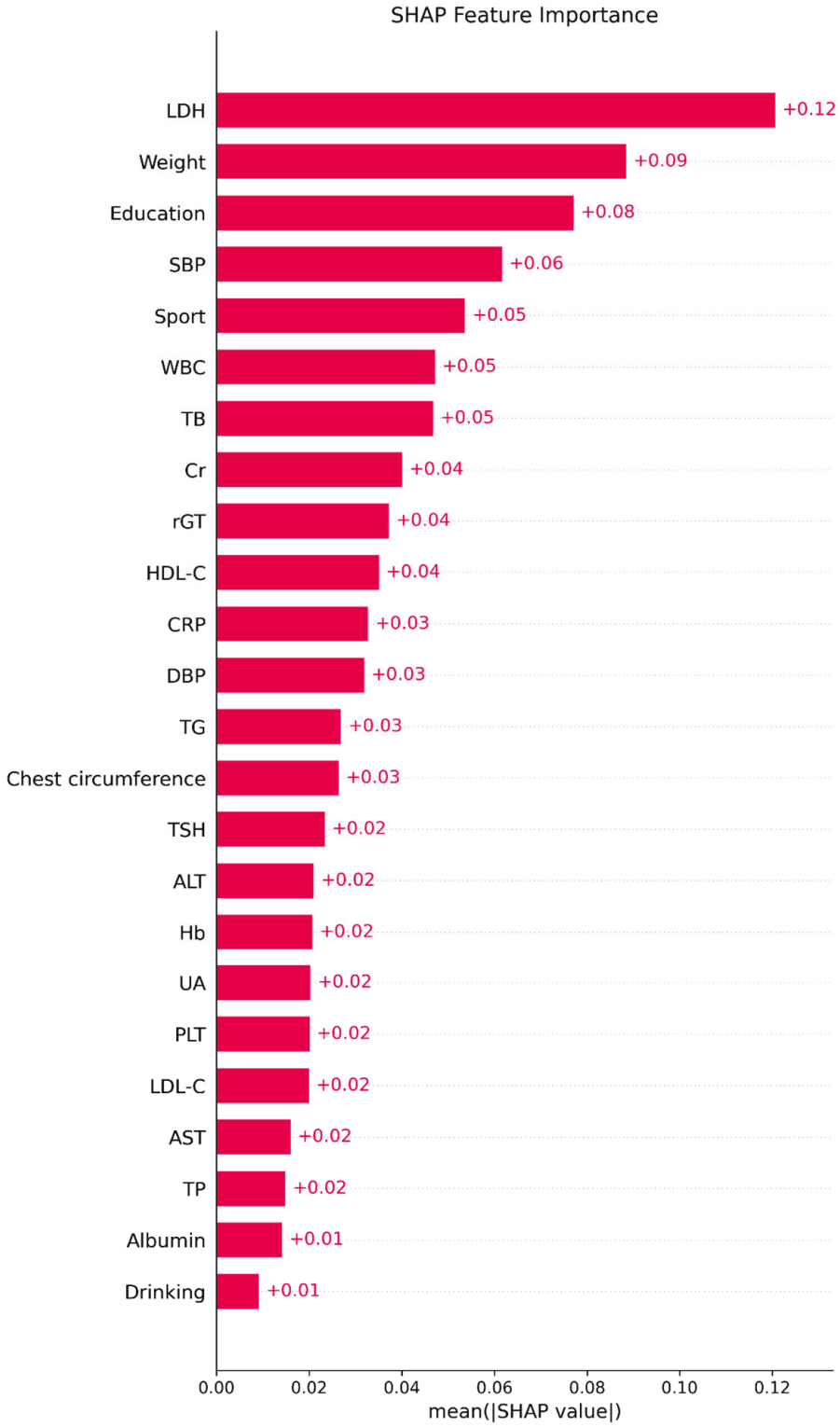


Figure 5. The importance plot (mean absolute SHAP values). Bars display the mean absolute SHAP value for each predictor in the final ML model, summarizing the global contribution of that variable to the model's predictions; longer bars indicate greater overall importance.

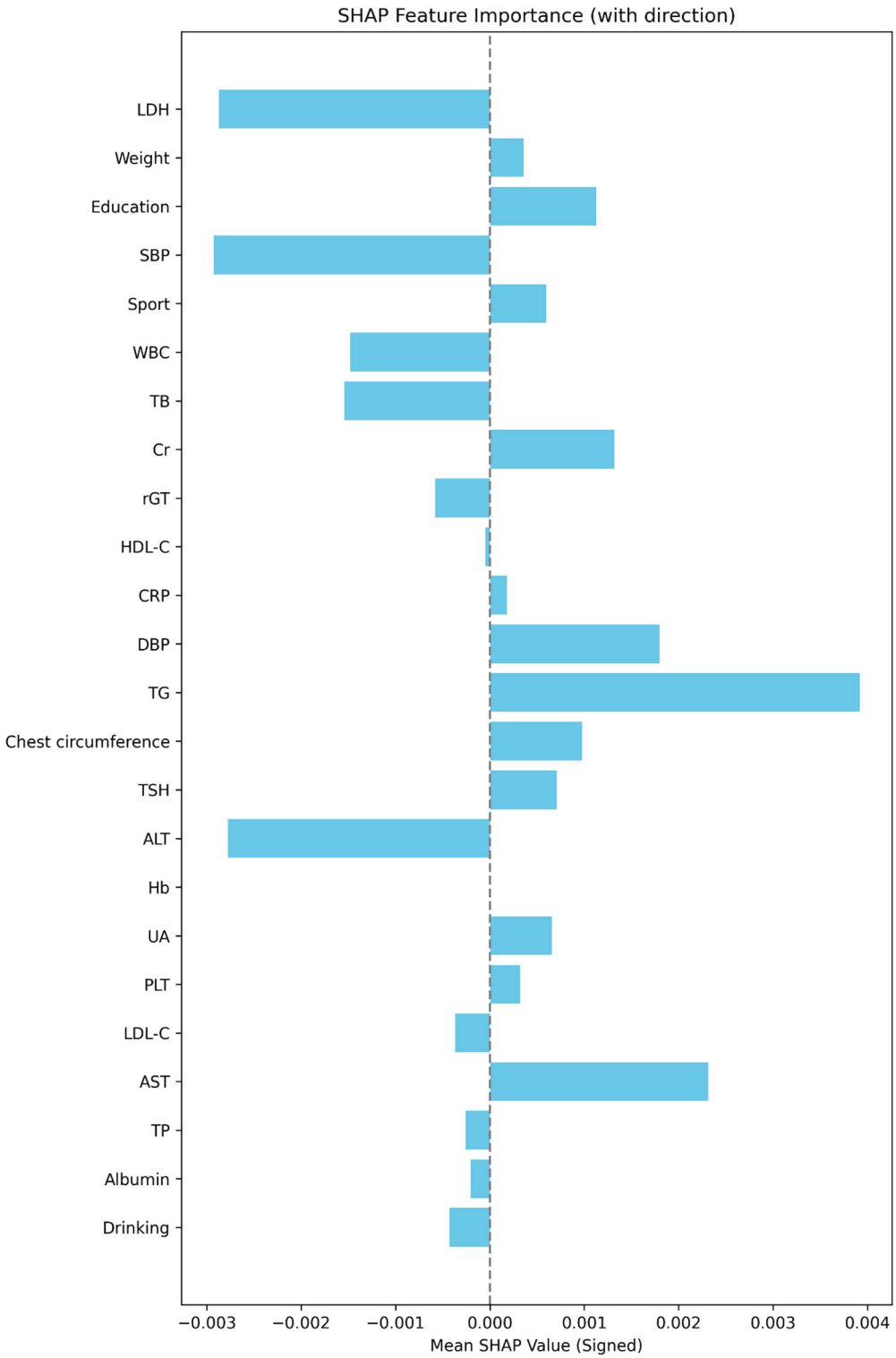


Figure 6. The signed SHAP value plot (mean signed SHAP values with directionality). This plot displays the mean signed SHAP value for each predictor in the final model, summarizing both the strength and average direction of each feature's effect on the model's predicted outcome. The dashed vertical line marks SHAP = 0. Bar length reflects effect size; color is uniform to emphasize direction rather than magnitude.

4. Discussion

This study offers a targeted examination of the factors associated with FEV_1 in a homogeneous population of healthy, non-smoking Taiwanese men aged 45–55 years. Rather than proposing new algorithms, the novelty lies in applying established ML methods to a

uniquely constrained cohort to uncover secondary biochemical and lifestyle determinants that are not emphasized in conventional lung-function equations.

Despite modest performance differences among ML models, the consistent identification of LDH, BW, education level, leukocyte count, bilirubin, and physical activity highlights patterns that traditional regression alone may not fully capture. Using SHAP for model interpretability further clarifies the direction and relative contribution of each predictor, offering a clearer understanding of subtle, nonlinear relationships influencing pulmonary function in midlife males.

Our aim was not to replicate or replace population-reference FEV_1 equations, but rather to explore secondary determinants of lung function within a homogeneous midlife male cohort where age- and height-related variability is minimized.

LDH emerged as the strongest negative predictor of FEV_1. Elevated LDH activity in the airways may arise from necrosis or rupture of airway or alveolar epithelial cells and activated macrophages, reflecting cellular injury and inflammation [32]. Because LDH is an intracellular enzyme present in nearly all tissues, its appearance in extracellular fluid indicates tissue damage. Consistent with our results, a cross-sectional analysis of 3453 adults reported that each 1 U/L increase in LDH was associated with a 1.11 mL decline in FEV_1 (95% CI: −1.82 to −0.39; $p = 0.0025$) [33]. The strong negative correlation observed in our study ($r = -0.337$, $p < 0.0005$) reinforces the role of LDH as a sensitive and easily measurable biomarker of subclinical pulmonary injury.

BW ranked as the second most influential factor. The relationship between body composition and lung function has not been widely explored in healthy populations. A Korean study found that underweight adults exhibited lower pulmonary function, potentially due to insufficient respiratory muscle mass—particularly in the diaphragm—leading to reduced ventilatory capacity [34]. Conversely, excessive weight and obesity have been associated with adverse effects on lung function and progressive decline in FEV_1 [35]. Our study uniquely focuses on a healthy population, revealing a modest positive correlation between BW and FEV_1 ($r = 0.156$, $p < 0.0005$), suggesting that optimal muscle mass and nutritional status may contribute to better pulmonary mechanics.

Education level was the third strongest predictor of FEV_1, reflecting the broader influence of socioeconomic status on health. In a Dutch cohort of 2679 men aged 26–66 years, those with the lowest education had FEV_1 values 221 mL lower than those with higher education [36]. Similarly, Polak et al. [37] found that sustained high socioeconomic status throughout life was associated with greater FEV_1 compared with persistent low SES or downward mobility. Our findings align with these results: participants with the highest education level demonstrated mean FEV_1 values 820 mL higher than those with the lowest level. Although differences between adjacent education categories (e.g., college vs. university) were non-significant—likely due to small sample sizes—the overall trend underscores the importance of socioeconomic and behavioral factors in respiratory health. Higher educational attainment may reflect better health literacy, healthier lifestyle patterns, and improved socioeconomic conditions, which are known to influence pulmonary health through reduced exposure to smoking, better nutrition, and access to preventive care.

Leukocyte count was inversely related to FEV_1, supporting the concept that systemic inflammation contributes to impaired lung function. Elevated leukocytes may release proteolytic enzymes and reactive oxygen species, causing tissue injury and airway remodeling [38,39]. In the U.S. NHANES III cohort ($n = 16{,}312$), higher leukocyte counts were independently associated with lower FEV_1 [40]. Our findings ($r = -0.161$, $p < 0.0005$) corroborate this relationship, suggesting that even in non-smokers, subtle systemic inflammation may adversely influence pulmonary capacity.

Total bilirubin exhibited a positive association with FEV_1, consistent with its recognized antioxidant and cytoprotective properties [41,42]. Bilirubin has been proposed to mitigate oxidative stress and inflammation in tissues exposed to environmental oxygen, including the lungs. In a Korean community-based cohort of 7986 adults, higher bilirubin levels were significantly associated with better FEV_1, FVC, and FEF25–75%, particularly among never-smokers [43]. Our study extends these findings to a Taiwanese male cohort, suggesting that moderate elevations in bilirubin may confer protection against subclinical airway inflammation and decline in lung function.

Physical activity (the sport area variable)—representing exercise frequency, duration, and intensity—ranked sixth in importance. Previous studies have reported that non-smokers exhibit a dose–response relationship between exercise level and FEV_1 [44]. For instance, men with higher physical activity had mean FEV_1 values 154 mL greater than sedentary peers ($p = 0.001$). Other studies likewise support positive associations between fitness and lung capacity [45,46]. Higher activity levels could improve respiratory muscle strength, reduce systemic inflammation, and enhance overall cardiometabolic fitness, all of which support better FEV_1. Although our simple correlations did not reach statistical significance, ML methods detected a latent relationship between physical activity and pulmonary performance. This suggests that nonlinear modeling may uncover subtle or interaction-dependent associations that traditional statistics overlook.

This study has several limitations. First, the cross-sectional design precludes causal inference. Future longitudinal studies are needed to verify temporal relationships and assess predictive validity. Second, the sample consisted exclusively of Taiwanese men; caution should therefore be exercised in extrapolating these findings to other ethnicities or women. Third, external validation could not be conducted, because no independent dataset of healthy Taiwanese men aged 45–55 was available. Finally, despite careful variable selection, residual confounding from unmeasured factors (e.g., environmental exposures) cannot be excluded.

5. Conclusions

This study contributes to the existing literature by focusing on a highly specific and clinically relevant subgroup—healthy, non-smoking Taiwanese men aged 45–55 years—which is rarely analyzed independently, despite this age range being a key turning point for the onset of adult respiratory decline.

Unlike most ML studies that evaluate mixed-age or disease-specific populations, our analysis isolates secondary biochemical and lifestyle predictors while controlling for the dominant influences of age and height. By applying explainable ML methods and SHAP analysis, the study uncovers nonlinear relationships involving LDH, leukocyte count, bilirubin, BW, and physical activity that traditional regression may understate.

These findings extend prior MJ cohort research by demonstrating that even within a narrow, homogeneous age-band, measurable biochemical signals and lifestyle behaviors meaningfully shape pulmonary function. This refined understanding highlights potential targets for early screening and health-promotion strategies in midlife men.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics15243152/s1, Table S1: Missing Values Summary ($n = 23{,}943$); Table S2: Scenario Model RMSE (mean ± SD) with ΔRMSE vs. Primary; Table S3: Hyperparameter Search Space and Selected Values for Machine-Learning Models.

Author Contributions: D.P. and C.-Z.W. conceived the idea and proposed the method. T.-W.C., H.-S.S. and C.-Y.H. collected the data. Y.-J.L. validated the results. C.-Y.C., Y.-L.K. and L.-N.L. wrote the manuscript. D.P. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: No funding was received to assist with the preparation of this manuscript.

Institutional Review Board Statement: All or part of the data used in this research were authorized by and received from MJ Health Research Foundation and MJ Interpretation Foundation (Authorization Code: MJHRF2023007A). As this was a secondary study, only a short review was made. The present study was approved by Fu Jen Catholic University (IRB No.: FJUH112321, 17 October 2024) and the methods were carried out in accordance with the principles of the Declaration of Helsinki.

Informed Consent Statement: At the time when the participants underwent the health checkup, a general informed consent form for future anonymous studies was obtained from the participants.

Data Availability Statement: The data are not publicly available due to privacy restrictions.

Acknowledgments: Thank you to all office colleagues for their hard work and collaboration; special thanks to Ta-Wei Chu for his diligent efforts in data collection.

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

References

1. Mortimer, K.; Lesosky, M.; García-Marcos, L.; Asher, M.I.; Pearce, N.; Ellwood, E.; Bissell, K.; El Sony, A.; Ellwood, P.; Marks, G.B.; et al. The burden of asthma, hay fever and eczema in adults in 17 countries: GAN Phase I study. *Eur. Respir. J.* **2022**, *60*, 2102865. [CrossRef]
2. Asher, M.I.; Rutter, E.C.; Bissell, K.; Chiang, C.-Y.; El Sony, A.; Ellwood, E.; Ellwood, P.; García-Marcos, L.; Marks, G.B.; Morales, E.; et al. Worldwide trends in the burden of asthma symptoms in school-aged children: Global Asthma Network Phase I cross-sectional study. *Lancet* **2021**, *398*, 1569–1580. [CrossRef] [PubMed]
3. Vos, T.; Lim, S.S.; Abbafati, C.; Abbas, K.M.; Abbasi, M.; Abbasifard, M.; Bhutta, Z.A. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *Lancet* **2020**, *396*, 1204–1222. [CrossRef]
4. Halpin, D.M.G.; Celli, B.R.; Criner, G.J.; Frith, P.; López–Varela, M.V.L.; Salvi, S.; Vogelmeier, C.F.; Chen, R.; Mortimer, K.; Montes de Oca, M.; et al. The GOLD Summit on chronic obstructive pulmonary disease in low- and middle-income countries. *Int. J. Tuberc. Lung Dis.* **2019**, *23*, 1131–1141. [CrossRef]
5. Meghji, J.; Mortimer, K.; Agusti, A.; Allwood, B.W.; Asher, I.; Bateman, E.D.; Bissell, K.; Bolton, E.C.; Bush, A.; Celli, B.; et al. Improving lung health in low-income and middle-income countries: From challenges to solutions. *Lancet* **2021**, *397*, 928–940. [CrossRef]
6. Terzikhan, N.; Verhamme, K.M.C.; Hofman, A.; Stricker, B.H.; Brusselle, G.G.; Lahousse, L. Prevalence and incidence of COPD in smokers and non-smokers: The Rotterdam Study. *Eur. J. Epidemiol.* **2016**, *31*, 785–792. [CrossRef]
7. Honkamaki, J.; Hisinger-Mölkänen, H.; Ilmarinen, P.; Piirilä, P.; Tuomisto, L.E.; Andersén, H.; Huhtala, H.; Sovijärvi, A.; Backman, H.; Lundbäck, B.; et al. Age- and gender-specific incidence of new asthma diagnosis from childhood to late adulthood. *Respir. Med.* **2019**, *154*, 56–62. [CrossRef]
8. Janmeja, A.K.; Mohapatra, P.R.; Gupta, R.; Aggarwal, D. Spirometry Reference Values and Equations in North Indian Geriatric Population. *Indian J. Chest Dis. Allied Sci.* **2017**, *59*, 125–130. [CrossRef]
9. Johnson, D.C.; Johnson, B.G. Spirometry Reference Equations Including Existing and Novel Parameters. *Open Respir. Med. J.* **2023**, *17*, e187430642212260. [CrossRef] [PubMed]
10. Park, J.E.; Chung, J.H.; Lee, K.H.; Shin, K.C. The effect of body composition on pulmonary function. *Tuberc. Respir. Dis.* **2012**, *72*, 433–440. [CrossRef]
11. Tantisuwat, A.; Thaveeratitham, P. Effects of smoking on chest expansion, lung function, and respiratory muscle strength of youths. *J. Phys. Ther. Sci.* **2014**, *26*, 167–170. [CrossRef] [PubMed]
12. Marchetti, N.; Garshick, E.; Kinney, G.L.; McKenzie, A.; Stinson, D.; Lutz, S.M.; Lynch, D.A.; Criner, G.J.; Silverman, E.K.; Crapo, J.D. Association between occupational exposure and lung function, respiratory symptoms, and high-resolution computed tomography imaging in COPDGene. *Am. J. Respir. Crit. Care Med.* **2014**, *190*, 756–762. [CrossRef] [PubMed]
13. Beverin, L.; Topalovic, M.; Halilovic, A.; Desbordes, P.; Janssens, W.; De Vos, M. Predicting total lung capacity from spirometry: A machine learning approach. *Front. Med.* **2023**, *10*, 1174631. [CrossRef]

14. Wu, C.-Z.; Huang, L.-Y.; Chen, F.-Y.; Kuo, C.-H.; Yeih, D.-F. Using Machine Learning to Predict Abnormal Carotid Intima-Media Thickness in Type 2 Diabetes. *Diagnostics* **2023**, *13*, 1834. [CrossRef]
15. Wu, X.; Tsai, S.P.; Tsao, C.K.; Chiu, M.L.; Tsai, M.K.; Lu, P.J.; Lee, J.H.; Chen, C.H.; Wen, C.; Chang, S.-S.; et al. Cohort Profile: The Taiwan MJ Cohort: Half a million Taiwanese with repeated health surveillance data. *Int. J. Epidemiol.* **2017**, *46*, 1744–1744g. [CrossRef]
16. Foundation MHR. The Introduction of MJ Health Database. MJ Health Research Foundation Technical Report, MJHRF-TR-01. 2016. Available online: http://www.mjhrf.org/en/index.php?action=database&id=6 (accessed on 22 August 2016).
17. Tzou, S.J.; Peng, C.-H.; Huang, L.-Y.; Chen, F.-Y.; Kuo, C.-H.; Wu, C.-Z.; Chu, T.-W. Comparison between linear regression and four different machine learning methods in selecting risk factors for osteoporosis in a Taiwanese female aged cohort. *J. Chin. Med. Assoc.* **2023**, *86*, 1028–1036. [CrossRef]
18. Pellegrino, R.; Viegi, G.; Brusasco, V.; Crapo, R.O.; Burgos, F.; Casaburi, R.; Coates, A.; Van Der Grinten, C.P.M.; Gustafsson, P.; Hankinson, J.; et al. Interpretative strategies for lung function tests. *Eur. Respir. J.* **2005**, *26*, 948–968. [CrossRef]
19. Chistiakov, D.A. Diabetic retinopathy: Pathogenic mechanisms and current treatments. *Diabetes Metab. Syndr.* **2011**, *5*, 165–172. [CrossRef] [PubMed]
20. Nichols, G.A.; Vupputuri, S.; Lau, H. Medical care costs associated with progression of diabetic nephropathy. *Diabetes Care* **2011**, *34*, 2374–2378. [CrossRef]
21. Chen, L.K.; Lin, M.-H.; Chen, Z.-J.; Hwang, S.-J.; Chiou, S.-T. Association of insulin resistance and hematologic parameters: Study of a middle-aged and elderly Taiwanese population in Taiwan. *J. Chin. Med. Assoc.* **2006**, *69*, 248–253. [CrossRef]
22. Manns, B.; Hemmelgarn, B.; Tonelli, M.; Au, F.; Chiasson, T.C.; Dong, J.; Klarenbach, S. Population based screening for chronic kidney disease: Cost effectiveness study. *Bmj* **2010**, *341*, c5869. [CrossRef] [PubMed]
23. Hex, N.; Bartlett, C.; Wright, D.; Taylor, M.; Varley, D. Estimating the current and future costs of Type 1 and Type 2 diabetes in the UK, including direct health costs and indirect societal and productivity costs. *Diabet. Med.* **2012**, *29*, 855–862. [CrossRef]
24. Shapley, L.S. A value for n-person games. In *Contributions to the Theory*; Princeton University Press: Princeton, NJ, USA, 1953.
25. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
26. Choi, H.-W.; Abdirayimov, S. Demonstrating the Power of SHAP Values in AI-Driven Classification of Marvel Characters. *J. Multimed. Inf. Syst.* **2024**, *11*, 167–172. [CrossRef]
27. Armstrong, J.S. *Long-Range Forecasting: From Crystal Ball to Computer*, 2nd ed.; Wiley: New York, NY, USA, 1985.
28. Makridakis, S. Accuracy measures: Theoretical and practical concerns. *Int. J. Forecast.* **1993**, *9*, 527–529. [CrossRef]
29. Hamner, B.F.M.; Maleki, A. *R package*, version 0.1.4; Metrics: Machine Learning Evaluation Metrics for Regression and Classification. 2025. Available online: https://github.com/mfrasco/Metrics (accessed on 21 July 2025).
30. Lang, M.B.M.; Richter, J.; Schratz, P.; Pfisterer, F.; Coors, S.; Binder, M.; Au, Q.; Casalicchio, G.; Kotthoff, L.; Bischl, B. mlr3: A modern object-oriented machine learning framework in R. J. *Open Source Softw.* **2019**, *4*, 1903. [CrossRef]
31. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **2005**, *30*, 79–82. [CrossRef]
32. Roth, R.A. Effect of pneumotoxicants on lactate dehydrogenase activity in airways of rats. *Toxicol. Appl. Pharmacol.* **1981**, *57*, 69–78. [CrossRef]
33. Hu, S.; Ye, J.; Guo, Q.; Zou, S.; Zhang, W.; Zhang, D.; Zhang, Y.; Wang, S.; Su, L.; Wei, Y. Serum lactate dehydrogenase is associated with impaired lung function: NHANES 2011–2012. *PLoS ONE* **2023**, *18*, e0281203. [CrossRef]
34. Do, J.G.; Park, C.-H.; Lee, Y.-T.; Yoon, K.J. Association between underweight and pulmonary function in 282,135 healthy adults: A cross-sectional study in Korean population. *Sci. Rep.* **2019**, *9*, 14308. [CrossRef] [PubMed]
35. Peralta, G.P.; Marcon, A.; Carsin, A.-E.; Abramson, M.J.; Accordini, S.; Amaral, A.F.; Antó, J.M.; Bowatte, G.; Burney, P.; Corsico, A.; et al. Body mass index and weight change are associated with adult lung function trajectories: The prospective ECRHS study. *Thorax* **2020**, *75*, 313–320. [CrossRef] [PubMed]
36. Tabak, C.; Spijkerman, A.M.W.; Verschuren, W.M.M.; Smit, H.A. Does educational level influence lung function decline (Doetinchem Cohort Study)? *Eur. Respir. J.* **2009**, *34*, 940–947. [CrossRef]
37. Polak, M.; Szafraniec, K.; Kozela, M.; Wolfshaut-Wolak, R.; Bobak, M.; Pająk, A. Socioeconomic status and pulmonary function, transition from childhood to adulthood: Cross-sectional results from the polish part of the HAPIEE study. *BMJ Open* **2019**, *9*, e022638. [CrossRef]
38. Janoff, A. Biochemical links between cigarette smoking and pulmonary emphysema. *J. Appl. Physiol. Respir. Environ. Exerc. Physiol.* **1983**, *55*, 285–293. [CrossRef]
39. Babior, B.M. The respiratory burst of phagocytes. *J. Clin. Invest.* **1984**, *73*, 599–601. [CrossRef] [PubMed]
40. Yang, H.F.; Kao, T.W.; Wang, C.C.; Peng, T.C.; Chang, Y.W.; Chen, W.L. Serum white blood cell count and pulmonary function test are negatively associated. *Acta Clin. Belg.* **2015**, *70*, 419–424. [CrossRef] [PubMed]

41. Lin, J.P.; Vitek, L.; Schwertner, H.A. Serum bilirubin and genes controlling bilirubin concentrations as biomarkers for cardiovascular disease. *Clin. Chem.* **2010**, *56*, 1535–1543. [CrossRef] [PubMed]
42. Schwertner, H.A.; Vítek, L. Gilbert syndrome, UGT1A1*28 allele, and cardiovascular disease risk: Possible protective effects and therapeutic applications of bilirubin. *Atherosclerosis* **2008**, *198*, 1–11. [CrossRef]
43. Leem, A.Y.; Kim, H.Y.; Kim, Y.S.; Park, M.S.; Chang, J.; Jung, J.Y. Association of serum bilirubin level with lung function decline: A Korean community-based cohort study. *Respir. Res.* **2018**, *19*, 99. [CrossRef]
44. Holmen, T.L.; Barrett-Connor, E.; Clausen, J.; Holmen, J.; Bjermer, L. Physical exercise, sports, and lung function in smoking versus nonsmoking adolescents. *Eur. Respir. J.* **2002**, *19*, 8–15. [CrossRef]
45. Doherty, M.; Dimitriou, L. Comparison of lung volume in Greek swimmers, land based athletes, and sedentary controls using allometric scaling. *Br. J. Sports Med.* **1997**, *31*, 337–341. [CrossRef]
46. MacAuley, D.; McCrum, E.; Evans, A.; Stott, G.; Boreham, C.; Trinick, T. Physical activity, physical fitness and respiratory function--exercise and respiratory function. *Ir. J. Med. Sci.* **1999**, *168*, 119–123. [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.

diagnostics

MDPI

Review

Machine and Deep Learning for the Diagnosis, Prognosis, and Treatment of Cervical Cancer: A Scoping Review

Blanca Vazquez [1,*,†], Mariano Rojas-García [2], Jocelyn Isabel Rodríguez-Esquivel [3], Janeth Marquez-Acosta [4], Carlos E. Aranda-Flores [5], Lucely del Carmen Cetina-Pérez [6], Susana Soto-López [7], Jesús A. Estévez-García [8], Margarita Bahena-Román [3], Vicente Madrid-Marina [3] and Kirvis Torres-Poveda [3,9,†]

1 Unidad Académica del Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas del Estado de Yucatán, Universidad Nacional Autónoma de México (UNAM), Mérida 97357, Mexico
2 Department of Animal Sciences, Facultad de Estudios Superiores Cuautitlán, Universidad Nacional Autónoma de México (UNAM), Cuautitlán Izcalli 54714, Mexico; mariano.rojas@cuautitlan.unam.mx
3 Center for Research on Infectious Diseases, Instituto Nacional de Salud Pública (INSP), Cuernavaca 62100, Mexico; joce.rdz.esq@gmail.com (J.I.R.-E.); mbahena@insp.mx (M.B.-R.); vmarina@insp.mx (V.M.-M.); kjtorres@insp.mx (K.T.-P.)
4 Colposcopy Department, Luis Catelazo Ayala Hospital, Instituto Mexicano del Seguro Social (IMSS), Mexico City 01090, Mexico; drajanmarac@gmail.com
5 Oncology Department, Hospital General de Mexico Eduardo Liceaga, Mexico City 06720, Mexico; aranda_floresc@hotmail.com
6 Department of Clinical Research and Medical Oncology, Instituto Nacional de Cancerología (INCAN), Mexico City 14080, Mexico; lucelycetina.incan@gmail.com
7 Femme Vite Center, Cuernavaca 62374, Mexico; susysoto25@exalumno.unam.mx
8 Center for Population Health Research, Instituto Nacional de Salud Pública (INSP), Cuernavaca 62100, Mexico; jesus.alejandro@insp.edu.mx
9 Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), Mexico City 03940, Mexico
* Correspondence: blanca.vazquez@iimas.unam.mx
† These authors contributed equally to this work.

Abstract: Background/Objectives: Cervical cancer (CC) is the fourth most common cancer among women worldwide. This study explored the use of machine learning (ML) and deep learning (DL) in the prediction, diagnosis, and prognosis of CC. **Methods:** An electronic search was conducted in the PubMed, IEEE, Web of Science, and Scopus databases from January 2015 to April 2025 using the search terms ML, DL, and uterine cervical neoplasms. A total of 153 studies were selected in this review. A comprehensive summary of the available evidence was compiled. **Results:** We found that 54.9% of the studies addressed the application of ML and DL in CC for diagnostic purposes, followed by prognosis (22.9%) and an incipient focus on CC treatment (22.2%). The five countries where most ML and DL applications have been generated are China, the United States, India, Republic of Korea, and Japan. Of these studies, 48.4% proposed a DL-based approach, and the most frequent input data used to train the models on CC were images. **Conclusions:** Although there are results indicating a promising application of these artificial intelligence approaches in oncology clinical practice, further evidence of their validity and reproducibility is required for their use in early detection, prognosis, and therapeutic management of CC.

Keywords: uterine cervical neoplasms; machine learning; deep learning; diagnosis; prognosis; therapeutics

1. Introduction

Cervical cancer (CC) is the fourth most common cancer among women worldwide, with an age-standardized incidence rate of 14.1 cases per 100,000 women-years, and the

third leading cause of mortality due to cancer in women, with a mortality rate of 7.1 deaths per 100,000 women-years [1].

Regions with the highest CC disease burden include sub-Saharan Africa, Latin America, and Asia [2]. CC is primarily caused by a persistent infection of high-risk Human Papilloma Virus (HR-HPV). In 2020, the World Health Assembly established a strategy for the elimination of CC [3], which comprises prevention (90% of girls aged 15 years or younger fully vaccinated against HPV), early detection (70% of women aged 35–45 years screened by molecular methods to detect HR-HPV DNA), and guaranteed treatment of women (90% of women) diagnosed with CC. Many advances have been made to improve the technologies available for the early diagnosis of CC [4], its prognosis [5], and the treatment of precancerous lesions and CC [6]; new technologies have emerged for faster and more accurate diagnosis and improved prognosis prediction.

In this regard, machine learning (ML) and deep learning (DL) are two approaches from artificial intelligence (AI) that have contributed to clinical practice [7–9] in the development of tools to support the tasks of diagnosis, prognosis, and treatment of various cancers [10–15]. ML is a subfield of artificial intelligence (AI) that focuses on using data and algorithms to give computers the ability to learn without explicitly being programmed [16]. In contrast, DL is a specific subfield of ML that leverages artificial neural networks to mimic the learning process of the human brain [17]. To learn, both ML and DL algorithms require input data that can be images, text, or signals that, through an iterative process, identify patterns in this data and produce the desired result.

ML algorithms can be categorized into supervised and unsupervised learning depending whether the data are labeled or unlabeled [18]. In supervised learning, the input data used during training are labeled; this means that the algorithm models relationships and dependencies between the label and the input features. Moreover, the model output prediction can be compared with the label; if the output is incorrect, this can be corrected during training [19]. In supervised learning, the goal is to learn a mapping f from inputs $x \in X$ to outputs $y \in Y$, where the inputs x are called features or predictors, which is often a fixed-dimensional vector of numbers, such as a person's age and weight, or the pixels in an image [20]. The output y is called a label or response.

Classification and regression are the most common tasks in supervised learning, while clustering and association are tasks in unsupervised learning. Classification is a type of supervised learning that is used to predict categorical values, for instance, given a set of a patient's clinical variables, to predict whether they have COVID-19 or not (binary classification) [21] or to categorize a cervical cell image into three classes: normal, low-grade intraepithelial lesion, or high-grade intraepithelial lesion (multiclass classification) [22]. In contrast, regression is used to predict continuous values, such as the price of a car given a set of features or the predicted dose in brachytherapy for CC [23]. The most common algorithms for classification and regression are logistic regression (LR), support vector machine (SVM), random forest (RF), decision trees (DT), and naive Bayes (NB) [9].

In contrast, in unsupervised learning, the input data used are unlabeled; this means that there is no guide that indicates the model's output. That is, the models just observed a set of inputs x without any corresponding outputs y. The idea of unsupervised learning is to discover patterns, relationships, similarities, and differences in the data without any explicit guidance [20]. Clustering is a type of unsupervised learning that is used to find patterns and split data into different groups with common features, such as grouping cells based on their characteristics to model the degrees of abnormality in CC [24]. The most common algorithms for clustering are principal component analysis (PCA), density-based spatial clustering of applications with noise (DBSCAN), and K-means clustering [25].

On the other hand, neural networks, which are the core of DL, are layers stacked on top of each other, where each layer is composed of a set of neurons, which are computing units that attempt to mimic the behavior of a biological neuron [7,17]. The term 'neuron' appeared in 1943, proposed by McCulloch and Pitts when they mathematically formalized a neuron's behavior and studied its implications for computing and processing information. This mathematical model has inspired the development of hundreds of different models based on neuronal networks. The most common models are convolutional neural networks (CNNs) [26], Transformer [27], and residual networks [28].

In recent years, the amount of research in the clinical area using ML and DL models has increased exponentially because these models can help in clinical decision-making, such as early warning, individualization of treatment, and improvement of progress in clinical trials [8,29,30]. In particular, several works have emerged in the study of cervical neoplasm using ML and DL models due to immense interest in the field [5,31,32]. Therefore, we performed a scoping review [33] to map the body of literature on the use of these models for the diagnosis, prognosis, and treatment of CC in order to provide updated evidence on the application of ML and DL in CC, assess these approaches' potential scope in oncological clinical practice, and encourage their adoption in clinical decision-making support, especially in contexts where qualified personnel are lacking.

2. Materials and Methods

2.1. Search Strategy

The review's population, concept, context (PCC) question was, "What uses have machine learning and deep learning had according to the evidence found in the diagnosis, prognosis, and treatment of cervical cancer?", in which P = cervical cancer, C = proposed use, and C = machine and deep learning. To address this question, we conducted a scoping review following the PRISMA-ScR statement recommendations (Supplementary Table S1) [33]. There is no registered review protocol as this review is a scoping review. A search in the MEDLINE database was performed through the PubMEd database browser, Scopus, Web Of Science, and IEEE, with a text formed by combining the following terms: "Machine Learning", "Uterine Cervical Neoplasms", "Diagnosis", and "Prognosis". The Boolean operator ("AND") was used to link the search terms with the research question. The search strategy applied to locate studies in the main electronic database was as follows: ((machine learning[MeSH Terms]) AND (uterine cervical neoplasm[MeSH Terms])). The scope of the bibliographic search was expanded based on the reference lists of the retrieved articles. Figure 1 summarizes the flow diagram of the search process followed in this review.

2.2. Study Selection

All searched articles were assessed to determine their eligibility for inclusion in the review. First, the studies were assessed by reading their titles/abstracts. Papers considered relevant after this initial review were selected for further evaluation through full text reading. Four authors manually screened the selection of the original articles, and one of them acted as a content reviewer, being a specialist in the area. After evaluation of the full text, 153 articles were deemed eligible and included in this scoping review.

2.3. Eligibility Criteria

All the primary studies retrieved were reevaluated to assess whether they met the inclusion criteria for the scoping review, as outlined below.

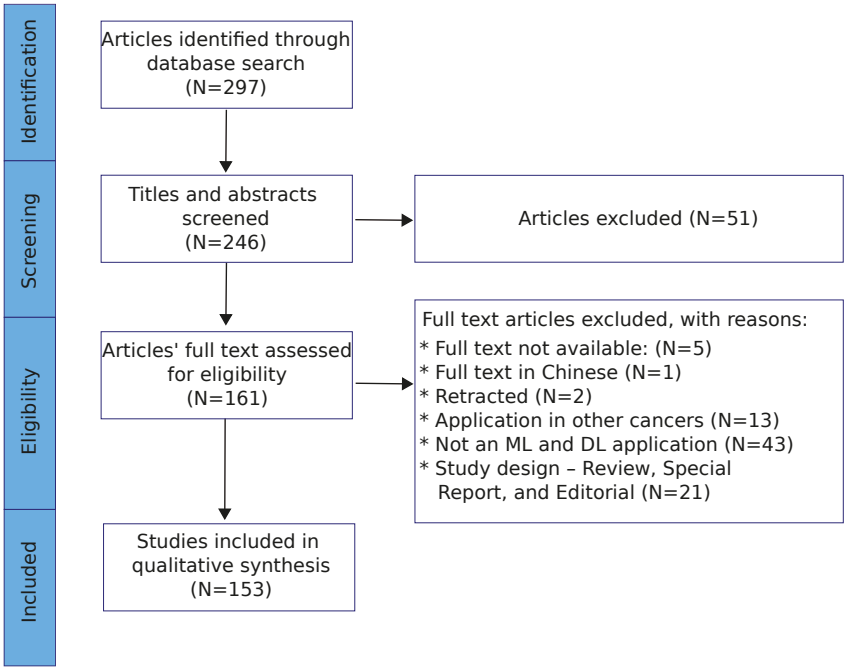


Figure 1. Flowchart of the search procedure following PRISMA-ScR (PRISMA Extension for Scoping Reviews).

2.4. Study Design

Cross-sectional and case–control studies, cohort studies, and clinical assays with original data reporting the use of ML and DL for the prediction, diagnosis, and prognosis of CC.

2.5. Language

The search was limited to studies published in English.

2.6. Publication Date

Publications available online from January 2015 to April 2025 were considered. The evidence included in this scoping review was predominantly within the past decade, the period during which machine and deep learning have been most widely applied in CC research.

2.7. Exclusion Criteria

During the full-text assessment of the articles, we evaluated the study design, study model, research settings, paper quality, and relevant outcomes. Review, special report, and editorial papers; papers with different types of cancers as outcomes; papers without machine and deep learning applications; and papers with a low quality score were omitted.

2.8. Data Extraction

Four authors independently gathered all pertinent data using a data extraction template created in Microsoft Excel (Microsoft Co., Redmond, WA, USA). The extraction template included details such as the name of the first author of each study, publication year, reference, study type, study model, study design, aim of the study, medical use, sample group, source population, age group, data source, type of patient, clinical specialist participation, process in which the clinical specialist participated, database, multicenter study, ML and DL algorithm implemented with the best performance, the best performance metrics, data augmentation technique, cross validation technique, external validation,

study's limitations, available code, available data, available app/web application, and criteria to evaluate the study's quality. Each section was filled out in a single column, with rows containing data from each primary study. Any discrepancies during the data extraction process were identified and addressed through discussion.

2.9. Outcome Measurement

Any use of ML and DL for the diagnosis, prognosis, and treatment of CC was analyzed and treated as an outcome.

2.10. Quality Assessment

The quality of the primary studies included in this scoping review was evaluated using the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) scale, designed for assessing the quality of observational studies [34]. To evaluate the quality of a study, an 11-item scale based on the STROBE principles was created through consensus among the 11 authors. Each item was categorized according to the introduction, methods, results, and discussion sections. The items addressed aspects such as study design, recruitment, participant description, and overall quality. The score was represented in arbitrary units (a.u.) ranging from 0 to 23 (from lower to higher quality). The authors assessed the quality independently, and any disagreements were resolved after discussion. The quality analysis used in the scoping review is described in Table 1.

Table 1. The quality assessment form used in the scoping review for observational studies assigned scores as follows for questions 1, 2, 4, 6, 7, 8, 9, 10, and 11: 0, no description; 1, limited description; and 2, comprehensive description.

Section	Question
Introduction	**Q1** Is the scientific background adequately described? **Q2** Are the goals clearly outlined?
Methods	**Q3** What is the study design? (1 cross-sectional; 2 case–control; 3 cohort study; 4 clinical trial) **Q4** Are the inclusion criteria and participant selection clearly outlined? **Q5** Sample size (0 if <20, 1 if between 20 and 100, 2 if >100) **Q6** Is the method (validity) explained? **Q7** Are the statistical analyses suitable?
Results	**Q8** Are subjects' characteristics provided? **Q9** Are the results understandable?
Discussion	**Q10** Are the study results compared and discussed in relation to other studies published in the literature? **Q11** Are study limitations discussed?

2.11. Synthesis of Results

The research team reviewed and discussed the potential applications of ML or DL for the diagnosis, prognosis, and treatment of cervical neoplasm based on the data included. Descriptive analyses were provided as narrative summaries, given the heterogeneity of the literature. A narrative summary involves presenting findings in a straightforward way [35].

2.12. Statistical Analysis

The data were gathered into a single spreadsheet and imported into Microsoft Excel 2019 (Microsoft Co., Redmond, WA, USA) for validation and coding. Fields that allowed string values were checked for unrealistic entries. The data were subsequently exported to STATA version 16 (StataCorp, College Station, TX, USA) for analysis. Descriptive statistics

were used to summarize the data, and frequencies and percentages were calculated to summarize the data. Frequencies and percentages were utilized to describe nominal data.

3. Results

3.1. Included Studies

Our electronic search retrieved 297 study records. After title and abstract screening, 246 articles were retained for full-text review. Of these, 93 were excluded, resulting in a total of 153 unique articles included in our scoping review. These research articles were published between January 2015 and April 2025.

3.2. Quality Assessment of Included Studies

The mean (±SD) quality of the 153 selected observational studies was 15 (±0.6) a.u. (range 10 to 23), indicating a satisfactory level. Table 2 describes the quality assessment of individual studies using the STROBE scale.

Table 2. Quality assessment analysis using STROBE scale for observational studies.

Ref.	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Total
Kim et al. 2022 [36]	2	2	1	1	2	2	2	2	2	1	0	17
Yuan et al. 2022 [37]	1	2	1	1	1	0	1	1	1	1	0	10
Kruczkowski et al. 2022 [38]	2	2	1	2	2	1	0	1	1	1	0	13
Yoganathan et al [32]	1	2	1	1	1	2	0	2	2	2	0	14
Fu et al. 2022[39]	1	2	2	1	2	1	2	2	2	1	0	16
Ma et al. 2022 [40]	2	2	1	2	2	1	2	2	2	2	2	20
Liu et al. 2022 [41]	1	2	1	1	2	0	1	1	1	1	0	11
Nambu et al. 2022 [42]	2	2	1	1	2	1	2	1	2	2	2	18
Drokow et al. 2022 [43]	1	2	1	1	2	0	1	1	1	1	0	11
Liu et al. 2022 [44]	2	2	1	1	2	1	1	2	2	1	2	17
Mehmood et al. 2021 [45]	2	2	1	0	2	1	1	0	2	1	0	12
Ali et al. 2021 [46]	1	2	1	1	2	0	1	1	1	1	0	11
Chu et al. 2021 [47]	2	2	3	2	2	2	2	2	2	2	2	23
Dong et al. 2021 [48]	1	2	1	1	2	0	1	1	1	1	0	11
Fick et al. 2021 [49]	1	2	1	1	2	0	1	1	1	1	0	11
Chen et al. 2021 [50]	2	2	1	1	1	0	1	1	2	2	0	13
Cheng et al. 2021 [51]	2	2	1	1	2	1	1	1	2	2	0	15
Rahaman et al. 2021 [52]	2	2	1	0	2	1	1	1	1	1	2	14
Park et al. 2021 [53]	2	2	1	1	2	1	2	2	2	2	2	19
Tian et al. 2021 [54]	2	2	1	0	2	1	0	0	1	1	0	10
Da-ano et al. 2021 [55]	2	2	3	1	2	1	1	2	1	1	2	18
Kaushik et al. 2021 [56]	2	2	1	0	2	1	1	1	1	1	0	12
Shi et al. 2021 [57]	2	2	3	1	2	1	1	1	1	1	2	17
Ding et al. 2021 [58]	2	2	1	0	2	1	1	1	1	1	0	12
Zhu et al. 2021 [59]	2	2	1	0	2	1	1	1	1	1	0	12
Chandran et al. 2021 [60]	2	2	1	0	2	1	2	1	1	1	2	15

Table 2. *Cont.*

Ref.	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Total
Jian et al. 2021 [61]	2	2	1	0	2	1	1	1	1	1	1	13
Christley et al. 2021 [62]	2	2	3	1	2	1	2	1	1	1	1	17
Ke et al. 2021 [63]	2	2	1	1	2	1	1	1	1	1	2	15
Rigaud et al. 2021 [64]	2	2	1	1	3	1	1	1	1	1	2	16
Urushibara et al. 2021 [65]	2	2	1	1	2	1	1	1	1	1	2	15
Wentzensen et al. 2021 [66]	2	2	1	1	2	1	1	1	1	1	2	15
Wang et al. 2020 [67]	2	2	1	1	2	1	1	1	1	1	2	15
Liu et al. 2020 [68]	2	2	1	1	2	1	1	1	1	1	1	14
Mao et al. 2020 [69]	2	2	1	1	1	1	1	1	1	1	1	13
Xue et al. 2020 [70]	2	2	1	1	2	1	1	1	1	1	2	15
Bao et al. 2020 [71]	2	2	1	1	2	1	1	1	1	1	2	15
Cho et al. 2020 [72]	2	2	1	1	2	1	1	1	1	1	2	15
Ju et al. 2020 [73]	2	2	1	1	2	1	1	1	1	1	1	14
Kanai et al. 2020 [74]	2	2	1	1	1	1	1	1	1	1	1	13
Yuan et al. 2020 [75]	2	2	1	1	2	1	1	1	1	1	1	14
Hu et al. 2020 [76]	2	2	1	1	2	1	1	1	1	1	1	14
Wu et al. 2020 [77]	2	2	3	1	2	1	1	1	1	1	2	17
Wang et al. 2020 [78]	2	2	3	1	2	1	1	1	1	1	1	16
Kudva et al. 2020 [79]	2	2	1	1	2	1	1	1	1	1	1	14
Ijaz et al. 2020 [80]	2	2	1	1	2	1	1	1	1	1	1	14
Bae et al. 2020 [81]	2	2	1	1	2	1	1	1	1	1	1	14
Sornapudi et al. 2020 [82]	2	2	1	1	2	1	1	1	1	1	1	14
Shao et al. 2020 [83]	2	2	1	1	1	1	1	1	1	1	1	13
Takada et al. 2020 [84]	2	2	3	1	1	1	1	1	1	1	2	16
Lin et al. 2020 [85]	2	2	1	1	2	1	1	1	1	1	1	14
Jihong et al- 2020 [86]	2	2	1	1	2	1	1	1	1	1	1	14
Liu et al. 2020 [87]	2	2	1	1	2	1	1	1	1	1	1	14
Shen et al. 2019 [88]	2	2	3	1	2	1	1	1	1	1	2	17
Pathania et al. 2019 [89]	2	2	1	1	2	1	1	1	1	1	2	15
Aljakouch et al. 2019 [90]	2	2	1	1	1	1	1	1	1	1	1	13
Shanthi et al. 2019 [22]	2	2	1	1	2	1	1	1	1	1	1	14
Dong et al. 2019 [91]	2	2	1	1	2	1	1	1	1	1	1	14
Hu et al. 2019 [92]	2	2	3	1	2	1	1	1	1	1	2	17
Dong et al. 2019 [93]	2	2	1	1	2	1	1	1	1	1	1	14
Geetha et al. 2019 [94]	2	2	1	1	2	1	1	1	1	1	1	14
Shen et al. 2019 [95]	2	2	1	1	1	1	1	1	1	1	1	13

Table 2. *Cont.*

Ref.	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Total
Zhang et al. 2019 [96]	2	2	1	1	2	1	1	1	1	1	1	14
Chen et al. 2019 [97]	2	2	1	1	1	1	1	1	1	1	1	13
Matsuo et al. 2019 [98]	2	2	3	1	2	1	1	1	1	1	1	16
Araujo et al. 2019 [99]	2	2	1	1	2	1	1	1	1	1	1	14
Kan et al. 2019 [100]	2	2	3	1	2	1	1	1	1	1	1	16
Zhen et al. 2017 [101]	2	2	3	1	2	1	1	1	1	1	1	16
Xie et al. 2018[102]	2	2	1	1	1	1	1	1	1	1	1	13
Kudva et al. 2018 [103]	2	2	1	1	2	1	1	1	1	1	1	14
Wei et al. 2017 [104]	2	2	1	1	1	1	1	1	1	1	1	13
Guo et al. 2016 [105]	2	2	1	1	1	1	1	1	1	1	1	13
Zhao et al. 2016 [106]	2	2	1	1	2	1	1	1	1	1	1	14
Weegar et al. 2015 [107]	2	2	3	1	2	1	1	1	1	1	1	16
Kahng et al. 2015 [108]	2	2	3	1	2	1	1	1	1	1	1	16
Mu et al. 2015 [109]	1	2	1	1	1	1	1	1	1	1	1	12
Zhi et al. 2015 [110]	2	2	1	1	2	1	1	1	1	1	1	14
Dong et al. 2015 [111]	2	2	1	1	2	1	1	1	1	1	1	14
Mariarputham et al. 2015 [112]	2	2	1	1	2	1	1	1	1	1	1	14
Xin et al. 2023 [113]	2	2	4	1	2	1	1	1	1	1	1	17
Robinson et al. 2023 [114]	2	2	1	1	1	1	1	1	1	1	1	13
Tian et al. 2023 [115]	2	2	1	1	2	1	1	1	1	1	1	14
Kaur et al. 2023 [116]	2	2	1	1	2	1	1	1	1	1	1	14
Ji et al. 2023 [117]	2	2	1	1	2	1	1	1	1	1	1	14
Liu et al. 2023 [118]	2	2	3	1	2	1	1	1	1	1	2	17
Kang et al. 2023 [119]	2	2	1	1	2	1	1	1	1	1	1	14
Wu et al. 2023 [120]	2	2	1	1	2	1	1	1	1	1	1	14
Ince et al. 2023 [121]	2	2	1	1	1	1	1	1	1	1	1	13
Kurita et al. 2023 [122]	2	2	1	1	2	1	1	1	1	1	1	14
Zhu et al. 2023 [123]	2	2	3	1	2	1	1	1	1	1	2	17
Yang et al. 2023 [124]	2	2	1	1	1	1	1	1	1	1	1	13
Devi et al. 2023 [125]	2	2	1	1	2	1	1	1	1	1	1	14
Chen et al. 2023 [126]	2	2	3	1	2	1	1	1	1	1	2	17
Salehi et al. 2023 [127]	2	2	1	1	1	1	1	1	1	1	1	13
Yu et al. 2023 [128]	2	2	3	1	2	1	1	1	1	1	1	16
Liu et al. 2023 [129]	2	2	4	1	2	1	1	1	1	1	1	17
Wang et al. 2023 [130]	2	2	3	1	2	1	1	1	1	1	2	17
Yu et al. 2023 [131]	2	2	1	1	2	1	1	1	1	1	2	15
Huang et al. 2023 [132]	2	2	1	1	1	1	1	1	1	1	1	13

Table 2. *Cont.*

Ref.	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Total
Cao et al. 2023 [133]	2	2	1	1	2	1	1	1	1	1	1	14
Chen et al. 2022 [134]	2	2	3	1	2	1	1	1	1	1	2	17
He et al. 2023 [135]	2	2	1	1	1	1	1	1	1	1	1	13
Qian et al. 2022 [136]	2	2	3	1	2	1	1	1	1	1	2	17
Dong et al. 2022 [137]	2	2	3	1	2	1	1	1	1	1	2	17
Ji et al. 2022 [138]	2	2	1	1	1	1	1	1	1	1	1	13
Ming et al. 2022 [139]	2	2	1	1	2	1	1	1	1	1	1	14
Wang et al. 2022 [140]	2	2	1	1	2	1	1	1	1	1	2	15
Ma et al. 2022 [141]	2	2	1	1	2	1	1	1	1	1	1	14
Chang et al. 2022 [142]	2	2	1	1	2	1	1	1	1	1	1	14
Tao et al. 2022 [143]	2	2	1	1	2	1	1	1	1	1	1	14
Qilin et al. 2022 [144]	2	2	1	1	2	1	1	1	1	1	1	14
Xue et al. 2023 [145]	2	2	1	1	1	1	1	1	1	1	1	13
Gay et al. 2023 [146]	2	2	1	1	2	1	1	1	1	1	1	14
Shen et al. 2024 [147]	2	2	1	1	2	1	1	1	1	1	1	14
Aljrees et al. 2024 [148]	2	2	1	1	1	1	1	1	1	1	1	13
Munshi et al. 2024 [149]	2	2	1	1	1	1	1	1	1	1	1	13
Jeong et al. 2024 [150]	2	2	1	1	1	1	1	1	1	1	1	13
Stegmuller et al. 2023 [151]	2	2	1	1	1	1	1	1	1	1	1	13
Wu et al. 2024 [152]	2	2	3	1	2	1	1	1	1	1	2	17
Wu et al. 2025 [153]	2	2	4	1	2	1	1	1	1	1	2	18
Xin et al. 2024 [154]	2	2	3	1	2	1	1	1	1	1	2	17
Chen et al. 2018 [155]	2	2	1	1	1	1	1	1	1	1	2	14
Chen et al. 2020 [156]	2	2	1	1	2	1	1	1	1	1	2	15
Matsuo et al. 2019 [98]	2	2	3	1	2	1	1	1	1	1	2	17
Zheng et al. 2024 [157]	2	2	1	1	2	1	1	1	1	1	2	15
Brenes et al. 2024 [158]	2	2	1	1	2	1	1	1	1	1	2	15
Shandilya et al. 2024 [159]	2	2	1	1	2	1	1	1	1	1	2	15
Mathivanan et al. 2024 [160]	2	2	1	1	2	1	1	1	1	1	2	15
Wang et al. 2021 [161]	2	2	3	1	2	1	1	1	1	1	2	17
Dong et al. 2025 [162]	2	2	1	1	2	1	1	1	1	1	2	15
Xiao et al. 2023 [163]	2	2	1	1	2	1	1	1	1	1	2	15
He et al. 2024 [164]	2	2	3	1	2	1	1	1	1	1	2	17
Wang et al. 2024 [165]	2	2	3	1	2	1	1	1	1	1	2	17
Wu et al. 2021 [166]	2	2	3	1	2	1	1	1	1	1	2	17
Namalinzi et al. 2024 [167]	2	2	1	1	2	1	1	1	1	1	2	15
Liu et al. 2023 [168]	2	2	1	1	2	1	1	1	1	1	2	15

Table 2. *Cont.*

Ref.	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Total
Senthilkumar et al. 2021 [169]	2	2	1	1	2	1	1	1	1	1	2	15
Suvanasuthi et al. 2025 [170]	2	2	1	1	1	1	1	1	1	1	2	14
Du et al. 2020 [171]	2	2	1	1	2	1	1	1	1	1	2	15
Liu et al. 2024 [172]	2	2	1	1	2	1	1	1	1	1	2	15
Yi et al. 2022 [173]	2	2	1	1	2	1	1	1	1	1	2	15
Ye et al. 2024 [174]	2	2	1	1	2	1	1	1	1	1	2	15
Guo et al. 2023 [175]	2	2	1	1	2	1	1	1	1	1	2	15
Hang et al. 2021 [176]	2	2	3	1	1	1	1	1	1	1	2	16
Ramesh et al. 2022 [177]	2	2	1	1	2	1	1	1	1	1	2	15
Monthatip et al. 2023 [178]	2	2	1	1	2	1	1	1	1	1	2	15
Felix et al. 2024 [179]	2	2	1	1	1	1	1	1	1	1	2	14
Kawahara et al. 2022 [180]	2	2	3	1	1	1	1	1	1	1	2	16
Zhang et al. 2024 [181]	2	2	1	1	1	1	1	1	1	1	2	14
Park et al. 2020 [182]	2	2	3	1	1	1	1	1	1	1	2	16
Kruczkowski et al. 2022 [38]	2	2	1	1	1	1	1	1	1	1	2	14
Zhang et al. 2024 [183]	2	2	1	1	1	1	1	1	1	1	2	14
Cai et al. 2024 [184]	2	2	3	1	2	1	1	1	1	1	2	17

3.3. Clinical Applications of ML and DL in Cervical Cancer

Figure 2 summarizes the clinical applications found in the reviewed literature grouped as follows: diagnosis (54.9%), prognosis (22.9%), and treatments (22.2%). For diagnostic applications, we found that most studies addressed the development of models to diagnose lesions in colposcopy images (e.g., [75,81,133]); we also grouped the studies that addressed diagnosis into six main targets: (i) states of CC (32.02%) [74,75,81,93], (ii) screening for CC (20.26%) [39,41,43], (iii) recurrence (cancer progression) (0.65%) [62], (iv) type prediction of Human Papilloma Virus (HPV) (0.65%) [54], (v) cancer progression (0.65%) [62], and (vi) segmentation of targets and organs at risk (OARs) (0.65%) [132].

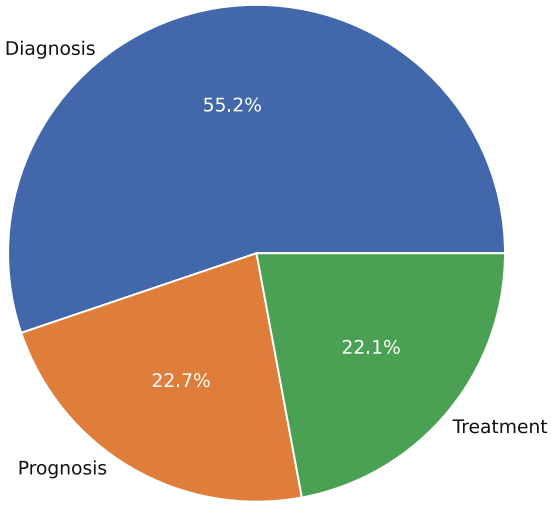


Figure 2. Distribution of the selected articles per clinical applications using ML and DL for CC.

For prognosis application, most studies focused on predicting cancer progression using CT and MRI (e.g., [77,78,156]); in particular, three main targets were identified in prognosis: (i) cancer progression (13.72%) [56,61,84], (ii) survival prediction (5.22%) [55,58,98], and (iii) recurrence (3.26%) [176] (e.g., [77,137,178]). For treatment applications, most of the studies focused on automatic segmentation of images to predict doses of treatments using CT (e.g., [68,131,141]). In this clinical application, four main targets were found: (i) therapeutic dose and planning (9.15%) [118,144,146], (ii) segmentation of targets/OARS (8.49%) [109,130,140], (iii) delineation of the clinical target volume (CTV) (3.26%) [115,141,153], and (iv) toxicity prediction in radiotherapy (0.65%) [101]. Figure 3 shows the distribution of targets for each clinical application based on ML and DL in CC.

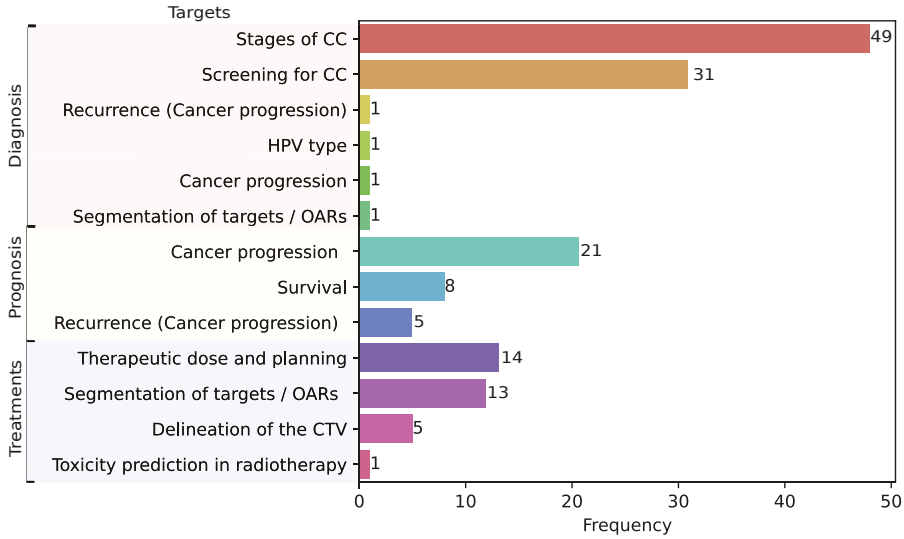


Figure 3. Distribution of targets grouped by clinical applications. The targets appear in order of highest to lowest frequency by clinical application.

3.4. Prediction Tasks in CC

Figure 4 displays the frequency of ML and DL objectives in the present selection study. Classification tasks were the most frequently employed, appearing in 108 studies (70.77%), followed by image classification in 34 articles (22.07%); 11 articles employed regression tasks (7.14%). For instance, one study [65] used DL classification to diagnose CC from magnetic resonance imaging (MRI) processing, the results of which were evaluated by experienced radiologists. After evaluation, the authors concluded that the performance of their algorithm was equivalent to that of experienced radiologists. Another study [64] trained DL models for automatic segmentation of OARS from computed tomography (CT) scans. The authors remarked that their algorithm is useful for automatic dose optimization for advanced radiation therapy strategies. Among the regression tasks, one study [98] developed an ML algorithm in the prediction of survival with CC from analyzing patient demographics, vital signs, laboratory test results, tumor characteristics, and treatment types. The authors concluded that their algorithm achieved a superior performance compared with a Cox proportional hazard regression model.

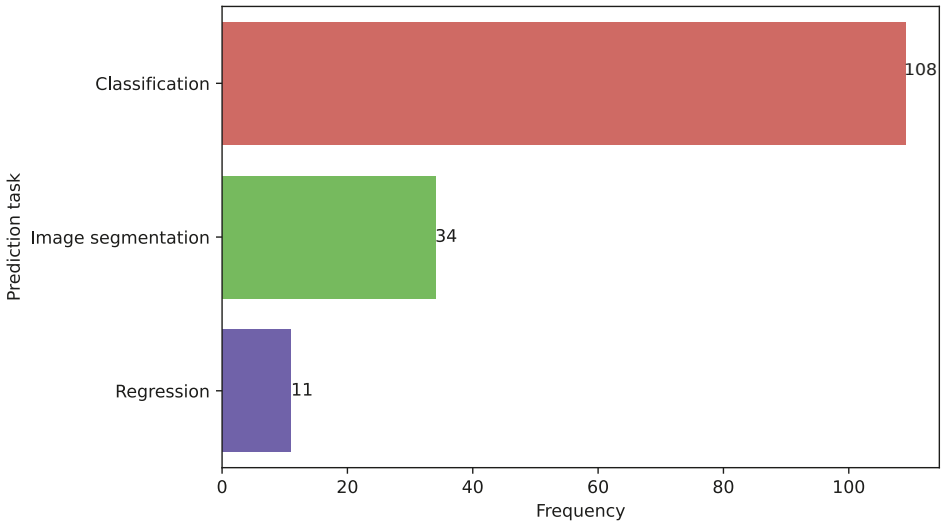


Figure 4. Distribution of the selected articles per prediction task in CC.

3.5. Models Trained in Classification, Segmentation, and Regression Tasks

Figure 5 presents the frequency of models trained using prediction task. Convolutional neural networks (CNNs) were the most frequently employed in the three prediction tasks analyzed. For classification tasks, CNN was the most used, with 40 articles [40,41,54], followed by support vector machine (SVM) [78,104,105], random forest (RF) [94,165,174], ResNet [52,53,72], and k-nearest neighbors (KNN) [81,138]. For instance, Liu et al. [41] described that an algorithm for cytopathology cell image classification based on CNN indicated that their algorithm was robust to changes in the aspect ratio of cells in cervical cytopathological images. Similarly, we found that CNN-based models were the most commonly used in image segmentation tasks, with 28 identified articles. For this task, Yoganathan et al. [32] presented an algorithm based on CNN for automatic segmentation of targets and OARs in MRI, achieving high performance for the automatic contouring, which could be useful in brachytherapy for CC. Finally, CNN-based models were the most common for regression tasks, with six articles. In this case, Yuan et al. [37] proposed a CNN algorithm for predicting doses in intensity-modulated radiation therapy (IMRT) plans. They concluded that there were no significant differences in dose parameters between automatic and real clinic plans.

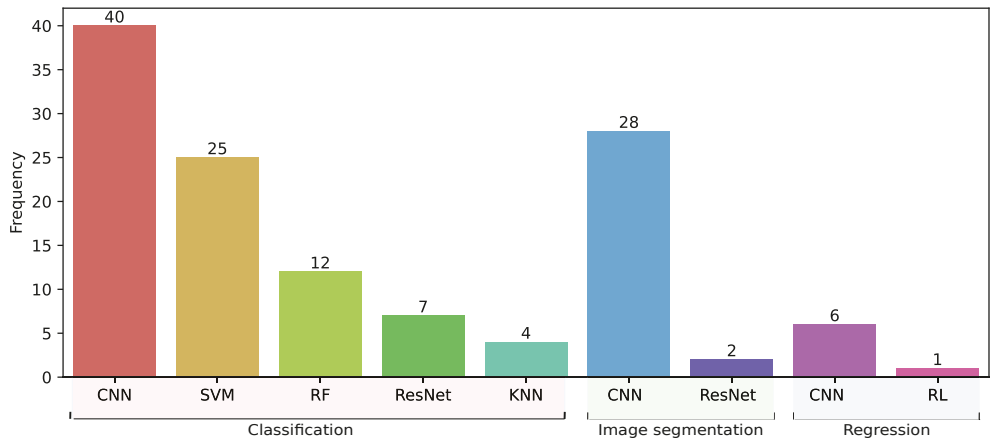


Figure 5. Distribution of the most common ML and DL models used per prediction task.

3.6. Temporal Analysis of ML and DL Implemented in CC

Figure 6 presents the distribution of models grouped by technique: ML, DL, and the fusion of ML and DL. Analysis of the 153 reviewed studies showed that the most common implemented technique was DL, with 61.2% (93 studies), whereas ML was implemented in 34.2% (53 studies), and the fusion of ML and DL was used in 4.6% (7 studies). In particular, fusion involves training models with different modalities separately (e.g., images, text, time series). After this training, the goal is to integrate the outputs of these models to improve performance and accuracy in the prediction task. For instance, Mathivanan et al. [160] proposed a hybrid methodology for feature extraction in cytology images based on DL, which was transferred to ML models for CC classification to obtain more accurate and timely interventions.

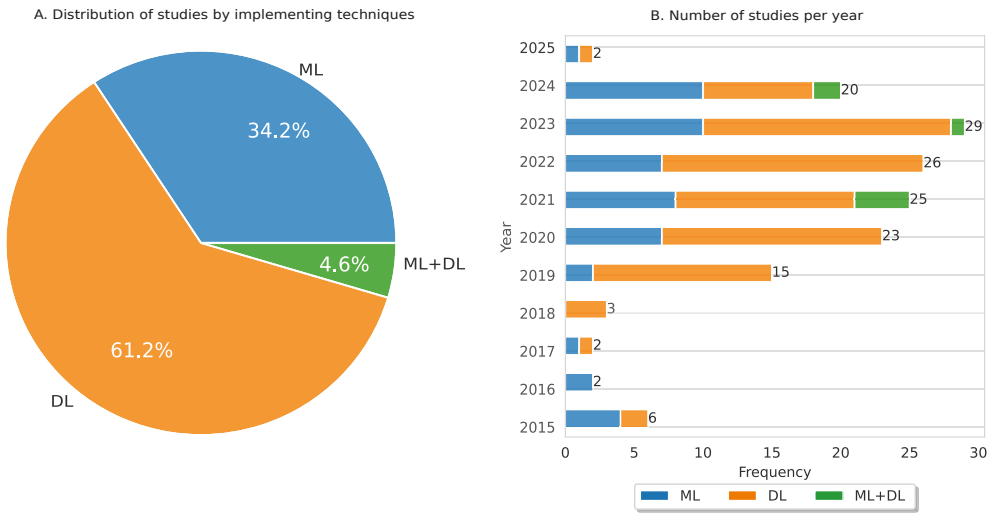


Figure 6. Distribution of studies grouped by ML, DL, and the fusion of ML and DL: (**A**) visualizes the distribution of studies according to implementation techniques; (**B**) lists the number of articles published per year.

Figure 6B describes the number of studies by year of publication. In particular, we observed that 2019 to 2024 had the highest number of publications (138 studies) compared to 2015 to 2018 (13 studies). We also identified that 2023 had the largest number of studies (with a total of 29) compared to the other years. In general, it is observed that the number of articles published using DL increases every year. Moreover, we noticed an increase in studies that merge DL and ML techniques in CC prediction.

3.7. Analysis of ML-Based Models Implemented in CC

A comparison table for CC prediction with ML techniques is given in Table 3. We found that in most studies, the accuracy was the metric most used to measure the performance of the models; we noted accuracies over 90% in most classification tasks [148,170]. In contrast, for image segmentation, the Dice similarity coefficient (DSC) was used to measure the similarity between two sets or samples. This metric is commonly used in image analysis, particularly for evaluating the performance of image segmentation models.

We also analyzed the validation strategies reported in the literature, namely, cross-validation (CV) with the following: (i) 3 folds, (ii) 5 folds, (iii) 10 folds, and (iv) leave-one-out cross-validation (LOOCV). From 53 studies that used ML models, 19 studies used 10 folds (35.84%), 12 studies used 5 folds (22.64%), and 2 studies used 3 folds and LOOCV, representing 3.77%. We also noted that 33.96% of the studies did not report any validation strategy. Finally, we noticed that of all the articles with ML (53), only 5 studies (9.43%) performed external validation (e.g., [113,166,172]).

Table 3. Summary of ML-based applications reported in the literature.

Ref.	Year	Clinical Application	Prediction Task	Target	Datasets	No. Folds for CV	Best Performance on Test Set		Extval
							Model	Metric	
[170]	2025	Diagnosis	Classification	Screening for CC	DNA	10	RF	ACC = 90.9%	-
[148]	2024	Diagnosis	Classification	Screening for CC	Clinical history	5	KNN	ACC = 99%	-
[149]	2024	Diagnosis	Classification	Screening for CC	Clinical history	5	SVM	ACC = 99%	-
[154]	2024	Prognosis	Classification	Survival	MRI images	-	RF	ACC = 86%	-
[164]	2024	Prognosis	Classification	Cancer progression	Clinical history	10	RF	ACC = 86%	-
[165]	2024	Treatment	Classification	Recurrence (Cancer progression)	DNA	10	RF	ACC = 84%	-
[167]	2024	Diagnosis	Classification	Screening for CC	Clinical history	-	RF	ACC = 90%	-
[172]	2024	Prognosis	Classification	Screening for CC	MRI images	10	SVM	AUC = 76%	Yes
[174]	2024	Treatment	Classification	Therapeutic dose and planning	DNA	10	RF	ACC = 96%	-
[179]	2024	Diagnosis	Classification	Stages of CC	Dose volume	5	SVM	ACC = 96%	-
[181]	2024	Prognosis	Classification	Stages of CC	DNA	-	Lightgbm	AUC = 98.7%	-
[113]	2023	Prognosis	Classification	Cancer progression	MRI images	5	SVM	ACC = 90%	Yes
[121]	2023	Diagnosis	Classification	Stages of CC	MRI images	5	SVM	ACC > 83%	-
[123]	2023	Diagnosis	Classification	Stages of CC	Spectral data	10	SVM	ACC = 95%	-
[124]	2023	Treatment	Classification	Therapeutic dose and planning	CT images	-	AAA	ACC = 73%	-
[125]	2023	Diagnosis	Classification	Stages of CC	Clinical history	5	LR DT	ACC > 88%	-
[128]	2023	Prognosis	Regression	Survival	Histopatology images Clinical history	5	XGBoost	AUC = 83%	-
[135]	2023	Diagnosis	Classification	Stages of CC	Histopathology images	-	SVM	ACC > 87%	-
[168]	2023	Prognosis	Classification	Cancer progression	MRI images Clinical history	5	MNB	ACC = 77%	-
[175]	2023	Treatment	Classification	Cancer progression	DNA	-	RF	-	-
[178]	2023	Prognosis	Classification	Cancer progression	CT images Clinical history	10	SVM	ACC = 90.1%	-
[38]	2022	Diagnosis	Classification	Stages of CC	Interferometry	3	NB	ACC = 92%	-
[39]	2022	Diagnosis	Classification	Screening for CC	Colposcopy images Cytology images HPV test	-	MLR	AUC = 92.1%	-
[138]	2022	Diagnosis	Classification	Stages of CC	Fluorescence images	-	KNN	SEN = 90%	-
[173]	2022	Prognosis	Classification	Cancer progression	Ultrasound	-	SVM	ACC = 85%	-
[177]	2022	Diagnosis	Classification	Stages of CC	DNA	-	SVM	ACC = 91.5%	-
[180]	2022	Prognosis	Classification	Recurrence (Cancer progression)	MRI images	5	LASSO	ACC = 93.1%	-
[38]	2022	Prognosis	Classification	Stages of CC	Dose volume	3	NB	ACC = 92%	-
[46]	2021	Diagnosis	Classification	Screening for CC	Biopsy results Cytology images	-	RT RF KNN	ACC = 95.5%	-
[47]	2021	Prognosis	Classification	Cancer progression	Histopathology images Clinical history	5	LR SVM	AUC = 88%	-
[55]	2021	Prognosis	Classification	Surv forest	MRI images CT images Clinical history	10	RF	AUC > 84%	-
[56]	2021	Prognosis	Classification	Cancer progression	DNA	5	Ridge	ACC = 84.7%	-
[58]	2021	Prognosis	Classification	Survival	DNA	10	SVM	AUC > 91%	-
[62]	2021	Diagnosis	Classification	Cancer progression	DNA	LOOCV	LR	ACC = 95%	-
[166]	2021	Diagnosis	Classification	Screening for CC	Ultrasound	-	LR	AUC = 91%	Yes
[176]	2021	Prognosis	Classification	Recurrence (Cancer progression)	DNA	-	SVM	ACC = 83%	Yes
[74]	2020	Diagnosis	Classification	Stages of CC	Cytology images	10	LSVM	ACC = 84%	-
[78]	2020	Prognosis	Classification	Cancer progression	MRI images	10	SVM	C-index 0.96	-
[80]	2020	Diagnosis	Classification	Screening for CC	Clinical history	10	RF	ACC > 95%	-

Table 3. *Cont.*

Ref.	Year	Clinical Application	Prediction Task	Target	Datasets	No. Folds for CV	Best Performance on Test Set		Extval
							Model	Metric	
[81]	2020	Diagnosis	Classification	Stages of CC	Colposcopy images	10	KNN	ACC = 80%	-
[156]	2020	Prognosis	Classification	Cancer progression	CT iimages	-	SVM	ACC = 76%	-
[171]	2020	Diagnosis	Classification	Screening for CC	DNA	10	DT	SEN = 88.6%	Yes
[182]	2020	Prognosis	Classification	Survival	MRI images	-	SF	AUC = 79.6%	-
[94]	2019	Diagnosis	Classification	Screening for CC	Clinical history	10	RF	ACC > 94%	-
[100]	2019	Prognosis	Classification	Cancer progression	MRI images Clinical history	-	SVM	AUROC = 75%	-
[104]	2017	Diagnosis	Classification	Stages of CC	Histopathology images	-	SVM	ACC > 90%	-
[105]	2016	Diagnosis	Classification	Stages of CC	Histopathology images	10	SVM	ACC = 88.5%	-
[106]	2016	Diagnosis	Classification	Screening for CC	Cytology images	10	SVM	ACC = 98%	-
[108]	2015	Prognosis	Classification	Cancer progression	Clinical history	10	SVM	ACC = 74%	-
[109]	2015	Treatment	Image segmentation	Segmentation of targets/OARs	CT images	LOOCV	SVM	DSC = 91.78	-
[111]	2015	Prognosis	Classification	Cancer progression	DNA	5	SVM	ACC = 80%	-
[112]	2015	Diagnosis	Classification	Stages of CC	Cytology images	10	SVM	Precision >87%	-

3.8. Analysis of DL-Based Models Implemented in CC

Table 4 presents a summary of the studies that used DL models for CC prediction. Related to performance, we observed a variety of metrics used, such as (i) accuracy, sensitivity, F1 score, precision, and area under the ROC curve (AUC) for classification tasks, (ii) DSC, Jaccard similarity, and accuracy for image segmentation, and (iii) mean absolute error (MAE), dose volume, and DSC for regression tasks. On the other hand, the most common validation strategy was 5 folds in 19 studies (20.43%), followed by 10 folds in 6 studies (6.45%). We also observed that 60 studies (64.51%) did not report any strategy of validation. Moreover, 13 studies (13.97%) did an external validation to evaluate the performance of their models (e.g., [116,118,153]).

Table 4. Summary of DL-based applications reported in the literature.

Ref.	Year	Clinical Application	Prediction Task	Target	Datasets	No. Folds for CV	Best Performance on Test Set		Externalval
							Model	Metric	
[147]	2024	Treatment	Classification	Therapeutic dose and planning	CT images Treatment plan Dose volume	-	U-Net	AUC = 94%	-
[150]	2024	Prognosis	Classification	Cancer progression	MRI images	5	CNN	ACC = 78%	-
[152]	2024	Prognosis	Regression	Cancer progression	CT images	5	DNN	ACC = 75%	-
[153]	2024	Treatment	Image Segmentation	Delineation of the CTV	CT images	-	ResCANet	DSC = 74.8	Yes
[157]	2024	Diagnosis	Classification	Stages of CC	DNA	-	EfficientNet DenseNet InceptionNet	ACC = 94.4%	-
[158]	2024	Diagnosis	Image Segmentation	Screening for CC	Colposcopy images	-	Efficient U-Net LSTM-Attention	ACC = 87%	-
[159]	2024	Diagnosis	Classification	Stages of CC	Cytology images	-	CNN	ACC = 99.11%	-
[183]	2024	Treatment	Classification	Therapeutic dose and planning	MRI images	10	ResNet101 MLP	AUC = 87%	-
[114]	2023	Diagnosis	Image Segmentation	Stages of CC	Interferometric Measurements	LOOCV	CNN	ACC = 81%	-
[115]	2023	Treatment	Image Segmentation	Delineation of the CTV	CT images	5	AFN	DSC > 88	-
[116]	2023	Diagnosis	Classification	Stages of CC	Cytology images	-	MLNet	ACC > 99%	Yes

Table 4. *Cont.*

Ref.	Year	Clinical Application	Prediction Task	Target	Datasets	No. Folds for CV	Best Performance on Test Set		Externalval
							Model	Metric	
[117]	2023	Diagnosis	Image Segmentation	Screening for CC	Cytology images	-	CNN	DSC = 0.94	-
[118]	2023	Treatment	Classification	Therapeutic dose and planning	Histopathology images	-	ViT RNN	ACC = 90%	Yes
[119]	2023	Diagnosis	Classification	Stages of CC	Spectral data	5	CNN	ACC = 94%	-
[120]	2023	Diagnosis	Classification	Stages of CC	Spectral data	LOOCV	DBN	ACC = 93%	-
[122]	2023	Diagnosis	Classification	Stages of CC	Cytology images	-	CNN	ACC = 87%	-
[126]	2023	Diagnosis	Classification	Stages of CC	Colposcopy images	-	Efficient Net GRU	ACC = 91%	-
[127]	2023	Treatment	Image Segmentation	Therapeutic dose and planning	CT images	-	CNN	Jaccard > 0.86	-
[129]	2023	Diagnosis	Classification	Screening for CC	Histopatology images	5	CNN	ACC = 76%	-
[130]	2023	Treatment	Image Segmentation	Segmentation of targets/OARs	CT images	-	CNN	DSC = 0.77	Yes
[131]	2023	Treatment	Regression	Therapeutic dose and planning	CT images	-	3DResUnet	Dose = 5%	-
[132]	2023	Diagnosis	Image Segmentation	Segmentation of targets/OARs	CT images	4	CNN	DSC = 0.80	-
[133]	2023	Diagnosis	Classification	Stages of CC	Colposcopy images	-	CNN	ACC > 88%	-
[145]	2023	Diagnosis	Classification	Stages of CC	Spectral data	LOOCV	DeepLabv3+ D-LinkNet	DSC = 0.95	-
[146]	2023	Treatment	Regression	Therapeutic dose and planning	CT images	-	U-Net	Score > 1%	-
[163]	2023	Diagnosis	Classification	Stages of CC	Clinical history	-	Stacking models	AUROC = 87%	-
[36]	2022	Diagnosis	Classification	Stages of CC	Colposcopy images	5	ResNet50	ACC = 81.3%	-
[37]	2022	Treatment	Regression	Therapeutic dose and planning	CT images	-	ResNet	DSC > 0.94	-
[32]	2022	Treatment	Image Segmentation	Segmentation of targets/OARs	MRI images	-	Inception ResNetv2	DSC = 0.72	-
[40]	2022	Treatment	Image Segmentation	Segmentation of targets/OARs	CT images	-	CNN	DSC > 0.70	-
[41]	2022	Diagnosis	Classification	Screening for CC	Cytology images	5	CNN	ACC = 95.4%	-
[42]	2022	Diagnosis	Classification	Screening for CC	Cytology images	-	YOLO ResNet	ACC = 90.5%	-
[43]	2022	Diagnosis	Classification	Screening for CC	Clinical history	5	Voting	ACC = 96.6%	-
[44]	2022	Diagnosis	Image Segmentation	Screening for CC	Colposcopy images	-	DeepLab V3+	ACC = 91.2%	-
[134]	2022	Prognosis	Regression	Survival	Histopathology images	-	CNN	AUC = 80%	-
[136]	2022	Prognosis	Regression	Cancer progression	MRI images PET images	5	CNN	AUC = 84%	-
[137]	2022	Prognosis	Classification	Cancer progression	Clinical history	-	MLP	AUC = 82%	-
[139]	2022	Diagnosis	Classification	Stages of CC	CT images	5	CNN	ACC > 60%	-
[140]	2022	Treatment	Image Segmentation	Segmentation of targets/OARs	CT images	-	CNN	DSC = 0.88	-
[141]	2022	Treatment	Image Segmentation	Delineation of the CTV	CT images	-	VN-Net	DSC = 0.81	-
[142]	2022	Diagnosis	Classification	Stages of CC	MRI images	5	CycleGAN	AUC = 89%	-
[143]	2022	Diagnosis	Classification	Stages of CC	Cytology images	-	CNN	SEN = 89%	-
[144]	2022	Treatment	Regression	Therapeutic dose and planning	Dose volume	-	U-Net	MAE = 2.4	Yes
[161]	2022	Treatment	Image Segmentation	Segmentation of targets/OARs	MRI images	-	CNN	Precision = 93%	-
[162]	2022	Diagnosis	Classification	Stages of CC	Colposcopy images	-	Dense-U-Net	ACC = 89%	Yes
[50]	2021	Diagnosis	Image Segmentation	Screening for CC	Cytology images	-	CNN	DSC = 0.92	-
[51]	2021	Diagnosis	Classification	Stages of CC	Cytology images	-	ResNet RNN CNN	SEN = 95.1%	Yes
[52]	2021	Diagnosis	Classification	Stages of CC	Cytology images	-	ResNet	ACC > 90%	Yes
[53]	2021	Diagnosis	Classification	Stages of CC	Colposcopy images	5	ResNet	AUC = 97%	-
[54]	2021	Diagnosis	Classification	HPV type	DNA	-	CNN	AUROC = 85%	Yes

Table 4. *Cont.*

Ref.	Year	Clinical Application	Prediction Task	Target	Datasets	No. Folds for CV	Best Performance on Test Set		Externalval
							Model	Metric	
[57]	2021	Treatment	Image Segmentation	Delineation of the CTV	CT images	3	U-Net CNN	DSC = 0.734	Yes
[59]	2021	Diagnosis	Image Segmentation	Screening for CC	Cytology images	10	YOLO v3	SEN = 92%	Yes
[60]	2021	Diagnosis	Classification	Screening for CC	Colposcopy images	-	CNN	ACC = 92%	-
[61]	2021	Prognosis	Classification	Cancer progression	MRI images	10	CNN	AUC = 91%	-
[63]	2021	Diagnosis	Classification	Screening for CC	Cytology images	5	CNN	ACC = 94%	-
[64]	2021	Treatment	Image Segmentation	Segmentation of targets/OARs	CT images	-	CNN	DSC = 0.85	Yes
[65]	2021	Diagnosis	Classification	Stages of CC	MRI images	-	XceptionNet	AUC = 93%	-
[66]	2021	Diagnosis	Classification	Screening for CC	Cytology images	5	CNN	AUC = 77%	-
[67]	2020	Treatment	Image Segmentation	Delineation of the CTV	CT images	-	CNN	DSC > 0.81	-
[68]	2020	Treatment	Image Segmentation	Therapeutic dose and planning	CT images	5	CNN	DSC > 0.82	-
[69]	2020	Treatment	Image Segmentation	Therapeutic dose and planning	CT images Dose volume	-	CNN	DVH = 0.73	-
[70]	2020	Diagnosis	Classification	Screening for CC	Colposcopy images	-	Faster R-CNN	AUC > 90%	-
[71]	2020	Diagnosis	Classification	Stages of CC	Cytology images	-	CNN	SEN = 100%	-
[72]	2020	Diagnosis	Classification	Stages of CC	Colposcopy images	10	Resnet	AUC = 78%	-
[73]	2020	Treatment	Image Segmentation	Segmentation of targets/OARs	CT images	-	CNN	DSC > 0.87	-
[75]	2020	Diagnosis	Image Segmentation	Stages of CC	Colposcopy images	-	CNN	ACC = 84%	-
[76]	2020	Diagnosis	Classification	Screening for CC	Colposcopy images	-	RetinaNet	AUC = 95%	-
[77]	2020	Prognosis	Image Segmentation	Cancer progression	MRI images	-	CNN	AUC = 93%	-
[79]	2020	Diagnosis	Classification	Screening for CC	Colposcopy images	-	CNN	ACC = 91%	-
[82]	2020	Diagnosis	Classification	Stages of CC	Cytology images	-	VGG-19	ACC = 95%	-
[83]	2020	Diagnosis	Classification	Stages of CC	MRI images Treatment plan	LOOCV	CNN	ACC = 94.3%	-
[85]	2020	Treatment	Image Segmentation	Therapeutic dose and planning	MRI images	5	UNet	SEN = 89%	-
[86]	2020	Treatment	Classification	Therapeutic dose and planning	Dose volume	-	CNN	$p < 0.017$	-
[87]	2020	Treatment	Image Segmentation	Segmentation of targets/OARs	CT images	-	UNet	DSC > 0.791	Yes
[88]	2019	Prognosis	Classification	Cancer progression	CT images	7	CNN	ACC = 89%	-
[89]	2019	Diagnosis	Classification	Stages of CC	Interferometric Measurements	-	CNN	SEN = 100%	-
[90]	2019	Diagnosis	Classification	Stages of CC	Spectral data	LOOCV	CNN	ACC = 100%	-
[22]	2019	Diagnosis	Classification	Stages of CC	Cytology images	-	CNN	ACC > 94%	-
[91]	2019	Diagnosis	Classification	Stages of CC	Cytology images	10	AGVFSM	ACC = 99%	-
[92]	2019	Diagnosis	Classification	Screening for CC	Colposcopy images	-	CNN	AUC = 91%	-
[93]	2019	Diagnosis	Classification	Stages of CC	CT images	-	CNN	ACC > 90%	-
[95]	2019	Treatment	Regression	Therapeutic dose and planning	Dose volume Histograms	-	Reinforcement learning	Score > 8.5%	Yes
[96]	2019	Diagnosis	Image Segmentation	Screening for CC	Cytology images	5	CNN	ACC = 91%	-
[97]	2019	Treatment	Image Segmentation	Segmentation of targets/OARs	CT images	5	CNN	DSC = 0.84	-
[98]	2019	Prognosis	Regression	Survival	Clinical history	5	FFNN	MAE = 29.3	-
[99]	2019	Diagnosis	Image Segmentation	Screening for CC	Cytology images	-	CNN	MAP = 0.936	-
[98]	2019	Prognosis	Regression	Survival	Clinical history	-	FFNN	MAE = 29.3	-

Table 4. *Cont.*

Ref.	Year	Clinical Application	Prediction Task	Target	Datasets	No. Folds for CV	Best Performance on Test Set		Externalval
							Model	Metric	
[102]	2018	Diagnosis	Classification	Stages of CC	Cytology images	-	CNN	Precision > 89%	Yes
[103]	2018	Diagnosis	Classification	Screening for CC	Colposcopy images	-	CNN	ACC = 100%	-
[155]	2018	Treatment	Image Segmentation	Segmentation of targets/OARs	CT images PET images	-	Graph cut	DSC = 0.83	-
[101]	2017	Treatment	Classification	Toxicity prediction in radiotherapy	CT images Treatment plan	10	VGG16	AUC = 89%	-
[107]	2015	Diagnosis	Classification	Screening for CC	Clinical history	-	NER	F1 = 67%	-
[110]	2015	Diagnosis	Image Segmentation	Stages of CC	Cytology images	-	GMM	DSC = 0.92	-

3.9. Analysis of Fusion of ML and DL Models Implemented in CC

Table 5 describes a comparison of studies that used ML and DL for CC. For classification tasks, we found the integration of ResNet (DL) with logistic regression (ML) [160], CNN (DL) with SVM (ML) [49], and recurrent neural networks (RNN—DL) with SVM [169], whereas ResNet with SVM [184] and U–Net with SVM [48] were identified by image segmentation. In all the studies in this group, the accuracy was used to measure the algorithm performance, and the validation strategy most used was 5 folds. We also did not find studies that evaluated their models with external data.

Table 5. Summary of ML and DL-based applications reported in the literature.

Ref.	Year	Clinical Application	Prediction Task	Target	Datasets	No. Folds for CV	Best Performance on Test Set		Externalval
							Model	Metric	
[160]	2024	Diagnosis	Classification	Stages of CC	Cytology images	-	ResNet152 LR	ACC = 98%	-
[184]	2024	Treatment	Image segmentation	Segmentation of targets/OARs	MRI images	-	SVM - RF ResNet50	ACC = 75%	-
[151]	2023	Diagnosis	Classification	Stages of CC	Cytology images	4	KNN ResNet50 ViT-S/16	ACC > 83%	-
[45]	2021	Diagnosis	Classification	Screening for CC	Clinical history	-	RF SNN	ACC = 93.6%	-
[48]	2021	Diagnosis	Image segmentation	Stages of CC	Histopathology images	5	U-Net SVM	ACC = 94.4%	-
[49]	2021	Diagnosis	Classification	Stages of CC	Histopatology images	10	CNN SVM	ACC = 97.4%	-
[169]	2021	Prognosis	Classification	Recurrence (Cancer progression)	DNA	10	SVM RNN	ACC = 92%	-

3.10. Evaluation Metrics for ML and DL in CC

Figure 7 presents the metrics used to evaluate the models of the selected studies. The accuracy metric was the most used, in 77 studies (50.32%), followed by area under the curve ROC (AUC) in 25 studies (16.33%), and DICE in 23 studies (15.03%). The DICE metric was used in the image segmentation task, while accuracy was used in classification tasks. Other less common metrics were precision (2.6%), mean absolute error—MAE (1.96%), and C–index with 0.65%. In particular, each study used a variety of metrics to measure their model performance in classification, image segmentation, and regression tasks using ML and DL.

3.11. Explainability

Figure 8 shows the several types of explainability tools employed in the reviewed studies. We found that the visualization of image segmentation was commonly used to explain the output of the models (21.2%), followed by Gini coefficients (6.0%), survival analysis (4.6%), Gradient-weighted Class Activation Mapping Grad-CAM (3.3%), and SHAP values with 2.6%. For instance, Ming et al. [139] showed the visualization of CC

detection between the prediction result and ground truth under different modal images. In contrast, Mehmood et al. [45] described the most important features associated with CC detection using Gini coefficients. Similarly, Matsuo et al. [98] listed the most important covariates based on the statistical significance associated with CC prediction, whereas Kurita et al. [122] presented a DL-based algorithm for classifying cervical cytological screening; using a Grad-CAM approach allowed visualization of the image components that contributed most to the classification. He et al. [164] utilized the SHAP values approach for the explainability analysis of their predictive model to rank variables in the prediction of cervical precancerous lesions. Interestingly, more than half (62.3%) do not use any explainability approach, which means that these works only presented probabilities without using any tool to open the "black-box" (e.g., [38,46]).

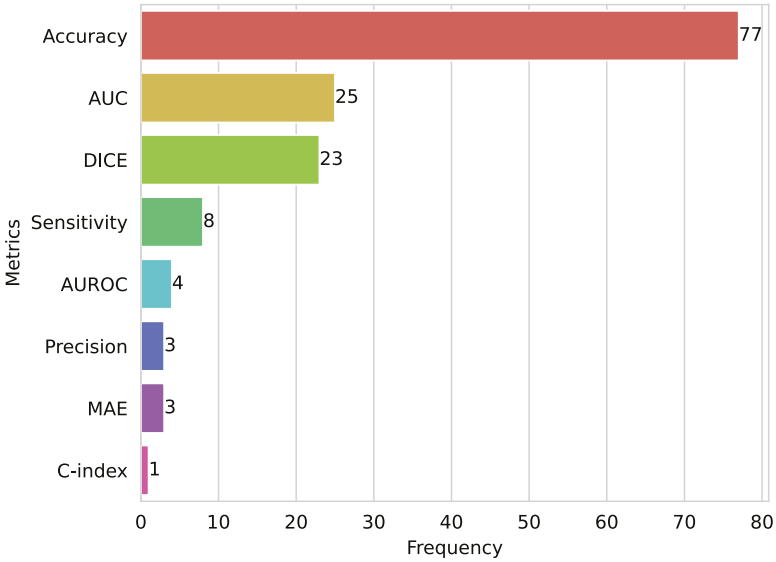


Figure 7. Distribution of the most common metrics used to measure model performance. AUC: area under the curve; DICE: Dice similarity coefficient; AUROC: area under the receiver operating characteristic curve; MAE: mean absolute error.

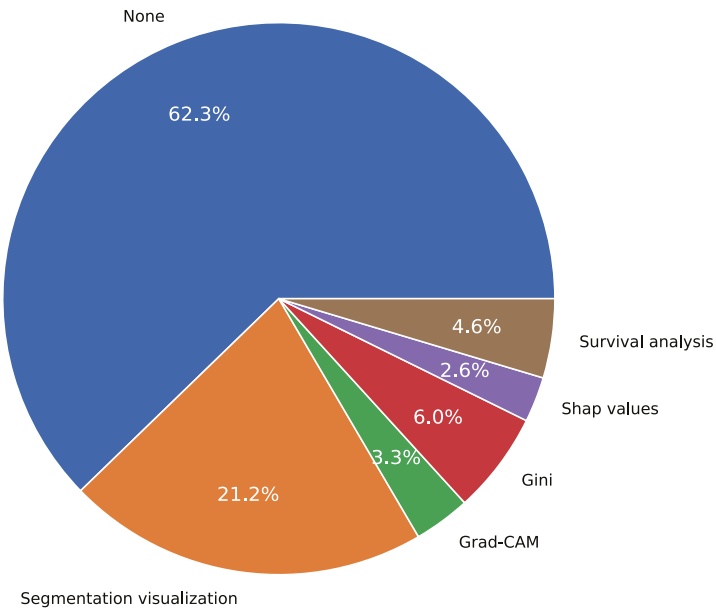


Figure 8. Distribution of explainability tools used in the literature.

3.12. Databases

Figure 9 highlights the databases used in the development of ML and DL models in CC. Cytology images were the most frequently employed, appearing in 33 studies [39,41], followed by CT images in 30 studies [153,156], MRI images [32,55] and clinical history [43,45] in 23 studies each, and colposcopy images [44,162] in 17 studies.

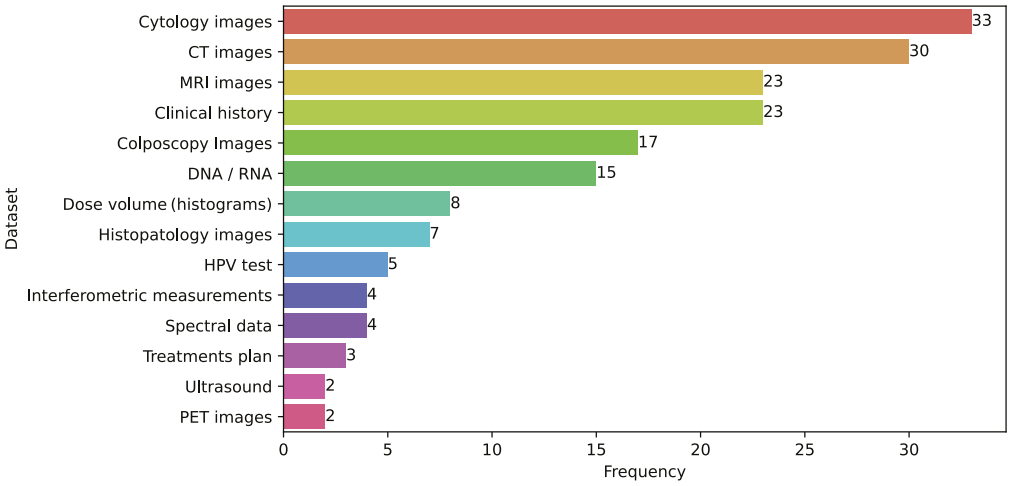


Figure 9. Distribution of the most common databases found in the literature.

In particular, the cytology dataset consisted of a set of images to examine abnormal cells that could indicate pre-cancerous changes or cancer [50,63,143]. The CT dataset included images to stratify advanced disease by evaluating lymph nodes [68,156]. For MRI, we found a set of images which were used to evaluate the tumor size, local invasion, and lymph node studies each, as well as colposcopy image involvement [172,182,183]. Related to clinical history, this group can include demographic details, lifestyle, historical medical records, risk factors, and family history of cancer [43,100,164]. Colposcopy images were used to predict and classify cervical lesions [53,70,162]. In the case of DNA/RNA datasets, they included genome sequences, genotyping of cytokines, and T-cell Receptor Sequencing [54,56,62]. Dose volume data refer to the clinical dose administrated to patients [95,124,144]. In histopathology, the dataset consisted of a set of whole-slide biopsy images used to predict lymph node metastasis [118,135].

The use of spectral and genomic data remains less explored in CC predictive modeling. However, spectral data has been primarily employed for diagnosis, while genomic data has been more frequently used for prognosis. Although research predominantly focused on diagnostic applications, images remained the most frequently used input data across all three medical applications. Specifically, 68.24% of diagnostic studies (58/85) used images, 45.71% of prognostic models (16/35) incorporated image-based data, and 70.58% of treatment models (24/34) utilized image inputs.

3.13. Limitations

Figure 10 summarizes the main limitations found in the reviewed literature. In particular, the limitation most commonly found was the limited number of patients/samples (48.6%) [38,182], followed by a lack of diversity in the dataset (11.4%) [144,156], lack of external validation (7.6%) [132,154], and data from a single data center (6.7%) [164,184]. The least common limitations were image noise [52,142] and uncertainty in contour delineation, each identified in 1% of all studies [59,87].

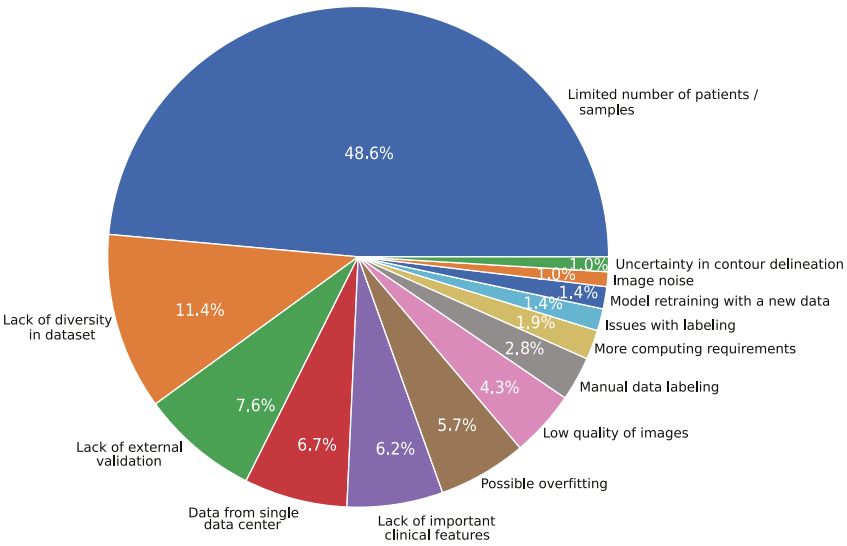


Figure 10. Most common limitations identified in the studies.

3.14. Reproducibility

Of the 153 studies reviewed, we found that only 6 studies (3.92%) have their programming codes available, compared to 145 studies (94.77%) whose codes are not available. For instance, Wentzensen et al. [66], Jeong et al. [150], and Cheng et al. [51] used a free platform to store and share the code of their DL-based models. In particular, these authors shared the steps to replicate their programming environment, the prerequisites list, requirements (hardware and software), implemented model code, and steps for training, evaluation, and inference. The code available in these studies promotes their reproducibility and can accelerate the research of other scientists.

We also analyzed the distribution of available data and found that only 20 studies (13%) mentioned that their data used during training and testing are freely available. In contrast, 35 studies (22.9%) stated that their data is available upon request, and 98 studies (64.1%) indicated their data is not available. In particular, the reason those studies mention that their data is available is because they used public databases, such as SIPaKMeD [41], Herlev [22,52,82], and Cytology Image Challenge [50], which are Pap Smear datasets. Figure 11 shows the distribution of studies with available code and data.

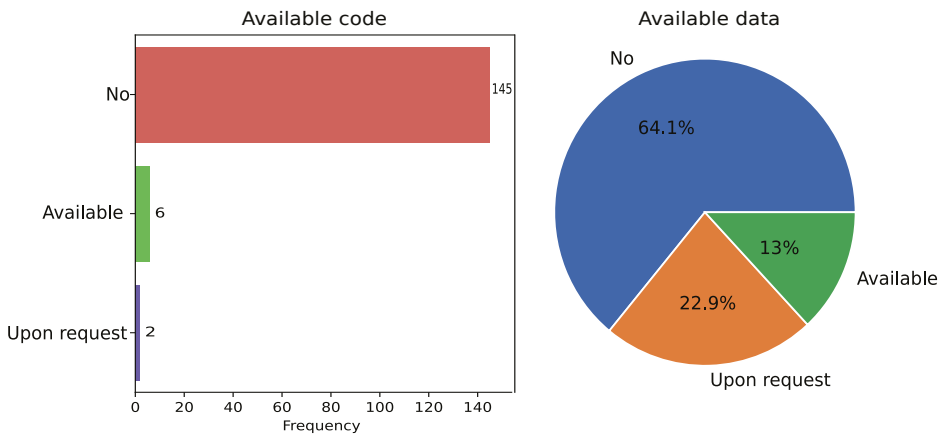


Figure 11. Distribution of available code and dataset.

3.15. Distribution of Publications by Country

In Figure 12, the distribution of publications by country, research origin, and medical purpose is presented. China had the highest number of studies, totaling 85 (40 focused

on diagnosis, 21 on prognosis, and 24 on treatment), followed by the United States with 30 (16 for diagnosis, 4 for prognosis, and 10 for treatment); India conducted 12 studies (10 on diagnosis and 2 on prognosis), while Republic of Korea carried out 9 studies (6 on diagnosis, 2 on prognosis, and 1 on treatment). We observed that countries such as Brazil, Germany, Portugal, Australia, Canada, Nigeria, Norway, Turkey, Saudi Arabia, Ethiopia, Rwanda, Pakistan, Bangladesh, Singapore, and New Zealand presented proposals that addressed only the diagnosis of CC. In contrast, proposals from Czech Republic, Finland, and Iraq addressed only the prognosis of CC. Finally, Sweden, Iran, and Qatar presented studies only for treatments.

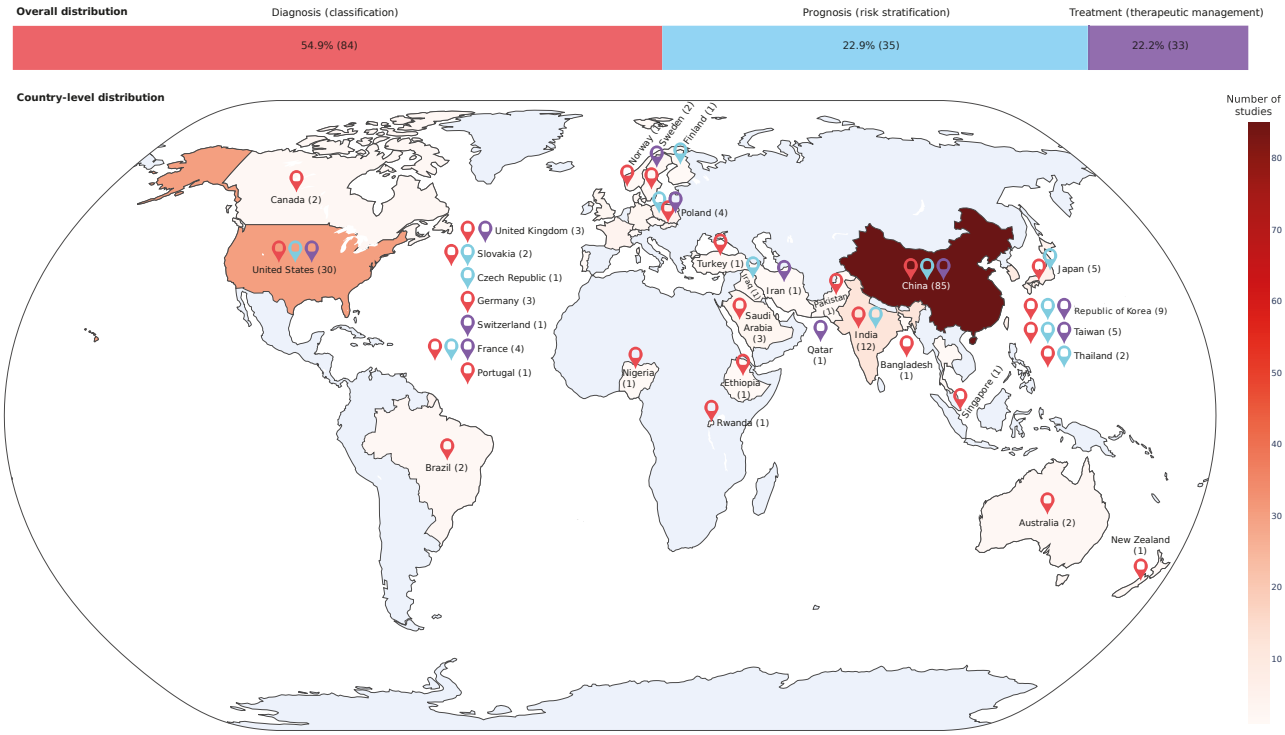


Figure 12. Applications of ML and DL in cervical cancer prediction across countries and targets.

Regarding nationality and predictive modeling objectives, only studies involving China, the United States, and France developed models for all three medical applications (diagnosis, prognosis, and treatment). Populations from Japan, Republic of Korea, India, and Thailand had models for diagnostic and prognostic purposes [36,42,56,178], while studies on the Taiwanese population focused on prognosis [88] and treatment [85]. Other populations were exclusively studied for the diagnosis of CC [82,99,148].

3.16. Study Design Characteristics

Regarding the characteristics of the reviewed studies, all of them were observational in nature. As summarized in Figure 13, the most common study design was cross-sectional (76.47%), followed by cohort studies (11.76%), retrospective cohort studies (11.11%), and a single case–control study (0.65%).

The majority of cross-sectional studies were used to develop diagnostic models for CC (66.67%), followed by models related to treatment (22.22%), and prognostic models (11.11%). For cohort and retrospective cohort studies, the most frequent medical application was prognostic (modeling predicting cancer progression to lymph node metastasis [77,100,136], recurrence [165,169,180], survival [55,58,98], or treatment response [113,175,184]), representing 50% and 88.24%, respectively. This is related to the longitudinal nature of medical

care and disease progression. These were followed by treatment-related models (dosage, therapeutic planning, toxicity), with 16.67% for cohort studies and 11.76% for retrospective cohort studies. Finally, 33% of cohort studies were focused on diagnosis, and the only case–control study was developed to predict diagnosis.

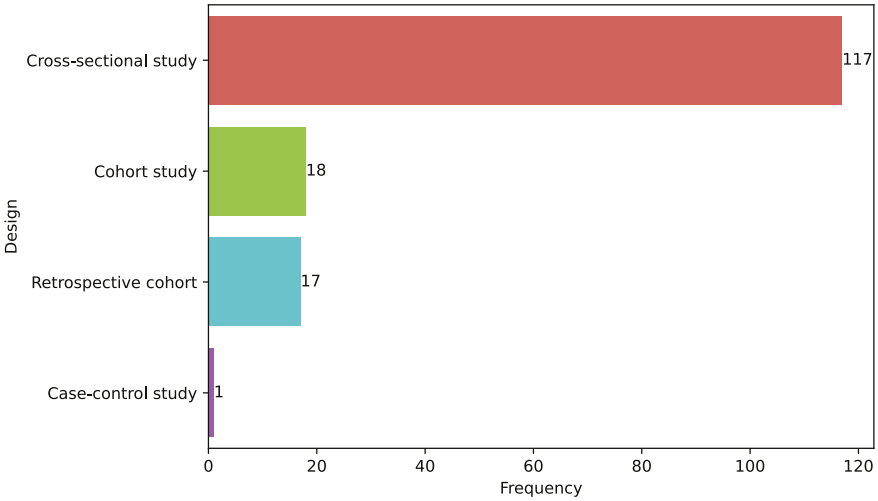


Figure 13. Most common study designs identified in the literature.

3.17. Study Population

A total of 14.38% of the reviewed studies did not report or provide information on the nationality and ethnicity of the study population, while 56.86% did not report the age of recruited patients [58,60,93,94]. We identified that the study populations in the reviewed works originated from 24 different nationalities. Notably, the Chinese population was the most studied, representing 43.14% of the analyses, followed by populations from the United States, Japan, Republic of Korea, and Denmark (Figure 14). The frequency with which some populations were studied is related to the availability of open access datasets. Some studies even used multiple open access databases containing data from populations of different nationalities (e.g., studies [52,82,151]). A relevant example is the Herlev database, which consists of cytological images from Danish patients collected by Herlev University Hospital, leading to limited data diversity across studies.

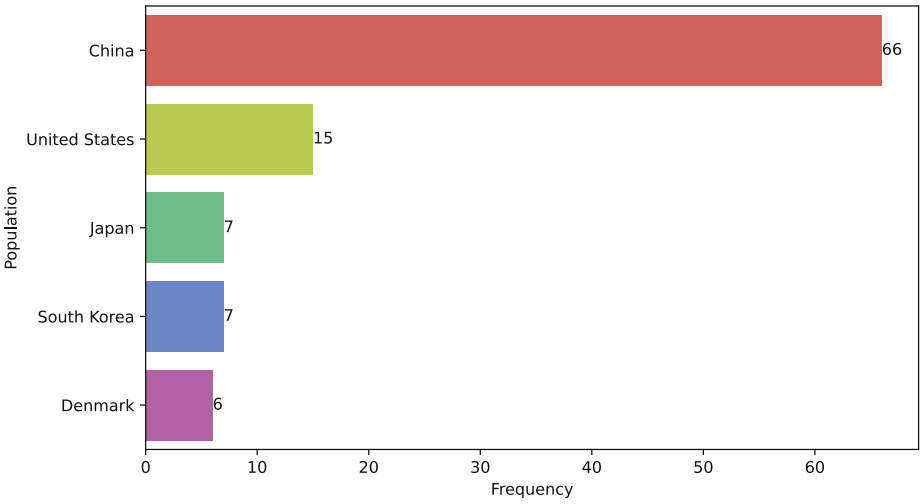


Figure 14. Most frequently studied populations.

3.18. Medical Specialist Involvement in Predictive Model Development

To assess the level of medical expert involvement, we explored their participation in study design, implementation, and evaluation. In 40.52% of the studies, there was no explicit reference to the participation of medical specialists. Some studies relied on secondary sources, classifying data based on medical records and notes (e.g., studies [48,91]). However, other studies using secondary data explicitly mentioned specialist involvement for validation or relabeling of patient-sample classifications. Examples include studies by Sornapudi et al. [82] and Shanthi [22], which reclassified cytological images with the assistance of a cytotechnologist and pathologist, respectively. Figure 15 highlights the most frequently referenced profiles across the reviewed studies. In particular, this figure provides a summary of the five most frequently involved medical specialties or specialists, categorizing their participation in different study phases, such as study design (patient recruitment and follow-up), label assignment (cancer staging, delineation, segmentation of key regions), and evaluation (prediction comparison). Some studies involved multiple specialists, such as two pathologists with different levels of experience, validating classification in the same study, or multiple medical profiles [65,83,86], such as an oncologist and radiologist, each performing different tasks within the same study [74,82,84].

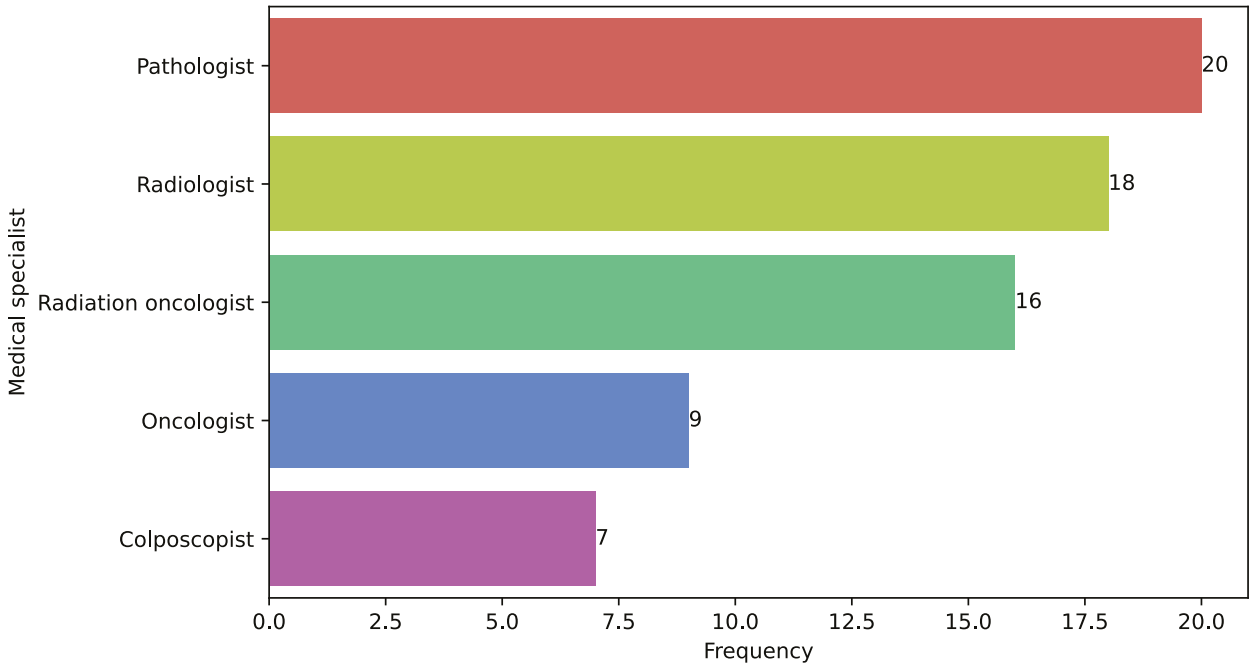


Figure 15. Most frequently involved medical specialists in studies.

4. Discussion

This scoping review identified that slightly more than half of the studies addressed the application of ML and DL in CC for diagnostic purposes, followed by prognosis and an incipient focus on CC treatment (Figure 16). Within the levels of cancer prevention, it is striking that AI in CC has been aimed at covering needs at the levels of secondary prevention (screening and timely diagnosis) and tertiary prevention (treatment); no study was found with an application approach at the level of primary prevention (HPV vaccines). Less than 20% corresponded to cohort or case–control studies, and another notable finding is that no studies were found that were carried out in Latin America or with a Latino population. To date, few reviews have analyzed the application of ML and DL in CC.

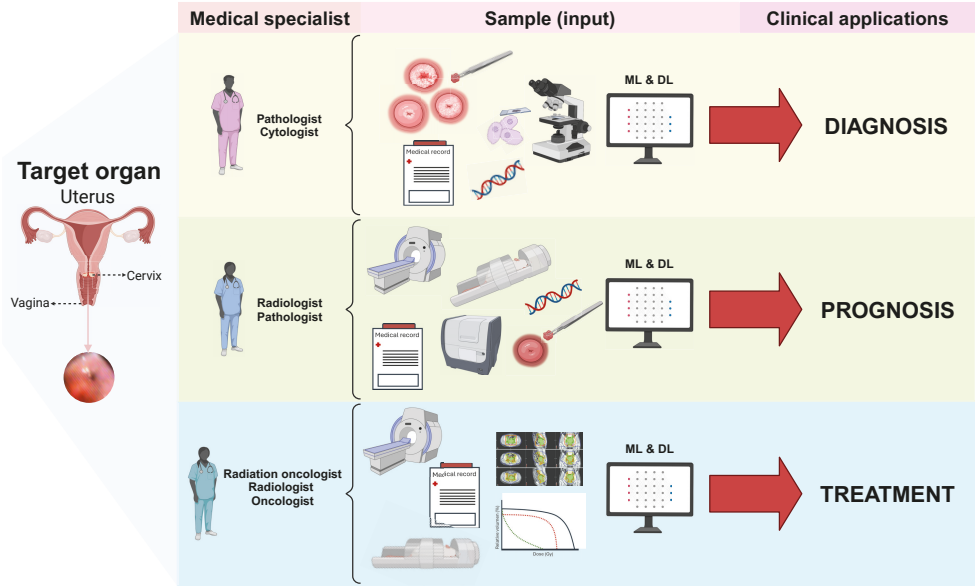


Figure 16. Overview of specialists, inputs, and outputs detected in the reviewed studies by clinical application. Created with BioRender.com.

William et al. presented a synthesis of 30 studies up to 2018 that focused on automated detection and classification of CC from pap smear images [185]. A high accuracy of over 90% was reported for existing models, especially for classification as normal or abnormal. However, low accuracy was reported for the classification of some cell classes that cannot be easily identified as normal or abnormal.

Since the manual analysis of pap smear images and microscopic biopsies by a trained cytologist is time-consuming and tedious, as cytotechnologists typically examine a large amount of data daily, the application of ML and DL for CC diagnosis has focused on optimizing the accuracy of CC detection and classification [185]. Similarly, DL models have been developed for dual staining of p16 and Ki-67 (markers that are closely related to cervical carcinogenesis) instead of cytological image identification, considering that the increase in the number of dual-stained cells correlates with increased severity of histopathology, with equal sensitivity and higher specificity compared to cytology and manual evaluation of double-stained slides and a reduction in referrals to colposcopy [66].

Another review published in 2022 focused on presenting the various ML models that had been used for CC prediction until November 2020. In this review, RF, DT, adaptive boosting, and gradient boosting models were reported to have 100% classification scores for CC prediction and 99% accuracy with SVM [186]. Another use of DL has been the segmentation of CC CT images. Radiation therapy is an effective way to improve the survival rate of patients with CC [67,187], especially for patients with locally advanced CC and those whose physical condition is not suitable for surgery. A systematic review and meta-analysis published in 2022 reported good accuracy in the automatic segmentation of CT images in CC with lower time consumption [188].

Manual segmentation of the CTV by a physician is still the standard, but it is a time-consuming and intensive task, taking an experienced physician at least 30 min [188]. Automatic segmentation has shown great potential in reducing physician burden, decreasing patient waiting time, and improving cancer treatment. However, for its use in future radiotherapy applications, high-quality public databases and large-scale research verification are required [188].

On the other hand, another review published in 2023 compiled the synthesis of 13 studies published until October 2022, which evaluated the use of ML to predict survival in patients with CC [189]. The range of the area under the curve reported for overall

survival was 0.40–0.99; for disease-free survival, 0.56–0.88; and for progression-free survival, 0.67–0.81. In addition, the combination of heterogeneous multidimensional data with ML techniques demonstrated potential in predicting CC survival. However, despite the benefits of ML, interpretability, explainability, and imbalanced datasets remain some of the biggest challenges. Therefore, for the use of ML models as a standard for survival prediction, further studies are required.

To reduce the risk of mortality from CC, faster diagnosis and accurate prognosis prediction are required. The incorporation of ML and DL models into the health-care system has shown great potential as a support tool in oncology clinical practice [7], and specifically in CC, these AI tools could help in the diagnosis and prediction of the prognosis of the disease as well as shortening the time consumed in the initiation of treatment. Specifically, in Mexico, AI models would be applied within colposcopy clinics where the colposcopic evaluation and biopsy are performed for the diagnosis of precursor lesions, which are the link to obtaining the result of the abnormal screening that will be used to refer patients to oncology clinics to receive treatment. The main challenges, limitations, and future directions identified in the reviewed literature are described below from two perspectives: computational (ML and DL) and clinical challenges.

4.1. Computational Challenges of Using ML and DL in CC

In recent years, ML and DL models have made significant progress in medicine and health care, primarily in the development of tools that can assist in clinical decision-making, such as diagnostics of diseases, targeted drug therapy, improvement of progress in clinical trials, and so on [190,191]. This progress has been brought about by three important milestones: (i) the advances in highly parallelizable graphics processing units (GPUs) that enable faster training and increased efficiency of DL models, (ii) development of programming frameworks (e.g., Keras, TensorFlow, PyTorch) that help abstract complex details in neural network implementation by providing libraries available to define network types (e.g., CNN, RNN) and common model architectures, and (iii) advances in DL architectures capable of analyzing multimodal data (e.g., images, text, signals, timeseries) [192–194]. Although these milestones mark significant progress, several challenges remain to be addressed to fully realize the potential of ML and DL in the diagnosis, prognosis, and treatment of CC. Table 6 summarizes these limitations, which are described below.

Table 6. Main computational challenges in cervical cancer using ML and DL models.

1. Data availability
2. Data leakage
3. Limited external validation
4. Limited evaluation of model performance
5. Complex data
6. Privacy issues
7. Lack of explainability
8. Lack of reproducibility

4.1.1. Data Availability

ML approaches, especially DL, typically require large datasets for training in order to learn patterns and achieve high performance [195,196]. A recent review describes DL techniques used in the field of pap smear whole slide imaging (WSI) classification analysis [31]. Most studies use supervised learning to extract features and train their models to identify normal, benign, and malignant cells from medical images. However, the main challenge is that it requires a large amount of labeled data that is not readily available.

Unfortunately, many applications in CC have few or limited data to train DL models. As observed in the literature review, the limited number of patients and samples was the main limitation in almost 50% of all studies reviewed. Some strategies to address this challenge include the following: (i) synthetic data generation based on DL models, such as variational autoencoders [197] and generative adversarial network [159], (ii) oversampling for the minority class based on Synthetic Minority Oversampling Technique (SMOTE) [94,167], and (iii) image transformations based on geometric, intensity, and spectral changes.

4.1.2. Data Leakage

Data leakage is a common problem, especially in applications based on ML and DL. This problem is caused by (i) inappropriate feature selection where the selected features are highly correlated with the target, (ii) errors during data preprocessing caused by incorrect data splitting, and (iii) filling missing data with the entire data set [198,199]. One of the main issues of data leakage is inflated performance metrics, such as high accuracy and precision [200]. Some strategies to avoid data leakage in ML and DL studies include [201]: applying data preprocessing to the training and test sets separately (e.g., scale, missing data), properly splitting the training and test sets, ensuring that data from the training set does not appear in the validation or test set, and using cross-validation.

Cross-validation is a technique for evaluating ML and DL models that is crucial to evaluate model performance and prevent overfitting. Overfitting occurs when a model overfits the training data, causing that model to perform extremely well on the data used for training but poorly on new, unknown data [202]. Splitting data into training, validation, and test sets helps to prevent overfitting, to tune hyperparameters, to select the best selection, and to improve the model generalization. Future research should adopt best practices for data splitting to avoid over-optimistic results, mainly in the clinical area.

4.1.3. Limited External Validation

External validation involves assessing whether the predicted model can be adopted in clinical practice [203]. According to Santos et al. [204], models should be evaluated on datasets separate and independent from the data on which the models were trained to increase confidence in their predictions and assess their clinical utility. In the literature review, we found that only 20 (13.07%) studies externally validated their models, meaning that these models may not be generalizable or may not perform well in populations other than those used in training. As noted, the lack of external validation is a major limitation of the reviewed studies; therefore, future research should include external validation to ensure the clinical relevance and generalization of the models.

4.1.4. Limited Evaluation of Model Performance

Performance metrics are important in evaluating the success or failure of developed models using ML and DL [205]. The metrics help to evaluate the predictive power of a developed model as well as to assess the reliability and rigor of a study [206]. In the reviewed studies, the most common metric used was accuracy, in 77 studies; however, in the clinical area, the confusion matrix is crucial for evaluating the performance of classification models, due to identifying the balance between false positives and false negatives [207]. It is important to highlight the importance of confusion matrices in classification tasks, which should be included in future research for both validation and test sets. If possible, confusion matrices for external validation sets should also be included. These matrices allow for detailed analysis of model performance, especially with respect to false positives and false negatives, which are crucial in the clinical setting.

The learning curve is another common metric for evaluating the performance of a machine learning algorithm. This curve assesses the model's convergence and diagnoses'

overfitting or underfitting during training [208]. Similar to the confusion matrix, the learning curves should be included in future studies to assess the convergence of the model, identify overfitting/underfitting, provide transparency, and allow readers to assess the generalization of the models.

4.1.5. Complex Data

One of the most common challenges in developing ML and DL models in CC is the complexity and high dimensionality of the data. As observed in Figure 9, the data can be represented as images, time series, spectral data, clinical history, laboratory tests, treatment plans, and so on. Each of these data contains valuable information; however, processing each type of data involves performing specific preprocessing tasks. For instance, the image processing can include resizing, normalization, noise reduction, and geometric transformations (for data augmentation) with the objective of improving the quality of the images [49,52,59,60]. In contrast, for clinical history, the processing tasks can involve feature selection, data standardization, missing data, and sometimes natural language processing in free-text notes [46,94,108]. Data in clinical settings are characterized by their high dimensionality and complexity; future research should adopt best practices to address this issue by applying specific preprocessing tasks to develop ML and DL-based solutions.

4.1.6. Privacy Issues

Clinic data contain highly sensitive and confidential information that identifies the patient and their medical condition or treatment [209,210]. However, several studies have pointed out that DL-based models can be subjected to reverse engineering techniques in order to extract sensitive information used during training [211,212]. The development of technologies that integrate cybersecurity tools is necessary to assess these risks and prevent them, ensuring patient privacy.

4.1.7. Lack of Explainability

In recent years, the performance of ML- and DL-based models has improved dramatically; however, they still often suffer from poor explainability, which limits their applicability in critical areas [213]. Explainability refers to the characteristic that ML and DL-based models must present to be interpreted by a human [213]. One of the main issues of DL-based models is that they are "black-box", which means they are non-intuitive and difficult for people to understand [214]. This challenge generates a barrier to the application of these models in clinical practice due to lack of interpretability, trust, and transparency [214,215]. In this review, we found that 62.3% of the studies do not use any explainability approach to clearly explain the model's decision-making (Figure 8), which could limit their use in clinical areas. Future work should consider the use of explainability approaches to increase the confidence and applicability of ML and DL-based models in real-world scenarios.

4.1.8. Lack of Reproducibility

According to Beam et al. [216], reproducibility is a prerequisite for the creation of new knowledge and scientific progress. A study is considered reproducible when an independent team, provided with the original data and code, is able to achieve the same findings as those reported in the initial research [216]. In the reviewed literature, we found that few studies share their code and data (Figure 11), which represents a challenge to reproducing these studies. Mislan et al. [217] mentioned that making code and data available allows analyses to be more easily reproduced and can accelerate the research of other scientists. It is important that future research promotes reproducibility and reuse of codes, which could improve the quality and accelerate the pace of CC research.

4.2. Clinical Implementation Challenges of Using ML and DL in CC

From the clinical perspective of the specialists who are the authors of this review, the application of ML and DL closest to CC is in diagnosis as a support tool for secondary prevention of the disease in the diagnosis of premalignant lesions. However, studies are required to demonstrate their reproducibility and internal and external validity for their long-term applicability in therapeutic management and for the monitoring and prediction of patients' prognoses. In summary, the use of ML and DL in oncology care is a turning point in advanced medical care, and their application in CC will have a deep impact on the lives of CC patients when their validity and reproducibility make them available for use in the clinical setting and data can be translated into therapy. This will especially be the case in more remote environments that are less favored in terms of care and face a shortage of personnel: cytologists, pathologists, and colposcopists. Table 7 describes the limitations for clinical implementation.

Table 7. Main challenges for clinical implementation.

1. Representativeness of clinical stages of CC in the training of deep learning-based models.
2. Privacy concerns and data security in health care.
3. Integration of AI with clinical workflows for real-time decision-making.
4. Ethical and regulatory considerations.
5. Public perspectives on using AI in their diagnoses and treatment decisions.

4.2.1. Representativeness of Clinical Stages of CC in the Training of DL-Based Models

Although molecular HPV detection testing is the recommended primary screening method for CC [218], and in some countries, cytology combined with screening for high-risk HPV subtypes at five-year intervals is recommended to improve the positive predictive value of screening, there are still excessive referrals for colposcopy due to the high prevalence of transient HPV infections [219]. In resource-limited settings, the use of alternative low-cost methodologies such as visual inspection with acetic acid (VIA) is indicated; however, it has low accuracy and reproducibility, so the development of automated visual assessment based on DL of cervical images could help improve the accuracy and reproducibility of VIA as an assistive technology. However, this technology requires some clinical and technical considerations for it to be widely used as a clinical test, including the training of deep learning-based models with representative images of each of the four distinct biological stages of CC: normal cervix, high-risk HPV infection, precancer, and invasive CC [220]. In this regard, Desai et al. described the essential clinical and technical considerations involved in building a DL-based automated visual evaluation (AVE) tool validated for broad use as a clinical test [220]. In particular, it emphasizes the need for a number of representative images with truth labels necessary to build an accurate but generalizable DL-based automated visual assessment algorithm for cervical images using a DL approach.

4.2.2. Privacy Concerns and Data Security in Health Care

As described above, ML and DL are AI applications that can help identify patterns and trends in health data to improve the diagnosis and treatment of CC. However, since a large amount of personal data is used to train AI models, there is a risk to patient privacy and security, as sensitive medical data are collected and stored, increasing the risk of cyberattacks and information breaches. Therefore, the guidelines for the responsible use of AI in public health developed by the World Health Organization (WHO) and FDA should be considered [221]. These guidelines include the need for a careful assessment of the risks

and benefits of AI in health care, transparency in the development and use of AI, and the need to ensure equity and nondiscrimination in health care.

4.2.3. Integration of AI with Clinical Workflows for Real-Time Decision-Making

According to the WHO, AI-assisted cervical cancer screening improves techniques involving the visual evaluation of digital images [222]. This is particularly relevant in low-resource areas with a shortage of specialized personnel. However, the challenge lies in access to computing resources and services for implementing ML and DL models on mobile devices or laptops in remote areas and in the willingness of health care professionals to adopt AI and make real-time clinical decisions when analyzing a specific case [223].

4.2.4. Ethical and Regulatory Considerations

Among the key ethical considerations in the clinical implementation of digital technologies for CC screening using ML and DL are safety monitoring, data privacy, traceability, accountability, and patient protection. These considerations are essential to uphold the principles of beneficence, nonmaleficence, autonomy, and justice [224]. Respecting patient autonomy supports the safeguarding of privacy, the maintenance of confidentiality, and the assurance of informed and valid consent to protect individuals' data. AI technologies should also be as explainable as possible and tailored to the level of understanding of their intended users. Moreover, these systems must undergo continuous, systematic, and transparent evaluation to ensure their effectiveness and appropriateness in the specific contexts in which they are deployed. From a regulatory perspective, ML and DL models often face challenges, such as limited generalizability beyond their training data and vulnerability to bias. Therefore, health care institutions must develop comprehensive strategies for AI implementation that address the management of technological infrastructure, cost considerations, and the integration of AI systems into existing clinical workflows [224].

4.2.5. Public Perspectives on Using AI in Diagnoses Decisions

Considering that the primary application of ML and DL in CC is in screening, and that current AI models achieve an estimated classification accuracy of 85% for cervical lesions, several critical needs must be addressed. These include ensuring algorithm accountability, improving model performance to increase specificity, and enhancing the accurate classification of premalignant cervical lesions to reduce the risk of misdiagnosis. Furthermore, robust validation of AI systems is essential to demonstrate their effectiveness at the population level [225].

4.3. Limitations

Our scoping review has some limitations. This scoping review is a qualitative review of the literature reported up to April 2025. Furthermore, this study focused only on ML and DL approaches. We conducted a search in four selected databases (IEEE, Web of Sciences, Scopus, PubMEd).

5. Conclusions

This scoping review gathered relevant scientific evidence on ML and DL applications in CC. AI in the prevention, diagnosis, and treatment process of CC would support timely medical care in hard-to-reach areas with less human capital and help doctors and the health system to reduce administrative burden, care time, professional burnout, and human error. It would also improve the correlation of histological study and the precision of diagnosis and therapeutic management. From 153 reviewed studies, we presented the main challenges and future works from two perspectives: computational and clinical challenges. From the computational view, we identified eight challenges: data availability,

data leakage, limited external validation, limited evaluation of model performance, complex data, privacy issues, lack of explainability, and lack of reproducibility. From the clinical view, we found five challenges: representative of clinical stages of CC, privacy concerns, integration of IA with clinical workflows, ethical considerations, and public perspective.

Although ML and DL's potential use in the diagnosis, prognosis, and treatment of CC has been reported with a significant future impact on oncology clinical practice that allows data to be translated into therapy, more evidence of validity and reproducibility is required for their use in the early detection, prognosis, and therapeutic management of CC. It is important to emphasize that the implementation of AI in public health must be careful and ethical, ensuring that patients' rights and privacy are respected.

Supplementary Materials: The following supporting information is available online at: https://www.mdpi.com/article/10.3390/diagnostics15121543/s1. Supplementary Table S1: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist.

Author Contributions: Conceptualization and methodology, K.T.-P. and B.V.; software, B.V.; validation, J.M.-A., C.E.A.-F., J.A.E.-G., S.S.-L. and L.d.C.C.-P.; formal analysis, investigation and data curation, M.R.-G., J.I.R.-E., B.V. and K.T.-P.; writing—original draft preparation, M.R.-G., B.V. and K.T.-P.; writing—review and editing, M.R.-G., B.V., K.T.-P., J.A.E.-G., M.B.-R., V.M.-M. and L.d.C.C.-P.; visualization, B.V.; supervision, project administration and funding acquisition, K.T.-P. and B.V. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Instituto Nacional de Salud Pública, Mexico, and grants from the Consejo Nacional de Humanidades Ciencias y Tecnologías (CONAHCYT, Mexico) to K.T.-P. (CONAHCYT-CBF 2023-2024 grant#3306) and V.M.-M. (grant CATEDRAS-2014-C01-#245520). This work was supported by the UNAM Postdoctoral Program (POSDOC) to B.V.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created.

Acknowledgments: This work was submitted in partial fulfillment of the requirements for the MPH degree of J.I.R.-E. from the Master Program in Public Health of the School of Public Health of Mexico.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ACC	Accuracy
AI	Artificial intelligence
AUC	Area Under (AUC) the Receiver Operating Characteristic (ROC) curve
CC	Cervical cancer
CLF	Classifier
CNN	Convolutional neural network
CT	Computed tomography
CTV	Clinical target volume
DCNN	Deep convolutional neural networks
DL	Deep learning
DRS	Diffuse reflectance spectroscopy
DT	Decision tree

FS	Feature selection
GPC	Gaussian process classifier
GMM	Gaussian mixture model
HPV	Human Papilloma Virus
KNN	K-Nearest neighbors
LDA	Linear discriminant analysis
LR	Logistic regression
Mask R-CNN	Mask regional CNN
MGI	Magnetic resonance imaging
ML	Machine Learning
MLP	Multi-layer perceptron
NB	Naive Bayes
OAR	Organ at risk
PCA	Principal component analysis
RF	Random forest
RNN	Recurrent neural network
SVM	Support vector machine
ViT	Vision transformer

References

1. International Agency for Research on Cancer. Cancer Today. Available online: https://gco.iarc.fr/today/en (accessed on 28 May 2024).
2. Momenimovahed, Z.; Mazidimoradi, A.; Maroofi, P.; Allahqoli, L.; Salehiniya, H.; Alkatout, I. Global, regional and national burden, incidence, and mortality of cervical cancer. *Cancer Rep.* **2022**, *6*, e1756. [CrossRef] [PubMed]
3. World Health Organization (WHO). *Global Strategy to Accelerate the Elimination of Cervical Cancer as a Public Health Problem*; World Health Organization (WHO): Geneva, Switzerland, 2020. Available online: https://www.who.int/publications-detail-redirect/9789240014107 (accessed on 28 May 2024).
4. Perkins, R.B.; Wentzensen, N.; Guido, R.S.; Schiffman, M. Cervical Cancer Screening: A Review. *JAMA* **2023**, *330*, 547–558. [CrossRef] [PubMed]
5. Jha, A.K.; Mithun, S.; Sherkhane, U.B.; Jaiswar, V.; Osong, B.; Purandare, N.; Kannan, S.; Prabhash, K.; Gupta, S.; Vanneste, B.; et al. Systematic review and meta-analysis of prediction models used in cervical cancer. *Artif. Intell. Med.* **2023**, *139*, 102549. [CrossRef]
6. Monk, B.J.; Enomoto, T.; Kast, W.M.; McCormack, M.; Tan, D.S.P.; Wu, X.; González-Martín, A. Integration of immunotherapy into treatment of cervical cancer: Recent data and ongoing trials. *Cancer Treat. Rev.* **2022**, *106*, 102385. [CrossRef]
7. Gaur, K.; Jagtap, M.M. Role of Artificial Intelligence and Machine Learning in Prediction, Diagnosis, and Prognosis of Cancer. *Cureus* **2022**, *14*, e31008. [CrossRef]
8. Lam, T.Y.T.; Cheung, M.F.K.; Munro, Y.L.; Lim, K.M.; Shung, D.; Sung, J.J.Y. Randomized Controlled Trials of Artificial Intelligence in Clinical Practice: Systematic Review. *J. Med. Internet Res.* **2022**, *24*, e37188. [CrossRef]
9. Arjmand, B.; Hamidpour, S.K.; Tayanloo-Beik, A.; Goodarzi, P.; Aghayan, H.R.; Adibi, H.; Larijani, B. Machine Learning: A New Prospect in Multi-Omics Data Analysis of Cancer. *Front. Genet.* **2022**, *13*, 824451. [CrossRef] [PubMed]
10. Gonzalez, R.; Nejat, P.; Saha, A.; Campbell, C.J.V.; Norgan, A.P.; Lokker, C. Performance of externally validated machine learning models based on histopathology images for the diagnosis, classification, prognosis, or treatment outcome prediction in female breast cancer: A systematic review. *J. Pathol. Inform.* **2024**, *15*, 100348. [CrossRef]
11. Li, G.X.; Chen, Y.P.; Hu, Y.Y.; Zhao, W.J.; Lu, Y.Y.; Wan, F.J.; Wu, Z.J.; Wang, X.Q.; Yu, Q.Y. Machine learning for identifying tumor stemness genes and developing prognostic model in gastric cancer. *Aging* **2024**, *16*, 6455–6477. [CrossRef]
12. Khalighi, S.; Reddy, K.; Midya, A.; Pandav, K.B.; Madabhushi, A.; Abedalthagafi, M. Artificial intelligence in neuro-oncology: Advances and challenges in brain tumor diagnosis, prognosis, and precision treatment. *NPJ Precis. Oncol.* **2024**, *8*, 80. [CrossRef]
13. Fiste, O.; Gkiozos, I.; Charpidou, A.; Syrigos, N.K. Artificial Intelligence-Based Treatment Decisions: A New Era for NSCLC. *Cancers* **2024**, *16*, 831. [CrossRef] [PubMed]
14. Huan, Q.; Cheng, S.; Ma, H.F.; Zhao, M.; Chen, Y.; Yuan, X. Machine learning-derived identification of prognostic signature for improving prognosis and drug response in patients with ovarian cancer. *J. Cell. Mol. Med.* **2024**, *28*, e18021. [CrossRef] [PubMed]
15. Ng, W.T.; But, B.; Choi, H.C.W.; de Bree, R.; Lee, A.W.M.; Lee, V.H.F.; López, F.; Mäkitie, A.A.; Rodrigo, J.P.; Saba, N.F.; et al. Application of Artificial Intelligence for Nasopharyngeal Carcinoma Management – A Systematic Review. *Cancer Manag. Res.* **2022**, *14*, 339–366. [CrossRef]
16. Mitchell, T. *Machine Learning*, 2nd ed.; McGraw Hill: Shelter Island, NY, USA, 1997.

17. Chollet, F. *Deep Learning with Python*, 2nd ed.; Manning Publications: Shelter Island, NY, USA, 2021;
18. Aurelien, G. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems | Guide Books | ACM Digital Library*, 2nd ed.; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2019.
19. Painuli, D.; Mishra, D.; Bhardwaj, S.; Aggarwal, M. Forecast and prediction of COVID-19 using machine learning. In *Data Science for COVID-19*; Academic Press: Cambridge, MA, USA, 2021. [CrossRef]
20. Murphy, K.P. *Probabilistic Machine Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2022.
21. Imagawa, K.; Shiomoto, K. Performance change with the number of training data: A case study on the binary classification of COVID-19 chest X-ray by using convolutional neural networks. *Comput. Biol. Med.* **2022**, *142*, 105251. [CrossRef]
22. Shanthi, P.B.; Faruqi, F.; Hareesha, K.S.; Kudva, R. Deep Convolution Neural Network for Malignancy Detection and Classification in Microscopic Uterine Cervix Cell Images. *Asian Pac. J. Cancer Prev.* **2019**, *20*, 3447–3456. [CrossRef]
23. Reijtenbagh, D.; Godart, J.; de Leeuw, A.; Seppenwoolde, Y.; Jürgenliemk-Schulz, I.; Mens, J.W.; Nout, R.; Hoogeman, M. Multi-center analysis of machine-learning predicted dose parameters in brachytherapy for cervical cancer. *Radiother. Oncol.* **2022**, *170*, 169–175. [CrossRef]
24. Gençtav, A.; Aksoy, S.; Önder, S. Unsupervised segmentation and classification of cervical cell images. *Pattern Recognit.* **2012**, *45*, 4151–4168. [CrossRef]
25. Zhang, H.; Chen, C.; Gao, R.; Yan, Z.; Zhu, Z.; Yang, B.; Chen, C.; Lv, X.; Li, H.; Huang, Z. Rapid identification of cervical adenocarcinoma and cervical squamous cell carcinoma tissue based on Raman spectroscopy combined with multiple machine learning algorithms. *Photodiagnosis Photodyn. Ther.* **2021**, *33*, 102104. [CrossRef] [PubMed]
26. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.U.; Polosukhin, I. Attention is All you Need. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 6000–6010.
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
29. Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; Dudley, J.T. Deep learning for healthcare: Review, opportunities and challenges. *Briefings Bioinform.* **2017**, *19*, 1236–1246. [CrossRef]
30. Adlung, L.; Cohen, Y.; Mor, U.; Elinav, E. Machine learning in clinical decision making. *Med* **2021**, *2*, 642–665. [CrossRef] [PubMed]
31. Sambyal, D.; Sarwar, A. Recent developments in cervical cancer diagnosis using deep learning on whole slide images: An Overview of models, techniques, challenges and future directions. *Micron* **2023**, *173*, 103520. [CrossRef]
32. Yoganathan, S.A.; Paul, S.N.; Paloor, S.; Torfeh, T.; Chandramouli, S.H.; Hammoud, R.; Al-Hammadi, N. Automatic segmentation of magnetic resonance images for high-dose-rate cervical cancer brachytherapy using deep learning. *Med. Phys.* **2022**, *49*, 1571–1584. [CrossRef]
33. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.J.; Horsley, T.; Weeks, L.; et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473. [CrossRef] [PubMed]
34. Vandenbroucke, J.; Von Elm, E.; Altman, D.; Gøtzsche, P.; Mulrow, C.; Pocock, S.; Poole, C.; Schlesselman, J.; Egger, M. Mejorar la comunicación de estudios observacionales en epidemiología (STROBE): Explicación y elaboración. *Gac. Sanit.* **2009**, *23*, 158e1–158e28. [CrossRef] [PubMed]
35. Dixon-Woods, M.; Agarwal, S.; Jones, D.; Young, B.; Sutton, A. Synthesising qualitative and quantitative evidence: A review of possible methods. *J. Health Serv. Res. Policy* **2005**, *10*, 45–53. [CrossRef]
36. Kim, Y.J.; Ju, W.; Nam, K.H.; Kim, S.N.; Kim, Y.J.; Kim, K.G. RGB Channel Superposition Algorithm with Acetowhite Mask Images in a Cervical Cancer Classification Deep Learning Model. *Sensors* **2022**, *22*, 3564. [CrossRef]
37. Yuan, Z.; Wang, Y.; Hu, P.; Zhang, D.; Yan, B.; Lu, H.M.; Zhang, H.; Yang, Y. Accelerate treatment planning process using deep learning generated fluence maps for cervical cancer radiation therapy. *Med. Phys.* **2022**, *49*, 2631–2641. [CrossRef]
38. Kruczkowski, M.; Drabik-Kruczkowska, A.; Marciniak, A.; Tarczewska, M.; Kosowska, M.; Szczerska, M. Predictions of cervical cancer identification by photonic method combined with machine learning. *Sci. Rep.* **2022**, *12*, 3762. [CrossRef]
39. Fu, L.; Xia, W.; Shi, W.; Cao, G.X.; Ruan, Y.T.; Zhao, X.Y.; Liu, M.; Niu, S.M.; Li, F.; Gao, X. Deep learning based cervical screening by the cross-modal integration of colposcopy, cytology, and HPV test. *Int. J. Med. Inform.* **2022**, *159*, 104675. [CrossRef]
40. Ma, C.Y.; Zhou, J.Y.; Xu, X.T.; Guo, J.; Han, M.F.; Gao, Y.Z.; Du, H.; Stahl, J.N.; Maltz, J.S. Deep learning-based auto-segmentation of clinical target volumes for radiotherapy treatment of cervical cancer. *J. Appl. Clin. Med. Phys.* **2022**, *23*, e13470. [CrossRef] [PubMed]

41. Liu, W.; Li, C.; Rahaman, M.M.; Jiang, T.; Sun, H.; Wu, X.; Hu, W.; Chen, H.; Sun, C.; Yao, Y.; et al. Is the aspect ratio of cells important in deep learning? A robust comparison of deep learning methods for multi-scale cytopathology cell image classification: From convolutional neural networks to visual transformers. *Comput. Biol. Med.* **2022**, *141*, 105026. [CrossRef]
42. Nambu, Y.; Mariya, T.; Shinkai, S.; Umemoto, M.; Asanuma, H.; Sato, I.; Hirohashi, Y.; Torigoe, T.; Fujino, Y.; Saito, T. A screening assistance system for cervical cytology of squamous cell atypia based on a two-step combined CNN algorithm with label smoothing. *Cancer Med.* **2022**, *11*, 520–529. [CrossRef] [PubMed]
43. Drokow, E.K.; Baffour, A.A.; Effah, C.Y.; Agboyibor, C.; Akpabla, G.S.; Sun, K. Building a predictive model to assist in the diagnosis of cervical cancer. *Future Oncol.* **2022**, *18*, 67–84. [CrossRef] [PubMed]
44. Liu, J.; Liang, T.; Peng, Y.; Peng, G.; Sun, L.; Li, L.; Dong, H. Segmentation of acetowhite region in uterine cervical image based on deep learning. *Technol. Health Care Off. J. Eur. Soc. Eng. Med.* **2022**, *30*, 469–482. [CrossRef]
45. Mehmood, M.; Rizwan, M.; Gregus Ml, M.; Abbas, S. Machine Learning Assisted Cervical Cancer Detection. *Front. Public Health* **2021**, *9*, 788376. [CrossRef]
46. Ali, M.M.; Ahmed, K.; Bui, F.M.; Paul, B.K.; Ibrahim, S.M.; Quinn, J.M.W.; Moni, M.A. Machine learning-based statistical analysis for early stage detection of cervical cancer. *Comput. Biol. Med.* **2021**, *139*, 104985. [CrossRef]
47. Chu, R.; Zhang, Y.; Qiao, X.; Xie, L.; Chen, W.; Zhao, Y.; Xu, Y.; Yuan, Z.; Liu, X.; Yin, A.; et al. Risk Stratification of Early-Stage Cervical Cancer with Intermediate-Risk Factors: Model Development and Validation Based on Machine Learning Algorithm. *Oncologist* **2021**, *26*, e2217–e2226. [CrossRef]
48. Dong, Y.; Wan, J.; Wang, X.; Xue, J.H.; Zou, J.; He, H.; Li, P.; Hou, A.; Ma, H. A Polarization-Imaging- Based Machine Learning Framework for Quantitative Pathological Diagnosis of Cervical Precancerous Lesions. *IEEE Trans. Med. Imaging* **2021**, *40*, 3728–3738. [CrossRef]
49. Fick, R.H.J.; Tayart, B.; Bertrand, C.; Lang, S.C.; Rey, T.; Ciompi, F.; Tilmant, C.; Farre, I.; Hadj, S.B. A Partial Label-Based Machine Learning Approach For Cervical Whole-Slide Image Classification: The Winning TissueNet Solution. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico, Mexico, 1–5 November 2021; Volume 2021, pp. 2127–2131. [CrossRef]
50. Chen, J.; Zhang, B. Segmentation of Overlapping Cervical Cells with Mask Region Convolutional Neural Network. *Comput. Math. Methods Med.* **2021**, *2021*, 3890988. [CrossRef]
51. Cheng, S.; Liu, S.; Yu, J.; Rao, G.; Xiao, Y.; Han, W.; Zhu, W.; Lv, X.; Li, N.; Cai, J.; et al. Robust whole slide image analysis for cervical cancer screening using deep learning. *Nat. Commun.* **2021**, *12*, 5639. [CrossRef] [PubMed]
52. Rahaman, M.M.; Li, C.; Yao, Y.; Kulwa, F.; Wu, X.; Li, X.; Wang, Q. DeepCervix: A deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques. *Comput. Biol. Med.* **2021**, *136*, 104649. [CrossRef] [PubMed]
53. Park, Y.R.; Kim, Y.J.; Ju, W.; Nam, K.; Kim, S.; Kim, K.G. Comparison of machine and deep learning for the classification of cervical cancer based on cervicography images. *Sci. Rep.* **2021**, *11*, 16143. [CrossRef]
54. Tian, R.; Zhou, P.; Li, M.; Tan, J.; Cui, Z.; Xu, W.; Wei, J.; Zhu, J.; Jin, Z.; Cao, C.; et al. DeepHPV: A deep learning model to predict human papillomavirus integration sites. *Briefings Bioinform.* **2021**, *22*, bbaa242. [CrossRef]
55. Da-ano, R.; Lucia, F.; Masson, I.; Abgral, R.; Alfieri, J.; Rousseau, C.; Mervoyer, A.; Reinhold, C.; Pradier, O.; Schick, U.; et al. A transfer learning approach to facilitate ComBat-based harmonization of multicentre radiomic features in new datasets. *PLoS ONE* **2021**, *16*, e0253653. [CrossRef] [PubMed]
56. Kaushik, M.; Chandra Joshi, R.; Kushwah, A.S.; Gupta, M.K.; Banerjee, M.; Burget, R.; Dutta, M.K. Cytokine gene variants and socio-demographic characteristics as predictors of cervical cancer: A machine learning approach. *Comput. Biol. Med.* **2021**, *134*, 104559. [CrossRef]
57. Shi, J.; Ding, X.; Liu, X.; Li, Y.; Liang, W.; Wu, J. Automatic clinical target volume delineation for cervical cancer in CT images using deep learning. *Med. Phys.* **2021**, *48*, 3968–3981. [CrossRef]
58. Ding, D.; Lang, T.; Zou, D.; Tan, J.; Chen, J.; Zhou, L.; Wang, D.; Li, R.; Li, Y.; Liu, J.; et al. Machine learning-based prediction of survival prognosis in cervical cancer. *BMC Bioinform.* **2021**, *22*, 331. [CrossRef]
59. Zhu, X.; Li, X.; Ong, K.; Zhang, W.; Li, W.; Li, L.; Young, D.; Su, Y.; Shang, B.; Peng, L.; et al. Hybrid AI-assistive diagnostic model permits rapid TBS classification of cervical liquid-based thin-layer cell smears. *Nat. Commun.* **2021**, *12*, 3541. [CrossRef]
60. Chandran, V.; Sumithra, M.G.; Karthick, A.; George, T.; Deivakani, M.; Elakkiya, B.; Subramaniam, U.; Manoharan, S. Diagnosis of Cervical Cancer based on Ensemble Deep Learning Network using Colposcopy Images. *BioMed Res. Int.* **2021**, *2021*, 5584004. [CrossRef]
61. Jiang, X.; Li, J.; Kan, Y.; Yu, T.; Chang, S.; Sha, X.; Zheng, H.; Luo, Y.; Wang, S. MRI Based Radiomics Approach With Deep Learning for Prediction of Vessel Invasion in Early-Stage Cervical Cancer. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *18*, 995–1002. [CrossRef] [PubMed]

62. Christley, S.; Ostmeyer, J.; Quirk, L.; Zhang, W.; Sirak, B.; Giuliano, A.R.; Zhang, S.; Monson, N.; Tiro, J.; Lucas, E.; et al. T Cell Receptor Repertoires Acquired via Routine Pap Testing May Help Refine Cervical Cancer and Precancer Risk Estimates. *Front. Immunol.* **2021**, *12*, 624230. [CrossRef] [PubMed]
63. Ke, J.; Shen, Y.; Lu, Y.; Deng, J.; Wright, J.D.; Zhang, Y.; Huang, Q.; Wang, D.; Jing, N.; Liang, X.; et al. Quantitative analysis of abnormalities in gynecologic cytopathology with deep learning. *Lab. Investig. J. Tech. Methods Pathol.* **2021**, *101*, 513–524. [CrossRef]
64. Rigaud, B.; Anderson, B.M.; Yu, Z.H.; Gobeli, M.; Cazoulat, G.; Söderberg, J.; Samuelsson, E.; Lidberg, D.; Ward, C.; Taku, N.; et al. Automatic Segmentation Using Deep Learning to Enable Online Dose Optimization During Adaptive Radiation Therapy of Cervical Cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **2021**, *109*, 1096–1110. [CrossRef]
65. Urushibara, A.; Saida, T.; Mori, K.; Ishiguro, T.; Sakai, M.; Masuoka, S.; Satoh, T.; Masumoto, T. Diagnosing uterine cervical cancer on a single T2-weighted image: Comparison between deep learning versus radiologists. *Eur. J. Radiol.* **2021**, *135*, 109471. [CrossRef]
66. Wentzensen, N.; Lahrmann, B.; Clarke, M.A.; Kinney, W.; Tokugawa, D.; Poitras, N.; Locke, A.; Bartels, L.; Krauthoff, A.; Walker, J.; et al. Accuracy and Efficiency of Deep-Learning-Based Automation of Dual Stain Cytology in Cervical Cancer Screening. *J. Natl. Cancer Inst.* **2021**, *113*, 72–79. [CrossRef] [PubMed]
67. Wang, Z.; Chang, Y.; Peng, Z.; Lv, Y.; Shi, W.; Wang, F.; Pei, X.; Xu, X.G. Evaluation of deep learning-based auto-segmentation algorithms for delineating clinical target volume and organs at risk involving data for 125 cervical cancer patients. *J. Appl. Clin. Med. Phys.* **2020**, *21*, 272–279. [CrossRef]
68. Liu, Z.; Liu, X.; Guan, H.; Zhen, H.; Sun, Y.; Chen, Q.; Chen, Y.; Wang, S.; Qiu, J. Development and validation of a deep learning algorithm for auto-delineation of clinical target volume and organs at risk in cervical cancer radiotherapy. *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.* **2020**, *153*, 172–179. [CrossRef]
69. Mao, X.; Pineau, J.; Keyes, R.; Enger, S.A. RapidBrachyDL: Rapid Radiation Dose Calculations in Brachytherapy Via Deep Learning. *Int. J. Radiat. Oncol. Biol. Phys.* **2020**, *108*, 802–812. [CrossRef]
70. Xue, Z.; Novetsky, A.P.; Einstein, M.H.; Marcus, J.Z.; Befano, B.; Guo, P.; Demarco, M.; Wentzensen, N.; Long, L.R.; Schiffman, M.; et al. A demonstration of automated visual evaluation of cervical images taken with a smartphone camera. *Int. J. Cancer* **2020**, *147*, 2416–2423. [CrossRef]
71. Bao, H.; Bi, H.; Zhang, X.; Zhao, Y.; Dong, Y.; Luo, X.; Zhou, D.; You, Z.; Wu, Y.; Liu, Z.; et al. Artificial intelligence-assisted cytology for detection of cervical intraepithelial neoplasia or invasive cancer: A multicenter, clinical-based, observational study. *Gynecol. Oncol.* **2020**, *159*, 171–178. [CrossRef] [PubMed]
72. Cho, B.J.; Choi, Y.J.; Lee, M.J.; Kim, J.H.; Son, G.H.; Park, S.H.; Kim, H.B.; Joo, Y.J.; Cho, H.Y.; Kyung, M.S.; et al. Classification of cervical neoplasms on colposcopic photography using deep learning. *Sci. Rep.* **2020**, *10*, 13652. [CrossRef]
73. Ju, Z.; Wu, Q.; Yang, W.; Gu, S.; Guo, W.; Wang, J.; Ge, R.; Quan, H.; Liu, J.; Qu, B. Automatic segmentation of pelvic organs-at-risk using a fusion network model based on limited training samples. *Acta Oncol.* **2020**, *59*, 933–939. [CrossRef] [PubMed]
74. Kanai, R.; Ohshima, K.; Ishii, K.; Sonohara, M.; Ishikawa, M.; Yamaguchi, M.; Ohtani, Y.; Kobayashi, Y.; Ota, H.; Kimura, F. Discriminant analysis and interpretation of nuclear chromatin distribution and coarseness using gray-level co-occurrence matrix features for lobular endocervical glandular hyperplasia. *Diagn. Cytopathol.* **2020**, *48*, 724–735. [CrossRef] [PubMed]
75. Yuan, C.; Yao, Y.; Cheng, B.; Cheng, Y.; Li, Y.; Li, Y.; Liu, X.; Cheng, X.; Xie, X.; Wu, J.; et al. The application of deep learning based diagnostic system to cervical squamous intraepithelial lesions recognition in colposcopy images. *Sci. Rep.* **2020**, *10*, 11639. [CrossRef]
76. Hu, L.; Horning, M.P.; Banik, D.; Ajenifuja, O.K.; Adepiti, C.A.; Yeates, K.; Mtema, Z.; Wilson, B.; Mehanian, C. Deep learning-based image evaluation for cervical precancer screening with a smartphone targeting low resource settings—Engineering approach. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; Volume 2020, pp. 1944–1949. [CrossRef]
77. Wu, Q.; Wang, S.; Zhang, S.; Wang, M.; Ding, Y.; Fang, J.; Wu, Q.; Qian, W.; Liu, Z.; Sun, K.; et al. Development of a Deep Learning Model to Identify Lymph Node Metastasis on Magnetic Resonance Imaging in Patients With Cervical Cancer. *JAMA Netw. Open* **2020**, *3*, e2011625. [CrossRef]
78. Wang, T.; Gao, T.; Guo, H.; Wang, Y.; Zhou, X.; Tian, J.; Huang, L.; Zhang, M. Preoperative prediction of parametrial invasion in early-stage cervical cancer with MRI-based radiomics nomogram. *Eur. Radiol.* **2020**, *30*, 3585–3593. [CrossRef]
79. Kudva, V.; Prasad, K.; Guruvare, S. Hybrid Transfer Learning for Classification of Uterine Cervix Images for Cervical Cancer Screening. *J. Digit. Imaging* **2020**, *33*, 619–631. [CrossRef]
80. Ijaz, M.F.; Attique, M.; Son, Y. Data-Driven Cervical Cancer Prediction Model with Outlier Detection and Over-Sampling Methods. *Sensors* **2020**, *20*, 2809. [CrossRef]
81. Bae, J.K.; Roh, H.J.; You, J.S.; Kim, K.; Ahn, Y.; Askaruly, S.; Park, K.; Yang, H.; Jang, G.J.; Moon, K.H.; et al. Quantitative Screening of Cervical Cancers for Low-Resource Settings: Pilot Study of Smartphone-Based Endoscopic Visual Inspection After Acetic Acid Using Machine Learning Techniques. *JMIR mHealth uHealth* **2020**, *8*, e16467.

82. Sornapudi, S.; Brown, G.T.; Xue, Z.; Long, R.; Allen, L.; Antani, S. Comparing Deep Learning Models for Multi-cell Classification in Liquid- based Cervical Cytology Image. *AMIA Annu. Symp. Proc.* **2020**, *2019*, 820–827. [PubMed]
83. Shao, J.; Zhang, Z.; Liu, H.; Song, Y.; Yan, Z.; Wang, X.; Hou, Z. DCE-MRI pharmacokinetic parameter maps for cervical carcinoma prediction. *Comput. Biol. Med.* **2020**, *118*, 103634. [CrossRef] [PubMed]
84. Takada, A.; Yokota, H.; Watanabe Nemoto, M.; Horikoshi, T.; Matsushima, J.; Uno, T. A multi-scanner study of MRI radiomics in uterine cervical cancer: prediction of in-field tumor control after definitive radiotherapy based on a machine learning method including peritumoral regions. *Jpn. J. Radiol.* **2020**, *38*, 265–273. [CrossRef] [PubMed]
85. Lin, Y.C.; Lin, C.H.; Lu, H.Y.; Chiang, H.J.; Wang, H.K.; Huang, Y.T.; Ng, S.H.; Hong, J.H.; Yen, T.C.; Lai, C.H.; et al. Deep learning for fully automated tumor segmentation and extraction of magnetic resonance radiomics features in cervical cancer. *Eur. Radiol.* **2020**, *30*, 1297–1305. [CrossRef]
86. Jihong, C.; Penggang, B.; Xiuchun, Z.; Kaiqiang, C.; Wenjuan, C.; Yitao, D.; Jiewei, Q.; Kerun, Q.; Jing, Z.; Tianming, W. Automated Intensity Modulated Radiation Therapy Treatment Planning for Cervical Cancer Based on Convolution Neural Network. *Technol. Cancer Res. Treat.* **2020**, *19*, 1533033820957002. [CrossRef]
87. Liu, Z.; Liu, X.; Xiao, B.; Wang, S.; Miao, Z.; Sun, Y.; Zhang, F. Segmentation of organs-at-risk in cervical cancer CT images with a convolutional neural network. *Phys. Medica* **2020**, *69*, 184–191. [CrossRef]
88. Shen, W.C.; Chen, S.W.; Wu, K.C.; Hsieh, T.C.; Liang, J.A.; Hung, Y.C.; Yeh, L.S.; Chang, W.C.; Lin, W.C.; Yen, K.Y.; et al. Prediction of local relapse and distant metastasis in patients with definitive chemoradiotherapy-treated cervical cancer by deep learning from [18F]-fluorodeoxyglucose positron emission tomography/computed tomography. *Eur. Radiol.* **2019**, *29*, 6741–6749. [CrossRef]
89. Pathania, D.; Landeros, C.; Rohrer, L.; D'Agostino, V.; Hong, S.; Degani, I.; Avila-Wallace, M.; Pivovarov, M.; Randall, T.; Weissleder, R.; et al. Point-of-care cervical cancer screening using deep learning-based microholography. *Theranostics* **2019**, *9*, 8438–8447. [CrossRef]
90. Aljakouch, K.; Hilal, Z.; Daho, I.; Schuler, M.; Krauß, S.D.; Yosef, H.K.; Dierks, J.; Mosig, A.; Gerwert, K.; El-Mashtoly, S.F. Fast and Noninvasive Diagnosis of Cervical Cancer by Coherent Anti-Stokes Raman Scattering. *Anal. Chem.* **2019**, *91*, 13900–13906. [CrossRef]
91. Dong, N.; Zhao, L.; Wu, A. Cervical cell recognition based on AGVF-Snake algorithm. *Int. J. Comput. Assist. Radiol. Surg.* **2019**, *14*, 2031–2041. [CrossRef]
92. Hu, L.; Bell, D.; Antani, S.; Xue, Z.; Yu, K.; Horning, M.P.; Gachuhi, N.; Wilson, B.; Jaiswal, M.S.; Befano, B.; et al. An Observational Study of Deep Learning and Automated Evaluation of Cervical Images for Cancer Screening. *J. Natl. Cancer Inst.* **2019**, *111*, 923–932. [CrossRef] [PubMed]
93. Dong, X.; Du, H.; Guan, H.; Zhang, X. Multiscale Time-Sharing Elastography Algorithms and Transfer Learning of Clinicopathological Features of Uterine Cervical Cancer for Medical Intelligent Computing System. *J. Med. Syst.* **2019**, *43*, 310. [CrossRef] [PubMed]
94. Geetha, R.; Sivasubramanian, S.; Kaliappan, M.; Vimal, S.; Annamalai, S. Cervical Cancer Identification with Synthetic Minority Oversampling Technique and PCA Analysis using Random Forest Classifier. *J. Med. Syst.* **2019**, *43*, 286. [CrossRef] [PubMed]
95. Shen, C.; Gonzalez, Y.; Klages, P.; Qin, N.; Jung, H.; Chen, L.; Nguyen, D.; Jiang, S.B.; Jia, X. Intelligent inverse treatment planning via deep reinforcement learning, a proof-of-principle study in high dose-rate brachytherapy for cervical cancer. *Phys. Med. Biol.* **2019**, *64*, 115013. [CrossRef]
96. Zhang, J.; Liu, Z.; Du, B.; He, J.; Li, G.; Chen, D. Binary tree-like network with two-path Fusion Attention Feature for cervical cell nucleus segmentation. *Comput. Biol. Med.* **2019**, *108*, 223–233. [CrossRef]
97. Chen, L.; Shen, C.; Zhou, Z.; Maquilan, G.; Albuquerque, K.; Folkert, M.R.; Wang, J. Automatic PET Cervical Tumor Segmentation by Combining Deep Learning and Anatomic Prior. *Phys. Med. Biol.* **2019**, *64*, 085019. [CrossRef]
98. Matsuo, K.; Purushotham, S.; Jiang, B.; Mandelbaum, R.S.; Takiuchi, T.; Liu, Y.; Roman, L.D. Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. *Am. J. Obstet. Gynecol.* **2019**, *220*, 381.e1–381.e14. [CrossRef]
99. Araújo, F.H.D.; Silva, R.R.V.; Ushizima, D.M.; Rezende, M.T.; Carneiro, C.M.; Campos Bianchi, A.G.; Medeiros, F.N.S. Deep learning for cell image segmentation and ranking. *Comput. Med. Imaging Graph. Off. J. Comput. Med. Imaging Soc.* **2019**, *72*, 13–21. [CrossRef]
100. Kan, Y.; Dong, D.; Zhang, Y.; Jiang, W.; Zhao, N.; Han, L.; Fang, M.; Zang, Y.; Hu, C.; Tian, J.; et al. Radiomic signature as a predictive factor for lymph node metastasis in early-stage cervical cancer. *J. Magn. Reson. Imaging JMRI* **2019**, *49*, 304–310. [CrossRef]
101. Zhen, X.; Chen, J.; Zhong, Z.; Hrycushko, B.; Zhou, L.; Jiang, S.; Albuquerque, K.; Gu, X. Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: A feasibility study. *Phys. Med. Biol.* **2017**, *62*, 8246–8263. [CrossRef]
102. Xie, Y.; Xing, F.; Shi, X.; Kong, X.; Su, H.; Yang, L. Efficient and robust cell detection: A structured regression approach. *Med. Image Anal.* **2018**, *44*, 245–254. [CrossRef] [PubMed]

103. Kudva, V.; Prasad, K.; Guruvare, S. Automation of Detection of Cervical Cancer Using Convolutional Neural Networks. *Crit. Rev. Biomed. Eng.* **2018**, *46*, 135–145. [CrossRef]
104. Wei, L.; Gan, Q.; Ji, T. Cervical cancer histology image identification method based on texture and lesion area features. *Comput. Assist. Surg.* **2017**, *22*, 186–199. [CrossRef]
105. Guo, P.; Banerjee, K.; Joe Stanley, R.; Long, R.; Antani, S.; Thoma, G.; Zuna, R.; Frazier, S.R.; Moss, R.H.; Stoecker, W.V. Nuclei-Based Features for Uterine Cervical Cancer Histology Image Analysis With Fusion-Based Classification. *IEEE J. Biomed. Health Inform.* **2016**, *20*, 1595–1607. [CrossRef] [PubMed]
106. Zhao, M.; Wu, A.; Song, J.; Sun, X.; Dong, N. Automatic screening of cervical cells using block image processing. *BioMed. Eng. Online* **2016**, *15*, 14. [CrossRef]
107. Weegar, R.; Kvist, M.; Sundström, K.; Brunak, S.; Dalianis, H. Finding Cervical Cancer Symptoms in Swedish Clinical Text using a Machine Learning Approach and NegEx. *AMIA Annu. Symp. Proc.* **2015**, *2015*, 1296–1305. [PubMed]
108. Kahng, J.; Kim, E.H.; Kim, H.G.; Lee, W. Development of a cervical cancer progress prediction tool for human papillomavirus-positive Koreans: A support vector machine-based approach. *J. Int. Med. Res.* **2015**, *43*, 518–525. [CrossRef]
109. Mu, W.; Chen, Z.; Liang, Y.; Shen, W.; Yang, F.; Dai, R.; Wu, N.; Tian, J. Staging of cervical cancer based on tumor heterogeneity characterized by texture features on 18F-FDG PET images. *Phys. Med. Biol.* **2015**, *60*, 5123. [CrossRef]
110. Zhi, L.; Carneiro, G.; Bradley, A.P. An improved joint optimization of multiple level set functions for the segmentation of overlapping cervical cells. *IEEE Trans. Image Process.* **2015**, *24*, 1261–1272. [CrossRef]
111. Dong, Y.; Kuang, Q.; Dai, X.; Li, R.; Wu, Y.; Leng, W.; Li, Y.; Li, M. Improving the Understanding of Pathogenesis of Human Papillomavirus 16 via Mapping Protein-Protein Interaction Network. *BioMed Res. Int.* **2015**, *2015*, 890381. [CrossRef]
112. Mariarputham, E.J.; Stephen, A. Nominated Texture Based Cervical Cancer Classification. *Comput. Math. Methods Med.* **2015**, *2015*, 586928. [CrossRef] [PubMed]
113. Xin, Z.; Yan, W.; Feng, Y.; Yunzhi, L.; Zhang, Y.; Wang, D.; Chen, W.; Peng, J.; Guo, C.; Chen, Z.; et al. An MRI-based machine learning radiomics can predict short-term response to neoadjuvant chemotherapy in patients with cervical squamous cell carcinoma: A multicenter study. *Cancer Med.* **2023**, *12*, 19383–19393. [CrossRef] [PubMed]
114. Robinson, D.; Hoong, K.; Kleijn, W.B.; Doronin, A.; Rehbinder, J.; Vizet, J.; Pierangelo, A.; Novikova, T. Polarimetric imaging for cervical pre-cancer screening aided by machine learning: Ex vivo studies. *J. Biomed. Opt.* **2023**, *28*, 102904. [CrossRef]
115. Tian, M.; Wang, H.; Liu, X.; Ye, Y.; Ouyang, G.; Shen, Y.; Li, Z.; Wang, X.; Wu, S. Delineation of clinical target volume and organs at risk in cervical cancer radiotherapy by deep learning networks. *Med. Phys.* **2023**, *50*, 6354–6365. [CrossRef]
116. Kaur, M.; Singh, D.; Kumar, V.; Lee, H.N. MLNet: Metaheuristics-Based Lightweight Deep Learning Network for Cervical Cancer Diagnosis. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 5004–5014. [CrossRef]
117. Ji, J.; Zhang, W.; Dong, Y.; Lin, R.; Geng, Y.; Hong, L. Automated cervical cell segmentation using deep ensemble learning. *BMC Med. Imaging* **2023**, *23*, 137. [CrossRef]
118. Liu, Q.; Jiang, N.; Hao, Y.; Hao, C.; Wang, W.; Bian, T.; Wang, X.; Li, H.; Zhang, Y.; Kang, Y.; et al. Identification of lymph node metastasis in pre-operation cervical cancer patients by weakly supervised deep learning from histopathological whole-slide biopsy images. *Cancer Med.* **2023**, *12*, 17952–17966. [CrossRef] [PubMed]
119. Kang, Z.; Liu, J.; Ma, C.; Chen, C.; Lv, X.; Chen, C. Early screening of cervical cancer based on tissue Raman spectroscopy combined with deep learning algorithms. *Photodiagnosis Photodyn. Ther.* **2023**, *42*, 103557. [CrossRef]
120. Wu, G.; Li, C.; Yin, L.; Wang, J.; Zheng, X. Compared between support vector machine (SVM) and deep belief network (DBN) for multi-classification of Raman spectroscopy for cervical diseases. *Photodiagnosis Photodyn. Ther.* **2023**, *42*, 103340. [CrossRef]
121. Ince, O.; Uysal, E.; Durak, G.; Onol, S.; Donmez Yilmaz, B.; Erturk, S.M.; Onder, H. Prediction of carcinogenic human papillomavirus types in cervical cancer from multiparametric magnetic resonance images with machine learning-based radiomics models. *Diagn. Interv. Radiol.* **2023**, *29*, 460–468. [CrossRef]
122. Kurita, Y.; Meguro, S.; Tsuyama, N.; Kosugi, I.; Enomoto, Y.; Kawasaki, H.; Uemura, T.; Kimura, M.; Iwashita, T. Accurate deep learning model using semi-supervised learning and Noisy Student for cervical cancer screening in low magnification images. *PLoS ONE* **2023**, *18*, e0285996. [CrossRef] [PubMed]
123. Zhu, J.; Yang, C.; Song, S.; Wang, R.; Gu, L.; Chen, Z. Classification of multiple cancer types by combination of plasma-based near-infrared spectroscopy analysis and machine learning modeling. *Anal. Biochem.* **2023**, *669*, 115120. [CrossRef] [PubMed]
124. Yang, B.; Liu, Y.; Chen, Z.; Wang, Z.; Zhou, Q.; Qiu, J. Tissues margin-based analytical anisotropic algorithm boosting method via deep learning attention mechanism with cervical cancer. *Int. J. Comput. Assist. Radiol. Surg.* **2023**, *18*, 953–959. [CrossRef]
125. Devi, S.; Gaikwad, S.R.; R, H. Prediction and Detection of Cervical Malignancy Using Machine Learning Models. *Asian Pac. J. Cancer Prev. APJCP* **2023**, *24*, 1419–1433. [CrossRef] [PubMed]
126. Chen, X.; Pu, X.; Chen, Z.; Li, L.; Zhao, K.N.; Liu, H.; Zhu, H. Application of EfficientNet-B0 and GRU-based deep learning on classifying the colposcopy diagnosis of precancerous cervical lesions. *Cancer Med.* **2023**, *12*, 8690–8699. [CrossRef]
127. Salehi, M.; Vafaei Sadr, A.; Mahdavi, S.R.; Arabi, H.; Shiri, I.; Reiazi, R. Deep Learning-based Non-rigid Image Registration for High-dose Rate Brachytherapy in Inter-fraction Cervical Cancer. *J. Digit. Imaging* **2023**, *36*, 574–587. [CrossRef]

128. Yu, W.; Lu, Y.; Shou, H.; Xu, H.; Shi, L.; Geng, X.; Song, T. A 5-year survival status prognosis of nonmetastatic cervical cancer patients through machine learning algorithms. *Cancer Med.* **2023**, *12*, 6867–6876. [CrossRef]
129. Liu, S.; Chu, R.; Xie, J.; Song, K.; Su, X. Differentiating single cervical cells by mitochondrial fluorescence imaging and deep learning-based label-free light scattering with multi-modal static cytometry. *Cytom. Part A J. Int. Soc. Anal. Cytol.* **2023**, *103*, 240–250. [CrossRef]
130. Wang, J.; Chen, Y.; Tu, Y.; Xie, H.; Chen, Y.; Luo, L.; Zhou, P.; Tang, Q. Evaluation of auto-segmentation for brachytherapy of postoperative cervical cancer using deep learning-based workflow. *Phys. Med. Biol.* **2023**, *68*, 055012. [CrossRef]
131. Yu, W.; Xiao, C.; Xu, J.; Jin, J.; Jin, X.; Shen, L. Direct Dose Prediction With Deep Learning for Postoperative Cervical Cancer Underwent Volumetric Modulated Arc Therapy. *Technol. Cancer Res. Treat.* **2023**, *22*, 15330338231167039. [CrossRef]
132. Huang, M.; Feng, C.; Sun, D.; Cui, M.; Zhao, D. Segmentation of Clinical Target Volume From CT Images for Cervical Cancer Using Deep Learning. *Technol. Cancer Res. Treat.* **2023**, *22*, 15330338221139164. [CrossRef] [PubMed]
133. Cao, Y.; Ma, H.; Fan, Y.; Liu, Y.; Zhang, H.; Cao, C.; Yu, H. A deep learning-based method for cervical transformation zone classification in colposcopy images. *Technol. Health Care Off. J. Eur. Soc. Eng. Med.* **2023**, *31*, 527–538. [CrossRef]
134. Chen, C.; Cao, Y.; Li, W.; Liu, Z.; Liu, P.; Tian, X.; Sun, C.; Wang, W.; Gao, H.; Kang, S.; et al. The pathological risk score: A new deep learning-based signature for predicting survival in cervical cancer. *Cancer Med.* **2022**, *12*, 1051–1063. [CrossRef]
135. He, S.; Xiao, B.; Wei, H.; Huang, S.; Chen, T. SVM classifier of cervical histopathology images based on texture and morphological features. *Technol. Health Care Off. J. Eur. Soc. Eng. Med.* **2023**, *31*, 69–80. [CrossRef]
136. Qian, W.; Li, Z.; Chen, W.; Yin, H.; Zhang, J.; Xu, J.; Hu, C. RESOLVE-DWI-based deep learning nomogram for prediction of normal-sized lymph node metastasis in cervical cancer: A preliminary study. *BMC Med. Imaging* **2022**, *22*, 221. [CrossRef] [PubMed]
137. Dong, T.; Wang, L.; Li, R.; Liu, Q.; Xu, Y.; Wei, Y.; Jiao, X.; Li, X.; Zhang, Y.; Zhang, Y.; et al. Development of a Novel Deep Learning-Based Prediction Model for the Prognosis of Operable Cervical Cancer. *Comput. Math. Methods Med.* **2022**, *2022*, 4364663. [CrossRef]
138. Ji, M.; Zhong, J.; Xue, R.; Su, W.; Kong, Y.; Fei, Y.; Ma, J.; Wang, Y.; Mi, L. Early Detection of Cervical Cancer by Fluorescence Lifetime Imaging Microscopy Combined with Unsupervised Machine Learning. *Int. J. Mol. Sci.* **2022**, *23*, 11476. [CrossRef]
139. Ming, Y.; Dong, X.; Zhao, J.; Chen, Z.; Wang, H.; Wu, N. Deep learning-based multimodal image analysis for cervical cancer detection. *Methods* **2022**, *205*, 46–52. [CrossRef] [PubMed]
140. Wang, J.; Chen, Y.; Xie, H.; Luo, L.; Tang, Q. Evaluation of auto-segmentation for EBRT planning structures using deep learning-based workflow on cervical cancer. *Sci. Rep.* **2022**, *12*, 13650. [CrossRef]
141. Ma, C.Y.; Zhou, J.Y.; Xu, X.T.; Qin, S.B.; Han, M.F.; Cao, X.H.; Gao, Y.Z.; Xu, L.; Zhou, J.J.; Zhang, W.; et al. Clinical evaluation of deep learning-based clinical target volume three-channel auto-segmentation algorithm for adaptive radiotherapy in cervical cancer. *BMC Med. Imaging* **2022**, *22*, 123. [CrossRef]
142. Chang, X.; Cai, X.; Dan, Y.; Song, Y.; Lu, Q.; Yang, G.; Nie, S. Self-supervised learning for multi-center magnetic resonance imaging harmonization without traveling phantoms. *Phys. Med. Biol.* **2022**, *67*, 145004. [CrossRef] [PubMed]
143. Tao, X.; Chu, X.; Guo, B.; Pan, Q.; Ji, S.; Lou, W.; Lv, C.; Xie, G.; Hua, K. Scrutinizing high-risk patients from ASC-US cytology via a deep learning model. *Cancer Cytopathol.* **2022**, *130*, 407–414. [CrossRef] [PubMed]
144. Qilin, Z.; Peng, B.; Ang, Q.; Weijuan, J.; Ping, J.; Hongqing, Z.; Bin, D.; Ruijie, Y. The feasibility study on the generalization of deep learning dose prediction model for volumetric modulated arc therapy of cervical cancer. *J. Appl. Clin. Med. Phys.* **2022**, *23*, e13583. [CrossRef]
145. Xue, Y.; Zheng, X.; Wu, G.; Wang, J. Rapid diagnosis of cervical cancer based on serum FTIR spectroscopy and support vector machines. *Lasers Med. Sci.* **2023**, *38*, 276. [CrossRef]
146. Gay, S.S.; Kisling, K.D.; Anderson, B.M.; Zhang, L.; Rhee, D.J.; Nguyen, C.; Netherton, T.; Yang, J.; Brock, K.; Jhingran, A.; et al. Identifying the optimal deep learning architecture and parameters for automatic beam aperture definition in 3D radiotherapy. *J. Appl. Clin. Med. Phys.* **2023**, *24*, e14131. [CrossRef]
147. Shen, Y.; Tang, X.; Lin, S.; Jin, X.; Ding, J.; Shao, M. Automatic dose prediction using deep learning and plan optimization with finite-element control for intensity modulated radiation therapy. *Med. Phys.* **2024**, *51*, 545–555. [CrossRef]
148. Aljrees, T. Improving prediction of cervical cancer using KNN imputer and multi-model ensemble learning. *PLoS ONE* **2024**, *19*, e0295632. [CrossRef] [PubMed]
149. Munshi, R.M. Novel ensemble learning approach with SVM-imputed ADASYN features for enhanced cervical cancer prediction. *PLoS ONE* **2024**, *19*, e0296107. [CrossRef]
150. Jeong, S.; Yu, H.; Park, S.H.; Woo, D.; Lee, S.J.; Chong, G.O.; Han, H.S.; Kim, J.C. Comparing deep learning and handcrafted radiomics to predict chemoradiotherapy response for locally advanced cervical cancer using pretreatment MRI. *Sci. Rep.* **2024**, *14*, 1180. [CrossRef]

151. Stegmüller, T.; Abbet, C.; Bozorgtabar, B.; Clarke, H.; Petignat, P.; Vassilakos, P.; Thiran, J.P. Self-supervised learning-based cervical cytology for the triage of HPV-positive women in resource-limited settings and low-data regime. *arXiv* **2023**, arXiv:2302.05195. [CrossRef]
152. Wu, K.C.; Chen, S.W.; Hsieh, T.C.; Yen, K.Y.; Chang, C.J.; Kuo, Y.C.; Chang, R.F.; Chia-Hung, K. Early prediction of distant metastasis in patients with uterine cervical cancer treated with definitive chemoradiotherapy by deep learning using pretreatment [18 F]fluorodeoxyglucose positron emission tomography/computed tomography. *Nucl. Med. Commun.* **2024**, *45*, 196–202. [CrossRef] [PubMed]
153. Wu, Z.; Wang, D.; Xu, C.; Peng, S.; Deng, L.; Liu, M.; Wu, Y. Clinical target volume (CTV) automatic delineation using deep learning network for cervical cancer radiotherapy: A study with external validation. *J. Appl. Clin. Med. Phys.* **2025**, *26*, e14553. [CrossRef] [PubMed]
154. Xin, W.; Rixin, S.; Linrui, L.; Zhihui, Q.; Long, L.; Yu, Z. Machine learning-based radiomics for predicting outcomes in cervical cancer patients undergoing concurrent chemoradiotherapy. *Comput. Biol. Med.* **2024**, *177*, 108593. [CrossRef] [PubMed]
155. Chen, L.; Shen, C.; Zhou, Z.; Maquilan, G.; Thomas, K.; Folkert, M.R.; Albuquerque, K.; Wang, J. Accurate segmenting of cervical tumors in PET imaging based on similarity between adjacent slices. *Comput. Biol. Med.* **2018**, *97*, 30–36. [CrossRef]
156. Chen, X.; Liu, W.; Thai, T.C.; Castellano, T.; Gunderson, C.C.; Moore, K.; Mannel, R.S.; Liu, H.; Zheng, B.; Qiu, Y. Developing a new radiomics-based CT image marker to detect lymph node metastasis among cervical cancer patients. *Comput. Methods Programs Biomed.* **2020**, *197*, 105759. [CrossRef]
157. Zheng, C.; Shen, Q.; Zhao, L.; Wang, Y. Utilising deep learning networks to classify ZEB2 expression images in cervical cancer. *Br. J. Hosp. Med.* **2024**, *85*, 1–13. [CrossRef]
158. Brenes, D.; Salcedo, M.P.; Coole, J.B.; Maker, Y.; Kortum, A.; Schwarz, R.A.; Carns, J.; Vohra, I.S.; Possati-Resende, J.C.; Antoniazzi, M.; et al. Multiscale Optical Imaging Fusion for Cervical Precancer Diagnosis: Integrating Widefield Colposcopy and High-Resolution Endomicroscopy. *IEEE Trans. Bio-Med. Eng.* **2024**, *71*, 2547–2556. [CrossRef]
159. Shandilya, G.; Gupta, S.; Almogren, A.; Bharany, S.; Altameem, A.; Rehman, A.U.; Hussen, S. Enhancing advanced cervical cell categorization with cluster-based intelligent systems by a novel integrated CNN approach with skip mechanisms and GAN-based augmentation. *Sci. Rep.* **2024**, *14*, 29040. [CrossRef]
160. Mathivanan, S.K.; Francis, D.; Srinivasan, S.; Khatavkar, V.; P, K.; Shah, M.A. Enhancing cervical cancer detection and robust classification through a fusion of deep learning models. *Sci. Rep.* **2024**, *14*, 10812. [CrossRef]
161. Wang, C.W.; Liou, Y.A.; Lin, Y.J.; Chang, C.C.; Chu, P.H.; Lee, Y.C.; Wang, C.H.; Chao, T.K. Artificial intelligence-assisted fast screening cervical high grade squamous intraepithelial lesion and squamous cell carcinoma diagnosis and treatment planning. *Sci. Rep.* **2021**, *11*, 16244. [CrossRef]
162. Dong, B.; Xue, H.; Li, Y.; Li, P.; Chen, J.; Zhang, T.; Chen, L.; Pan, D.; Liu, P.; Sun, P. Classification and diagnosis of cervical lesions based on colposcopy images using deep fully convolutional networks: A man-machine comparison cohort study. *Fundam. Res.* **2025**, *5*, 419–428. [CrossRef] [PubMed]
163. Xiao, T.; Wang, C.; Yang, M.; Yang, J.; Xu, X.; Shen, L.; Yang, Z.; Xing, H.; Ou, C.Q. Use of Virus Genotypes in Machine Learning Diagnostic Prediction Models for Cervical Cancer in Women With High-Risk Human Papillomavirus Infection. *JAMA Netw. Open* **2023**, *6*, e2326890. [CrossRef] [PubMed]
164. He, S.; Zhu, G.; Zhou, Y.; Yang, B.; Wang, J.; Wang, Z.; Wang, T. Predictive models for personalized precision medical intervention in spontaneous regression stages of cervical precancerous lesions. *J. Transl. Med.* **2024**, *22*, 686. [CrossRef] [PubMed]
165. Wang, Y.; Wang, T.; Yan, D.; Zhao, H.; Wang, M.; Liu, T.; Fan, X.; Xu, X. Vaginal microbial profile of cervical cancer patients receiving chemoradiotherapy: The potential involvement of Lactobacillus iners in recurrence. *J. Transl. Med.* **2024**, *22*, 575. [CrossRef]
166. Wu, Z.; Li, T.; Han, Y.; Jiang, M.; Yu, Y.; Xu, H.; Yu, L.; Cui, J.; Liu, B.; Chen, F.; et al. Development of models for cervical cancer screening: Construction in a cross-sectional population and validation in two screening cohorts in China. *BMC Med.* **2021**, *19*, 197. [CrossRef]
167. Namalinzi, F.; Galadima, K.R.; Nalwanga, R.; Sekitoleko, I.; Uwimbabazi, L.F.R. Prediction of precancerous cervical cancer lesions among women living with HIV on antiretroviral therapy in Uganda: A comparison of supervised machine learning algorithms. *BMC Women's Health* **2024**, *24*, 393. [CrossRef]
168. Liu, S.; Zhou, Y.; Wang, C.; Shen, J.; Zheng, Y. Prediction of lymph node status in patients with early-stage cervical cancer based on radiomic features of magnetic resonance imaging (MRI) images. *BMC Med. Imaging* **2023**, *23*, 101. [CrossRef]
169. Senthilkumar, G.; Ramakrishnan, J.; Frnda, J.; Ramachandran, M.; Gupta, D.; Tiwari, P.; Shorfuzzaman, M.; Mohammed, M.A. Incorporating Artificial Fish Swarm in Ensemble Classification Framework for Recurrence Prediction of Cervical Cancer. *IEEE Access* **2021**, *9*, 83876–83886. [CrossRef]
170. Suvanasuthi, R.; Therasakvichya, S.; Kanchanapiboon, P.; Promptmas, C.; Chimnaronk, S. Analysis of precancerous lesion-related microRNAs for early diagnosis of cervical cancer in the Thai population. *Sci. Rep.* **2025**, *15*, 142. [CrossRef]

171. Du, S.; Zhao, Y.; Lv, C.; Wei, M.; Gao, Z.; Meng, X. Applying Serum Proteins and MicroRNA as Novel Biomarkers for Early-Stage Cervical Cancer Detection. *Sci. Rep.* **2020**, *10*, 9033. [CrossRef]
172. Liu, Y.; Dong, T.f.; Li, P.j.; Chen, L.b.; Song, T. MRI-based radiomics features for the non-invasive prediction of FIGO stage in cervical carcinoma: A multi-center study. *Magn. Reson. Imaging* **2024**, *110*, 170–175. [CrossRef] [PubMed]
173. Yi, J.; Lei, X.; Zhang, L.; Zheng, Q.; Jin, J.; Xie, C.; Jin, X.; Ai, Y. The Influence of Different Ultrasonic Machines on Radiomics Models in Prediction Lymph Node Metastasis for Patients with Cervical Cancer. *Technol. Cancer Res. Treat.* **2022**, *21*, 15330338221118412. [CrossRef]
174. Ye, Y.; Li, M.; Pan, Q.; Fang, X.; Yang, H.; Dong, B.; Yang, J.; Zheng, Y.; Zhang, R.; Liao, Z. Machine learning-based classification of deubiquitinase USP26 and its cell proliferation inhibition through stabilizing KLF6 in cervical cancer. *Comput. Biol. Med.* **2024**, *168*, 107745. [CrossRef]
175. Guo, L.; Wang, W.; Xie, X.; Wang, S.; Zhang, Y. Machine learning-based models for genomic predicting neoadjuvant chemotherapeutic sensitivity in cervical cancer. *Biomed. Pharmacother.* **2023**, *159*, 114256. [CrossRef]
176. Ma, J.H.; Huang, Y.; Liu, L.Y.; Feng, Z. An 8-gene DNA methylation signature predicts the recurrence risk of cervical cancer. *J. Int. Med. Res.* **2021**, *49*, 03000605211018443. [CrossRef] [PubMed]
177. Ramesh, K.; Agarwal, P.; Ahuja, V.; Mir, B.A.; Yuriy, S.; Altuwairiqi, M.; Nuagah, S.J. Biomedical Application of Identified Biomarkers Gene Expression Based Early Diagnosis and Detection in Cervical Cancer with Modified Probabilistic Neural Network. *Contrast Media Mol. Imaging* **2022**, *2022*, 4946154. [CrossRef] [PubMed]
178. Monthatip, K.; Boonnag, C.; Muangmool, T.; Charoenkwan, K. A machine learning-based prediction model of pelvic lymph node metastasis in women with early-stage cervical cancer. *J. Gynecol. Oncol.* **2023**, *35*, e17. [CrossRef]
179. Félix, M.M.; Tavares, M.V.; Santos, I.P.; Batista de Carvalho, A.L.M.; Batista de Carvalho, L.A.E.; Marques, M.P.M. Cervical Squamous Cell Carcinoma Diagnosis by FTIR Microspectroscopy. *Molecules* **2024**, *29*, 922. [CrossRef]
180. Kawahara, D.; Nishibuchi, I.; Kawamura, M.; Yoshida, T.; Koh, I.; Tomono, K.; Sekine, M.; Takahashi, H.; Kikuchi, Y.; Kudo, Y.; et al. Radiomic Analysis for Pretreatment Prediction of Recurrence Post-Radiotherapy in Cervical Squamous Cell Carcinoma Cancer. *Diagnostics* **2022**, *12*, 2346. [CrossRef]
181. Zhang, X.; Yan, W.; Jin, H.; Yu, B.; Zhang, H.; Ding, B.; Chen, X.; Zhang, Y.; Xia, Q.; Meng, D.; et al. Transcriptional and post-transcriptional regulation of CARMN and its anti-tumor function in cervical cancer through autophagic flux blockade and MAPK cascade inhibition. *J. Exp. Clin. Cancer Res.* **2024**, *43*, 305. [CrossRef]
182. Park, S.H.; Hahm, M.H.; Bae, B.K.; Chong, G.O.; Jeong, S.Y.; Na, S.; Jeong, S.; Kim, J.C. Magnetic resonance imaging features of tumor and lymph node to predict clinical outcome in node-positive cervical cancer: A retrospective analysis. *Radiat. Oncol.* **2020**, *15*, 86. [CrossRef] [PubMed]
183. Zhang, X.F.; Wu, H.y.; Liang, X.W.; Chen, J.L.; Li, J.; Zhang, S.; Liu, Z. Deep-learning-based radiomics of intratumoral and peritumoral MRI images to predict the pathological features of adjuvant radiotherapy in early-stage cervical squamous cell carcinoma. *BMC Women's Health* **2024**, *24*, 182. [CrossRef]
184. Cai, Z.; Li, S.; Xiong, Z.; Lin, J.; Sun, Y. Multimodal MRI-based deep-radiomics model predicts response in cervical cancer treated with neoadjuvant chemoradiotherapy. *Sci. Rep.* **2024**, *14*, 19090. [CrossRef]
185. William, W.; Ware, A.; Basaza-Ejiri, A.H.; Obungoloch, J. A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. *Comput. Methods Programs Biomed.* **2018**, *164*, 15–22. [CrossRef]
186. Al Mudawi, N.; Alazeb, A. A Model for Predicting Cervical Cancer Using Machine Learning Algorithms. *Sensors* **2022**, *22*, 4132. [CrossRef]
187. Rhee, D.J.; Jhingran, A.; Rigaud, B.; Netherton, T.; Cardenas, C.E.; Zhang, L.; Vedam, S.; Kry, S.; Brock, K.K.; Shaw, W.; et al. Automatic contouring system for cervical cancer using convolutional neural networks. *Med. Phys.* **2020**, *47*, 5648–5658. [CrossRef] [PubMed]
188. Yang, C.; Qin, L.H.; Xie, Y.E.; Liao, J.Y. Deep learning in CT image segmentation of cervical cancer: A systematic review and meta-analysis. *Radiat. Oncol.* **2022**, *17*, 175. [CrossRef] [PubMed]
189. Rahimi, M.; Akbari, A.; Asadi, F.; Emami, H. Cervical cancer survival prediction by machine learning algorithms: A systematic review. *BMC Cancer* **2023**, *23*, 341. [CrossRef]
190. Chakraborty, C.; Bhattacharya, M.; Pal, S.; Lee, S.S. From machine learning to deep learning: Advances of the recent data-driven paradigm shift in medicine and healthcare. *Curr. Res. Biotechnol.* **2024**, *7*, 100164. [CrossRef]
191. Rahman, A.; Debnath, T.; Kundu, D.; Khan, M.S.I.; Aishi, A.A.; Sazzad, S.; Sayduzzaman, M.; Band, S.S.; Rahman, A.; Debnath, T.; et al. Machine learning and deep learning-based approach in smart healthcare: Recent advances, applications, challenges and opportunities. *AIMS Public Health* **2024**, *11*, 58–109.
192. Pandey, M.; Fernandez, M.; Gentile, F.; Isayev, O.; Tropsha, A.; Stern, A.C.; Cherkasov, A. The transformational role of GPU computing and deep learning in drug discovery. *Nat. Mach. Intell.* **2022**, *4*, 211–221. [CrossRef]
193. Wang, Z.; Liu, K.; Li, J.; Zhu, Y.; Zhang, Y. Various Frameworks and Libraries of Machine Learning and Deep Learning: A Survey. *Arch. Comput. Methods Eng.* **2024**, *31*, 1–24. [CrossRef]

194. Mienye, I.D.; Swart, T.G. A Comprehensive Review of Deep Learning: Architectures, Recent Advances, and Applications. *Information* **2024**, *15*, 755. [CrossRef]
195. Alzubaidi, L.; Bai, J.; Al-Sabaawi, A.; Santamaría, J.; Albahri, A.S.; Al-dabbagh, B.S.N.; Fadhel, M.A.; Manoufali, M.; Zhang, J.; Al-Timemy, A.H.; et al. A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications. *J. Big Data* **2023**, *10*, 46. [CrossRef]
196. Hakami, A. Strategies for overcoming data scarcity, imbalance, and feature selection challenges in machine learning models for predictive maintenance. *Sci. Rep.* **2024**, *14*, 9645. [CrossRef] [PubMed]
197. Vazquez, B.; Hevia-Montiel, N.; Perez-Gonzalez, J.; Haro, P. Weighted–VAE: A deep learning approach for multimodal data generation applied to experimental T. cruzi infection. *PLoS ONE* **2025**, *20*, e0315843. [CrossRef] [PubMed]
198. Apicella, A.; Isgrò, F.; Prevete, R. Don't Push the Button! Exploring Data Leakage Risks in Machine Learning and Transfer Learning. *arXiv* **2024**. arXiv:2401.13796. [CrossRef]
199. Yoshizawa, R.; Yamamoto, K.; Ohtsuki, T. Investigation of Data Leakage in Deep-Learning-Based Blood Pressure Estimation Using Photoplethysmogram/Electrocardiogram. *IEEE Sens. J.* **2023**, *23*, 13311–13318. [CrossRef]
200. Tampu, I.E.; Eklund, A.; Haj-Hosseini, N. Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. *Sci. Data* **2022**, *9*, 580. [CrossRef]
201. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4*, 100804. [CrossRef]
202. Xu, Y.; Goodacre, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J. Anal. Test.* **2018**, *2*, 249–262. [CrossRef] [PubMed]
203. Oeding, J.F.; Krych, A.J.; Pearle, A.D.; Kelly, B.T.; Kunze, K.N. Medical Imaging Applications Developed Using Artificial Intelligence Demonstrate High Internal Validity Yet Are Limited in Scope and Lack External Validation. *Arthrosc. J. Arthrosc. Relat. Surg.* **2025**, *41*, 455–472. [CrossRef] [PubMed]
204. Santos, C.S.; Amorim-Lopes, M. Externally validated and clinically useful machine learning algorithms to support patient-related decision-making in oncology: A scoping review. *BMC Med. Res. Methodol.* **2025**, *25*, 45. [CrossRef] [PubMed]
205. Shaharuddin, S.; Abdul Maulud, K.N.; Syed Abdul Rahman, S.A.F.; Che Ani, A.I.; Pradhan, B. The role of IoT sensor in smart building context for indoor fire hazard scenario: A systematic review of interdisciplinary articles. *Internet Things* **2023**, *22*, 100803. [CrossRef]
206. Woldaregay, A.Z.; Årsand, E.; Walderhaug, S.; Albers, D.; Mamykina, L.; Botsis, T.; Hartvigsen, G. Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes. *Artif. Intell. Med.* **2019**, *98*, 109–134. [CrossRef]
207. Lawan, A.A.; Cavus, N.; Yunusa, R.; Abdulrazak, U.I.; Tahir, S. Chapter 12—Fundamentals of machine-learning modeling for behavioral screening and diagnosis of autism spectrum disorder. In *Neural Engineering Techniques for Autism Spectrum Disorder*; El-Baz, A.S., Suri, J.S., Eds.; Academic Press: Cambridge, MA, USA, 2023; pp. 253–268. [CrossRef]
208. Viering, T.; Loog, M. The Shape of Learning Curves: A Review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 7799–7819. [CrossRef]
209. Taylor, M.J. Legal Bases For Disclosing Confidential Patient Information for Public Health: Distinguishing Between Health Protection And Health Improvement. *Med. Law Rev.* **2015**, *23*, 348–374. [CrossRef] [PubMed]
210. Sharif, M.I.; Mehmood, M.; Uddin, M.P.; Siddique, K.; Akhtar, Z.; Waheed, S. Federated Learning for Analysis of Medical Images: A Survey. *J. Comput. Sci.* **2024**, *20*, 1610–1621. [CrossRef]
211. Liu, X.; Xie, L.; Wang, Y.; Zou, J.; Xiong, J.; Ying, Z.; Vasilakos, A.V. Privacy and Security Issues in Deep Learning: A Survey. *IEEE Access* **2021**, *9*, 4566–4593. [CrossRef]
212. Golla, G. Security and Privacy Challenges in Deep Learning Models. *arXiv* **2023**, arXiv:2311.13744. [CrossRef]
213. Meng, H.; Wagner, C.; Triguero, I. Explaining time series classifiers through meaningful perturbation and optimisation. *Inf. Sci.* **2023**, *645*, 119334. [CrossRef]
214. Teng, Q.; Liu, Z.; Song, Y.; Han, K.; Lu, Y. A survey on the interpretability of deep learning in medical diagnosis. *Multimed. Syst.* **2022**, *28*, 2335–2355. [CrossRef] [PubMed]
215. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]
216. Beam, A.L.; Manrai, A.K.; Ghassemi, M. Challenges to the Reproducibility of Machine Learning Models in Health Care. *JAMA* **2020**, *323*, 305–306. [CrossRef]
217. Mislan, K.A.S.; Heer, J.M.; White, E.P. Elevating The Status of Code in Ecology. *Trends Ecol. Evol.* **2016**, *31*, 4–7. [CrossRef] [PubMed]
218. Torres-Poveda, K.; Piña-Sánchez, P.; Vallejo-Ruiz, V.; Lizano, M.; Cruz-Valdez, A.; Juárez-Sánchez, P.; Garza-Salazar, J.d.l.; Manzo-Merino, J. Molecular Markers for the Diagnosis of High-Risk Human Papillomavirus Infection and Triage of Human Papillomavirus-Positive Women. *Rev. Investig. Clín.* **2020**, *72*, 198–212. [CrossRef]

219. Goodman, A. Deep-Learning-Based Evaluation of Dual Stain Cytology for Cervical Cancer Screening: A New Paradigm. *J. Natl. Cancer Inst.* **2021**, *113*, 1451–1452. [CrossRef]
220. Desai, K.T.; Befano, B.; Xue, Z.; Kelly, H.; Campos, N.G.; Egemen, D.; Gage, J.C.; Rodriguez, A.C.; Sahasrabuddhe, V.; Levitz, D.; et al. The development of "automated visual evaluation" for cervical cancer screening: The promise and challenges in adapting deep-learning for clinical testing: Interdisciplinary principles of automated visual evaluation in cervical screening. *Int. J. Cancer* **2022**, *150*, 741–752. [CrossRef]
221. U.S. Food and Drug Administration. Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products. Available online: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological (accessed on 28 May 2025).
222. World Health Organization (WHO). Atlas of Visual Inspection of the Cervix with Acetic Acid for Screening, Triage, and Assessment for Treatment. Available online: https://screening.iarc.fr/atlasvia.php (accessed on 28 May 2025).
223. Li, J.; Hu, P.; Gao, H.; Shen, N.; Hua, K. Classification of cervical lesions based on multimodal features fusion. *Comput. Biol. Med.* **2024**, *177*, 108589. [CrossRef]
224. Mennella, C.; Maniscalco, U.; De Pietro, G.; Esposito, M. Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. *Heliyon* **2024**, *10*, e26297. [CrossRef]
225. Wu, T.; Lucas, E.; Zhao, F.; Basu, P.; Qiao, Y. Artificial intelligence strengthens cervical cancer screening - present and future. *Cancer Biol. Med.* **2024**, *21*, 864–879. [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.

diagnostics

MDPI

Article

Diagnosis of Mesothelioma Using Image Segmentation and Class-Based Deep Feature Transformations

Siyami Aydın [1,*], Mehmet Ağar [1], Muharrem Çakmak [1] and Mesut Toğaçar [2]

[1] Department of Thoracic Surgery, Faculty of Medicine, Fırat University, Elazığ 23119, Türkiye; md.mehmetagar@gmail.com (M.A.); g.c.dr.ckmk@gmail.com (M.Ç.)
[2] Department of Management Information Systems, Faculty of Economics and Administrative Sciences, Fırat University, Elazığ 23119, Türkiye; mtogacar@firat.edu.tr
* Correspondence: siyamiaydin51@gmail.com

Abstract: Background/Objectives: Mesothelioma is a rare and aggressive form of cancer that primarily affects the lining of the lungs, abdomen, or heart. It typically arises from exposure to asbestos and is often diagnosed at advanced stages. Limited datasets and complex tissue structures contribute to delays in diagnosis. This study aims to develop a novel hybrid model to improve the accuracy and timeliness of mesothelioma diagnosis. **Methods**: The proposed approach integrates automatic image segmentation, transformer-based model training, class-based feature extraction, and image transformation techniques. Initially, CT images were processed using the segment anything model (SAM) for region-focused segmentation. These segmented images were then used to train transformer models (CaiT and PVT) to extract class/type-specific features. Each class-based feature set was transformed into an image using Decoder, GAN, and NeRV techniques. Discriminative score and class centroid analysis were then applied to select the most informative image representation for each input. Finally, classification was performed using a residual-based support vector machine (SVM). **Results**: The proposed hybrid method achieved a classification accuracy of 99.80% in diagnosing mesothelioma, demonstrating its effectiveness in handling limited data and complex tissue characteristics. **Conclusions**: The results indicate that the proposed model offers a highly accurate and efficient approach to mesothelioma diagnosis. By leveraging advanced segmentation, feature extraction, and representation techniques, it effectively addresses the major challenges associated with early and precise detection of mesothelioma.

Keywords: mesothelioma disease; mesothelioma detection; transformer-based segmentation; class-based feature; deep generative; discriminative score

1. Introduction

Malignant mesothelioma (MM) is a rare and aggressive tumor that originates in the mesothelial cells that line the pleura, peritoneum, pericardium, or tunica vaginalis. Malignant pleural mesothelioma (MPM) constitutes nearly 80% of all cases and is primarily associated with prolonged asbestos exposure [1]. Despite global regulations limiting asbestos use, the disease burden persists due to its long latency period, which is often 20 to 50 years [2].

MM presents insidiously with symptoms such as progressive dyspnea, pleuritic chest pain, chronic cough, fatigue, and unintentional weight loss. These nonspecific symptoms often result in delayed diagnoses; more than 60% of patients are diagnosed at advanced stages. Physical examination findings may include decreased breath sounds and

dullness to percussion [3]. Paraneoplastic manifestations, such as thrombocytosis or digital clubbing, may also be observed. Diagnosis typically requires a combination of imaging, histopathology, and molecular testing. Chest CT is the initial imaging modality and often reveals unilateral pleural thickening or effusion. PET-CT can aid in staging. However, tissue biopsy remains the gold standard for diagnosis. Immunohistochemistry is essential for distinguishing MM from other malignancies, particularly metastatic adenocarcinoma [4]. Recent advances in artificial intelligence (AI)-assisted imaging analysis show promise in improving diagnostic accuracy by identifying radiographic patterns predictive of MM. This could allow for earlier and more reliable detection [5].

MPM is pathologically classified into three subtypes: epithelioid, sarcomatoid, and biphasic. The epithelioid subtype has a better prognosis, while the sarcomatoid and biphasic variants are more resistant to therapy. Molecular profiling is playing an increasingly important role in guiding treatment, particularly in identifying candidates for targeted therapies and immunotherapies. Historically, the first-line treatment for unresectable MPM has been a combination of cisplatin and pemetrexed, which offers limited survival benefit. However, the therapeutic landscape has changed significantly in recent years. In 2020, the FDA approved the combination of the immune checkpoint inhibitors nivolumab and ipilimumab as a first-line treatment for unresectable MPM. The pivotal CheckMate 743 trial demonstrated a median overall survival of 18.1 months with immunotherapy versus 14.1 months with chemotherapy [6].

Further therapeutic advancements include using ADI-PEG20, an enzyme that depletes arginine, to treat ASS1-deficient mesothelioma. A recent randomized phase 2/3 trial (ATOMIC-Meso) reported a significant improvement in overall survival with the addition of ADI-PEG20 to pemetrexed and cisplatin [7].

Experimental studies are investigating the utility of AI in treatment planning and prognostication. Algorithms trained on multimodal data (histopathology, genomics, and imaging) may soon enable clinicians to personalize treatment regimens more effectively and predict responses to immunotherapy more accurately [8]. Malignant mesothelioma remains a devastating disease with a poor prognosis and limited treatment options. Nevertheless, recent progress in immunotherapy, targeted agents, and AI-assisted diagnostics offers hope for earlier diagnosis and improved outcomes. Continued interdisciplinary research is essential to translating these innovations into routine clinical practice. We will examine some recent studies in the literature that have implemented AI-based mesothelioma diagnosis.

Kitajima et al. [9] designed a three-dimensional deep convolutional neural network (3D-CNN) based on PET/CT data and successfully distinguished MPM diagnosis from benign types with an overall accuracy of 77.3%. Gill et al. [10] used a limited number of biological and radiological parameter data (pleural effusion, pH, C-reactive protein, etc.) for mesothelioma diagnosis and employed gradient-boosted trees (GBTs), support vector machines (SVMs), and logistic regression (LR) methods. They achieved 100% overall accuracy in their study. Cheng et al. [11] analyzed biological and radiological parameter data for mesothelioma diagnosis using machine learning methods. In their study, they obtained the best results with the random forest (RF) method, achieving an overall accuracy of 86.54%. Kidd et al. [12] used CT images for mesothelioma diagnosis. They performed segmentation using the U-Net model and then classification using the CNN model they designed. They achieved an overall accuracy of 66.67% in the binary classification process. A. Choudhury [13] used clinical and biomarker data to diagnose mesothelioma. In the experimental analysis, he used various machine learning methods and obtained the best result with the AdaBoost method. With the AdaBoost method, he achieved an overall accuracy of 71.29%.

Unlike previous studies, this study proposes a hybrid model that combines new approaches for MPM diagnosis. The proposed model involves the following steps: segmentation; model training; extraction of class-based feature sets from the models; conversion of these feature sets into image features; and classification. The contributions of the proposed approach to the literature are as follows:

- Use of the segment anything model (SAM) for region-focused segmentation of CT images.
- Integration of transformer-based models for class-based feature extraction.
- New use of class-based image transformation through Decoder, GAN, and NeRV techniques.
- Application of a discriminative score-based selection to determine the most informative image representations.
- Achieving high diagnostic accuracy through residual-based SVM classification.

The remaining sections of the article are summarized as follows: materials and methods are presented in Section 2; experimental analyses are presented in Section 3; comparison and discussion of the analysis results are presented in Section 4; and conclusions and information about future studies are presented in Section 5.

2. Materials and Methodology

2.1. CT Image Dataset

The dataset is not publicly accessible and was created by specialists in the field. The images included in the dataset were obtained from Fırat University Research Hospital. The retrospective study included a total of 172 patients, comprising 86 patients diagnosed with mesothelioma based on radiological and pathological findings between 2012 and 2024, and 86 patients investigated for suspected mesothelioma but who did not receive a mesothelioma diagnosis. A total of 1008 sections were included in the study, with an average of 5.8 chest CT images obtained from each group (mesothelioma and non-mesothelioma). Of the patients included in the study, 98 were male and 74 were female. The average age was 56.8 ± 14.6 years. In the staging, stage 3 was the most common, and the most common pathological diagnosis was epithelioid-type mesothelioma.

Asbestos exposure was present in 131 (76.2%) of the patients, and a history of smoking was present in 97 (56.4%) of the patients. Of the patients, 102 (59.3%) reported chest wall or chest pain, and 115 (66.9%) had weight loss. Detailed information is provided in Table 1. The dataset consists of two classes. The number of images from patients diagnosed with mesothelioma is 504. The number of patients without a mesothelioma diagnosis but with images suggestive of mesothelioma is 504. The non-mesothelioma group in the dataset includes only benign pleural diseases; other subgroups, such as metastatic pleural diseases and pleural effusion, are not included in this assessment. Benign pleural diseases include pleural plaques, infection-related conditions, inflammatory diseases, and findings related to trauma or surgery. This dataset was created to compare individuals diagnosed with mesothelioma with individuals who initially had suspected mesothelioma but were ultimately diagnosed with benign pleural disease after further investigation. The images in the dataset are equally classified and consist of a total of 1008 images. All CT slices corresponding to a single patient were kept in the same subset (either training or test), ensuring no overlap of patient data across sets. Since CT imaging inherently provides multiple slices at different depths and angles, all slices belonging to the same patient, regardless of orientation, depth, or reconstruction parameters, were grouped within the same subset. This patient-based separation ensured the independence of training and test datasets, eliminating the risk of data leakage. A sample image set is shown in Figure 1.

Table 1. Demographic, clinical, and imaging characteristics of the CT dataset used for mesothelioma and non-mesothelioma patients.

Characteristic	Category/Unit	Patients (N, %)
Gender	Male	98
	Female	74
Age	Year, mean ± SD	56.8 ± 14.6
Stage	1	15 (8.7%)
	2	42 (24.4%)
	3	68 (39.5%)
	4	47 (27.3%)
Histological type	Epithelioid	133 (77.3%)
	Biphasic	16 (9.3%)
	Sarcomatoid	15 (8.7%)
	Not determined	8 (4.7%)
Asbestos exposure	No	41 (23.8%)
	Yes	131 (76.2%)
Smoking	No	75 (43.6%)
	Yes	97 (56.4%)
Pain	No	70 (40.7%)
	Yes	102 (59.3%)
Weight loss	No	57 (33.1%)
	Yes	115 (66.9%)

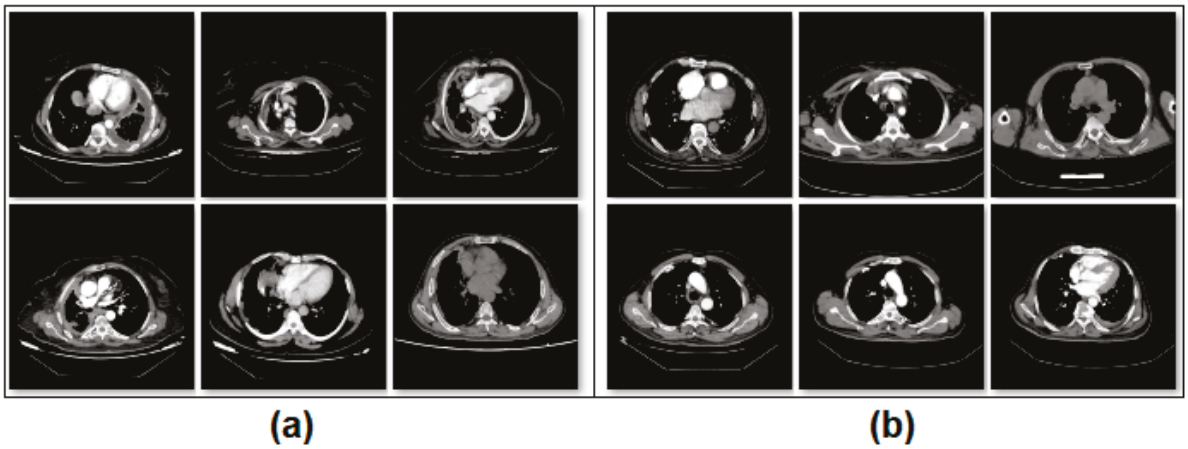


Figure 1. Class-based sample images of the dataset; (**a**) mesothelioma, (**b**) non-mesothelioma.

2.2. Segment Anything Model

The segment anything model (SAM) is a segmentation approach that distinguishes objects in images and processes them using a masking technique. The SAM can focus on different types of commands, such as single or multiple points or textual regions, within an image. These commands are specified before running the model. One of the SAM's key advantages is its ability to perform meaningful segmentation on image sets, even when commands are ambiguous. This success is underpinned by the model's prior learning of object concepts. This feature enables successful results to be achieved even on untrained test examples [14]. During SAM training, pre-trained large data files (e.g., "sam vit-h") are used. The core modules in the model's architectural structure perform the following tasks:

- Image Encoder: Processes the input image and extracts visual information. This information is converted into a high-dimensional and deep visual representation. The resulting representation is prepared for use in the next stage.

- Prompt Encoder: Processes commands provided by the user (e.g., a single point, multiple points, etc.). These commands determine which part of the image the model should focus on. The commands are converted into vectors that the model can understand.
- Mask Decoder: Combines the image representations and command vectors obtained from the previous two stages to create a segmentation mask for the focused area. In short, it predicts which pixel belongs to which object. The SAM can generate multiple masks probabilistically; however, the mask with the highest probability is selected during the process [15].

The reasons for choosing the SAM in this study are as follows: its ability to operate using command-based inputs (such as points, text, boxes, etc.), its capability to distinguish even previously undefined objects, its ability to differentiate between various types of objects, its ability to process an image only once to produce faster results, and its provision of different masking options to apply the most suitable/likely mask to the image.

2.3. Transformer Models

Transformer models were first developed for text-based applications. However, in recent years, they have been adapted to visual-based datasets due to their successful results in image processing. At the core of transformer models lies the division of an image into small patches. The relationship between each patch and the others is modeled through multi-head self-attention (MHSA). Visual features present in the patches are extracted by the model and passed on to the next layer as a sequence. The extracted information is processed using feedforward neural networks (FFNNs) to perform feature extraction [16]. These steps are summarized in Figure 2.

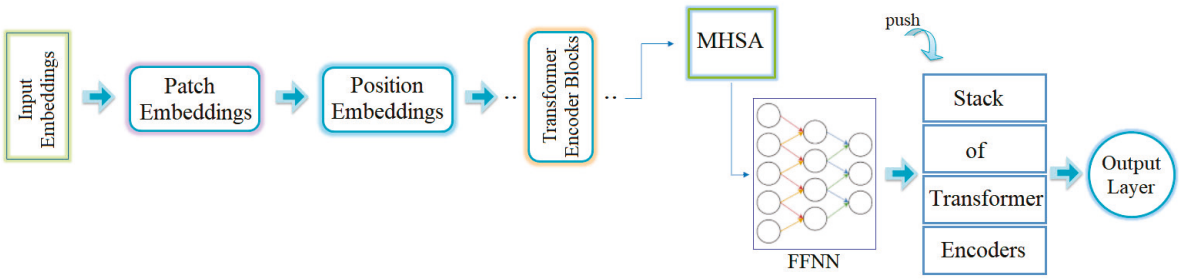


Figure 2. General workflow process of the transformer model [16].

The Class-Attention in Image Transformers (CaiT) model was developed to address challenges in training the ViT model more deeply. It uses a unique attention module architecture called "Class-Attention." This module ensures that attention is focused only on class-specific features. Unlike ViT models, CaiT uses the "LayerScale" module, which normalizes input values by applying a different mathematical operation to them. CaiT achieves better results in deep architectures [17].

The Pyramid Vision Transformer (PVT-v2) is a transformer-based approach designed to perform various object detection tasks with dense segmentation. It serves as a multi-modal backbone. Its layers process the input image from low to high in a pyramid structure at different resolution levels. This structure enables more efficient results with fewer parameters. The PVT-v2 architecture incorporates a "spatial-reduction attention" module that reduces computational costs and resource usage. The model has demonstrated high performance in classification and segmentation tasks [18].

The reasons for choosing two different transformer models in this study can be summarized as follows: CaiT enables more efficient training of the model thanks to its deep-layer architecture. In addition, it improves model performance by focusing directly on target object features through attention mechanisms specific to class information. Furthermore,

approaches such as "LayerScale" have made the training process of the CaiT model more stable. PVT-v2, on the other hand, offers the advantage of capturing both the general structure and detailed features by processing input images at variable resolutions, unlike many other transformers. This feature reduces training costs and enables the creation of a lighter and faster model. Both models have a complementary structure thanks to their unique features in addition to their common architectural structure. Another reason for preferring multiple models in hybrid approaches is the high potential for deficiencies in one model to be compensated for by other models.

2.4. Deep Generative Approaches

Deep generative approaches are techniques that stand out in AI-based technologies. These methods aim to produce realistic input data using their own architectures. These approaches can process features extracted from complex, high-dimensional images to generate examples similar to the original image [19].

2.4.1. DecoderMLP

These approaches, which are often used in generative architectures, focus on solving latent representations through deep processing. The result is a reconstructed image that represents the original image. During reconstruction, multi-layer perceptrons (MLPs) recreate the spatial and temporal characteristics of the input data. The decoder component decodes the information encoded in the input data and converts it into an image that represents the original. This technique can also be applied to features obtained from the final fully connected (FC) layers of deep learning models or transformer architectures. This enables the processing of more meaningful features, thereby improving the accuracy of the representative images [20]. The operation performed in the Lth layer of the Decoder-MLP technique is formulated in Equation (1). In this equation, $h^{(L)}$ represents the activation output in layer L. The other parameters in Equation (1) represent the hidden vector values $h^{(L-1)}$, the weight parameter W, the bias setting b^L, and the activation function σ. In Equation (2), z: latent representation, x: pixel vector output in an image, f_{dec}: MLP-based decoder function, and S: total number of layers.

$$h^{(L)} = \sigma\left(W^{(L)}h^{(L-1)} + b^L\right) \tag{1}$$

$$x = f_{dec}(z) = h^{(S)} \tag{2}$$

The Decoder-MLP architecture consists of multiple fully connected layers, each followed by an activation function, designed to transform latent feature representations into image outputs. The operation of the L-th layer is formally defined in Equations (1) and (2). In Equation (1), ReLU is used for hidden layers to mitigate vanishing gradients and accelerate convergence, while Tanh is applied in the final layer to scale outputs to $[-1, 1]$. Each layer progressively improves the spatial and temporal qualities of the input features, enabling the reconstruction of images that accurately reflect the original data.

2.4.2. GAN Generator

Generative Adversarial Networks (GANs) create new, realistic data by using parts of existing data that machines have learned. A GAN architecture consists of two main components: a generator and a discriminator. The generator takes a randomly selected noise vector and meaningfully distributes it in the target image space, thereby producing fake images similar to the target images. This process operates through a learned function, with deep learning architectures, such as CNNs and MLPs, playing an active role. The generator's primary goal is to produce realistic images that can fool the discriminator.

During GAN training, the generator and discriminator take on opposing yet complementary roles: while the generator produces fake but realistic data, the discriminator attempts to distinguish whether this data is real or fake [21,22]. Through this adversarial learning process, feature sets extracted from the FC layers of transformer or CNN models can be converted into 2D GAN images. These fake images can then serve as more efficient representations of each example in the original dataset and be transferred to the output layer.

2.4.3. NeRV-like

The Neural Representations for Videos (NeRV) method has a unique architecture that processes video-based data directly with neural networks. Unlike traditional pixel-based video processing methods, NeRV processes each frame by considering the time index of the extracted frames through a continuous function. This enables the generation of new image frames. In this structure, the decoder component uses structural components, such as FC layers, MLPs, and convolutional encoder–decoders, to generate output similar to GAN architectures. A key feature distinguishing NeRV from similar approaches is its ability to detect temporal continuity in frames and convert it into relationships between patterns using vector representation spaces. In NeRV, rather than storing each frame separately, the architecture encodes each frame in a reproducible manner based on its time index. This approach ensures preservation of image quality with less resource usage. NeRV can be effectively used in various video processing tasks such as compression, reconstruction, editing, and data transfer. Its architecture consists of a temporal encoder and a temporal decoder. The temporal encoder converts the information obtained from the input data into high-dimensional feature vectors. These vectors are then processed by the decoder and reproduced in a sequential structure for each frame. Another important advantage of NeRV is its ability to generate frames at specific points in time through the neural network during operation. This capability fills in missing areas of frames and prevents resolution errors [23,24]. In recent years, NeRV-like methods have been combined with new approaches and have begun to be used in areas such as generative image creation.

The NeRV-like architecture includes a temporal encoder that converts input video frames into high-dimensional vectors. These vectors are then processed by a decoder consisting of convolutional and fully connected layers, which enable the sequential reconstruction of each frame. This structure enables efficient frame generation while preserving temporal continuity, without the need to store each frame separately.

2.5. Discriminative Score Method

The discriminative score (DS) is a method developed in fields such as machine learning, deep learning, and pattern recognition, which are subfields of artificial intelligence, to select the most discriminating feature among image sets or features representing these images. This method consists of two basic steps. First, feature vectors are extracted from examples belonging to each class, and then the class centers (average vectors) of these vectors are calculated. In the second step, the distances between each example and its own class center and other class centers are calculated. For effective discrimination, it is expected that the distance of a sample from its own class center is minimum, while the distances from other class centers are maximum. The DS identifies the most discriminative sample by considering these differences. The Euclidean distance is generally preferred for this measurement. The sample with the lowest DS is the one closest to its own class and farthest from other classes, and is therefore considered the best representative sample [25,26]. The general formula used in calculating the DS is given in Equation (3). In Equation (3), m_i represents the sample to be examined, C_{poz} represents the center vector of the class in

which m_i is located (positive class), and C_{neg} represents the center vector of the class in which m_i is not located but is the closest class (negative class). "d" calculates the distance between the two values using the Euclidean formula.

$$DS(m_i) = \frac{d(m_i, C_{neg})}{d(m_i, C_{poz})} \tag{3}$$

This method is particularly useful for selecting images created using generative techniques or for measuring the discriminative power between variants. This enables the selection of more efficient images, thereby contributing to the improvement of the training performance of machine learning or deep learning models.

2.6. Proposed Hybrid Approach

Mesothelioma is a rare, rapidly progressing type of pleural cancer. In its early stages, it can be difficult to diagnose and is often mistaken for other respiratory diseases by specialists. This makes it difficult to diagnose the disease in a timely manner. Ultimately, many cases are detected at an advanced stage, when treatment options are limited. In this context, there is an increasing need for digital solutions that enable early and reliable diagnoses while remaining applicable in clinical settings. This study proposes an advanced, AI-driven, multi-layered, hybrid model to support the mesothelioma diagnosis process. The proposed model consists of three steps.

The first step (step #1) involves processing the original dataset using the SAM approach and performing regional segmentation on the CT images. After segmentation, each region is processed to increase transparency by at least 50%, revealing the original background image. This method allows the segmented area and the tissue to which it belongs to be displayed simultaneously, enabling distinctive features to be highlighted more clearly.

In the second step (step #2), the SAM-segmented dataset is trained using transformer-based CaiT and PVT-v2 models, known as next-generation models. Under normal conditions, 384 features are extracted from the FC layer (final layer) of the CaiT model for each image; similarly, 512 features are extracted from the FC layer of the PVT-v2 model for each image. The fact that these two models produce different numbers of feature columns stems from their architectural design. However, by adding a new layer (logits) to the final layer of both models (CaiT and PVT-v2), a feature set of size [number of images $\times$ number of classes] is created. Since the dataset used in this study consists of two classes (mesothelioma and non-mesothelioma), 2 features are extracted from the logits layer for each image during model training. We refer to this process as "class-based feature extraction." Class-based feature extraction typically produces more efficient features compared to the FC layer. In this study, the "F1" and "F2" feature columns were extracted from the logits layer for the CaiT model, while the "F3" and "F4" feature columns were extracted for the PVT-v2 model. In the CaiT model, the "F1" feature column represents the probability of being class #1, while the "F2" feature column represents the probability of being class #2. Similarly, in the PVT-v2 model, the "F3" feature column represents the probability of being class #1, while the "F4" feature column represents the probability of being class #2.

In the third step (step #3), the F1, F2, F3, and F4 feature columns are processed using deep generative approaches such as Decoder, GAN, and NeRV techniques, and a 2D image is obtained for each feature value. For example, the "$F1_1$" probability value in the F1 column for the first sample in the dataset is represented by three different images based on Decoder, GAN, and NeRV. Similarly, the "$F1_n$" value in the F1 column for the last sample in the dataset is also visualized using the same three methods (for this study, n = 1008, which is the total number of data). This process is carried out similarly for the F2, F3, and F4 columns. Since the CaiT model is represented by columns F1 and F2, the best representation

is determined among the three images (Decoder, GAN, NeRV) that visually represent each possibility in column F1. For this purpose, the "discriminative score" method is used, and the best representative image is selected for each sample using this method. For example, for the first feature in the F1 column (class #1: mesothelioma) of the CaiT model, the image that provides the best representation according to the discriminative score method may be Decoder-based; the other two images (GAN and NeRV-based) are eliminated. This process is repeated for all samples and applied separately for each model (CaiT, PVT-v2). Subsequently, the best image set representing the F1 and F2 columns obtained with the CaiT model is classified using the residual-based SVM method. Similarly, the image set representing the F3 and F4 columns obtained with the PVT-v2 model is also classified using the same method. The classification performances obtained from the two models are compared, and the overall success of the proposed model is accepted. The general design of the proposed model, showing the pipeline, is presented in Figure 3. The proposed model aims to successfully diagnose mesothelioma by combining innovative approaches in a hybrid manner.

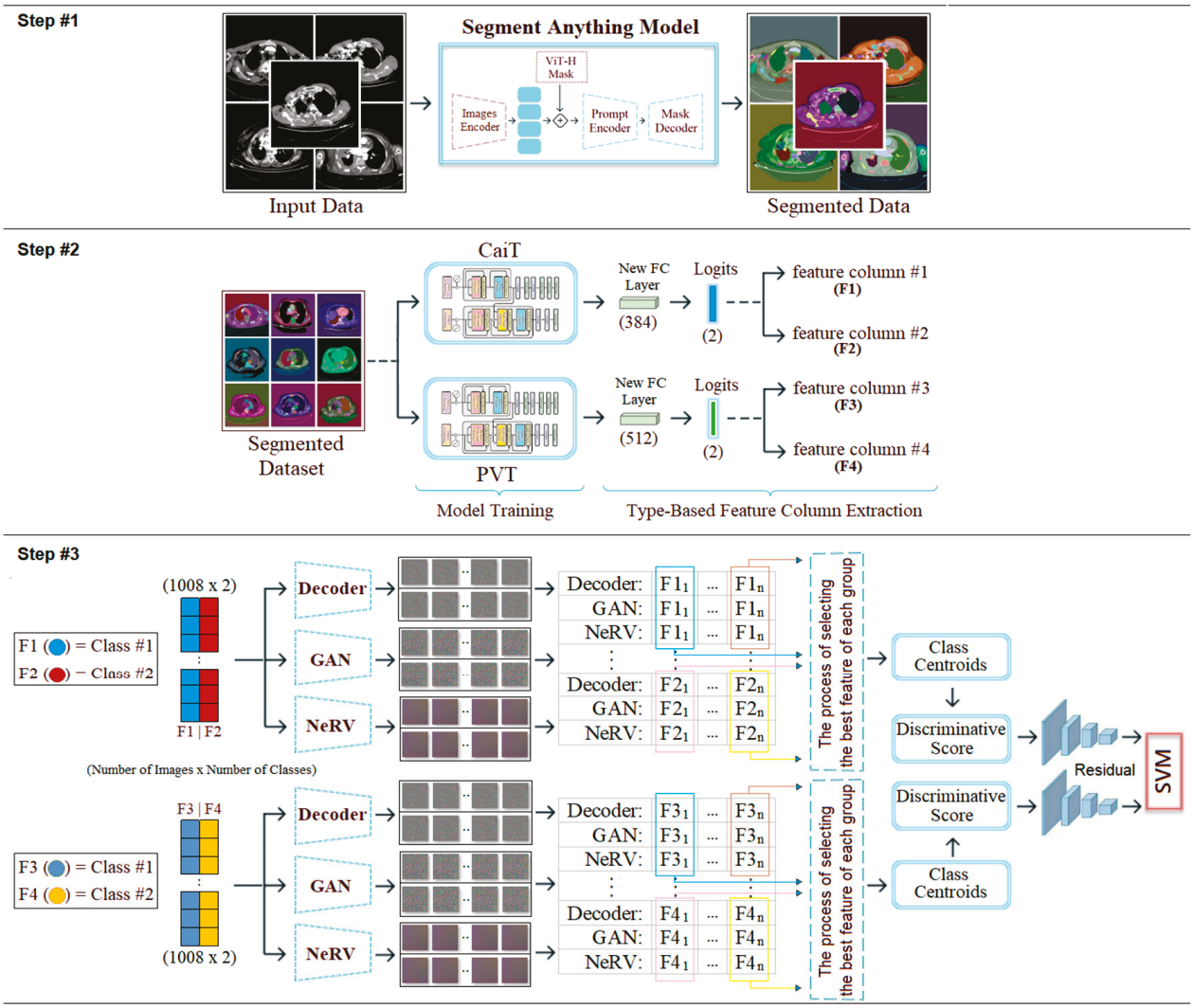


Figure 3. Pipeline steps and general design of the proposed hybrid model.

3. Experimental Analysis

3.1. Experimental Settings

Experimental analyses were performed using the Python 3.12.3 programming language via the Jupyter Notebook interface. The Google Colab server was used during the analysis process. Some important parameters of the models and methods used in the recommended hybrid approach and their values are presented in Table 2. Default parameter values were used for other approaches not specified. In this study, 70% of the dataset was separated as training data and 30% as test data.

Table 2. Model/method parameters and values used in the proposed hybrid approach.

Model (Type)/Method	Parameter	Preference/Value
CaiT and PVT-v2	Loss function	Cross Entropy
	Learning rate	10^{-4}
	CaiT model type	cait_s24_224
	PVT-v2 model type	pvt_v2_b3
	Optimization	Adam
	Input size	224 × 224
	Epoch	15
	Mini-batch	32
	Random Seed	42 (fixed)
	Learning rate schedule	Constant (no decay)
	Weight decay	Default value (0.0)
	Early stopping	Not applied
	Preprocessing	RandomResizedCrop (224), RandomHorizontalFlip (), ToTensor (), Normalize (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225])
	Training rate: testing rate	0.7:0.3
SVM	Kernel function	Cubic
	Kernel scale	Auto
	Box constraint level	1
	Multiclass method	One-vs-One
SAM	Model type	vit_h
ResNet-18 (used with SVM)	Loss function	Cross Entropy
	Optimization	Adam
	Pre-trained weights	ImageNet
	Fine-tuning	Last FC layers
	Epoch	15
	Mini-batch	32
	Input size	224 × 224

The confusion matrix, which is commonly used in classification processes, was used to validate the analyses performed in this study. The metrics associated with the confusion matrix were calculated using the formulas given in Equations (4)–(8) [27–29]. The terms

used in these equations include positive (p), negative (n), true (t), and false (f). The abbreviations for the metrics obtained are as follows: accuracy (acc), f-score (f-scr), specificity (sp), sensitivity (se), and precision (pre). The acc metric, in particular, stands out as a frequently preferred criterion in evaluations performed on balanced datasets [30–33].

$$se = \frac{tp}{fn + tp} \tag{4}$$

$$sp = \frac{tn}{fp + tn} \tag{5}$$

$$pre = \frac{tp}{fp + tp} \tag{6}$$

$$f - scr = \frac{2 \times tp}{2 \times tp + fn + fp} \tag{7}$$

$$acc = \frac{tn + tp}{tn + tp + fn + fp} \tag{8}$$

3.2. Experimental Results

In the first step of the experimental analysis, the original dataset was processed using the SAM approach to create a new segmentation-based dataset. The purpose of this step was to ensure that the model training to be performed in the next step would focus on the regional segments obtained from the CT images and that unnecessary areas would be left in the background by the model. In the CT images where segmentation was applied, a 50% transparency was applied to keep the original regions visible. The internal architecture of the SAM is composed of an Image Encoder, Prompt Encoder, and Mask Decoder. The Image Encoder extracts high-dimensional visual features from the input image, while the Prompt Encoder converts user-specified commands (points, boxes, text) into vectors guiding the segmentation. The Mask Decoder combines these representations to predict segmentation masks for target regions. The processing steps performed using the SAM are shown in Figure 4. Sample images obtained as a result of the segmentation process of the original dataset are shown in Figure 5.

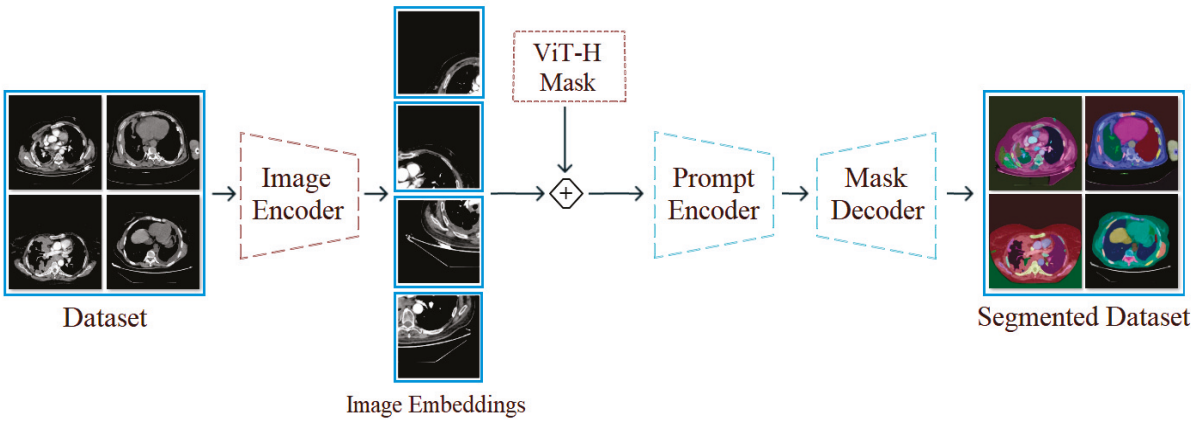


Figure 4. Design showing the general workflow steps of the SAM and its application to a sample dataset.

The second step of the experiment involves training transformer-based models (CaiT and PVT-v2) using the original dataset and the segmented dataset. The purpose of this step is to observe whether the segmented dataset contributes to model performance compared to the original dataset. In this regard, both the original and segmented datasets were trained sequentially with the CaiT and PVT-v2 models. The overall accuracy graphs of

the models are presented in Figure 6. The confusion matrices obtained by the models as a result of training with the datasets are presented in Figure 7. The macro average metric results obtained from the confusion matrices are presented in Table 3. When examining Figures 6 and 7 and Table 3, it can be observed that the highest success was achieved with the segmented dataset. When the segmented dataset was used, the CaiT model achieved a general accuracy success rate of 93.40%, while the PVT-v2 model achieved 94.39%. Among the transformer models used in the proposed approach, the PVT-v2 model yielded the best results. The CaiT model also demonstrated performance close to that of the PVT-v2 model.

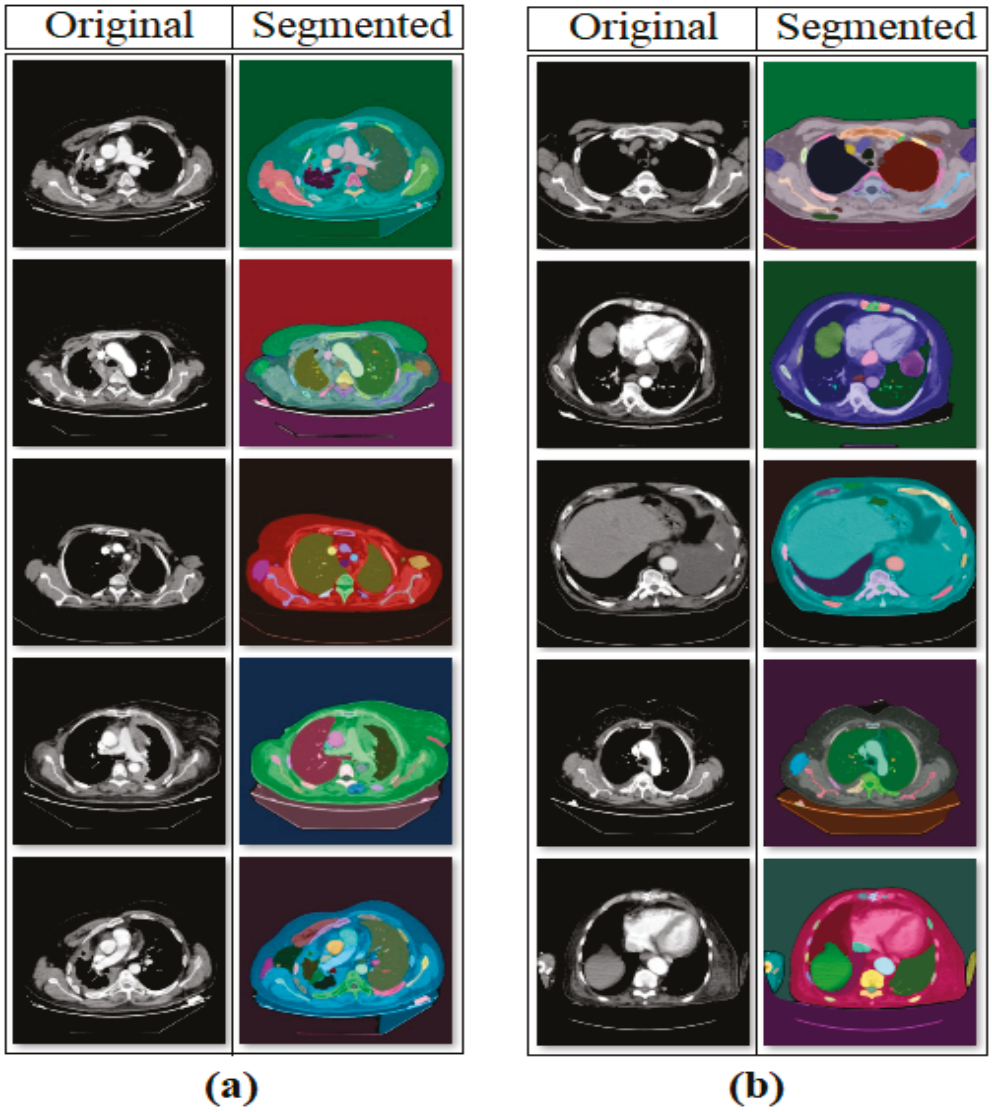


Figure 5. Sample set of segmentation-based data obtained with the SAM: (**a**) mesothelioma, (**b**) non-mesothelioma.

Table 3. Macro average metric results from the confusion matrix of the transformer models (%).

Transformer Model	Dataset	Se	Sp	Pre	F-Scr	Acc
CaiT	Original	85.47	85.47	85.50	85.47	85.48
PVT-v2	Original	85.81	85.81	85.82	85.81	85.81
CaiT	Segmented	93.40	93.40	93.43	93.40	93.40
PVT-v2	Segmented	94.39	94.39	94.39	94.39	94.39

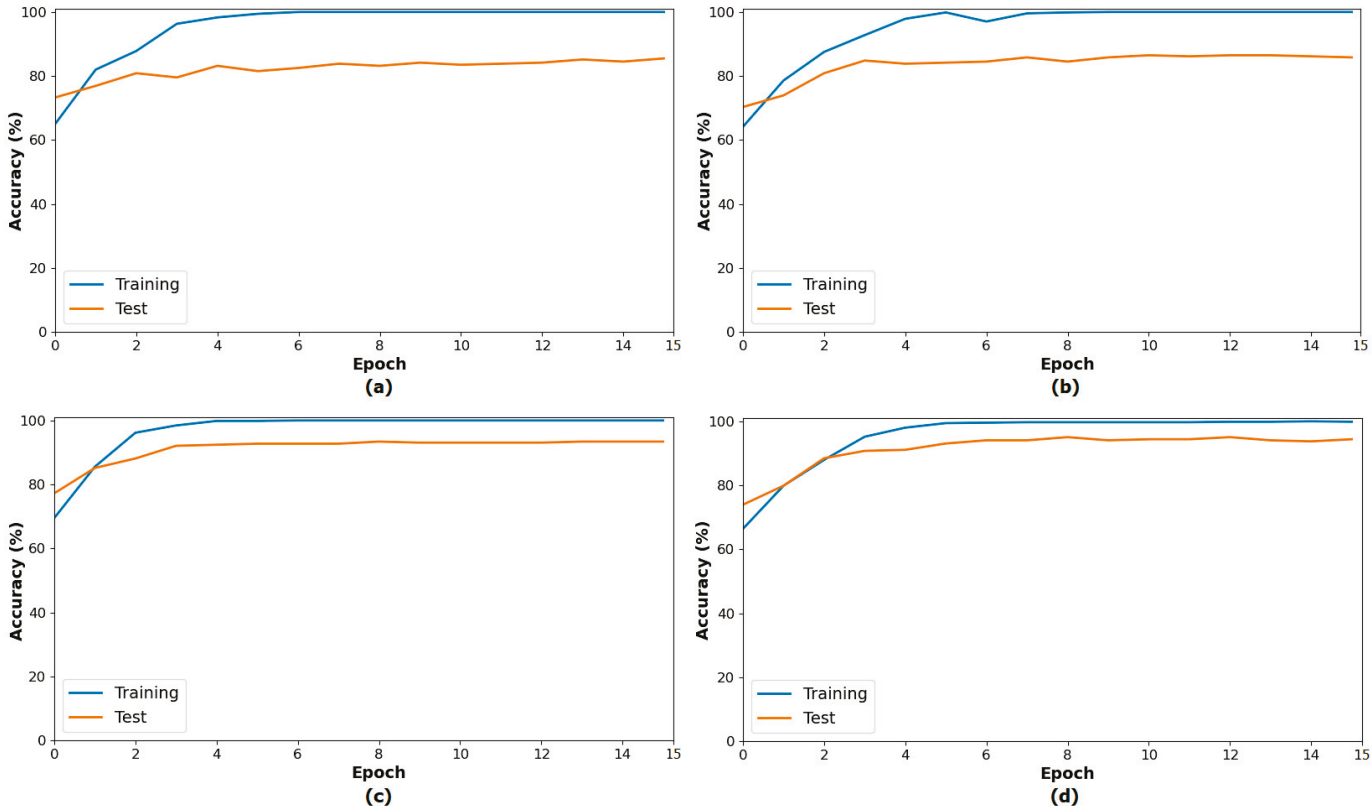


Figure 6. Performance graphs of the CaiT (**a**) and PVT-v2 (**b**) models trained on the original dataset, and the CaiT (**c**) and PVT-v2 (**d**) models trained on the segmented dataset.

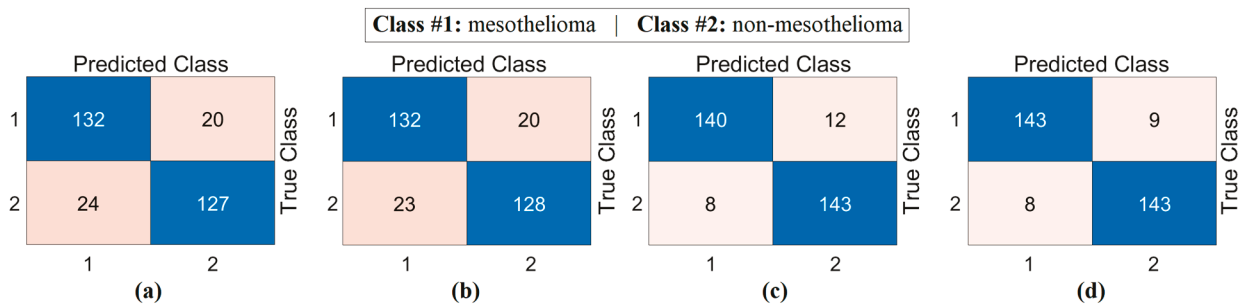


Figure 7. Confusion matrices of the CaiT (**a**) and PVT-v2 (**b**) models trained on the original dataset, and the CaiT (**c**) and PVT-v2 (**d**) models trained on the segmented dataset.

In the final stage of the second step, class-based feature sets were extracted from models trained with segmented datasets (logits layer). The aim at this stage was to extract feature columns equal to the number of classes for each sample, rather than extracting multi-column features from the models. F1 and F2 feature columns were extracted for the CaiT model, and F3 and F4 feature columns were extracted for the PVT-v2 model.

In the third step of the experimental analysis, deep generative techniques (Decoder, GAN, NeRV) were used. Two features representing a sample from the class-based feature sets (F1, F2, F3, F4) obtained from the previous step were converted into images based on Decoder, GAN, and NeRV, respectively. A total of 1008 samples were available, and 1008 decoder, 1008 GAN, and 1008 NeRV-based image sets were obtained for the CaiT model. Similarly, 1008 decoder, 1008 GAN, and 1008 NeRV-based image sets were obtained for the PVT-v2 model. The model-based sample image set is shown in Figure 8.

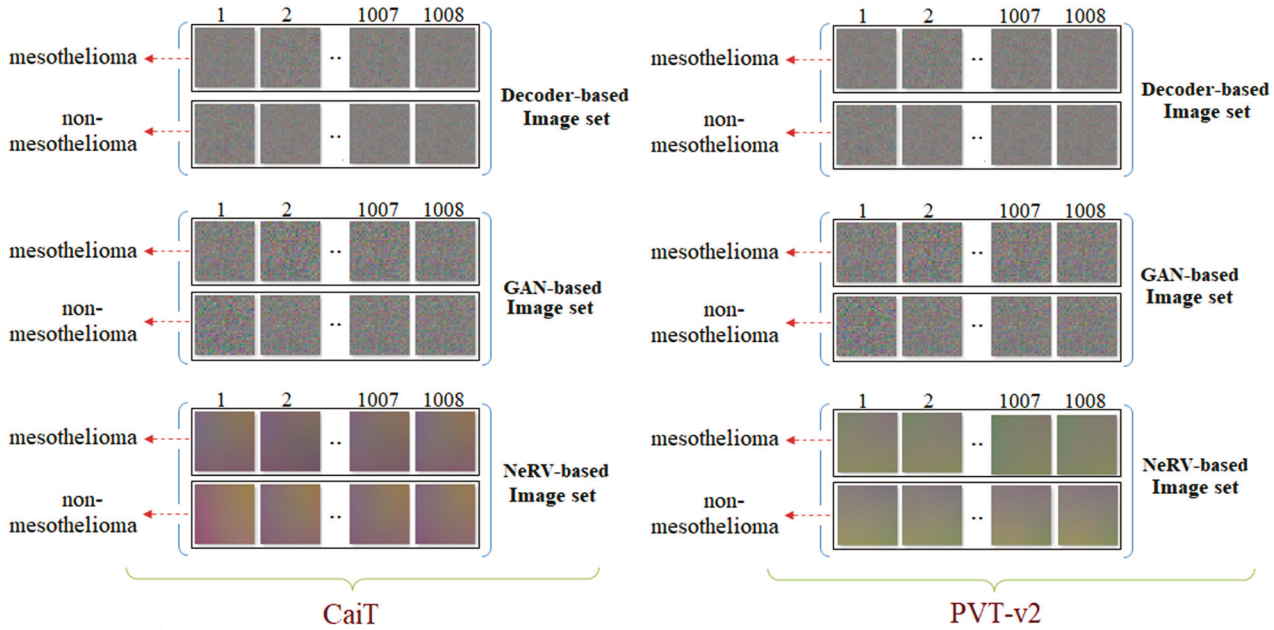


Figure 8. Example images generated from class-based features using deep generative techniques (Decoder, GAN, and NeRV) for the CaiT and PVT-v2 models.

Figure 8 shows sample images generated from class-based feature sets obtained from the CaiT and PVT-v2 models. These images provide an intermediate visualization of how latent features can be projected onto a 2D image space using deep generative techniques (Decoder, GAN, NeRV). Although these images are not directly used for human interpretation, they serve as a standardized input for the subsequent SVM classification stage. The figure demonstrates that generative methods can preserve the discriminative information in the original features, as evidenced by the high classification accuracies presented in Table 4. Therefore, Figure 8 supports the contribution of the generative transformation step in enhancing the performance of the hybrid model.

Table 4. Macro average metric results of confusion matrices obtained by classifying image sets generated using deep generative techniques with a residual-based SVM method.

Model	Image Set	Se	Sp	Pre	F-Scr	Acc
CaiT	Generated using Decoder	95.38	95.38	95.41	95.38	95.38
	Generated using GAN	94.05	94.05	94.35	94.05	94.06
	Generated using NeRV	95.37	95.37	95.46	95.38	95.38
PVT-v2	Generated using Decoder	94.37	94.37	94.97	94.37	94.39
	Generated using GAN	95.38	95.38	95.39	95.38	95.38
	Generated using NeRV	95.38	95.38	95.39	95.38	95.38

At this stage, image sets created using deep generative techniques from class-based features obtained with both CaiT and PVT-v2 models were classified using a residual-based SVM method. At this stage, since the resolutions of the obtained images were quite low, the deep learning-based ResNet-18-supported SVM method was preferred instead of transformer models (which can make training difficult due to patch separation). The purpose of this process is to determine which imaging technique performed best and to evaluate whether deep generative techniques contributed to classification success. The confusion matrices obtained as a result of the analyses performed at this stage are presented in Figure 9, and the macro average metric results for the relevant matrices are presented in

Table 4. Table 4 shows that the results obtained are close to each other and demonstrate high performance. The best results based on CaiT were obtained from both Decoder-based and NeRV-based image sets, with a general accuracy rate of 95.38%. Similarly, the best results based on PVT-v2 were also obtained from both GAN-based and NeRV-based image sets, with a general accuracy rate of 95.38%. In addition, higher accuracy rates were achieved in this stage through analyses performed after the segmentation process. The analyses revealed that the Decoder, GAN, and NeRV techniques contributed positively to the classification process overall.

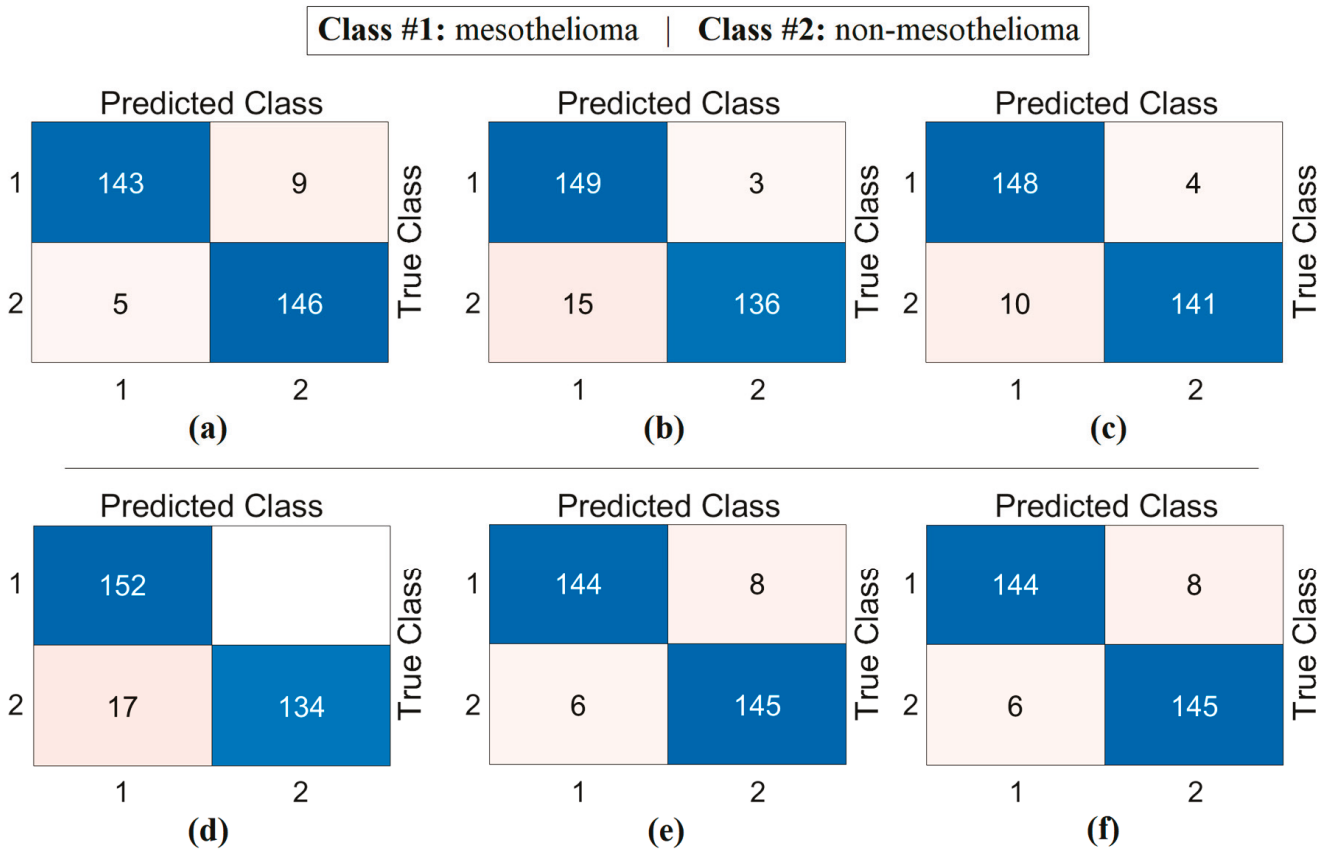


Figure 9. Confusion matrices obtained by classifying images generated from class-based features using deep generative techniques with a residual-based SVM: (**a**) CaiT-Decoder, (**b**) CaiT-GAN, (**c**) CaiT-NeRV, (**d**) PVT-v2-Decoder, (**e**) PVT-v2-GAN, (**f**) PVT-v2-NeRV.

The workflow abstract of the third step performed in the experimental analysis is shown in Figure 10. The class-based feature of each sample was represented using three different visualization methods (Decoder, GAN, and NeRV). Among these visualization types, the image that best represents the relevant sample was determined using the discriminative score method (second stage of the third step). This process was performed separately for both the CaiT model and the PVT-v2 model. The analysis results obtained are presented in Table 5. After determining the best representative image for each sample using the discriminative score method (see Table 5), the classification process was performed using the residual-based SVM method. In the classification process performed using the SVM method, as in previous analyses, 70% of the dataset was allocated for training and 30% for testing. To validate the obtained results and enhance the generalizability of the proposed hybrid approach, the data was further partitioned using the cross-validation method (k = 5) and reclassified using the SVM method. The confusion matrices obtained at this stage are presented in Figure 11, while the relevant metric results are provided in Table 6. When Table 6 is examined, it can be seen that a general accuracy rate of 99.67% was obtained

in the best representative image set processed using the holdout technique. On the other hand, a general accuracy rate of 99.80% was achieved in the best representative image set processed using the cross-validation technique (k = 5).

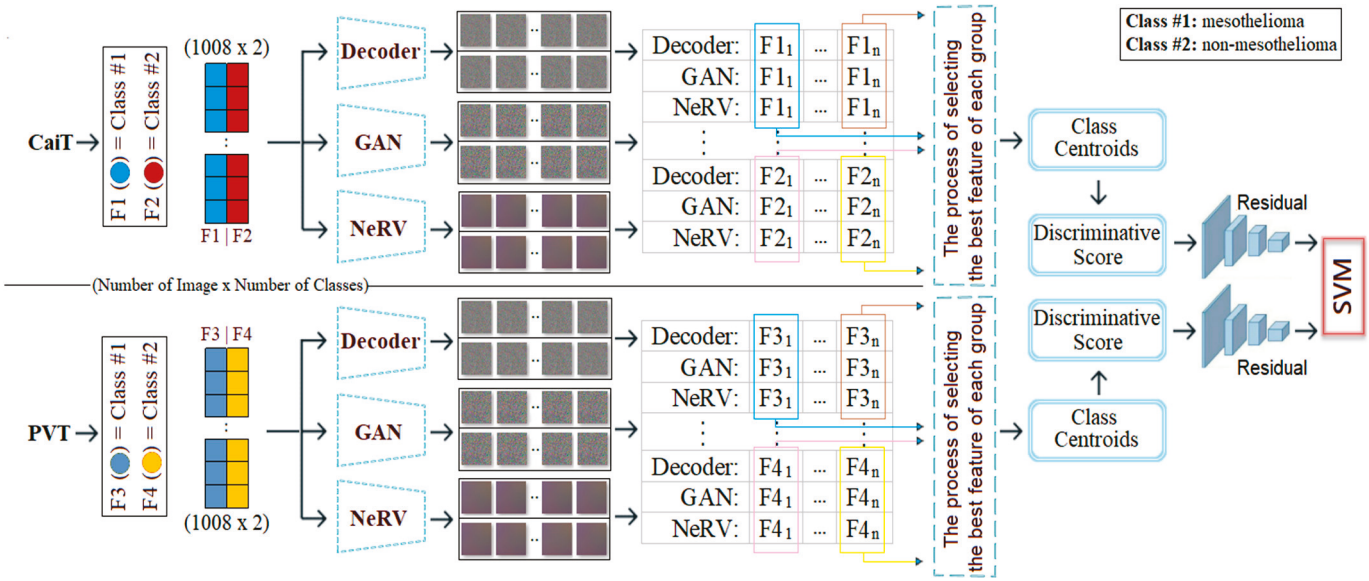


Figure 10. Workflow of the third step: feature visualization, representative image selection, and SVM classification for CaiT and PVT-v2 models.

Table 5. Selection list of images represented by discriminative scores in CaiT and PVT-v2 models.

Selection List of the Image Represented by the Discriminative Score in the CaiT Model.	**Selection List of the Image Represented by the Discriminative Score in the PVT-v2 Model.**
Index,Selected_Method,Label,Image_Path	**Index,Selected_Method,Label,Image_Path**
0,decoder,Diseased,. . .\decoder\Diseased\img_0.png	0,nerv,Diseased,. . .\nerv\Diseased\img_0.png
1,decoder,Diseased,. . .\decoder\Diseased\img_1.png	1,nerv,Diseased,. . .\nerv\Diseased\img_1.png
2,gan,Diseased,. . .\gan\Diseased\img_2.png	2,nerv,Diseased,. . .\nerv\Diseased\img_2.png
3,decoder,Diseased,. . .\decoder\Diseased\img_3.png	3,nerv,Diseased,. . .\nerv\Diseased\img_3.png
4,gan,Diseased,. . .\gan\Diseased\img_4.png	4,nerv,Diseased,. . .\nerv\Diseased\img_4.png
5,decoder,Diseased,. . .\decoder\No disease\img_5.png	5,nerv,Diseased,...\nerv\No disease\img_5.png
6,gan,Diseased,. . .\gan\No disease\img_6.png	6,nerv,Diseased,. . .\nerv\No disease\img_6.png
7,gan,Diseased,. . .\gan\No disease\img_7.png	7,nerv,Diseased,. . .\nerv\No disease\img_7.png
8,decoder,Diseased,. . .\decoder\Diseased\img_8.png	8,nerv,Diseased,. . .\nerv\Diseased\img_8.png
9,decoder,Diseased,. . .\decoder\No disease\img_9.png	9,nerv,Diseased,. . .\nerv\No disease\img_9.png
10,decoder,Diseased,. . .\decoder\No disease\img_10.png	10,nerv,Diseased,. . .\nerv\No disease\img_10.png
11,decoder,Diseased,. . .\decoder\Diseased\img_11.png	11,nerv,Diseased,. . .\nerv\Diseased\img_11.png
12,decoder,Diseased,. . .\decoder\No disease\img_12.png	12,nerv,Diseased,. . .\nerv\No disease\img_12.png
13,decoder,Diseased,. . .\decoder\No disease\img_13.png	13,nerv,Diseased,. . .\nerv\No disease\img_13.png
14,decoder,Diseased,. . .\decoder\No disease\img_14.png	14,nerv,Diseased,. . .\nerv\No disease\img_14.png
15,decoder,Diseased,. . .\decoder\Diseased\img_15.png	15,nerv,Diseased,. . .\nerv\Diseased\img_15.png
16,decoder,Diseased,. . .\decoder\Diseased\img_16.png	16,nerv,Diseased,. . .\nerv\Diseased\img_16.png
17,decoder,Diseased,. . .\decoder\No disease\img_17.png	17,nerv,Diseased,. . .\nerv\No disease\img_17.png
18,decoder,Diseased,. . .\decoder\No disease\img_18.png	18,nerv,Diseased,. . .\nerv\No disease\img_18.png
19,decoder,Diseased,. . .\decoder\No disease\img_19.png	19,nerv,Diseased,. . .\nerv\No disease\img_19.png
20,decoder,Diseased,. . .\decoder\No disease\img_20.png	20,nerv,Diseased,. . .\nerv\No disease\img_20.png
21,decoder,Diseased,. . .\decoder\No disease\img_21.png	21,nerv,Diseased,. . .\nerv\No disease\img_21.png
22,decoder,Diseased,. . .\decoder\Diseased\img_22.png	22,nerv,Diseased,. . .\nerv\Diseased\img_22.png
23,decoder,Diseased,. . .\decoder\No disease\img_23.png	23,nerv,Diseased,. . .\nerv\No disease\img_23.png
24,decoder,Diseased,. . .\decoder\No disease\img_24.png	24,nerv,Diseased,. . .\nerv\No disease\img_24.png
25,decoder,Diseased,. . .\decoder\Diseased\img_25.png	25,nerv,Diseased,. . .\nerv\Diseased\img_25.png
. . .	. . .

Table 5. *Cont.*

Selection List of the Image Represented by the Discriminative Score in the CaiT Model.	Selection List of the Image Represented by the Discriminative Score in the PVT-v2 Model.
Index,Selected_Method,Label,Image_Path	**Index,Selected_Method,Label,Image_Path**
993,nerv,No disease,. . .\nerv\No disease\img_993.png	993,gan,No disease,. . .\gan\No disease\img_993.png
994,nerv,No disease,. . .\nerv\Diseased\img_994.png	994,gan,No disease,. . .\gan\Diseased\img_994.png
995,nerv,No disease,. . .\nerv\No disease\img_995.png	995,gan,No disease,. . .\gan\No disease\img_995.png
996,nerv,No disease,. . .\nerv\Diseased\img_996.png	996,gan,No disease,. . .\gan\Diseased\img_996.png
997,nerv,No disease,. . .\nerv\No disease\img_997.png	997,gan,No disease,. . .\gan\No disease\img_997.png
998,nerv,No disease,. . .\nerv\No disease\img_998.png	998,gan,No disease,. . .\gan\No disease\img_998.png
999,nerv,No disease,. . .\nerv\No disease\img_999.png	999,gan,No disease,. . .\gan\No disease\img_999.png
1000,nerv,No disease,. . .\nerv\Diseased\img_1000.png	1000,gan,No disease,...\gan\Diseased\img_1000.png
1001,nerv,No disease,. . .\nerv\Diseased\img_1001.png	1001,gan,No disease,. . .\gan\Diseased\img_1001.png
1002,nerv,No disease,. . .\nerv\Diseased\img_1002.png	1002,gan,No disease,. . .\gan\Diseased\img_1002.png
1003,nerv,No disease,. . .\nerv\No disease\img_1003.png	1003,gan,No disease,. . .\gan\No disease\img_1003.png
1004,nerv,No disease,. . .\nerv\No disease\img_1004.png	1004,gan,No disease,. . .\gan\No disease\img_1004.png
1005,nerv,No disease,. . .\nerv\Diseased\img_1005.png	1005,gan,No disease,. . .\gan\Diseased\img_1005.png
1006,nerv,No disease,. . .\nerv\Diseased\img_1006.png	1006,gan,No disease,. . .\gan\Diseased\img_1006.png
1007,nerv,No disease,. . .\nerv\Diseased\img_1007.png	1007,gan,No disease,. . .\gan\Diseased\img_1007.png

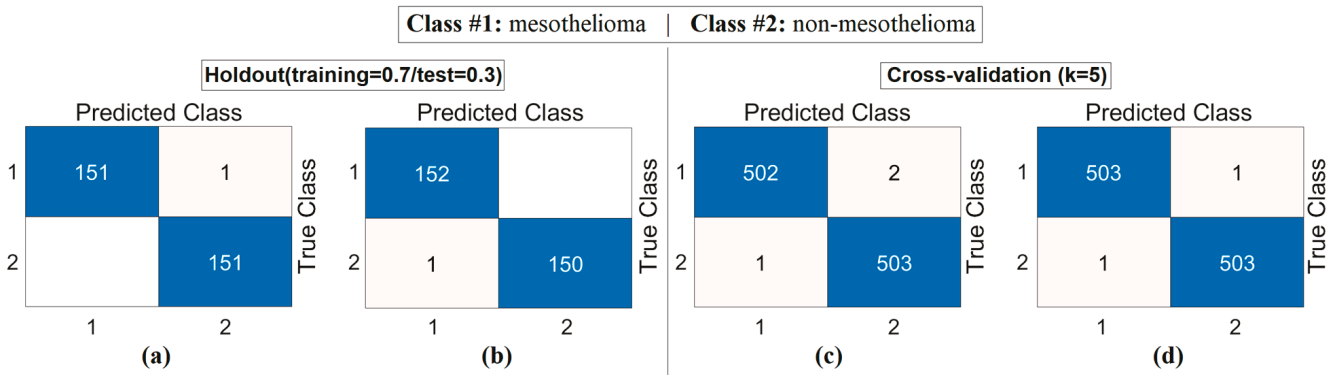


Figure 11. Confusion matrices obtained by classifying images selected from generative image sets using the discriminative score method with the SVM method: (**a**) CaiT-based (holdout), (**b**) PVT-v2-based (holdout), (**c**) CaiT-based (cross-validation), (**d**) PVT-v2-based (cross-validation).

Table 6. Macro average metric results of confusion matrices obtained by classifying images selected from productive image sets using the discriminative score method with the SVM method (%).

Method	Dataset Technique	Which Model Image Set Was Used?	Se	Sp	Pre	F-Scr	Acc
Discriminative Score	Holdout (training: 0.7/test: 0.3)	CaiT-based	99.67	99.67	99.67	99.67	99.67
		PVT-v2-based	99.67	99.67	99.67	99.67	99.67
	Cross-validation (k = 5)	CaiT-based	99.70	99.70	99.70	99.70	99.70
		PVT-v2-based	99.80	99.80	99.80	99.80	99.80

The analyses conducted in the final stage revealed that the discriminative score method works as an effective selection mechanism and thus improves overall performance. The findings confirm the validity of the proposed hybrid model and demonstrate that the model shows high success in the diagnosis of mesothelioma.

The performance of the proposed approach was comprehensively evaluated across multiple model and dataset configurations. As reported in Table 7 and Figures 12–16, the models consistently achieved excellent results, with AUC values ranging from 0.997 to 0.998, accuracy between 0.997 and 0.998, and both sensitivity and precision exceeding 0.994

across all scenarios. Specificity values were similarly high, ranging from 0.993 to 1.000. Bootstrap-based 95% confidence intervals were narrow for all metrics, further supporting the statistical reliability and robustness of the method.

Table 7. Performance metrics of the proposed approach for different model and dataset configurations, including 95% confidence intervals (CIs).

Model/Dataset Configuration	AUC (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)
CaiT-based (holdout)	0.998 (0.995–1.000)	0.997 (0.990–1.000)	1.000 (1.000–1.000)	0.993 (0.979–1.000)	0.993 (0.979–1.000)
PVT-v2-based (holdout)	0.997 (0.991–1.000)	0.997 (0.990–1.000)	0.994 (0.978–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
CaiT-based (cross-validation)	0.997 (0.993–1.000)	0.997 (0.993–1.000)	0.998 (0.994–1.000)	0.996 (0.990–1.000)	0.996 (0.990–1.000)
PVT-v2-based (cross-validation)	0.998 (0.995–1.000)	0.998 (0.995–1.000)	0.998 (0.994–1.000)	0.998 (0.994–1.000)	0.998 (0.994–1.000)

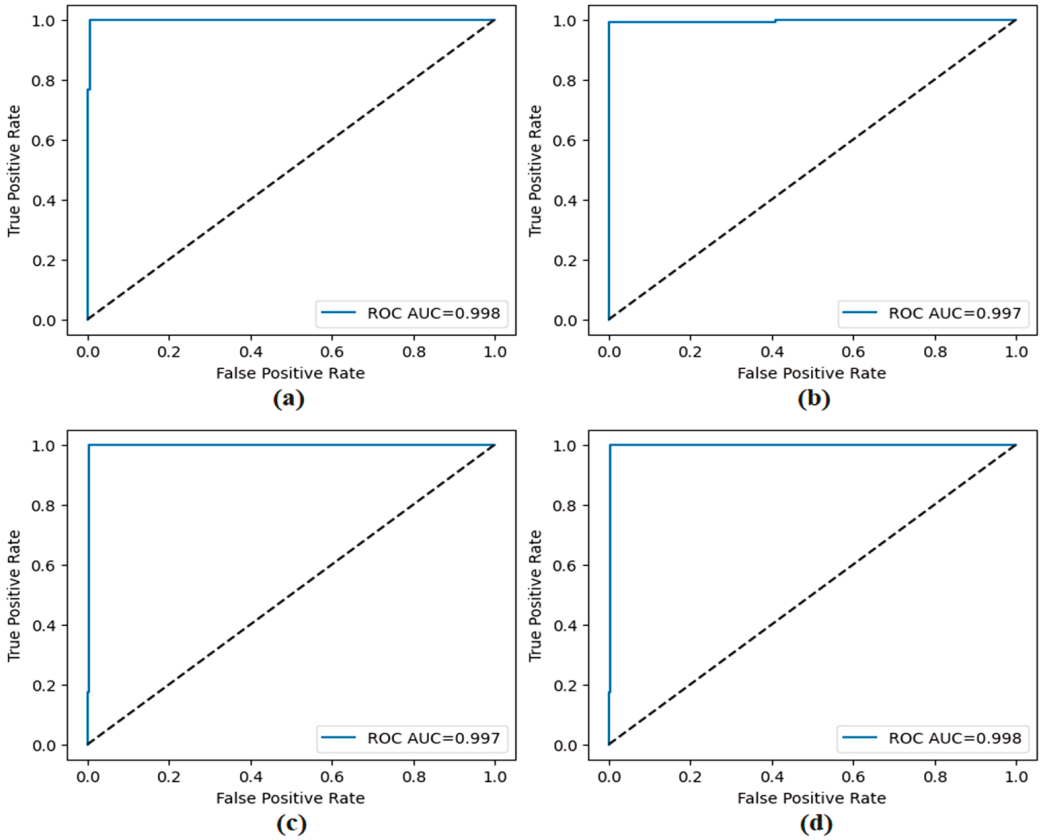


Figure 12. ROC curves obtained by the proposed approach; (**a**) CaiT-based (holdout), (**b**) PVT-v2-based (holdout), (**c**) CaiT-based (cross-validation), (**d**) PVT-v2-based (cross-validation).

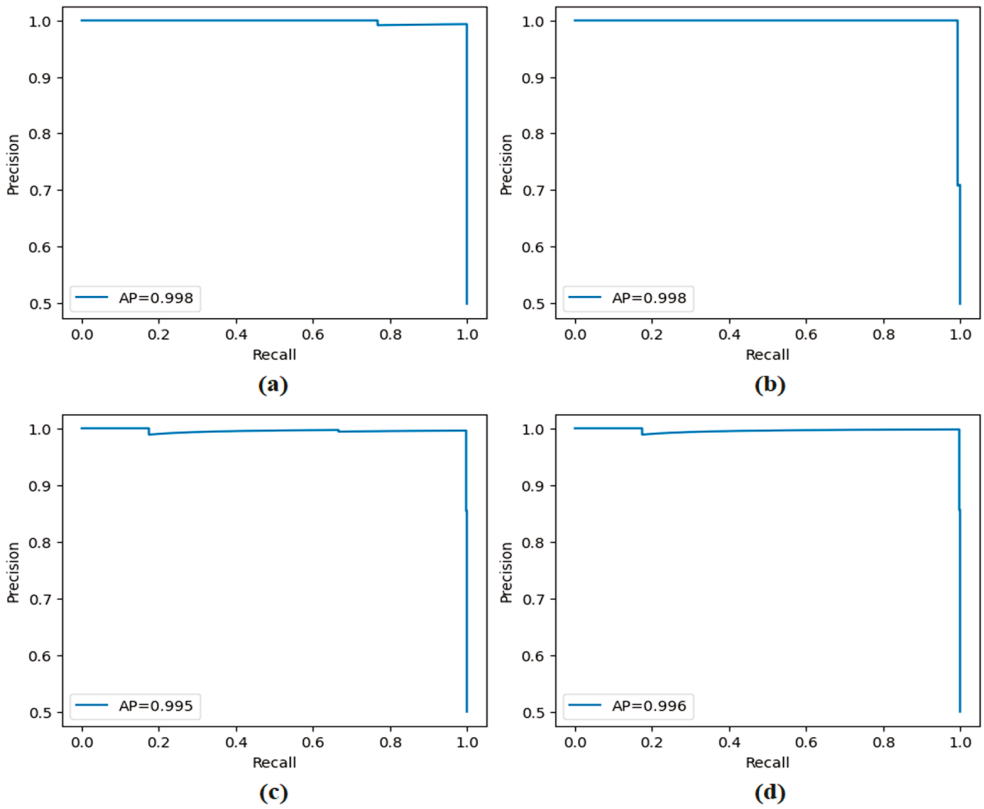


Figure 13. Precision–recall curves obtained by the proposed approach; (**a**) CaiT-based (holdout), (**b**) PVT-v2-based (holdout), (**c**) CaiT-based (cross-validation), (**d**) PVT-v2-based (cross-validation).

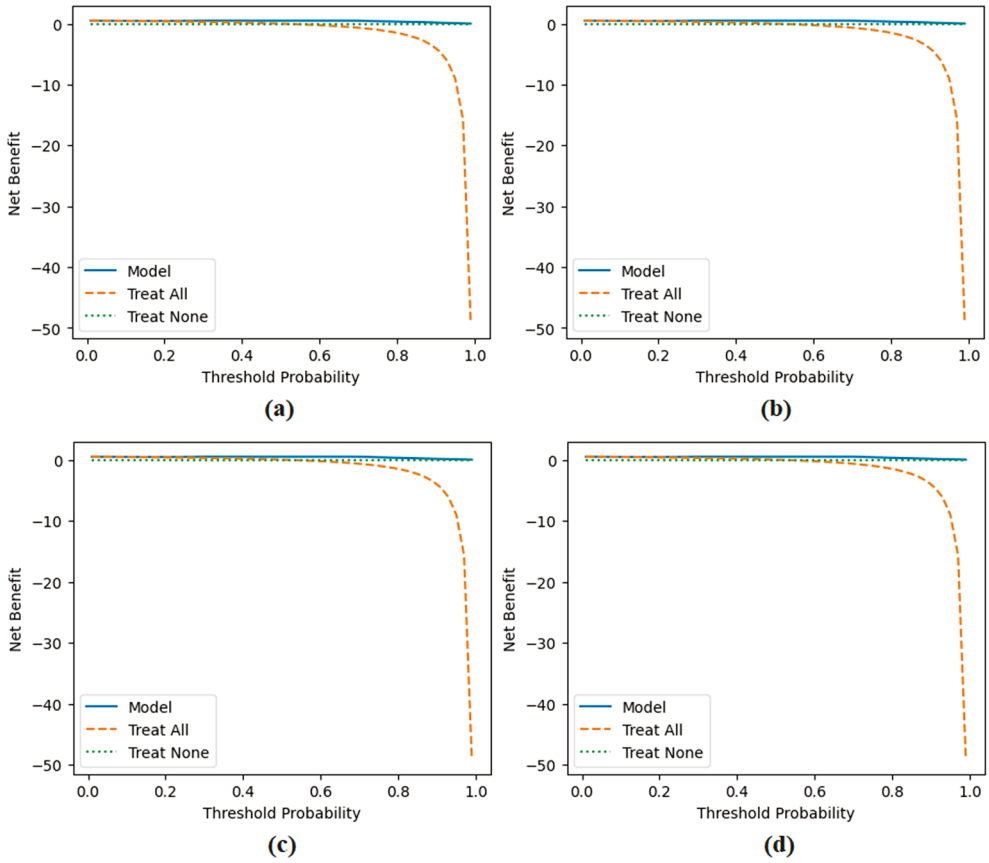


Figure 14. Decision curve analysis results obtained by the proposed approach; (**a**) CaiT-based (holdout), (**b**) PVT-v2-based (holdout), (**c**) CaiT-based (cross-validation), (**d**) PVT-v2-based (cross-validation).

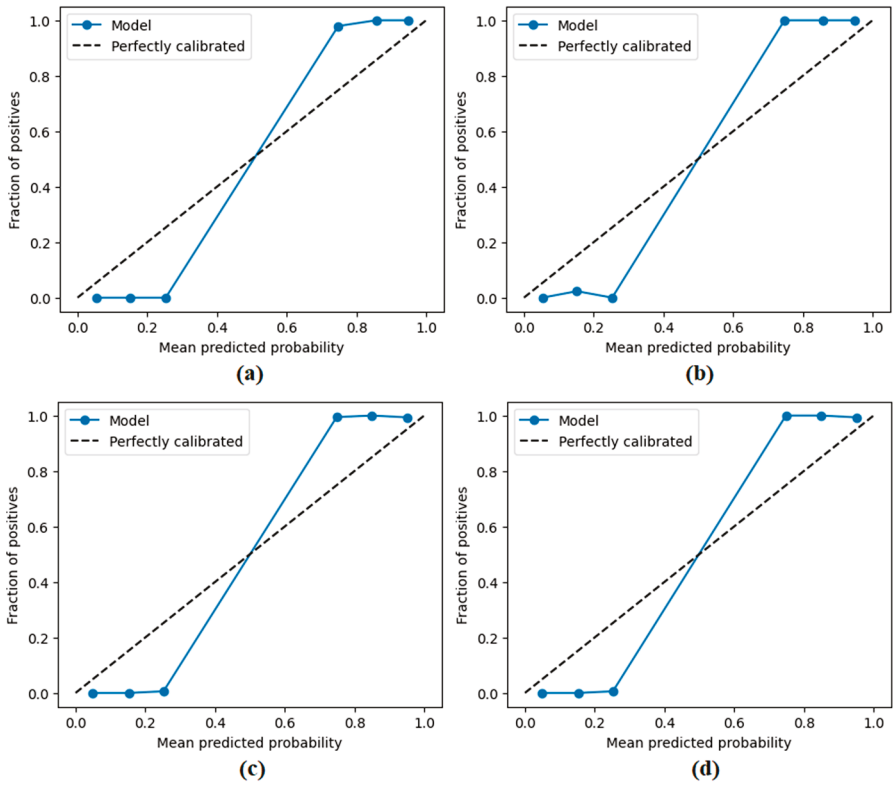


Figure 15. Calibration plots (reliability diagrams) obtained by the proposed approach; (**a**) CaiT-based (holdout), (**b**) PVT-v2-based (holdout), (**c**) CaiT-based (cross-validation), (**d**) PVT-v2-based (cross-validation).

Figure 11 presents the overall confusion matrices, while Figure 16 provides fold-wise confusion matrices, demonstrating stability and consistency across cross-validation folds. Complementary evaluation curves—including ROC (Figure 12), precision–recall (Figure 13), decision curve analysis (Figure 14), and calibration plots (Figure 15)—confirm the high discriminative ability, reliable probability estimation, and practical utility of the proposed models. To elaborate further, precision–recall curves provide a significant advantage in accurately identifying the positive class (mesothelioma) by simultaneously maintaining high sensitivity and high precision. Calibration curves demonstrate that the probability values predicted by the model are consistent with actual observations and provide reliable decision support. Decision curve analysis shows that the model provides higher net benefit compared to the "Treat All" and "Treat None" strategies across all threshold values, thus supporting clinical benefit. Both CaiT-based and PVT-v2-based models, under holdout and cross-validation schemes, maintained near-perfect performance, indicating strong generalizability and reproducibility of the proposed approach.

To examine the contribution of each component of the proposed approach in greater detail, ablation studies were conducted on the PVT-v2-based architecture that yielded the best results. Figure 17 presents the confusion matrices obtained in scenarios where different components were removed, while Table 8 summarizes the accuracy rates for these scenarios. Removing the generative rendering step and directly feeding the logit values into the SVM resulted in a significant drop in accuracy rates (93.73% holdout and 95.05% cross-validation), highlighting the importance of generative representation learning in extracting discriminative features. Replacing DS-based feature selection with a simpler variance thresholding method yielded higher accuracy compared to removing the generative rendering step; however, it fell short of the results obtained with all model components (95.71% holdout and 96.13% cross-validation). This supports the success of the DS method

in identifying the most meaningful features. Furthermore, when logistic regression, a linear method, was used instead of the SVM classifier, a decrease in performance was observed (95.05% retention), indicating that a residual-based SVM is critical for robust classification. The obtained data generally confirms that the combination of generative rendering, discriminative feature selection, and a residual SVM enables the proposed hybrid structure to achieve high performance, reaching 99.80% in cross-validation.

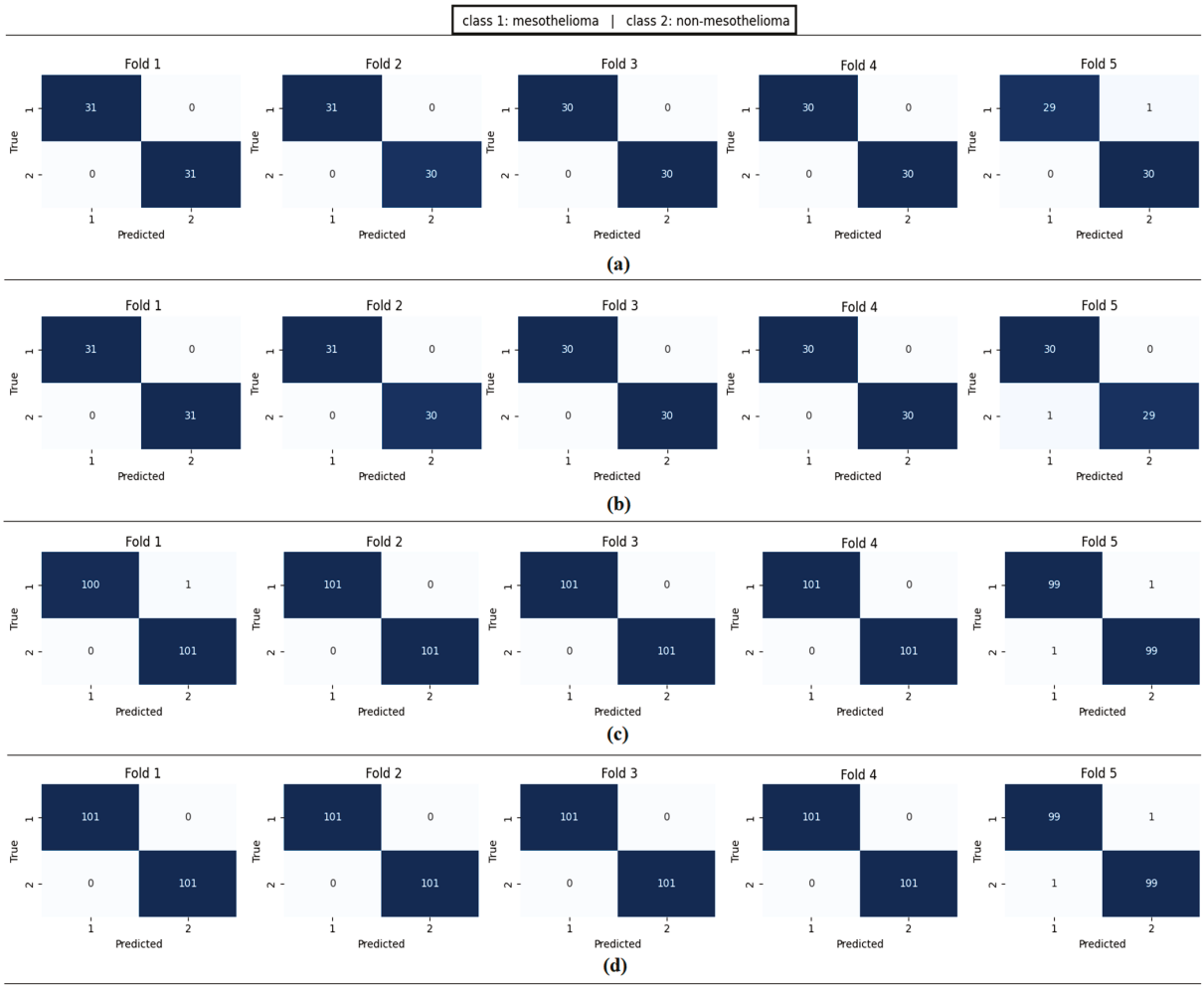


Figure 16. Confusion matrices per fold obtained by the proposed approach; (**a**) CaiT-based (holdout), (**b**) PVT-v2-based (holdout), (**c**) CaiT-based (cross-validation), (**d**) PVT-v2-based (cross-validation).

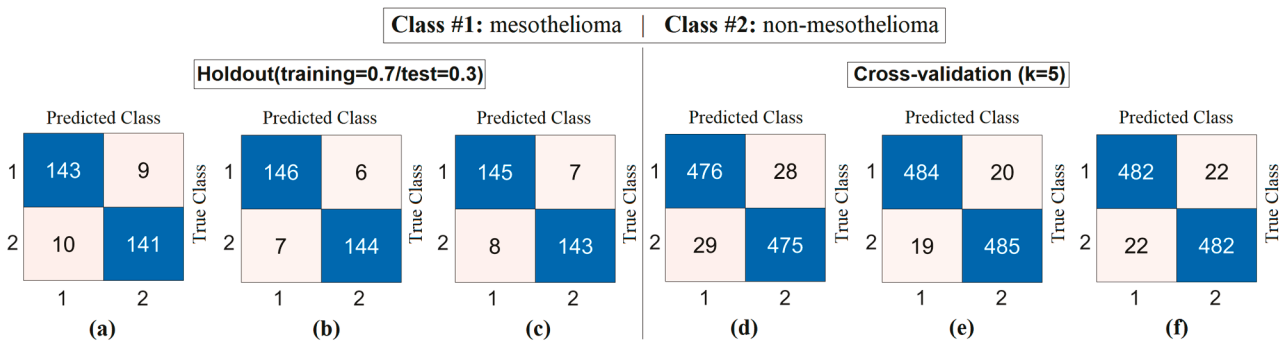


Figure 17. Confusion matrices from ablation analyses on the PVT-v2 model: (**a**) without generative rendering (holdout), (**b**) simpler feature selector (holdout), (**c**) replacing the SVM with a linear classifier (holdout), (**d**) without generative rendering (cross-validation), (**e**) simpler feature selector (cross-validation), (**f**) replacing the SVM with a linear classifier (cross-validation).

Table 8. Ablation analyses on different components of the proposed hybrid approach performed on the PVT-v2 backbone. (The PVT-v2 model was chosen as it achieved the highest performance in the proposed approach.)

Experiment Setup	Dataset Technique	Acc (%)
Proposed Hybrid Model (SAM + generative rendering + DS + SVM)	Holdout (CaiT-based)	99.67
	Holdout (PVT-v2-based)	99.67
	Cross-val. (CaiT-based)	99.70
	Cross-val. (PVT-v2-based)	99.80
Without Generative Rendering (direct logits → SVM)	Holdout	93.73
	Cross-val.	95.05
Simpler Feature Selector (variance threshold instead of DS)	Holdout	95.71
	Cross-val.	96.13
Replacing SVM with Linear Classifier (logistic regression)	Holdout	95.05
	Cross-val.	95.63

4. Discussion

The experimental results demonstrate that the hybrid AI model developed in this study offers an effective approach to overcoming the current challenges in mesothelioma diagnosis. The system's subcomponents were examined individually to evaluate their impact on diagnostic performance. The results of the study highlight the superiority of AI-based modern techniques in addressing the shortcomings of traditional diagnostic methods. Processes such as image segmentation, deep learning-based feature extraction, and data generation using generative techniques have emerged as critical factors significantly enhancing the model's diagnostic success. The balanced nature of the dataset ensured that the model's learning process was conducted in a more robust manner. In the analysis conducted on the proposed hybrid model, a cross-validation technique was applied to the dataset, resulting in an overall accuracy of 99.80%. Similarly, when the holdout technique was applied, an overall accuracy of 99.67% was achieved. The contribution of each stage to the proposed hybrid model was verified by experimental analysis. The contributions and limitations of the proposed approach are as follows:

- Meaningful regions were highlighted in CT images using the SAM method; deep learning models were enabled to focus on these regions and leave meaningless regions in the background.
- Instead of extracting a large number of feature columns from FC layers using recently popular transformer architectures (CaiT, PVT-v2), a new final layer was added that produces feature columns as efficient as the number of classes. This eliminates unnecessary feature columns from the hybrid model.
- Numerical values in feature columns are represented as 2D images using generative techniques such as Decoder, GAN, and NeRV. This method allows approaches that operate on images to extract more diverse and rich numerical content from their architectural structures.
- Alternative image sets were created using generative techniques (Decoder, GAN, and NeRV). The best set was selected using a discriminative scoring method to prevent the model from considering unnecessary images.
- Maintenance and update difficulty: The model's structure includes many components, which may make it difficult to maintain and develop in the long term.

- The complexity of the decision-making process: Hybrid systems consisting of multiple layers and components can make interpreting the model's outputs difficult. Such complex structures may also result in additional time costs during the training and inference processes.
- Computational resource requirements: Integrating different architectures, such as the SAM, transformer, GAN, Decoder, and NeRV, may require high processing power and memory in real-time applications.
- The proposed hybrid model is limited by the use of a single-center CT dataset that is not publicly accessible due to ethical constraints. The absence of external validation on independent datasets may limit the generalizability of the findings. Future studies should aim to evaluate the performance of the model on multicenter or publicly available datasets of mesothelioma or other related diseases to validate its broader applicability.

In terms of computational efficiency, applying the most computationally intensive process, SAM-based segmentation, only once during preprocessing provides a significant advantage. The subsequent model training and inference processes were completed within reasonable timeframes using standard GPU hardware. The classification stage provided response times fast enough for clinical applications. Training transformer-based models took approximately two and a half hours in total. Despite their complex hybrid architecture, these times demonstrate the feasibility of the system with commonly available GPU resources. Further optimization for real-time use is planned for future work.

Table 9 summarizes similar studies that recently conducted analyses using CT-based datasets for mesothelioma diagnosis. The datasets used in these studies were obtained with ethical approval and are not publicly available. Studies on CT-based mesothelioma disease diagnosis are limited in the literature.

Table 9. Similar studies in the literature using CT images for the detection of mesothelioma.

Paper	Year	Number of Patients	Number of Classes	Model/Method	Acc (%)
Kitajima et al. [9]	2021	875	2	3D-CNN	77.3
Ye Li et al. [34]	2022	397	2	Multivariate logistic regression	95.5
This study	2025	1008	2	SAM and transformers and class-based feature extraction and deep generative and discriminative score	99.80 (cross-val.) 99.67 (holdout)

Kitajima et al. [9] used both PET- and CT-based datasets to diagnose mesothelioma. Although the 3D-CNN model they designed achieved limited success, they could have improved their results by integrating approaches such as preprocessing, feature extraction, and feature selection into their proposed approach. Ye Li et al. [34] developed a diagnostic model to distinguish between MPM and metastatic pleural disease (MPD). In their study, they applied a multivariate logistic regression analysis using 397 CT images and specific clinical features derived from these images, achieving an overall accuracy rate of 95.5%. However, the limited number of CT images used and the lack of model diversity restricted the generalizability of the model's success and limited performance improvements. The proposed hybrid model offers an architectural structure that integrates both advanced AI techniques and new-generation analysis approaches in CT imaging data. Thanks to this integrated design, our model has demonstrated higher accuracy, segmentation support,

and robust classification performance compared to a limited number of similar studies in the literature. The results obtained demonstrate that the proposed model contributes to the clinical decision-making process for mesothelioma.

Although the proposed hybrid model demonstrates high accuracy rates and strong performance, certain limitations associated with the dataset used must be considered. This study was conducted using a single-center CT dataset obtained from Fırat University Research Hospital; in accordance with the ethics committee decision, this dataset cannot be made publicly available. Furthermore, external validation could not be performed due to the lack of an accessible dataset related to CT-based mesothelioma. This situation limits the evaluation of the model's performance on different data sources. Nevertheless, to maintain the transparency of the study, the methods followed, patient group information, and analysis processes are presented in detail. In future research, it is planned to evaluate the model using multicenter datasets or publicly available CT data for similar diseases to test its validity across different patient groups.

Despite incorporating various sophisticated AI modules, the system's analyses were conducted using a conventional GPU setup, with inference times kept within a range suitable for clinical applications. Thanks to its modular design, the system allows for future enhancements through the use of lightweight models and compression strategies, potentially lowering computational requirements and supporting real-time use in hospital environments.

Also, the hybrid encoder–decoder architecture we propose, which utilizes class-based feature sets and deep generative techniques, offers a promising approach for diagnosing COVID-19 variants from lung CT images. The study conducted by Fekri-Ershad and Dehkordi proposes a multi-channel deep network that combines tissue and spatial information to detect new COVID-19 variants. In this approach, tissue-based and spatial data are processed in separate channels, which increases diagnostic accuracy [35]. Integrating similar multi-channel architectures into our current framework could further enrich feature representations and improve classification performance. Future work could aim to capture fine details in CT images more effectively by integrating advanced tissue analysis methods, such as local binary patterns (LBPs), into our encoder–decoder architecture. Furthermore, adapting our system to process multiple imaging modalities such as CT and X-ray could make the model more robust and generalizable across different modalities. Such improvements could contribute to the development of more effective diagnostic tools for the early and reliable diagnosis of COVID-19 variants.

5. Conclusions

This study emphasizes the current challenges in diagnosing and treating mesothelioma, and shows that a multidisciplinary approach can contribute to an earlier diagnosis. Traditional diagnostic methods have limitations that make it difficult to detect the disease in its early stages, which can delay effective treatment. The proposed hybrid model addresses this issue by integrating new-generation, AI-based approaches, offering a more effective solution. The model aims to enable more accurate and precise mesothelioma diagnoses, particularly through the use of CT images. The hybrid model consists of the following stages: meaningful CT segmentation; training with transformer-based models; class-based feature extraction from models; conversion of each feature into an image using deep generative techniques (Decoder, GAN, and NeRV); and determination of the image that best represents each data sample using a discriminative score method. As a result of these steps, the hybrid model recommended for mesothelioma diagnosis achieved an overall accuracy rate of 99.80%.

The model proposed in this paper provides an innovative contribution to the literature by hybridizing approaches such as meaningful image segmentation, feature extraction using class-based transformer models, and representative image generation using generative AI techniques. The results obtained with the model's high accuracy rates are at a level that can provide significant support to physicians in clinical decision-making processes. The model's detailed analysis capability is particularly noteworthy in image-based diagnosis processes. The model's prominent features are as follows:

- Increased interpretability through visualization of numerical data.
- Production of stable and transparent results thanks to a simplified feature selection mechanism.
- Flexible architecture that can adapt to different clinical scenarios and treatment protocols.
- Elimination of negativity in the decision-making processes of expert physicians; objective decision-making mechanism.
- Prevention of confusion with different disease types in the early diagnosis process.
- These features demonstrate that the model has a broad application potential not only in mesothelioma diagnosis but also in other medical imaging fields.

In future studies, the hybrid model developed in this study will be tested on different types of cancer and various medical imaging modalities to better evaluate its generalization ability. In addition, improving real-time performance and exploring lightweight architectures or model compression techniques will help reduce computational cost and facilitate practical deployment in clinical environments. Another important step will be integrating the model with a user-friendly interface to enhance accessibility for healthcare professionals. Furthermore, the use of explainable artificial intelligence (XAI) techniques will be prioritized to improve interpretability, enabling clinicians and patients to better understand the model's outputs. Such developments are expected to significantly enhance the effectiveness, efficiency, and reliability of the model in real-world clinical applications. In addition to mesothelioma, the modular design of the proposed hybrid model suggests its potential applicability to other rare cancers that similarly suffer from limited datasets and challenging imaging characteristics. Extending the model to other cancers will be an important direction in our future work.

Author Contributions: Conceptualization, S.A. and M.A.; methodology, S.A., M.A. and M.T.; software, M.T.; validation, S.A., M.Ç. and M.T.; investigation, M.Ç.; data curation, S.A., M.A. and M.Ç.; writing—original draft preparation, S.A., M.A. and M.T.; writing—review and editing, M.A., S.A., M.Ç. and M.T.; visualization, S.A. and M.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: The study was approved by the Scientific Research and Publication Ethics Board of Fırat University (Meeting No: 2025/09, Decision No: 26, Approval Date: 3 July 2025).

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the 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.

References

1. Popat, S.; Baas, P.; Faivre-Finn, C.; Girard, N.; Nicholson, A.G.; Nowak, A.K.; Opitz, I.; Scherpereel, A.; Reck, M. Malignant Pleural Mesothelioma: ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-Up. *Ann. Oncol.* **2022**, *33*, 129–142. [CrossRef] [PubMed]

2. Odgerel, C.-O.; Takahashi, K.; Sorahan, T.; Driscoll, T.; Fitzmaurice, C.; Yoko-o, M.; Sawanyawisuth, K.; Furuya, S.; Tanaka, F.; Horie, S.; et al. Estimation of the Global Burden of Mesothelioma Deaths from Incomplete National Mortality Data. *Occup. Environ. Med.* **2017**, *74*, 851–858. [CrossRef] [PubMed]
3. Metintas, M.; Hillerdal, G.; Metintas, S. Malignant Mesothelioma Due to Environmental Exposure to Erionite: Follow-Up of a Turkish Emigrant Cohort. *Eur. Respir. J.* **1999**, *13*, 523–526. [CrossRef]
4. Ahmadzada, T.; Reid, G.; Kao, S. Biomarkers in Malignant Pleural Mesothelioma: Current Status and Future Directions. *J. Thorac. Dis.* **2018**, *10*, S1003–S1007. [CrossRef] [PubMed]
5. Courtiol, P.; Maussion, C.; Moarii, M.; Pronier, E.; Pilcer, S.; Sefta, M.; Manceron, P.; Toldo, S.; Zaslavskiy, M.; Le Stang, N.; et al. Deep Learning-Based Classification of Mesothelioma Improves Prediction of Patient Outcome. *Nat. Med.* **2019**, *25*, 1519–1525. [CrossRef]
6. Baas, P.; Scherpereel, A.; Nowak, A.K.; Fujimoto, N.; Peters, S.; Tsao, A.S.; Mansfield, A.S.; Popat, S.; Jahan, T.; Antonia, S.; et al. First-Line Nivolumab Plus Ipilimumab in Unresectable Malignant Pleural Mesothelioma (CheckMate 743): A Multicentre, Randomised, Open-Label, Phase 3 Trial. *Lancet* **2021**, *397*, 375–386. [CrossRef]
7. Szlosarek, P.W.; Creelan, B.C.; Sarkodie, T.; Nolan, L.; Taylor, P.; Olevsky, O.; Grosso, F.; Cortinovis, D.; Chitnis, M.; Roy, A.; et al. Pegargiminase Plus First-Line Chemotherapy in Patients with Nonepithelioid Pleural Mesothelioma. *JAMA Oncol.* **2024**, *10*, 475. [CrossRef]
8. Luu, V.P.; Fiorini, M.; Combes, S.; Quemeneur, E.; Bonneville, M.; Bousquet, P.J. Challenges of Artificial Intelligence in Precision Oncology: Public-Private Partnerships Including National Health Agencies as an Asset to Make It Happen. *Ann. Oncol.* **2024**, *35*, 154–158. [CrossRef]
9. Kitajima, K.; Matsuo, H.; Kono, A.; Kuribayashi, K.; Kijima, T.; Hashimoto, M.; Hasegawa, S.; Murakami, T.; Yamakado, K. Deep Learning with Deep Convolutional Neural Network Using FDG-PET/CT for Malignant Pleural Mesothelioma Diagnosis. *Oncotarget* **2021**, *12*, 1187–1196. [CrossRef]
10. Gill, T.S.; Shirazi, M.A.; Zaidi, S.S.H. Early Detection of Mesothelioma Using Machine Learning Algorithms. In Proceedings of the IEEC 2023, Karachi, Pakistan, 20 September 2023; MDPI: Basel, Switzerland, 2023; p. 6.
11. Cheng, T.S.Y.; Liao, X. Binary Classification of Malignant Mesothelioma: A Comparative Study. *J. Data Sci.* **2023**, *21*, 205–224. [CrossRef]
12. Kidd, A.C.; Anderson, O.; Cowell, G.W.; Weir, A.J.; Voisey, J.P.; Evison, M.; Tsim, S.; Goatman, K.A.; Blyth, K.G. Fully Automated Volumetric Measurement of Malignant Pleural Mesothelioma by Deep Learning AI: Validation and Comparison with Modified RECIST Response Criteria. *Thorax* **2022**, *77*, 1251–1259. [CrossRef]
13. Choudhury, A. Predicting Cancer Using Supervised Machine Learning: Mesothelioma. *Technol. Health Care* **2021**, *29*, 45–58. [CrossRef]
14. Mazurowski, M.A.; Dong, H.; Gu, H.; Yang, J.; Konz, N.; Zhang, Y. Segment Anything Model for Medical Image Analysis: An Experimental Study. *Med. Image Anal.* **2023**, *89*, 102918. [CrossRef]
15. Wu, J.; Wang, Z.; Hong, M.; Ji, W.; Fu, H.; Xu, Y.; Xu, M.; Jin, Y. Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation. *Med. Image Anal.* **2025**, *102*, 103547. [CrossRef] [PubMed]
16. Aydın, S.; Ağar, M.; Çakmak, M.; Koç, M.; Toğaçar, M. Detection of Aspergilloma Disease Using Feature-Selection-Based Vision Transformers. *Diagnostics* **2025**, *15*, 26. [CrossRef] [PubMed]
17. Wang, Y.; Deng, Y.; Zheng, Y.; Chattopadhyay, P.; Wang, L. Vision Transformers for Image Classification: A Comparative Survey. *Technologies* **2025**, *13*, 32. [CrossRef]
18. Dinh, M.-T.; Choi, D.-J.; Lee, G.-S. DenseTextPVT: Pyramid Vision Transformer with Deep Multi-Scale Feature Refinement Network for Dense Text Detection. *Sensors* **2023**, *23*, 5889. [CrossRef]
19. Yüksel, N.; Börklü, H.R. A Generative Deep Learning Approach for Improving the Mechanical Performance of Structural Components. *Appl. Sci.* **2024**, *14*, 3564. [CrossRef]
20. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *arxiv* **2020**, arXiv:2003.08934. [CrossRef]
21. Aggarwal, A.; Mittal, M.; Battineni, G. Generative Adversarial Network: An Overview of Theory and Applications. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100004. [CrossRef]
22. Abedi, M.; Hempel, L.; Sadeghi, S.; Kirsten, T. GAN-Based Approaches for Generating Structured Data in the Medical Domain. *Appl. Sci.* **2022**, *12*, 7075. [CrossRef]
23. Bai, Y.; Dong, C.; Wang, C.; Yuan, C. PS-NeRV: Patch-Wise Stylized Neural Representations for Videos. In Proceedings of the 2023 IEEE International Conference on Image Processing (ICIP), Kuala Lumpur, Malaysia, 8 October 2023; IEEE: New York, NY, USA, 2023; pp. 41–45.
24. Ji, J.; Fu, S.; Man, J. I-NeRV: A Single-Network Implicit Neural Representation for Efficient Video Inpainting. *Mathematics* **2025**, *13*, 1188. [CrossRef]
25. Ali, A.; Khan, Z.; Aldahmani, S. Centroid Decision Forest. *arxiv* **2020**, arXiv:2503.19306.

26. Ghosh, T.; Kirby, M. Nonlinear Feature Selection Using Sparsity-Promoted Centroid-Encoder. *Neural Comput. Appl.* **2023**, *35*, 21883–21902. [CrossRef]
27. Gür, Y.E.; Toğaçar, M.; Solak, B. Integration of CNN Models and Machine Learning Methods in Credit Score Classification: 2D Image Transformation and Feature Extraction. *Comput. Econ.* **2025**, *65*, 2991–3035. [CrossRef]
28. Aslan, S. A Deep Learning-Based Sentiment Analysis Approach (MF-CNN-BILSTM) and Topic Modeling of Tweets Related to the Ukraine–Russia Conflict. *Appl. Soft Comput.* **2023**, *143*, 110404. [CrossRef]
29. Başaran, E.; Çelik, Y. Skin Cancer Diagnosis Using CNN Features with Genetic Algorithm and Particle Swarm Optimization Methods. *Trans. Inst. Meas. Control* **2024**, *46*, 2706–2713. [CrossRef]
30. Çalışkan, A. Detecting Human Activity Types from 3D Posture Data Using Deep Learning Models. *Biomed. Signal Process. Control* **2023**, *81*, 104479. [CrossRef]
31. Yildirim, M.; Cengil, E.; Eroglu, Y.; Cinar, A. Detection and Classification of Glioma, Meningioma, Pituitary Tumor, and Normal in Brain Magnetic Resonance Imaging Using Deep Learning-Based Hybrid Model. *Iran J. Comput. Sci.* **2023**, *6*, 455–464. [CrossRef]
32. Şener, A.; Doğan, G.; Ergen, B. A Novel Convolutional Neural Network Model with Hybrid Attentional Atrous Convolution Module for Detecting the Areas Affected by the Flood. *Earth Sci. Inform.* **2024**, *17*, 193–209. [CrossRef]
33. Aktas, A.; Cap, T.; Serbes, G.; Ilhan, H.O.; Uzun, H. Advanced Multi-Level Ensemble Learning Approaches for Comprehensive Sperm Morphology Assessment. *Diagnostics* **2025**, *15*, 1564. [CrossRef] [PubMed]
34. Li, Y.; Cai, B.; Wang, B.; Lv, Y.; He, W.; Xie, X.; Hou, D. Differentiating Malignant Pleural Mesothelioma and Metastatic Pleural Disease Based on a Machine Learning Model with Primary CT Signs: A Multicentre Study. *Heliyon* **2022**, *8*, e11383. [CrossRef] [PubMed]
35. Fekri-Ershad, S.; Dehkordi, K.B. A Flexible Multi-Channel Deep Network Leveraging Texture and Spatial Features for Diagnosing New COVID-19 Variants in Lung CT Scans. *Tomography* **2025**, *11*, 99. [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.

diagnostics

MDPI

Article

Ensemble Algorithm Based on Gene Selection, Data Augmentation, and Boosting Approaches for Ovarian Cancer Classification

Zne-Jung Lee [1], Jing-Xun Cai [2,*], Liang-Hung Wang [3,*] and Ming-Ren Yang [4]

[1] School of Advanced Manufacturing, Fuzhou University, Quanzhou 362200, China; johnlee@fzu.edu.cn
[2] Graduate School of New Generation Electronic Information Engineer, School of Advanced Manufacturing, Fuzhou University, Quanzhou 362200, China
[3] Department of Microelectronics, College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China
[4] Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei 235, Taiwan
* Correspondence: jingxuncai321@163.com (J.-X.C.); eetommy@fzu.edu.cn (L.-H.W.)

Abstract: Background: Ovarian cancer is a difficult and lethal illness that requires early detection and precise classification for effective therapy. Microarray technology has permitted the simultaneous assessment of hundreds of genes' expression levels, yielding important insights into the molecular pathways driving ovarian cancer. To reduce computational complexity and improve accuracy, choosing the most likely differential genes to explain the impacts of ovarian cancer is necessary. Medical datasets, including those related to ovarian cancer, are often limited in size due to privacy concerns, data collection challenges, and the rarity of certain conditions. Data augmentation allows researchers to expand the dataset, providing a larger and more diverse set of examples for model training. Recent advances in machine learning and bioinformatics have shown promise in improving ovarian cancer classification based on gene information. **Methods:** In this paper, we present an ensemble algorithm based on gene selection, data augmentation, and boosting approaches for ovarian cancer classification. In the proposed approach, the initial genetic data were first subjected to feature selection. **Results:** The target genes were screened and combined with data augmentation and ensemble boosting algorithms. From the results, the chosen ten genes could accurately classify ovarian cancer at 98.21%. **Conclusions:** We further show that the proposed algorithm based on clustering approaches is effective for real-world ovarian cancer data, with 100% accuracy and strong performance in distinguishing between distinct ovarian cancer subtypes. The proposed algorithm may help doctors identify ovarian cancer patients early and develop individualized treatment plans.

Keywords: ovarian cancer; microarray data; gene selection; data augmentation; classification; boosting algorithm

1. Introduction

The lifetime risk of developing ovarian cancer is 2.7% among females, representing the sixth highest risk among all types of cancer [1]. The incidence of ovarian cancer is 2.5% of all female cancers; however, 5% of malignant cases result in mortality due to a low survival rate. There is a strong correlation between late-stage diagnosis and the absence of early symptoms [2]. Therefore, the timely identification of ovarian cancer is crucial, given that 70% of women diagnosed with epithelial ovarian carcinoma are only identified once the disease has advanced to the upper abdomen [3].

The application of machine learning, which employs algorithms and data, enables computers to learn and enhance their capabilities autonomously. Its principal strength lies in its capacity to process extensive, intricate, and multi-dimensional data, thereby improving the precision and efficacy of decision making. The potential of machine learning-based

approaches for constructing predictive models for platinum reactions in ovarian cancer has been the subject of investigation by researchers. The medical field has witnessed a surge in the adoption of machine learning in recent years, largely due to the growing accessibility of medical data and the continual advancements in computer analysis capabilities [4]. Akazawa and Hashimoto utilized the Extreme Gradient Boosting (XGBoost) machine learning algorithm, achieving an impressive accuracy of 0.80 in classifying ovarian cancer [5]. Zhang et al. demonstrated the effectiveness of deep learning in accurately predicting definitive diagnoses of ovarian cancers using color ultrasound tests [6]. Conversely, only a few studies have incorporated tumor markers or blood parameters. Kawakami's research revealed that through employing 32 clinical parameters from a blood examination, machine learning could predict diagnoses with an accuracy of 0.92 [7]. Gu et al. employed machine learning to forecast the postprandial change in serum CA125 [8]. Additionally, Lee et al. proposed the use of the most likely differentially expressed genes to circumvent higher computational complexity in the context of ovarian cancer [9]. Machine learning offers the advantage of guiding clinicians and patients in decision making based on ovarian cancer predictions. However, ovarian cancer research often involves the analysis of high-dimensional data, which may be constrained in size, presenting challenges in training effective machine learning models. Consequently, it remains uncertain whether the purported classification accuracy will translate into enhanced performance in clinical settings. Addressing the dimensionality challenge in ovarian cancer research requires a fusion of contemporary machine learning algorithms and a deep understanding of the underlying data characteristics. Efficient feature selection, dimensionality reduction, and data augmentation techniques are pivotal in developing robust and interpretable models for ovarian cancer. Presently, a comprehensive overview of the predictive efficacy of machine learning for classification, incorporating gene selection and data augmentation in the context of ovarian cancer, is lacking. Hence, this study was undertaken to introduce an ensemble algorithm tailored for ovarian cancer analysis. The algorithm under consideration integrates gene selection, data augmentation, and boosting approaches for classification purposes.

The following sections of this work are structured as follows. Section 2 delineates the materials and methods used. Section 3 provides a detailed elaboration of the proposed algorithm. Section 4 presents a comprehensive account of the results and discussions, with a detailed comparison between the proposed algorithm and other existing approaches. Finally, Section 5 presents the conclusions derived from this study.

2. Materials and Methods

Ovarian cancer refers to the proliferation of malignant cells in the ovaries, integral components of the female reproductive system. Ovarian cancer can develop from a variety of cell types in the ovaries and has several subtypes, each with unique traits and behaviors. It is staged according to the magnitude of the sickness, and this helps evaluate the cancer's severity and spread. Patients from China Medical University Hospital provided samples of their ovarian, vaginal, cervical, and myometrium tissues for this study. All microarray operations were carried out in a controlled laboratory setting at the same institution. This study utilized tissue samples from 5 benign ovarian tumors (OVT), 5 ovarian cancers stage I (OVC_I), and 5 stage III ovarian cancers (OVC_III). The pathological and clinical characteristics of OVC_I and OVC_III are detailed in Table 1 [10]. The human cDNA library, comprising 9600 validated human cDNA clones, was provided by the National Health Research Institute of Taiwan. These clones were sourced from the Minimum Information About a Microarray Experiment collaboration libraries through an authorized distributor [11].

Table 1. The pathological and clinical characteristics of OVC_III and OVC_I.

Age	Pathology	Stage
44	Mucinous cystadenoma and borderline malignancy	OVC_I
39	Serous cystadenoma and borderline malignancy	OVC_I
46	Mucinous cystadenoma and borderline malignancy	OVC_I
58	Clear cell carcinoma	OVC_I
36	Mucinous cystadenoma and borderline malignancy	OVC_I
26	Mucinous cystadenoma and borderline malignancy	OVC_III
51	Serous cystadenocarcinoma	OVC_III
62	Serous cystadenocarcinoma	OVC_III
53	Clear cell carcinoma and endometrioid adenocarcinoma	OVC_III
61	Serous cystadenocarcinoma	OVC_III

In this paper, we introduce an ensemble algorithm that relies on gene selection, data augmentation, and boosting approaches for ovarian cancer classification. Our proposed technique integrates ensemble recursive feature elimination (RFE), data augmentation, and boosting algorithms to identify the most informative genes and enhance classification performance. RFE is a feature selection technique in machine learning. It operates by iteratively removing the least significant features based on model performance to select the most important features. The process involves training the model iteratively and eliminating the least important features until the desired number of features is obtained [12]. RFE works by first training the model on all features and assigning weights to each one based on its relevance. It then removes the least important feature(s) and retrains the model using the remaining features. This technique is continued iteratively until the desired number of features is achieved. RFE rates features according to their impact on the model's performance. Features with higher ranks are considered more vital, whereas those with lower levels are deemed less relevant. RFE aids in the development of more accurate models that are less prone to overfitting by focusing on the most relevant features. RFE's chosen subset of features produces a more interpretable model, making it easier to grasp the key reasons driving the predictions. The RFE flowchart is illustrated in Figure 1. The pseudocode is described as follows:

Input dataset with multiple features;
While the number of selected features is not reached, perform the following:

Train the model on all features and assign importance scores to each feature;
Rank features based on importance scores and identify the least significant feature(s);
Remove the least significant feature(s);
Evaluate the model's performance metrics;

End the process;
Output the subset of selected features.

Data augmentation artificially extends a dataset by generating modified versions of existing data. This technique is frequently utilized in machine learning and deep learning to optimize model performance, particularly in contexts where labeled data are scarce. The process of data augmentation utilizing Generative Adversarial Networks (GANs) entails the exploitation of GAN models to generate synthetic data samples that exhibit a high degree of similarity to the original dataset. A GAN comprises two neural networks, namely a generator and a discriminator, which are trained in an adversarial manner to produce synthetic data that are realistic in appearance. In the context of data augmentation, GANs can be employed to generate supplementary training instances for machine learning models, particularly when the original dataset is limited. After training a GAN on the original dataset, the generator network can learn to produce synthetic data samples that

closely resemble authentic data [13]. These synthetic data can then be combined with the actual dataset to provide more diverse instances for training machine learning models. The structure of a GAN is seen in Figure 2.

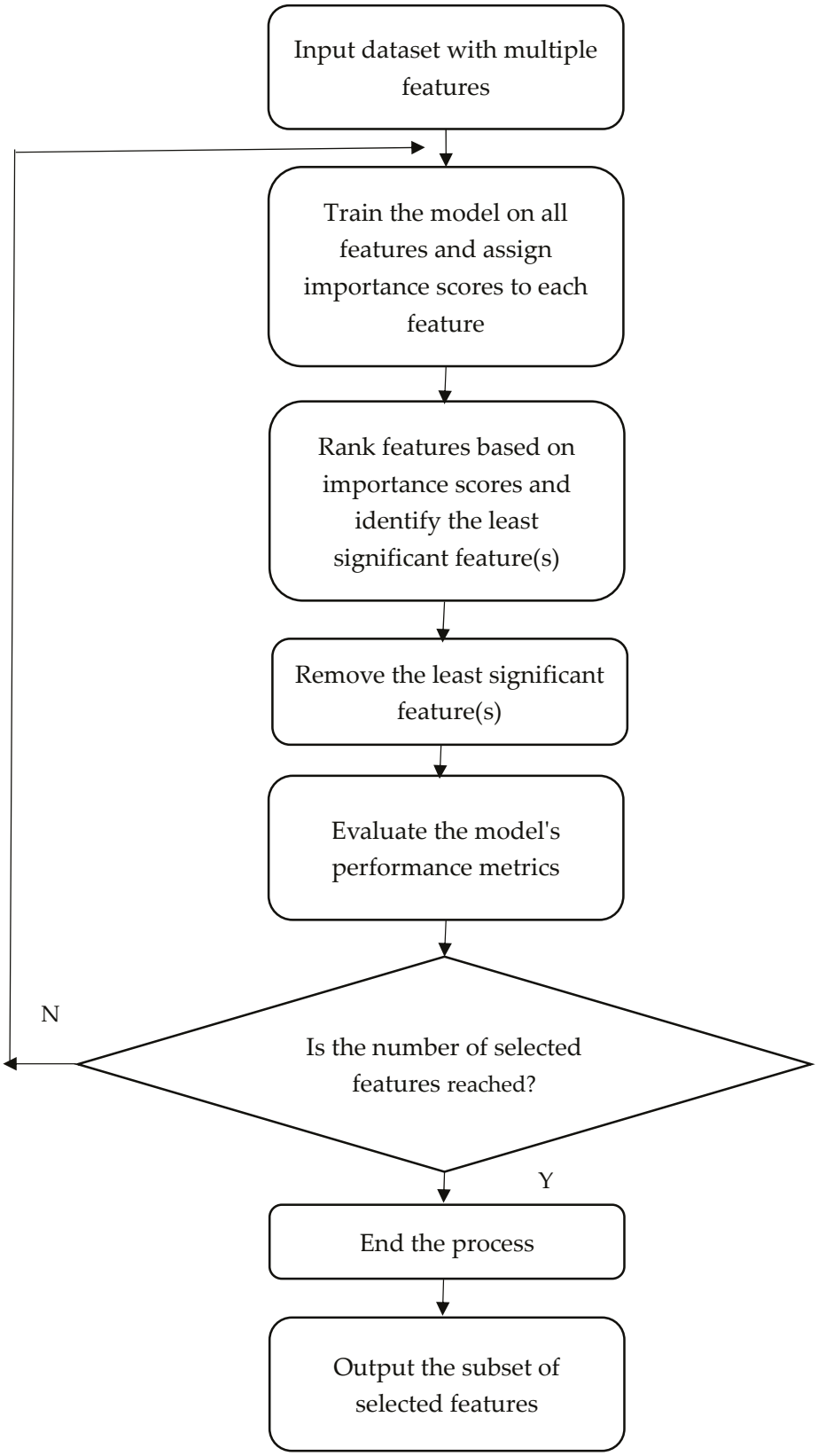


Figure 1. The flowchart of RFE.

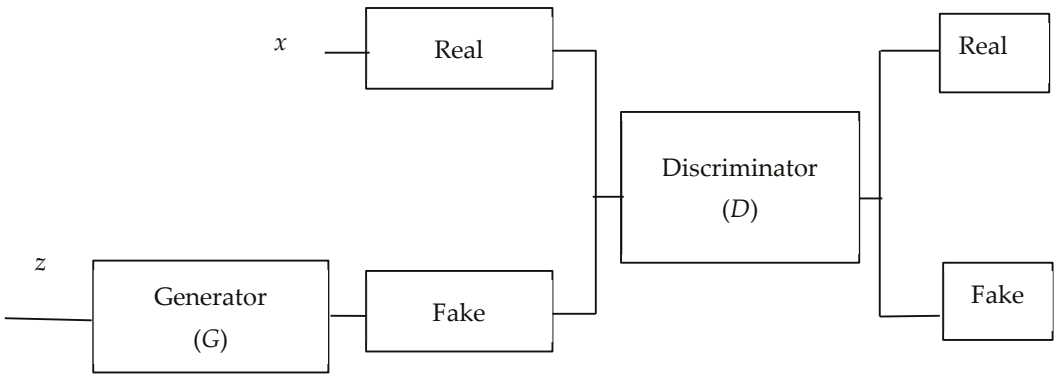


Figure 2. The structure of a GAN.

Boosting is a machine learning technique that aggregates the predictions of multiple weak learners, typically simple models, to create a powerful learner with enhanced predictive capabilities. Boosting algorithms work by sequentially training a sequence of models, with each succeeding model focusing on cases that the prior models misclassified. The final prediction is obtained by averaging the predictions of the individual models. Boosting algorithms, including Gradient Boosting Decision Tree (GBDT), CatBoost, and XGBoost, are well known for their ability to handle complex data relationships while maintaining high predicted accuracy. The foundation of GBDT is the notion that, when it comes to a complicated problem, the consensus reached by a group of experts is superior to the opinion of a single expert. Let the dataset be represented as $D = \{(x_1, y_1),\ (x_2, y_2), \cdots, (x_n, y_n)\}$. The GBDT algorithm can be seen as an additive model consisting of m trees.

$$\hat{y}_i = \sum_{j=1}^{m} f_j(x_i),\ f_j \in F \tag{1}$$

where F represents the function space comprising all trees, with each tree corresponding to a function f. In contrast to typical machine learning algorithms, this additive model does not optimize weights in a high-dimensional space; instead, it directly learns a collection of functions. The cost function of the additive model is defined as

$$Cost^{(t)} = \sum_{i=1}^{n} loss(y_i, \hat{y}_i^t) + \pi(f_t) \tag{2}$$

where π represents the complexity of the decision tree, and $loss(y_i, \hat{y}_i^t)$ is the loss function between y_i and $\hat{y}_i^t$ [14]. CatBoost is a gradient boosting framework that utilizes a distinctive tree-based learning algorithm. It is intended to handle categorical features seamlessly and contains built-in support for categorical data [15]. Its fundamental formulation is based on GBDT, a strong machine learning technique that blends gradient boosting and decision tree models. The XGBoost algorithm is a well-known tree-learning technique. Its core component is the second-order Taylor expansion of the cost function, which XGBoost utilizes to ensure model robustness [16]. This algorithm repeatedly splits a new tree to align with the residuals of the previous prediction. By randomly traversing selected features, XGBoost enhances the random sampling ratio and reduces overfitting resulting from examining all features. The development of the XGBoost model is conducted following the aforementioned process. In the initial stage, a preliminary tree is constructed using the training data, and the discrepancy between the anticipated and actual model values is determined. Subsequently, at each iteration, a new tree is constructed to fit the residuals predicted by the model, until the learning process is complete. This iterative process generates a series of residual trees, which are subsequently integrated into diverse tree models.

3. The Proposed Algorithm

We applied ensemble RFE, data augmentation, and boosting algorithms to identify the most informative genes and ameliorate classification performance in this paper. The

proposed algorithm's flowchart is depicted in Figure 3. In the proposed algorithm, we apply RFE with GBDT, CatBoost, and XGBoost to select the most important 30 features from 9600 features for each algorithm. This will result in three subsets of features, each selected by a different algorithm. Then, we identify the features that appear in at least two subsets out of the three. These features are potential candidates for the selected features in the ensemble. We select the most important 10 important features from the three subsets to create the ensemble subset of features. These features appear in at least two subsets and are considered important by multiple algorithms.

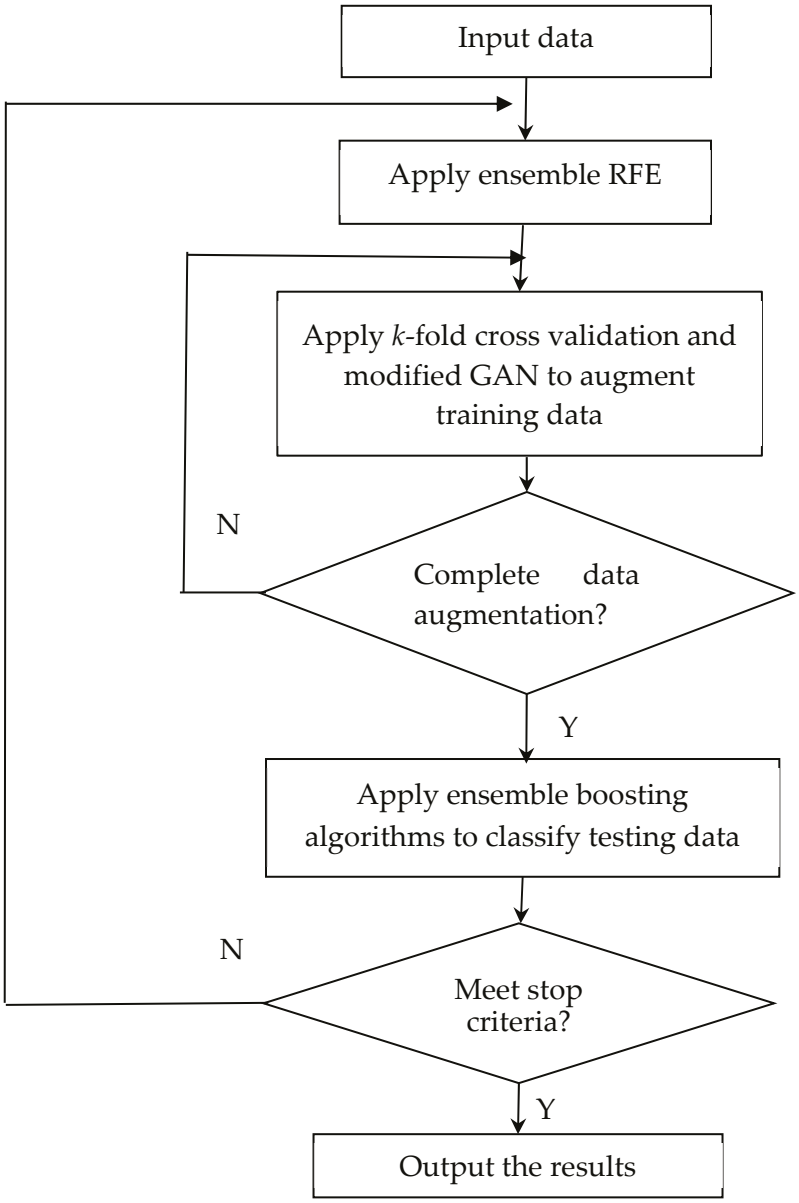


Figure 3. The flowchart of the proposed method.

Obtaining a vast array of diverse training data for ovarian cancer is not a viable option. Data augmentation is the solution to this limitation. It generates synthetic variations while preserving privacy and security. In this research, we applied a modified GAN to augment data after ensemble RFE. In evaluating the performance of the modified GAN, we employed k-fold cross-validation (k = 3). This technique partitions the training data into k subsets or folds. The model undergoes training and evaluation k times, with each iteration utilizing a different fold as the validation set and the remaining $k - 1$ folds as the training set. During

each training iteration within the *k*-fold cross-validation process, the modified GAN is trained on the training folds containing the chosen features. It is noteworthy that the generated data are exclusively employed for training purposes. The testing data utilized in this study were derived solely from the original dataset, ensuring that the model's evaluation and validation were based on authentic, real-world data. Figure 4 illustrates the modified GAN's ability to produce synthetic data that closely mirror the original training data. It demonstrates the integration of the generative strength of GAN with the discriminative capabilities of boosting algorithms like GBDT, CatBoost, and XGBoost. Data augmentation using GAN and boosting approaches involves generating synthetic data samples that resemble real data instances. The specific modifications applied during the data augmentation process can vary based on the nature of the data and the objectives of the task. Through applying these modifications and validating the generated data through various techniques, data augmentation can effectively enhance the diversity and quantity of the training data, leading to improved model performance and robustness in machine learning tasks. The core network structure of the modified GAN retains a generator and discriminator. The generator takes input comprising both classification details (c) and random noise (z). In leveraging the boosting algorithm, data classification accuracy is enhanced, effectively tackling the challenge of limited data availability for analysis. In the loss function $(Loss_x)$ of the modified GAN, x denotes the raw data. The discriminator's goal is to maximize $Loss_x + Loss_c$, while the generator's goal is to minimize $Loss_x + Loss_c$.

$$Loss_x = E[\log P(x = real|X_{real})] + E\left[\log P\left(x = fake\middle|X_{fake}\right)\right] \tag{3}$$

$$Loss_C = E[\log P(c = real|X_{real})] + E\left[\log P\left(c = fake\middle|X_{fake}\right)\right] \tag{4}$$

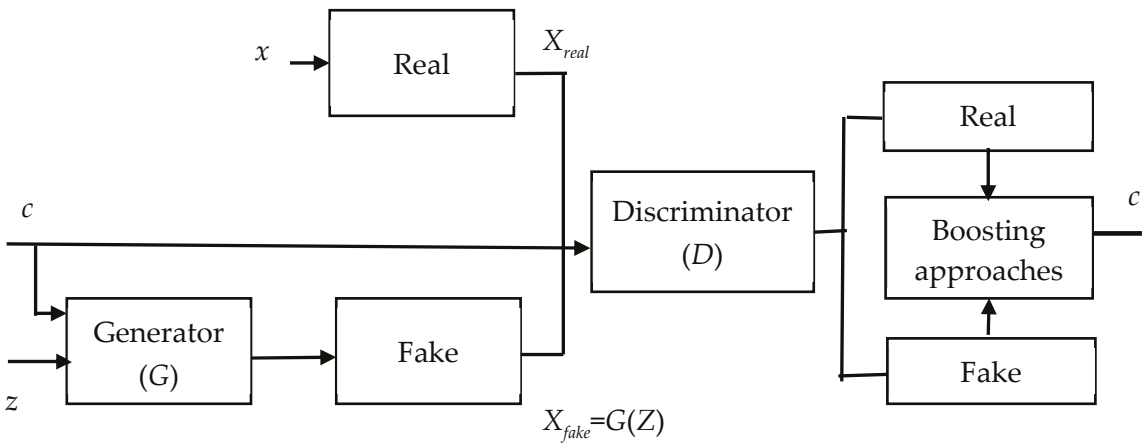


Figure 4. The modified GAN for data augmentation.

After training the modified GAN, we created augmented data incorporating the chosen features for every fold within the dataset. This augmented data served to enrich the original training dataset with the selected features, thereby enhancing both the diversity and scale of the training data. Utilizing the augmented training data containing the selected features enabled us to train boosting algorithms on each fold and assess their performance on the respective dataset. Once the *k*-fold cross-validation was complete, we could combine the results from each fold to obtain a clear picture of the model's overall performance, including the benefits gained from the augmented data generated by the modified GAN. The algorithm continues to run until the stop criteria are satisfied, at which point it produces the final results. By following these steps, we effectively created an ensemble of RFE with GBDT, CatBoost, and XGBoost, and then selected the most important features that appear in multiple subsets. Compared to existing methods, the approach described as the proposed algorithm emphasizes a combination of diversity, feature importance, and

cross-validation to identify the most informative and reliable genes for the ensemble. This comprehensive approach aims to address potential limitations of traditional gene selection methods by leveraging the strengths of multiple algorithms, ensuring robustness and enhancing the predictive power of the ensemble model. This approach is the most effective way to combine the strengths of different feature selection algorithms and machine learning models to create a robust ensemble.

4. Results and Discussions

To ensure equitable comparisons, maintaining consistent parameters across different methods is crucial. The experiments utilized the *k*-fold cross-validation method with $k = 3$. This research selected 10 features from the proposed algorithm, as shown in Table 2. The target genes were selected using the proposed algorithm, with their cardinalities and F scores all exceeding 10. The gene index of 8683 had the highest F score (19.68) of all the selected genes. These 10 genes were Caenorhabditis elegans, small inducible cytokine A2, KIAA1140 protein, mitochondrial ribosomal protein L23, DKFZp434A115, myxovirus resistance 2, retinoic acid receptor responder 1, protocadherin alpha 5, myeloid differentiation primary response gene, and EGF-containing fibulin.

Table 2. The 10 selected genes from the proposed algorithm.

Index	Gene Name	F Score
8683	Caenorhabditis elegans	19.68
6760	Small inducible cytokine A2	18.17
1720	KIAA1140 protein	16.32
3958	Mitochondrial ribosomal protein L23	15.09
6269	DKFZp434A115	12.90
5954	Myxovirus resistance 2	12.85
3699	Retinoic acid receptor responder 1	12.84
4201	Protocadherin alpha 5	12.10
6324	Myeloid differentiation primary response gene	11.28
3390	EGF-containing fibulin	10.97

A detailed comparison of the various methodologies employed to validate the proposed algorithm's performance is provided in Figure 5 and Table 3. The results demonstrate that the proposed algorithm enhances the classification accuracy to 98.21% when compared to traditional machine learning techniques, including GBDT, CatBoost, XGBoost, SVM, DT, and IA. The Improved Algorithm (IA) employs a scatter search to identify the optimal parameter configurations for DT. The DT algorithm employs the partition information entropy minimization algorithm to recursively divide the dataset into smaller subsets, which are then represented as tree structures. SVM is a learning algorithm that employs a hypothesis space of linear functions within a high-dimensional feature set. It is also noteworthy that GBDT, CatBoost, and XGBoost incorporate feature selection techniques. The criteria used for feature selection with GBDT is the feature importance, such as for weight, that provides a measure of how each feature contributes to the predictive performance of the model. For CatBoost, the used criteria for feature selection were the prediction value change, which highlights features with higher prediction value change, considering them more significant for the model. As for XGBoost, the used criteria for feature selection were the gain importance, which identify features with higher gain importance, marking them as more influential in the model's decision-making process.

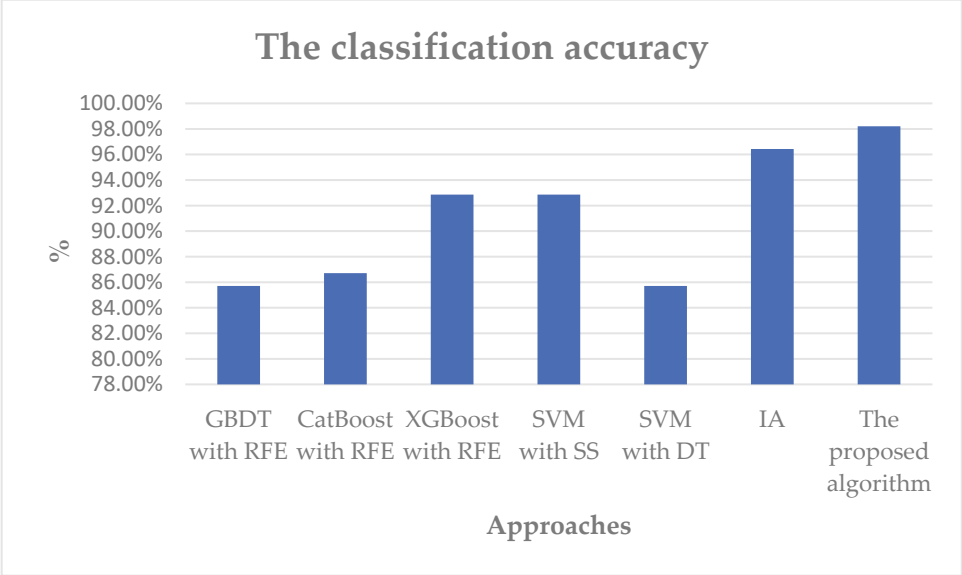


Figure 5. The classification accuracy for various approaches.

Table 3. The classification accuracy for ovarian cancer.

Method	Classification Accuracy	Selected Genes
GBDT with RFE	85.71%	30
CatBoost with RFE	86.71%	30
XGBoost with RFE	92.86%	30
SVM with SS [17]	92.86%	12
SVM with DT [17]	85.71%	14
IA [17]	96.43%	5
The proposed algorithm	98.21%	10

For index 8683, the gene is weakly similar to a hypothetical protein encoded by the T12A2.1 gene in the caenorhabditis elegans organism. Caenorhabditis elegans is a widely studied model organism in biology, particularly in the field of developmental biology and genetics. Caenorhabditis elegans is a nematode and may seem unrelated to human cancer; it is possible that specific genetic elements or pathways in this organism could have analogs or functional equivalents in human biology, including cancer-related processes [18]. The small inducible cytokine A2 is a protein that in humans is encoded by the CCL2 gene. It plays a crucial role in the recruitment and activation of monocytes, memory T cells, and dendritic cells to sites of injury or infection. The small inducible cytokine A2 plays a crucial role in the regulation of immune responses and inflammation. In the context of ovarian cancer, the presence of specific cytokines may be indicative of tumor microenvironment characteristics, immune cell infiltration, or tumor progression [19]. The KIAA1140 protein is a human protein. The function of the KIAA1140 protein is not yet fully understood, and it is classified as an uncharacterized protein [20]. Mitochondrial ribosomal protein L23 is a protein. It is a component of the large subunit of the mitochondrial ribosome, which is responsible for protein synthesis within the mitochondria, the energy-producing organelles within cells. It plays a crucial role in the structure and function of the mitochondrial ribosome, contributing to the process of translation, where genetic information is used to synthesize proteins [21]. The DKFZp434A115 appears to be a description of a specific mRNA transcript originating from a cDNA clone [22]. Myxovirus resistance 2 is a human protein that is involved in the innate immune response to viral infections. It is particularly recognized for its role in restricting the replication of HIV-1, as well as other retroviruses, by targeting the viral capsid

and preventing the nuclear import of viral DNA [23]. Retinoic acid receptor responder 1 is a gene that encodes a protein involved in various cellular processes, including cell growth, differentiation, and apoptosis. It has been implicated in the regulation of cell proliferation and has been suggested to function as a tumor suppressor in certain contexts. The retinoic acid receptor is known to be involved in cell differentiation, apoptosis, and cell cycle control. In ovarian cancer, the dysregulation of retinoic acid signaling pathways could impact tumor cell behavior and response to treatment [24]. Protocadherin alpha 5 is a protein-coding gene that belongs to the protocadherin family, a group of cell adhesion molecules that play key roles in the development and maintenance of the nervous system. These proteins are particularly important in the regulation of neuronal connectivity, including axon guidance, synapse formation, and neuronal circuit assembly [25]. The myeloid differentiation primary response gene is an essential adapter protein involved in the signaling pathways of the innate immune system. It plays a critical role in the transmission of signals from various toll-like receptors and interleukin-1 receptors, which are key components of the innate immune response to pathogens. In the context of ovarian cancer, its expression levels may relate to the tumor microenvironment, immune cell recruitment, and, potentially, response to immunotherapies [26]. EGF-containing fibulin is a gene that encodes a protein involved in the formation and maintenance of the extracellular matrix, a complex network of molecules that provides structural support and influences cellular behavior in various tissues. In ovarian cancer, the dysregulation of genes involved in extracellular matrix interactions can influence tumor invasion and metastasis. It has been implicated in various pathological conditions, including cancer, where its dysregulation has been linked to tumor progression and metastasis in certain types of cancer [27]. Insights from these genes could potentially provide avenues for research in ovarian cancer, such as biological pathways, therapeutic targets, and biomarker development. By delving deeper into the pathways and networks associated with these genes, researchers may gain a better understanding of their roles in ovarian cancer progression and response to therapy. The genes identified in this study hold potential as targets for therapeutic interventions, including targeted therapies and immunotherapies. Furthermore, unraveling the expression patterns and functional significance of these genes could play a pivotal role in the development of prognostic or predictive biomarkers for ovarian cancer.

The correlation of the 10 selected target genes is shown in Figure 6. In Figure 6, the gene of mitochondrial ribosomal protein L23 has the highest positive correlation (0.76) with the label (Group). The dysregulation of mitochondrial function, including alterations in mitochondrial ribosomal proteins, could potentially contribute to cancer development and progression, including ovarian cancer. The gene of small inducible cytokine A2 has the highest negative correlation (−0.87) with the label. Studies have suggested a potential correlation with ovarian cancer, and it may contribute to the vascularization of ovarian tumors.

Moreover, these selected genes were verified by Kmeans Clustering, Agglomerative Clustering, and Hierarchical Clustering [28,29]. The simulation results are tabulated in Tables 4 and 5 and Figure 7. From Tables 4 and 5, the results show that these gene markers verified by Kmeans Clustering and Hierarchical Clustering can classify OVT, OVC_III, and OVC_I successfully. Figure 7 shows that Hierarchical Clustering has the same performance as Kmeans Clustering. The proposed algorithm demonstrates promising results in classifying ovarian cancer subtypes. The classification model achieved high accuracy and specificity in distinguishing between different cancer subtypes based on gene expression patterns. Comparative analysis with existing methods showed the superior performance of our approach in terms of classification accuracy and robustness.

Figure 6. The correlation of the 10 selected target genes.

Table 4. The stage results of Kmeans Clustering and Agglomerative Clustering with 10 selected genes.

Stage	Kmeans Clustering #	Agglomerative Clustering #
OVT	0	0
OVT	0	0
OVT	0	0
OVT	0	0
OVT	0	0
OVC_IA	1	1
OVC_IA	1	1
OVC_IA	1	1
OVC_IC	1	1
OVC_IA	1	1
OVC_IIIB	2	2
OVC_IIIC	2	2

Table 4. *Cont.*

Stage	Kmeans Clustering #	Agglomerative Clustering #
OVC_IIIC	2	2
OVC_IIIC	2	2
OVC_IIIC	2	2

Table 5. The clustering results of Kmeans Clustering and Agglomerative Clustering with 10 selected genes.

Cluster #	OVT	OVC				Total
		IA	IC	IIIB	IIIC	
0	5	0	0	0	0	5
1	0	4	1	0	0	5
2	0	0	0	4	1	5
Total	5	4	1	4	1	15

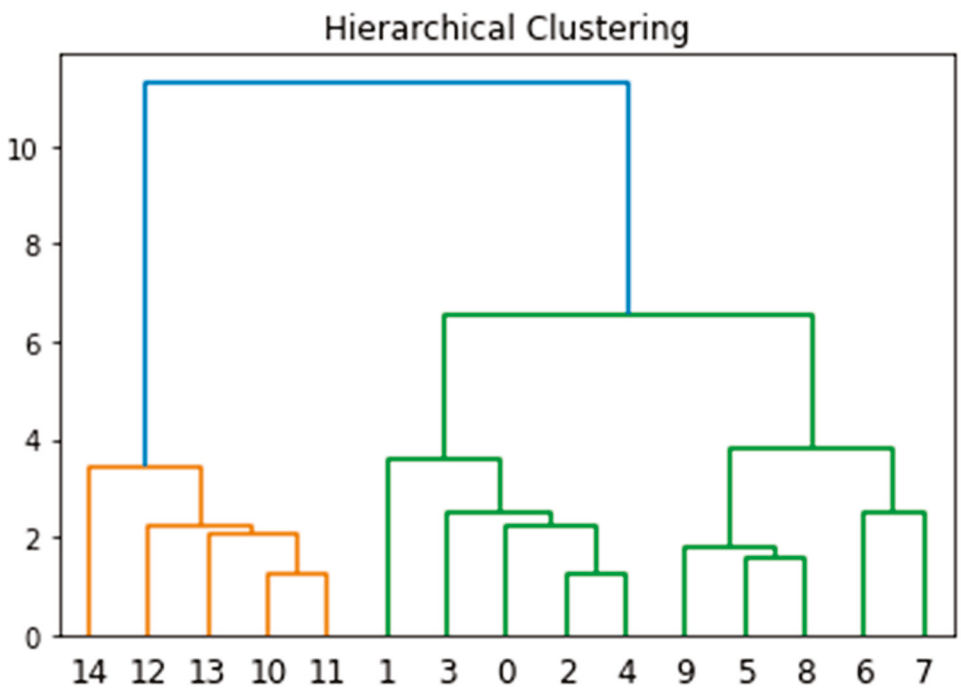


Figure 7. The results of Hierarchical Clustering.

5. Conclusions and Future Work

In this paper, we present an ensemble algorithm based on gene selection, data augmentation, and boosting approaches for ovarian cancer classification. This is the most effective method for analyzing this type of data. The integration of gene selection techniques with ensemble boosting algorithms and data augmentation proved to be an effective method for enhancing the classification of ovarian cancer subtypes. The proposed algorithm leveraging boosting approaches showcases significant performance advantages when contrasted with traditional models like SVM or DT. Boosting approaches, including GBDT, CatBoost, and XGBoost, consistently outperform SVM or DT in terms of classification accuracy. Their effectiveness is particularly pronounced when handling large and intricate datasets, leading to superior accuracy. Furthermore, boosting approaches possess inherent capabilities for

capturing intricate non-linear relationships within data, rendering them well suited for addressing a wide array of real-world challenges. In contrast, SVM and DT, while effective in specific scenarios, may encounter difficulties in managing non-linear relationships without the supplemental usage of kernel tricks. Our approach improved the interpretability and predictive power of the classification model by identifying key genes associated with specific cancer subtypes and leveraging the ensemble learning capabilities of the boosting algorithms. These results demonstrate the potential of this intelligent approach to aid clinicians in early diagnosis and personalized treatment strategies for ovarian cancer patients. The outcomes validate the efficacy of this approach in precisely classifying ovarian cancer subtypes and offering significant insights into the fundamental molecular mechanisms of the disease.

To thoroughly evaluate the algorithm's performance in real-world clinical settings and its ability to generalize to diverse populations, it is imperative to conduct rigorous validation using diverse and representative clinical datasets, accompanied by comprehensive evaluation metrics. Furthermore, assessing the algorithm's performance across various population groups is crucial for its real-world applicability and ethical deployment. Insights from these selected genes could potentially provide avenues for future research in ovarian cancer, such as biological pathways, therapeutic targets, and biomarker development. By delving deeper into the pathways and networks associated with these genes, researchers may gain a better understanding of their roles in ovarian cancer progression and response to therapy. Moreover, the genes identified in this study hold potential as targets for therapeutic interventions, including targeted therapies and immunotherapies. Furthermore, unraveling the expression patterns and functional significance of these genes could play a pivotal role in the development of prognostic or predictive biomarkers for ovarian cancer. Exploring the integration of multi-omics data, including genomics, transcriptomics, proteomics, and metabolomics, could enrich the model's understanding of the molecular landscape of ovarian cancer. This integration may provide a more comprehensive view of the disease and potentially lead to the discovery of novel biomarkers or therapeutic targets.

Author Contributions: Conceptualization, Z.-J.L. and M.-R.Y.; methodology, Z.-J.L. and L.-H.W.; formal analysis, M.-R.Y. and J.-X.C.; investigation, M.-R.Y. and L.-H.W.; resources, M.-R.Y.; data curation, M.-R.Y.; writing—review and editing, M.-R.Y. and J.-X.C.; supervision, Z.-J.L. 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: The utilized code and data will be supplied upon request with the approval of the healthcare facility.

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

References

1. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424. [CrossRef] [PubMed]
2. Ahamad, M.M.; Aktar, S.; Uddin, M.J.; Rahman, T.; Alyami, S.A.; Al-Ashhab, S.; Moni, M.A. Early-stage detection of ovarian cancer based on clinical data using machine learning approaches. *J. Pers. Med.* **2022**, *12*, 1211. [CrossRef] [PubMed]
3. Lee, Z.J. An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer. *Artif. Intell. Med.* **2008**, *42*, 81–93. [CrossRef] [PubMed]
4. Wang, Q.; Chang, Z.; Liu, X.; Wang, Y.; Feng, C.; Ping, Y.; Feng, X. Predictive value of machine learning for platinum chemotherapy responses in ovarian cancer: Systematic review and meta-analysis. *J. Med. Internet Res.* **2024**, *26*, e48527. [CrossRef]
5. Akazawa, M.; Hashimoto, K. Artificial intelligence in ovarian cancer diagnosis. *Anticancer Res.* **2020**, *40*, 4795–4800. [CrossRef]
6. Zhang, L.; Huang, J.; Liu, L. Improved deep learning network based in combination with cost-sensitive learning for early detection of ovarian cancer in color ultrasound detecting system. *J. Med. Syst.* **2019**, *43*, 251. [CrossRef] [PubMed]

7. Kawakami, E.; Tabata, J.; Yanaihara, N.; Ishikawa, T.; Koseki, K.; Iida, Y.; Saito, M.; Komazaki, H.; Shapiro, J.S.; Goto, C.; et al. Application of artificial intelligence for preoperative diagnostic and prognostic prediction in epithelial ovarian cancer based on blood biomarkers. *Clin. Cancer Res.* **2019**, *25*, 3006–3015. [CrossRef] [PubMed]
8. Gu, Z.; He, Y.; Zhang, Y.; Chen, M.; Song, K.; Huang, Y.; Li, Q.; Di, W. Postprandial increase in serum CA125 as a surrogate biomarker for early diagnosis of ovarian cancer. *J. Transl. Med.* **2018**, *16*, 114. [CrossRef] [PubMed]
9. Lee, Z.J.; Lin, S.W.; Hsu CC, V.; Huang, Y.P. Gene extraction and identification tumor/cancer for microarray data of ovarian cancer. In Proceedings of the TENCON 2006-2006 IEEE Region 10 Conference, Hong Kong, China, 14–17 November 2006; IEEE: Piscataway, NJ, USA, 2006; pp. 1–3.
10. Huang, G.S.; Chen, A.; Hung, Y.C.; Hong, M.Y. Microarray analysis of ovarian cancer. In Proceedings of the 2005 IEEE International Conference on Systems, Man and Cybernetics, Waikoloa, HI, USA, 12 October 2005; IEEE: Piscataway, NJ, USA, 2005; Volume 2, pp. 1036–1041.
11. Brazma, A.; Hingamp, P.; Quackenbush, J.; Sherlock, G.; Spellman, P.; Stoeckert, C.; Vingron, M. Minimum information about a microarray experiment (MIAME)—Toward standards for microarray data. *Nat. Genet.* **2001**, *29*, 365–371. [CrossRef]
12. Priyatno, A.M.; Widiyaningtyas, T. A Systematic Literature Review: Recursive Feature Elimination Algorithms. *JITK J. Ilmu Pengetah. Dan Teknol. Komput.* **2024**, *9*, 196–207. [CrossRef]
13. Shatnawi, M.A.; Bani Younes, M. An enhanced model for detecting and classifying emergency vehicles using a generative adversarial network (GAN). *Vehicles* **2024**, *6*, 1114–1139. [CrossRef]
14. Dong, X.; Guo, M.; Wang, S. GBDT-based multivariate structural stress data analysis for predicting the sinking speed of an open caisson foundation. *Georisk Assess. Manag. Risk Eng. Syst. Geohazards* **2024**, *18*, 333–345. [CrossRef]
15. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. Advances in neural information processing systems. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1–11.
16. Hoque, R.; Das, S.; Hoque, M.; Haque, E. Breast Cancer Classification using XGBoost. *World J. Adv. Res. Rev.* **2024**, *21*, 1985–1994. [CrossRef]
17. Lin, S.W.; Lee, C.Y.; Chou, S.T.; Lee, Z.J. An Improved Algorithm with Gene Selection and Decision Rules for Ovarian Cancer. *Adv. Comput. Sci. Its Appl.* **2012**, *1*, 26–31.
18. Santoro, A.L.; Vendramini, P.H.; Braga, D.L.; Martins, L.G.; Dos Santos, G.F.; Lima, G.S.; Mori, M.A. *Caenorhabditis elegans* (*C. elegans*) sample preparation for metabolomics and lipidomics analysis–A review. *TrAC Trends Anal. Chem.* **2023**, *169*, 117374. [CrossRef]
19. Rollins, B.J.; Morton, C.C.; Ledbetter, D.H.; Eddy, R.L., Jr.; Shows, T.B. Assignment of the human small inducible cytokine A2 gene, SCYA2 (encoding JE or MCP-1), to 17q11. 2–12: Evolutionary relatedness of cytokines clustered at the same locus. *Genomics* **1991**, *10*, 489–492. [CrossRef] [PubMed]
20. Dan, Z.; Xiujing, H.; Ting, L.; Xiaorong, Z.; Hong, Z.; Jiqiao, Y.; Jing, J. Long non-coding RNA BTG3-7: 1 and JUND co-regulate C21ORF91 to promote triple-negative breast cancer progress. *Front. Mol. Biosci.* **2021**, *7*, 605623. [CrossRef] [PubMed]
21. Yin, A.; Feng, M.; Zhang, L.; Cheng, Z.; Li, Y.; Qian, L. Identification of a novel native peptide derived from 60S ribosomal protein L23a that translationally regulates p53 to reduce myocardial ischemia-reperfusion. *Pharmacol. Res.* **2022**, *175*, 105988. [CrossRef]
22. Zhang, Z.Y.; Yang, Y.H.; Ding, H.; Wang, D.; Chen, W.; Lin, H. Design powerful predictor for mRNA subcellular location prediction in Homo sapiens. *Brief. Bioinform.* **2021**, *22*, 526–535. [CrossRef]
23. Zeng, M.; Chen, S.; Wang, M.; Jia, R.; Zhu, D.; Liu, M.; Cheng, A. Molecular identification and comparative transcriptional analysis of myxovirus resistance GTPase (Mx) gene in goose (*Anser cygnoide*) after H9N2 AIV infection. *Comp. Immunol. Microbiol. Infect. Dis.* **2016**, *47*, 32–40. [CrossRef] [PubMed]
24. Lo, K.W.; Chung, G.T. RARRES1 (Retinoic Acid Receptor Responder (Tazarotene Induced) 1). 2007. Available online: http://AtlasGeneticsOncology.org,121 (accessed on 5 December 2024).
25. Hirayama, T.; Yagi, T. The role and expression of the protocadherin-alpha clusters in the CNS. *Curr. Opin. Neurobiol.* **2006**, *16*, 336–342. [CrossRef] [PubMed]
26. Das, S.J.; Dutta, S.; Tiwari, D.; Basumatary, T.K.; Kashyap, N.; Kalita, M.P.; Bose, P.D. Prognostic value of myeloid differentiation primary response protein 88 in type II diabetes mellitus in non-obese NAFLD: A case-control study from Assam. *Hum. Gene* **2024**, *39*, 201246. [CrossRef]
27. Livingstone, I.; Uversky, V.N.; Furniss, D.; Wiberg, A. The pathophysiological significance of fibulin-3. *Biomolecules* **2020**, *10*, 1294. [CrossRef] [PubMed]
28. Rizalde, A.R.; Mubarak, H.A.; Ramadhan, G.; Fatan, M.A. Comparison of K-Means, BIRCH and Hierarchical Clustering Algorithms in Clustering OCD Symptom Data. *Public Res. J. Eng. Data Technol. Comput. Sci.* **2024**, *1*, 102–108. [CrossRef]
29. Siet, S.; Peng, S.; Ilkhomjon, S.; Kang, M.; Park, D.S. Enhancing sequence movie recommendation system using deep learning and kmeans. *Appl. Sci.* **2024**, *14*, 2505. [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.

diagnostics

MDPI

Article

CAD-EYE: An Automated System for Multi-Eye Disease Classification Using Feature Fusion with Deep Learning Models and Fluorescence Imaging for Enhanced Interpretability

Maimoona Khalid [1], Muhammad Zaheer Sajid [1], Ayman Youssef [2], Nauman Ali Khan [1], Muhammad Fareed Hamid [1] and Fakhar Abbas [3,*]

[1] Department of Computer Software Engineering, Military College of Signals, National University of Science and Technology, Islamabad 44000, Pakistan; mkhalid.msse22mcs@student.nust.edu.pk (M.K.); msajid.msse-27mcs@student.nust.edu.pk (M.Z.S.); nauman@mcs.edu.pk (N.A.K.); fareedhamid@hotmail.com (M.F.H.)

[2] Department of Computers and Systems, Electronics Research Institute, Cairo 11843, Egypt; aymanmahgoub@eri.sci.eg

[3] Centre for Trusted Internet and Community, National University of Singapore, Singapore 119077, Singapore

* Correspondence: fakhar.5@nus.edu.sg

Abstract: Background: Diabetic retinopathy, hypertensive retinopathy, glaucoma, and contrast-related eye diseases are well-recognized conditions resulting from high blood pressure, rising blood glucose, and elevated eye pressure. Later-stage symptoms usually include patches of cotton wool, restricted veins in the optic nerve, and buildup of blood in the optic nerve. Severe consequences include damage of the visual nerve, and retinal artery obstruction, and possible blindness may result from these conditions. An early illness diagnosis is made easier by the use of deep learning models and artificial intelligence (AI). **Objectives:** This study introduces a novel methodology called CAD-EYE for classifying diabetic retinopathy, hypertensive retinopathy, glaucoma, and contrast-related eye issues. **Methods:** The proposed system combines the features extracted using two deep learning (DL) models (MobileNet and EfficientNet) using feature fusion to increase the diagnostic system efficiency. The system uses fluorescence imaging for increasing accuracy as an image processing algorithm. The algorithm is added to increase the interpretability and explainability of the CAD-EYE system. This algorithm was not used in such an application in the previous literature to the best of the authors' knowledge. The study utilizes datasets sourced from reputable internet platforms to train the proposed system. **Results:** The system was trained on 65,871 fundus images from the collected datasets, achieving a 98% classification accuracy. A comparative analysis demonstrates that CAD-EYE surpasses cutting-edge models such as ResNet, GoogLeNet, VGGNet, InceptionV3, and Xception in terms of classification accuracy. A state-of-the-art comparison shows the superior performance of the model against previous work in the literature. **Conclusions:** These findings support the usefulness of CAD-EYE as a diagnosis tool that can help medical professionals diagnose an eye disease. However, this tool will not be replacing optometrists.

Keywords: multi-eye disease; feature fusion; deep learning

1. Introduction

The eye is considered a key organ that is essential for functioning vision. Illnesses affecting the eyes have the potential to cause irreparable damage to the retina, which might lead to vision impairment or even blindness.

These eye-related disorders pose obstacles to fundamental activities such as reading, recognizing faces, and driving. Consequently, vision impairment will have a great impact on a person's quality of life. It is estimated that there is a minimum of 2.2 billion people struggling with various forms of visual impairment worldwide. The economic costs of

vision impairment are significant; they are projected to be USD 411 billion globally per year. Statistics highlight a substantial disparity in treatment access. Globally, only 36% of individuals experiencing vision impairment from refractive errors and a mere 17% of those affected by cataracts have undergone necessary and suitable treatments. These small percentages prove the pressing requirement for enhanced global initiatives aimed at addressing and preventing visual impairments, with a focus on the potential positive impact on both society and the economy [1]. The early identification and detection of eye diseases such as glaucoma play a crucial role in preventing irreversible vision loss. This is very important as the condition can lead to permanent visual impairment. One disease that affects the eye is glaucoma. Traditional methods for glaucoma detection have often shown limited accuracy [2]. However, a novel approach has been developed, enabling faster and more effective disease detection by analyzing the characteristics of the optic disk in retinal images. An essential factor in glaucoma is intraocular pressure (IOP), analogous to blood pressure but specific to eye pressure, and can be used for the classification of this issue. Elevated IOP can cause damage to the optic nerve, resulting in symptoms such as blurred vision and eventual blindness over time [3–5]. Because manual processing of eye pictures takes a lot of time and is prone to errors depending on the examiner's skill level, automated techniques for glaucoma identification are essential. These days, an analysis of 50 automated computerized retinal images has become a useful screening technique for detecting a range of eye conditions and hazards. Glaucoma manifests in two forms, chronic (called open-angle glaucoma) and acute (closed-angle glaucoma), both of which can increase intraocular pressure. Notably, during the early stages of glaucoma, symptoms cannot be noticed in patients, posing a challenge to early detection. Detecting glaucoma at its early stages is crucial to prevent progression to irreversible blindness. In recent years, digital retinal images have emerged as a valuable tool for conducting 57 glaucoma screenings. Various techniques and procedures have been developed to identify retinal abnormalities associated with glaucoma, enabling early detection and intervention for this vision-threatening condition. A non-invasive medical treatment called fundoscopy, often known as ophthalmoscopy, is used to look at the inner parts of the eye (retinal blood vessels, retina, etc.). This examination, which is usually carried out by medical specialists like ophthalmologists, internists, neurologists, or pediatricians, offers an in-depth examination of the eye's central nerve system, arteries, and veins.

Its main function is to diagnose and track down a range of eye diseases, including glaucoma, retinal hemorrhages, diabetic retinopathy, and macular degeneration [6]. In fundoscopic examination, eye drops must be used. The patient's pupils are dilated with these topical and short-acting eye drops. Following this, an ophthalmoscope, a specialized optical device, is employed to illuminate the retina through the pupil. This illumination enables the formation of an image of the retina, visible through the pupil, offering valuable insights into overall eye health and the presence of any abnormalities or diseases [6]. These images can be captured using a camera that is integrated into an ophthalmoscope that focuses on the retina or other eye structures [7].

The collected fundoscopic images serve as a valuable data source for the development of artificial intelligence (AI) systems designed to support various eye illnesses in terms of the diagnosis and treatment. Widely recognized open ophthalmological imaging datasets, including MESSIDOR, EyePACS, DRIVE, and E-ophtha, have been used by a large number of machine learning (ML) researchers in their research [8–11]. Another way to detect such diseases is AI systems, which can be used to recognize and evaluate patterns related to certain eye disorders by training AI models using a large dataset of fundoscopic pictures.

This saves time and improves accessibility to care. Also, AI systems have the ability to standardize treatment steps and enhance overall treatment outcomes for the patient.

Diabetic retinopathy is another eye issue that is a complication of the diabetes condition. Eye diseases are a major cause of blindness in elderly and middle-aged people because they affect the eyes. DR is considered a major global health problem [12]. According to estimates, the prevalence of vision-threatening diabetic retinopathy (VTDR) and diabetic

retinopathy (DSR) in individuals with diabetes in the US is 4.4% and 28.5%, respectively [13]. Projections of diabetic retinopathy suggest that by 2045, DR and VTDR will affect around 160.50 million and 44.82 million people, respectively, globally. People living in the Middle East, North Africa, and Western Pacific will be more affected by these conditions [14]. Diabetic retinopathy can be identified by recognizing certain features in the retina. Clinical classification can classify the case to mild, moderate, and severe levels of diabetic retinopathy [12]. It has been demonstrated [15] that in countries with developed healthcare systems, like the United Kingdom, blindness can be effectively decreased through the early detection of vision-threatening conditions and timely treatment of them. Another exam that is used in the diagnosis of diabetic retinopathy is retinal photography. This screening method for diabetic retinopathy has proved superior performance in certain cases compared to an in-person eye exam [16]. Most low- and middle-income countries have difficulties in implementing systematic diabetic retinopathy screening programs. Hence, it is important to build cost-effective automated methods for screening and treatment of diabetic retinopathy [15]. Deep learning (DL) models are powerful machine learning models that have the capability to learn and extract important features from images. These models, when trained on a large dataset, can generalize for classification problems [10]. DL models are proven to have impressive results in implementing automated systems for an image analysis of eye images. In recent years, DL models were used widely for medical imaging classification, especially eye diseases. In [17], a deep learning-based convolutional neural network is proposed for the classification of eye diseases. In [18], a deep learning framework for retinal disease classification is proposed. All these research papers prove the importance of DL in medical imaging classification.

Hence, integrating deep neural networks into the screening process for diabetic retinopathy has importance. DL models can help with the classification of diabetic retinopathy lesions with high accuracy [15].

Another eye disease is called glaucoma. This disease can cause vision loss and blindness if left untreated. Globally, glaucoma is the primary cause of irreversible blindness and can affect the quality of a person's life [19]. A class of progressive visual neuropathies known as glaucoma is defined by changes in the optic nerve head brought on by the degradation of retinal ganglion cells and retinal nerve fiber layers [20]. The condition is associated with damage to the optic nerve brought on by high intraocular pressure (IOP), which results in the degeneration of retinal ganglion cells [21]. Glaucoma is more commonly observed in adults. It is estimated that in 2020, over 76 million people were affected by glaucoma and it is anticipated that the number of adults affected by glaucoma will reach 111.8 million by 2040 [22]. The most common glaucoma type is primary open-angle glaucoma (POAG). This type is estimated to be affecting 2.2% globally. This means that approximately 57.5 million people are affected by it worldwide. In Europe, the POAG percentage is estimated to be 2.51%. This means that it affects around 7.8 million individuals.

Primary angle-closure glaucoma (PACCG) is another type of glaucoma, which is less common and affects only 0.17% of individuals younger than 40 years. Currently, the diagnosis of glaucoma needs eye examination and medical expertise, which might be costly or not available [23]. Furthermore, there is a significant amount of similarity in ocular characteristics between those without glaucoma and people who have the disease in its early stages [24,25].

Important instruments for the diagnosis of glaucoma include color fundus photography with red-free imaging, Heidelberg retinal tomography (HRT), and visual field (VF) tests [26]. In recent years, there is an increased interest in building AI systems for medical diagnoses, which can help with glaucoma screening and diagnoses. "DYSEO" is a web-based telemedicine program designed to test for glaucoma. The system combines Optical Coherence Tomography (OCT) and retinography in the detection of the glaucoma condition. The screening program accuracy was reported as medium to high [27]. The integration of artificial intelligence (AI) and machine learning (ML) algorithms into telemedicine screen-

ing programs can enhance the diagnostic accuracy. Much of the works in the literature related to glaucoma classification have faced challenges like a small number of images in datasets and lack of variations in images. All these challenges contribute to limiting the robustness and generalizability of developed models [28,29]. The creation of a model that can handle photos taken in a variety of environmental settings is essential to overcoming this limitation. In this work, a novel DL model that combines the features extracted from two of the most efficient DL models is proposed. The proposed model is built to classify four different eye diseases. A large dataset that is collected from different public and private sources is used to train the proposed model. The main objective of this robust model and huge dataset was enhancing the accuracy and reliability of retinal classification systems. The proposed system is evaluated against other DL models and other systems in the literature.

1.1. Research Motivation

In spite of the progress in different methodologies for classifying eye diseases from digital images, encompassing normal cases, diabetic retinopathy, hypertensive retinopathy, glaucoma, and contrast-related conditions, substantial challenges persist. Figure 1 shows images of different eye disorders.

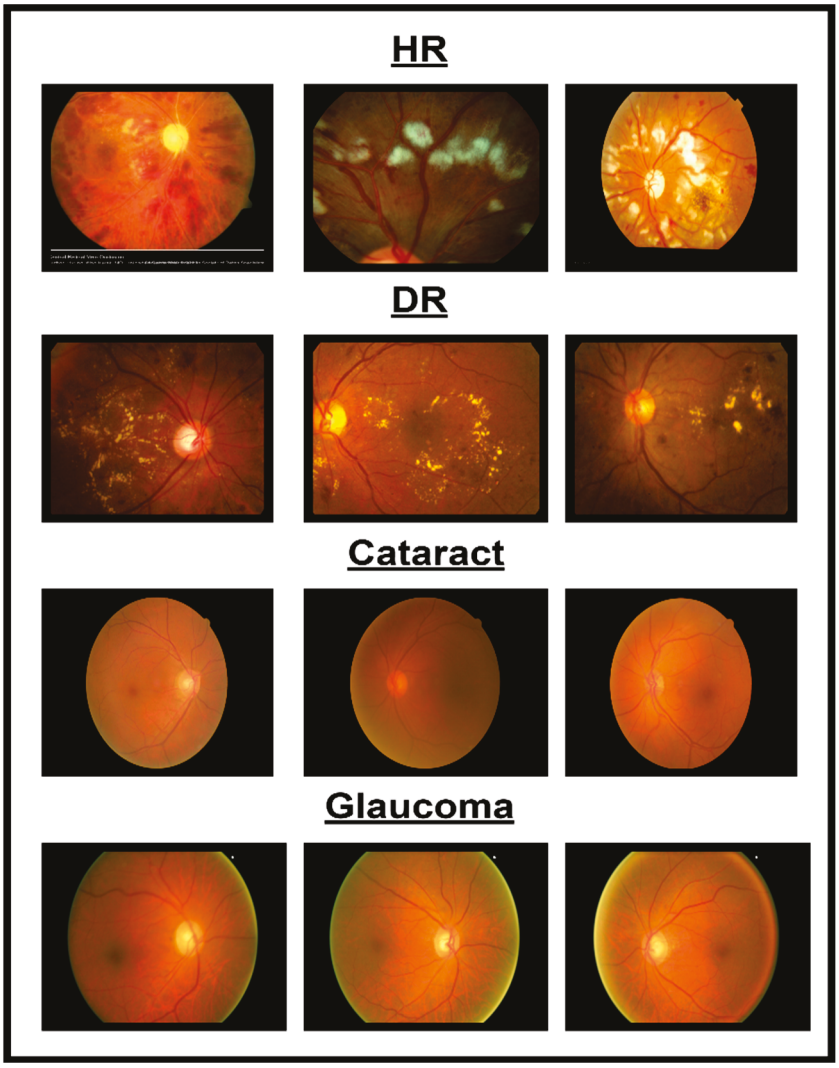


Figure 1. Illustration of eye disorders.

- Even with the utilization of sophisticated image processing technologies, accurately delineating features from images related to normal cases, diabetic retinopathy, hypertensive retinopathy, glaucoma, and contrast remains challenging. The challenges involved in precisely locating and extracting features associated with eye diseases contribute to this difficulty.
- A dataset combining photos from various eye illnesses such as hypertensive retinopathy, diabetic retinopathy, glaucoma, and contrast is not publicly available. Because of this, it is hard to implement an automated system to classify different eye diseases.

This study aims to develop an automatic system for the classification of different eye diseases. To achieve this goal, a first step is creating a dataset for four different diseases (diabetic retinopathy, hypertensive retinopathy, glaucoma, and contrast-related diseases) and a normal eye. The second step is to create a multi-layer deep learning (DL) model that can generalize to different eye diseases classes. The two steps combined produce a CAD-EYE system with multi-layered architecture. Through extensive training, this CAD-EYE system becomes proficient in classifying different eye diseases. Regarding the diseases, the analysis of the eye was performed by a skilled ophthalmologist.

1.2. Research Contribution

This work proposes a novel deep learning (DL) model to address the problem of recognizing four different eye diseases. Also, in this work, a new dataset was collected from reputable internet websites and a private collected dataset from previous research. The following steps summarize the significant contributions of the CAD-EYE system:

- In this work, the researchers compiled a substantial dataset consisting of 10,000 photos sourced from reputable internet platforms and supplemented by private datasets from previous studies. This large dataset was important since it enabled the model to achieve remarkably high classification accuracy.
- This work is the first to introduce a Fluorescence Imaging Simulation algorithm into multi-eye classification research. The image processing algorithm along with the proposed feature fusion model allowed the model to achieve higher accuracies than reported systems while classifying four different eye diseases.
- In this work, the feature fusion technique was incorporated to combine the features from two DL models to build the CAD-EYE system. The approach resulted in the creation of a multi-layered model that was successful in solving the classification problem.
- Additional layers (custom layers including dense layers following the feature fusion process, which help refine features extracted from both models) are added to the design of the CAD-EYE model to enable the model to classify a number of eye-related diseases. The convolutional neural network (CNN) models (MobileNet, EfficientNet) are used to extract features associated with eye disorders, and these features are subsequently combined through the feature fusion approach.
- The authors claim that this is the first effort to develop an automated system that is superior to current approaches in the identification of five different eye classes (normal, diabetic retinopathy, hypertensive retinopathy, glaucoma, and contrast-related eye disorders), as illustrated in Figure 1.
- Our systems exhibited superior performance compared to the approaches proposed in the available research, achieving a remarkably higher accuracy percentage of 98%.

1.3. Paper Organization

The structure of this paper is as follows: Section 2 includes the literature survey. Section 3 outlines the proposed model design. Section 4 presents the experimental results. In Section 5, our findings are compared with those of contemporary studies in the field. Section 6 conducts a comprehensive discussion of the research findings. Finally, Section 7 presents the study's conclusions.

2. Literature Survey

Much of the papers in the literature include the extraction of blood vessels or the partitioned location of lesions, and the instruments and strategies utilized are different, contributing to the complexity of framework design [30,31].

Daniel Shu and Wei Ting et al. emphasize that deep learning has gained considerable attention in recent years [32]. Despite that deep learning models were invented a long time ago, their impact on picture identification, natural language processing, and speech recognition started to appear mainly in the healthcare industry [33,34]. The application of deep learning (DL) in ophthalmology, specifically in fundus pictures, has facilitated the classification of conditions such as a glaucoma-like disk, macular edema, and age-related macular degeneration. In primary care and community settings, DL in ocular imaging and telemedicine enables the screening, identification, and follow-up of patients with serious eye problems [35,36]. However, there are still some challenges in using deep learning models in ophthalmology. These challenges include problems with clinical implementation, difficulty explaining the results for doctors and patients, and difficulties convincing doctors and patients to receive a diagnose from "black-box" models.

Ophthalmologists' manual diagnosis of DR retina fundus images is expensive, time-consuming, and prone to errors. Deep learning, particularly convolutional neural networks, has shown enhanced performance in the classification of medical images, offering a promising avenue for the diagnosis of DR and other medical conditions [37]. In research paper [38], novel deep learning models for diabetic retinopathy classification are investigated and examined. The analysis also includes an assessment of datasets specific to DR in color fundus retinas. Additionally, certain complex challenges have been identified, calling for further research. The proposed deep learning approach significantly enhances the detection of the disk and cup. However, because medium-sized cups were so common in the model's training, it has a tendency to overestimate tiny cups and underestimate large ones. The conventional approach to glaucoma identification is measuring increased intraocular pressure (IOP). In an earlier work [39], binary distinctions between participants with and without glaucoma were made using features that were calculated at the image level utilizing picture attributes. Three major layers are commonly seen in CNN designs, or convolutional neural networks. Numerous research studies have explored the identification of eye diseases, employing diverse methodologies and technologies crucial for enhancing the precision and efficiency of disease diagnoses in the field of ophthalmology. In a research paper published by Malik et al. [40], the emphasis was on applying artificial intelligence techniques in healthcare systems, particularly for fast diagnoses. They made use of machine learning models including a Random Forest, a decision tree, Naïve Bayes, and a neural network. These models were trained with varied data, including patient information, age, disease history, and clinical observations. In the work, it was found that Random Forest and decision tree algorithms demonstrated high accuracy rates that were better in comparison to previous research papers.

In [41], the researchers investigated the development of computer-aided diagnostic (CAD) systems for glaucoma detection. In the study, a convolutional neural network (CNN) was used to extract important features from retinal images. In the work, the objective was a binary classification between glaucoma and non-glaucoma using a dataset, utilizing and comprising 1200 retinal images. The model achieved a high accuracy of 99%, proving its ability to enhance glaucoma diagnoses. In [42], the authors built a model to differentiate individuals with eye disease and healthy people. The authors were able to train a model on datasets containing both patient and healthy individuals' retinal fundus images. The model results in high accuracy ranging from 96.5% to 99.7%, proving the effectiveness of their approach. In [43], authors addressed the critical problem of retinal disease, concentrating on the early diagnosis of diseases. The authors of the work used deep learning models incorporating CNN-based models. These models were used to classify different types of eye disorders. The model achieved average accuracy ranging from 81% to 94%, along with an average F1-score of 0.96 for normal retinas. Another work [44] focused on diabetic

eye disease (DED). The authors applied CNN methods for the classification of retinal eye diseases. The proposed model is trained to classify a range of DED classes; also, an ophthalmologist carefully examined a large dataset of retinal fundus images to verify the model. The overall accuracy (ACC) reached 81.33%, accompanied by 100% sensitivity and 100% specificity for multiclass classification, showcasing the potential of CNNs in detecting a range of diabetic eye diseases. Umer et al. [45] investigated the application of Optical Coherence Tomography (OCT) in the computer-aided detection and classification of retinal eye disorders. Currently, ophthalmologists rely on manually examining OCT images, a process prone to inaccuracies and subjectivity. To address this, the study introduced various methods to automate disease detection, making use of a four-class retinal eye disorder dataset that was made available to the public. To extract deep feature vectors, modified versions of the AlexNet and ResNet-50 models were used. With a remarkable overall average accuracy (ACC) rating of over 99.95%, the suggested strategy for identifying retinal eye illnesses demonstrated the potential for accurate and objective disease detection. Gargeya et al. [46] addressed the computer-aided diagnosis of diabetic retinopathy (DR) using image processing techniques and CNN models. Their study involved a substantial dataset comprising 75,137 available retinal eye images from different patients. The assessment of model performance uses the area under the receiver operating characteristic curve (AUC) as an evaluation metric, employing five-fold cross-validation. Notably, the model's AUC for diagnosing DR was 97%, while its sensitivity and specificity were 94% and 98%, respectively, highlighting its effectiveness in automated disease diagnoses. In summary, these studies collectively showcase diverse approaches and technologies in the field of ophthalmology for detecting and diagnosing a wide spectrum of eye diseases. They illustrated how automated diagnostic technologies, deep learning, and artificial intelligence might improve illness diagnoses' efficiency, objectivity, and accuracy. The authors of [47] tackle the crucial challenge of detecting and classifying diabetic retinopathy (DR), a significant concern for diabetic patients vulnerable to severe visual impairment. They introduce an innovative automated system named DR-NASNet, which employs advanced techniques, including preprocessing methods such as Ben Graham and CLAHE, data augmentation to address class imbalance, and the combination of dense blocks within the NASNet architecture. In addition to achieving impressively accurate state-of-the-art outcomes, the system keeps its model size small and complexity low. Through the utilization of combined datasets and a linear SVM classifier, DR-NASNet proficiently categorizes DR images into five severity levels. This breakthrough helps build automatic systems to help ophthalmologists and offers an efficient tool for the early classification of DR. This helps the early detection of eye diseases and to prevent vision loss. Study [48] also has a nice contribution in the early detection of hypertensive retinopathy (HR). In the work, authors introduced the Incept-HR system combined with the collection of the Pak-HR dataset, providing a new contribution to the application of deep learning models in healthcare. The Incept-HR system proved its efficiency and impressive performance, making it an important diagnostic tool. The proposed model was compared with established models like VGG19 and VGG16 and outperformed their performance. The work is considered a step forward in achieving more eye disease classification models. On the other hand, Ref. [49] introduced a novel diagnostic model for hypertensive retinopathy classification called "Mobile-HR". The model is based on fine tuning a MobileNet model using transfer learning. HR is a critical eye disease caused due to elevated blood pressure, resulting in various visual symptoms. Experimental results prove the effectiveness of the model in classifying HR disease in eye images.

The classification and investigation of cataracts necessitate a comprehensive understanding of their diverse manifestations within the eye lens. Both subjective observation and objective measurement techniques continue to be widely employed in this field. Notably, techniques such as a Scheimpflug slit image analysis, among the latter, offer a more precise means of identifying early transparency breakdowns. Objective approaches are pivotal in epidemiological research as they facilitate the accurate monitoring of risk variables, including UV-B radiation exposure, and their potential role in cataract development. This

significance arises from the fact that age-related changes in lens transparency occur before apparent opacifications. Longitudinal cohort studies, involving repeated examinations, are imperative for gaining deeper insights into the multifactorial processes associated with cataracts. Subjective evaluations alone may fall short in detecting minor changes in transparency. In summary, objective techniques for categorizing cataracts are indispensable for advancing our understanding of this vision-impairing disorder and the linked risk factors [50]. As a substantial contributor to visual impairment and a critical public health issue, cataracts are addressed through an automated identification method utilizing retinal image categorization rooted in computer science. Recognizing the crucial role of an early diagnosis in preventing blindness, our approach involves employing deep learning networks to extract distinctive characteristics from infected images, coupled with preprocessing using the maximum entropy approach. Subsequently, the automated identification of four classes of cataract photos—normal, mild, medium, and severe—is carried out using conventional classification techniques, specifically SVM and Softmax. Noteworthy in our findings is the demonstrated efficacy and utility of our study, emphasizing that the features obtained through deep learning and categorized by Softmax exhibit superior accuracy. Our overarching objective is to enhance the prospects for timely intervention and improved visual outcomes by advancing the early detection of cataracts through the fusion of computer science and medical imaging [51]. The healthcare challenge presented by uveitis, with its potential for blindness and societal impact, has become more addressable with advancements in machine learning technologies. Through our extensive examination, encompassing detection, screening, and the standardization of uveitis nomenclature, we have identified prospective roles for AI in uveitis research. It is crucial to acknowledge, however, that the current state of AI in uveitis diagnoses faces challenges, including overall subpar model performance, limited datasets, insufficient validation studies, and a lack of publicly available data and codes. Despite these obstacles, artificial intelligence holds significant potential in identifying and diagnosing ocular abnormalities associated with uveitis. Further research endeavors are essential, along with the creation of large, representative datasets for training and validation, to unlock its full potential and ensure its reliability and fairness [52]. From this literature survey, it is clear that there is a research gap in the area of the classification of eye diseases. In this work, a novel model is proposed for multi-eye disease classification. Also, a new dataset combined from previous datasets is introduced. This work is considered one step towards filling the research gap explored in this literature survey. Table 1 shows an overview of previous research articles related to the classification of eye illnesses.

Table 1. A comparison of the present literature.

Reference	Methodology	Datasets	Models
[29]	Classification was made by convolutional neural network	MESSIDOR, DIARETDB, STARE	CNN
[30]	Preprocessing of image then passing it to a convolutional neural network to predict if the patient is diabetic	High-Resolution Fundus (HRF)	CNN
[31]	Gray level co-occurrence matrices (GLCMs), classifiers like Random Forests, Support Vector Machine (SVM), gradient boost, AdaBoost, Naïve Bayes, and Gaussian	MESSIDOR	GLCM
[35]	Disk-aware ensemble network	SCES, SINDI	CNN
[37]	Random Forest classification using different features such as the perimeter and area of the blood vessels and hemorrhages	STARE	Filters and Random Forests

Table 1. *Cont.*

Reference	Methodology	Datasets	Models
[43]	CNN architecture	DiaretDB1	CNN
[44]	Deep belief network (DBN) and convolutional neural network (CNN)	DRIONS-DB, sjchoi86-HRF	CNN
[45]	Multiclass DED, auto-mated classification framework	MESSIDOR	CNN
[46]	Retinal illness classification based on selection utilizing Modified-Alexnet and ResNet-50 networks	-	AlexNet, ResNet50
[47]	DR severity detection and classification method based on enhanced pretrained NASNet Model	APTOS-2019, PAK-DR (Private)	NASNet
[48]	Using inceptionV3 model to detect and classify HR eye disease	PAK-HR (Private)	InceptionV3
[49]	A novel model is built based on the pretrained MobileNet architecture through the addition of dense blocks to make the network more efficient	PAK-HR (Private), DRIVE, DiaRetDB0	MobileNet

3. Proposed Approach

In this work, a novel system named CAD-EYE, combining the strengths of EfficientNet and MobileNet, is introduced. The CAD-EYE approach is deployed to categorize images of eye diseases, discerning between diabetic retinopathy, hypertensive retinopathy, glaucoma, contrast-related issues, and normal cases. Within the CAD-EYE system, the feature fusion methodology of EfficientNet and MobileNet is utilized to extract valuable features, employing transfer learning for training on eye-related abnormalities. The CAD-EYE system incorporates essential mechanisms for detecting images depicting eye diseases and recognizing the mentioned issues. The system steps are presented in Figure 2. Throughout the training phase, the parameters are continuously refined by fusing the characteristics from EfficientNet and MobileNet. To combine features using element-wise multiplication, a feature transform layer is added. Lastly, the application of an XGBoost classifier increases the classification results.

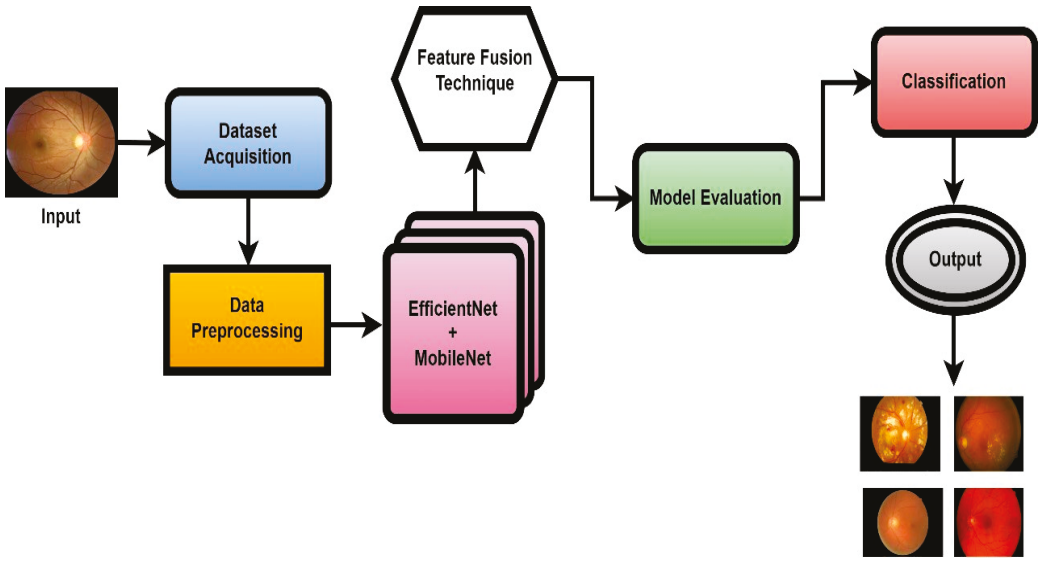


Figure 2. The systematic flow chart of the CAD-EYE system of eye disease classification.

A noteworthy advancement in the areas of machine learning and computer vision is the creation of the CAD-EYE model. This inventive methodology leverages the capabilities of two well-established architectural paradigms, EfficientNet and MobileNet, culminating in a unique hybrid model with the potential to surpass its predecessors. The proposed model combines the effective feature extraction of EfficientNet with the strong training stability and transferability of MobileNet; this combination has the potential to increase the model performance and can be applied to a variety of applications. This model can be adapted to different applications; this is considered as a distinctive feature of the model. Another advantage of this model is its ability to work in different applications with medium resources, making it suitable for limited-resource applications. This proposed model is used in different applications and proved superior performance over traditional methods. The feature extraction of both models is enhanced because components such as MB-Conv layers and SE blocks in EfficientNet are combined with inverted residual blocks in MobileNet.

3.1. Data Collection and Preprocessing

The proposed model (CAD-EYE) was trained and tested using the Multiple_EYE dataset that was collected especially for this work. This dataset contains 65,871 photos. These images were collected from different private and public sources. The sources included internet platforms and ophthalmic treatment facilities. In the collection processes, the Maestro2 OCT-Fundus Camera system is used, which is a user-friendly device that performs automatic alignment, focus, and capture with a single touch. This system enables a detailed analysis of the macula, optic disk, and anterior segment, with reports that can be automatically exported, printed, or integrated with EMR systems in common file formats. Explicit consent to use the data in research activities was acquired from patients and their doctors, with a promise of confidentiality of the data. This ensures accessibility of the data for other researchers while protecting the patients' privacy. The collected dataset (Multiple-Eye) contains eye images pertaining to different eye conditions like diabetic retinopathy, hypertensive retinopathy, glaucoma, and contrast, and normal images. The training dataset was divided into normal class and eye disease photos by qualified ophthalmologists. The ophthalmologists' experience in detecting eye-related traits was quite helpful in dividing the dataset. It used the Fluorescence Imaging Simulation preprocessing approach to enhance the image and make the feature extraction process more accurate. The technique strengthens the green channel of the image, imitating fluorescence, which is critical in some pathologies of eye fundus images. Fluorescence Imaging Simulation increases the visibility of these regions to clearly explain where the model is looking while deciding. More generally, the integration of the Fluorescence Imaging Simulation technique reflects a continued effort to increase transparency and thus reliability of deep learning models for improved accessibility and trustworthiness in numerous applications, as shown in Figure 3.

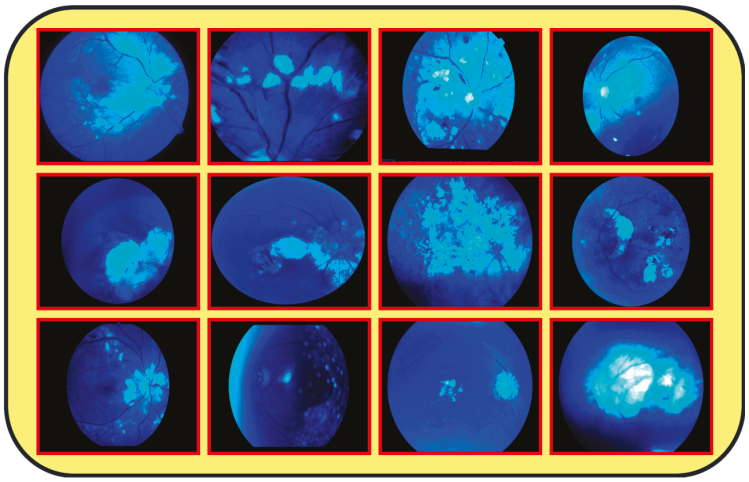


Figure 3. Results from image processing algorithm.

Figure 1: A detailed breakdown of the 65,871 eye photos involved in the research study. Table 3: Three datasets used to create the training and testing sets of fundus images, where different configurations are involved for every dimension. All images used in the experiments were uniformly resized to 700 × 600 pixels with their corresponding binary labels generated. This image size was chosen to balance the level of details needed from each image and the computational cost needed by deep learning models. In total, there were 65,871 photos within the dataset, and 9393 of the photos were used in order to evaluate the system. To make the dataset fair, proper classification into different classes was performed, and the number of images in each class before and after the onset of disease was balanced. The photos were preprocessed before feeding into the algorithm devised for the CAD-EYE model, which included resizing to 700 × 600 pixels. Standardization was further used to reduce variation across data points. The CAD-EYE system was trained and tested with the help of the dataset called Multiple-EYE, including images that were originally saved at a resolution of 1125 × 1264 pixels. Table 2 effectively simulates fluorescence in eye fundus images, outlining areas of interest and hence improving the interpretability of medical imaging. The images, sourced from three distinct origins, were resized to standard dimensions of 700 × 600 pixels for the purpose of the simplification and standardization of the dataset. To enhance image features and eliminate interference, the Grad-Cam technique was applied to preprocess the images, as depicted in Figure 3. Employing Grad-Cam on eye fundus images facilitated the identification of important areas and their importance in identifying the presence of eye diseases. This method helped identify distinguishing features that had a major impact on CNN's predictions on the diagnosis of glaucoma, diabetic retinopathy, hypertensive retinopathy, and contrast images.

Table 2. Fluorescence Imaging Simulation.

Step	Explanation
Input Image Acquisition	Load the RGB image of the eye fundus from the dataset. All the images are normally captured with a high resolution and full color, for instance, 1125 × 1264 pixels.
Channel Separation	Separate the RGB image into its three color channels: red, green, and blue. Each channel is considered to be an individual gray-scale image representing the intensity of that color.
Green Channel Enhancement	Strengthen the green channel by scaling the intensity values with the formula $G' = \alpha \times G$. Here, G is the source of the green channel, and α the scaling factor, which often takes a value of 2.0 to amplify the green's intensity.
Image Reconstruction	Reconstruct the RGB image by combining the original red and blue with the enhanced green channel. I'' is given by I' = merge (R, G', B), where R and B are the original red and blue channels.
Blue Channel Suppression	This will turn off the blue channel to B = 0, which might allow the fluorescence effect to be more apparent because of the loss of the blue component but strongly showing the green fluorescence.
Colormap Application (Optional)	Just apply a colormap—for example, 'HOT'—to emphasize the fluorescence more: it will just stretch the color palette of the image towards the warm end to show strongly bright green areas.
Output Image	Save or show the final processed image. This will shift the color scheme of the outputted image to accentuate green fluorescence and therefore pinpoint the areas of interest.

3.2. Data Augmentation

It is clear from Table 3 that there is a biased distribution of data. This bias will cause the classification model to favor a certain class over others. Investigating different approaches for balancing the data is critical to achieve more generalized models. To address this challenge, usually "data augmentation" techniques can be applied on the data to balance the data. Data augmentation generates extra data points from existing data,

hence enhancing the dataset's diversity. Alongside a data augmentation technique, other techniques may be applied to solve the data balancing problem. These techniques include resampling methods like oversampling and undersampling. Also, other procedures can be applied like boosting, bagging, or GANs. Each of these procedures has its own advantages and can be tuned to the specific application. Also, other techniques like bagging, boosting, and GANs may be considered. Each of these techniques has its own advantages and can be tuned to the specific dataset used. Data augmentation is usually performed using a simple algorithm that does not require a deep learning mode or tuning of the hyperparameters. This algorithm or procedure involves different transformation steps that include rotation, flipping, and cropping in order to create new samples to be added to the training dataset. The data augmentation procedure usually creates balanced classes with no majority and minority classes. In this work, we apply data augmentation to the training dataset only; this is to enhance the model generalization without affecting the evaluation of the model.

Table 3. Collected Datasets of Eye disorders used to train the CAD-EYE system.

Ref.	Datasets	Normal	Diabetic Retinopathy	Hypertensive Retinopathy	Glaucoma	Contrast	Total
[53]	Eyepacs, Aptos, Messidor	23,125	23,125	-	-	-	46,250
[54]	Eye disease dataset	-	-	-	50	100	150
[55]	Eye disease classification	250	250	-	250	250	1000
[56]	Dataset for different eye diseases	1637	-	-	1637	1637	4101
[57]	DiaRetDB1	100	100	-	-	-	200
Private	PAK-HR	3000	-	3000	-	-	6000
Private	DR-Insight	1000	-	4000	-	-	5000
Private	Imam-HR	1130	-	2040	-	-	3170
-	-	30,242	23,475	9040	1937	1987	65,871

By using the AutoAugment approach to automatically enhance b images during the training phase, DL models can be made more generic. The first step in this AutoAugment technique is creating the transformation pipeline and choosing the transformation policy. This transformation pipeline includes image resizing, horizontal flipping, and the chosen AutoAugment policy. The next step in this AutoAugment technique is loading the dataset that is organized in another directory. The Dataloader is used to load the dataset that needs augmentation. On the other hand, the ImageFolder class is used to apply the AutoAugment transformation to each image. The DataLoader is used during training to provide batches of augmented photos from the custom dataset. This step increases the performance of the model. The steps of the Auto Augmentation algorithm are shown in Table 4.

Table 4. Auto Augmentation Algorithm steps.

Step	Number
1	Importing the necessary python libraries (albumentations, torchvision, and torch).
2	Write the code for the get_autoaugment_transform() function. The AutoAugment policy and other augmentation settings are set up by this function.
3	The main code will have the following functionalities: a. Load the dataset for augmentation. b. After the dataset has been loaded, apply the already defined AutoAugment transformation.
4	This step uses torch.utils.data to create the DataLoader class. During training, the shuffled batches of the augmented dataset will be generated by the DataLoader.

Table 4. *Cont.*

Step	Number
5	Define a basic CNN model class. With nn.Module, SimpleCNN is used. For the model's training, configure the optimizer (like SGD or Adam) and loss function (like CrossEntropyLoss). Indicate the batch size and learning rate, among other hyperparameters, as well as the quantity of training epochs.
6	The following steps are repeated to train the model for a given number of epochs: a. Put the model in the training mode first. b. Call the DataLoader to generate batches of images from the augmented images along with their labels. c. Run the model in the forward pass mode. d. A defined loss function is used to compute the difference between reference labels and expected outputs. e. The computation of the model's parameter gradients with respect to the loss is performed using a backward pass. f. Tweak the model's parameters with the selected optimizer in step number one and use computed gradients. g. The model can be evaluated now on a separate validation dataset during training.
7	The last step is the evaluation of the model on a new dataset.

4. Proposed Architecture

The proposed architecture combines the advantages of MobileNetV2 and EfficientNet using a well-designed feature fusion process. This leads to a final layer that is compatible with the XGBoost classifier. This means that the pretrained convolutional neural networks (MobilNetV2 and EfficientNet), which are known for their advantageous characteristics, serve as components in this novel model. In this study, features were extracted from the final convolutional layer of both MobileNetV2 and EfficientNetB0 models. This choice was made because the final convolutional layer provides high-level feature representations that are well suited for the classification task. The proposed fusion methodology extracts and harmonizes features from both architectures. This fusion methodology combines global average pooling and concatenation. In this work, element-wise summation was used to combine features extracted from both models as it effectively combines features while preserving spatial consistency. The subsequent dense layers further process the features and extract relationships. In this work, all features that were extracted from the final convolutional layer of MobileNetV2 and EfficientNetB0 were utilized as is, without additional feature ranking or selection. This is carried out to preserve all features to contribute to the classification accuracy. The final element in the architecture is the XGBoost classifier, which extends the model beyond CNNs. The proposed hybrid architecture combines deep learning models, offering a sophisticated ensemble approach for image classification tasks. The model benefits from combining the advantages of both deep learning models and high accuracy of the XGBoost model to achieve accurate classification results as illustrated in Figure 4.

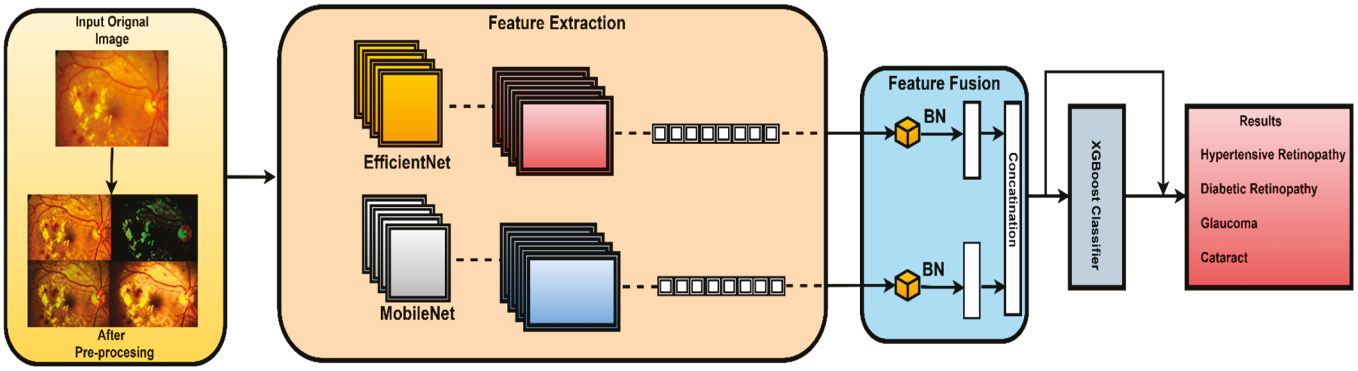


Figure 4. Architecture of the CAD-EYE system.

Let M represent the MobileNetV2 model, and E represent the EfficientNet model.

$$MobileNetV2\ features : X_M = M(Input) \quad (1)$$

$$EfficientNetB0\ features : X_E = E(Input) \tag{2}$$

Make both models' layers non-trainable.

$$X_M^{fiozen} = Freeze(X_M) \tag{3}$$

$$X_E^{frozen} = Freeze(X_E) \tag{4}$$

Let Dense () stand in for a dense layer and GlobalAveragePooling2D() for the global average pooling technique.

$$X_M^{Pooled} = Global\ AveragePooling2D(X_M^{frozen}) \tag{5}$$

$$X_E^{Pooled} = Global\ AveragePooling2D(X_E^{frozen}) \tag{6}$$

$$X_{concat} = Concatenate\ (X_M^{Pooled},\ X_M^{Pooled}) \tag{7}$$

$$X_{dense1} = Dense\ (128, activation = 'relu')\ (X_{Concat}) \tag{8}$$

$$X_{dense2} = Dense\ (num_{classes}, activation = 'softmax')\ (X_{dense1}) \tag{9}$$

$$fusion_{model} = Model\ (inputs = [M.input, E.input], outputs = X_{dense2}) \tag{10}$$

The proposed model architecture starts by using the pretrained MobileNetV2 and EfficientNet models, respectively, to extract features from the input images. After that, in order to preserve their learnt weights, all layers in both models are made non-trainable. The frozen features are then subjected to global average pooling, yielding X_MPˆooled and X_EPˆooled.

These features are then concatenated to form X_Concat. This vector uses, as a base, a dense layer with rectified linear unit (ReLU) activation, producing two new vectors, X_dense1 and X_dense2. Finally, the fusion model is defined with two inputs (from MobileNetV2 and EfficientNet) and one X_dense2. Subsequent to model definition, the architecture uses Adam as an optimizer and categorical cross-entropy loss. The model uses the SoftMax algorithm to transform output scores into probabilities; this is essential for the XGBoost model to work effectively. The architecture also uses accuracy as the monitoring metric. The fusion model is fitted to the training dataset for a specified number of epochs. Training yields a finely tuned model that incorporates the advantages of both MobileNetV2 and EfficientNet for image classification tasks. Figures 5–8 show the infection images of the CAD-EYE system. Algorithm of the model is presented in Table 5.

Table 5. Steps of the CAD-EYE feature map extraction Algorithm.

Step	Operation		Explanations
1	Load trained Models	Pre-	Extract features from images using EfficientNetB0 and MobileNetV2.
2	Freeze ers	Lay-	Freeze the models trained by weight and transform the models to non-trainable ones.
3	Global Average Pooling		$X_M^{Pooled} = Global\ AveragePooling\ 2D(X_M^{frozen})$ Apply global average pooling to both MobileNetV2 features and EfficientNetB0 features.
4	Concatenation		$X_{Concat} = Concatenate\ (X_M^{Poolea}, X_E^{Poolea})$ Concatenate pooled features.
5	Dense ers	Lay-	$X_{dense1} = Dense(128, activation = 'relu')(X_{Concat})$. Apply the dense layer with ReLU activation. $X_{dense2} = Dense(num_{classes}, activation = 'softmax')(X_{dense1})$. The final dense layer for classification.
6	Fusion Model Definition		Combine the features extracted from MobileNetV2 and EfficientNet models to train the model.
7	Compile the Model		Using the compile function, we compiled the model. The function parameters include the metrics used in evaluating (accuracy) and the optimizer (Adam).
8	Train Model	the	Using the fit function, the model is trained for 10 epochs. In this instruction, we specify for the fit function the training, validation, and test data arrays.

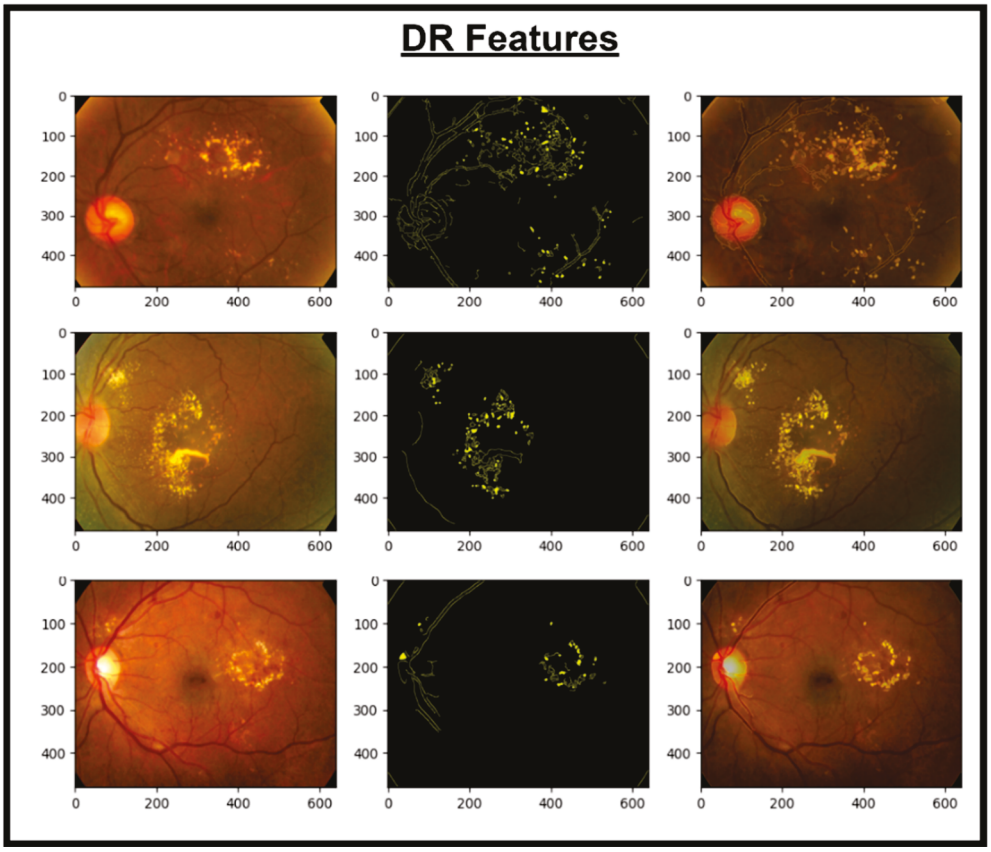


Figure 5. Images and DR extracted features.

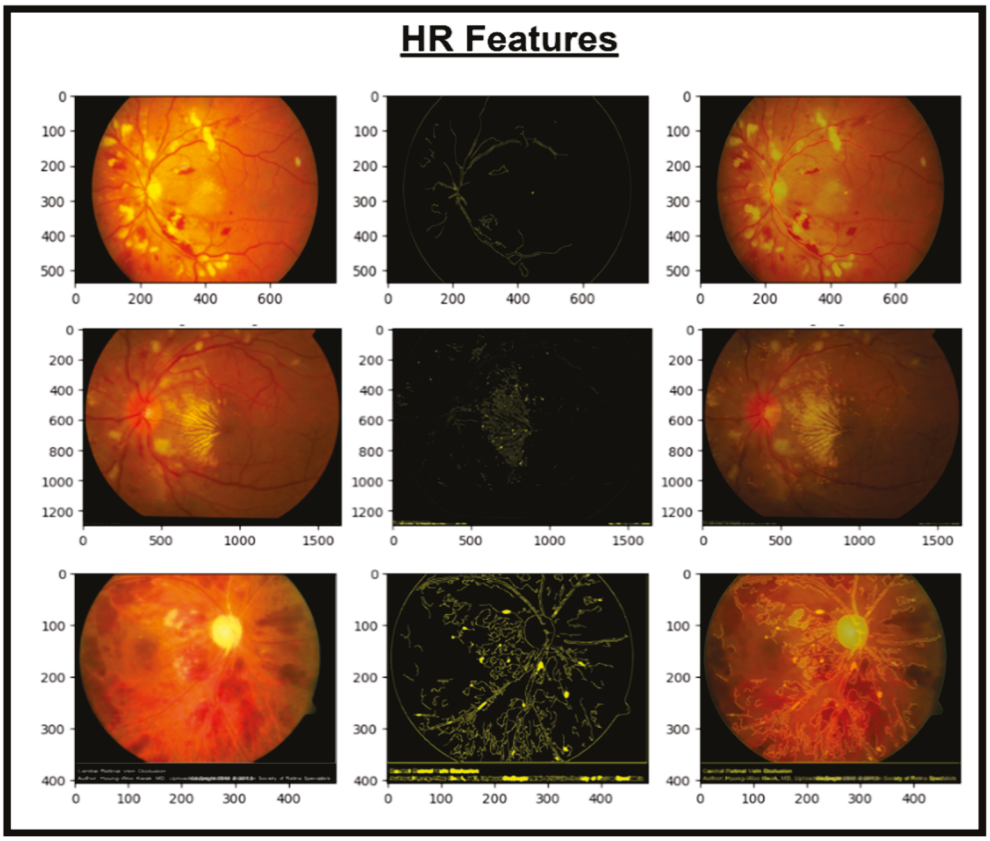


Figure 6. Images and HR extracted features.

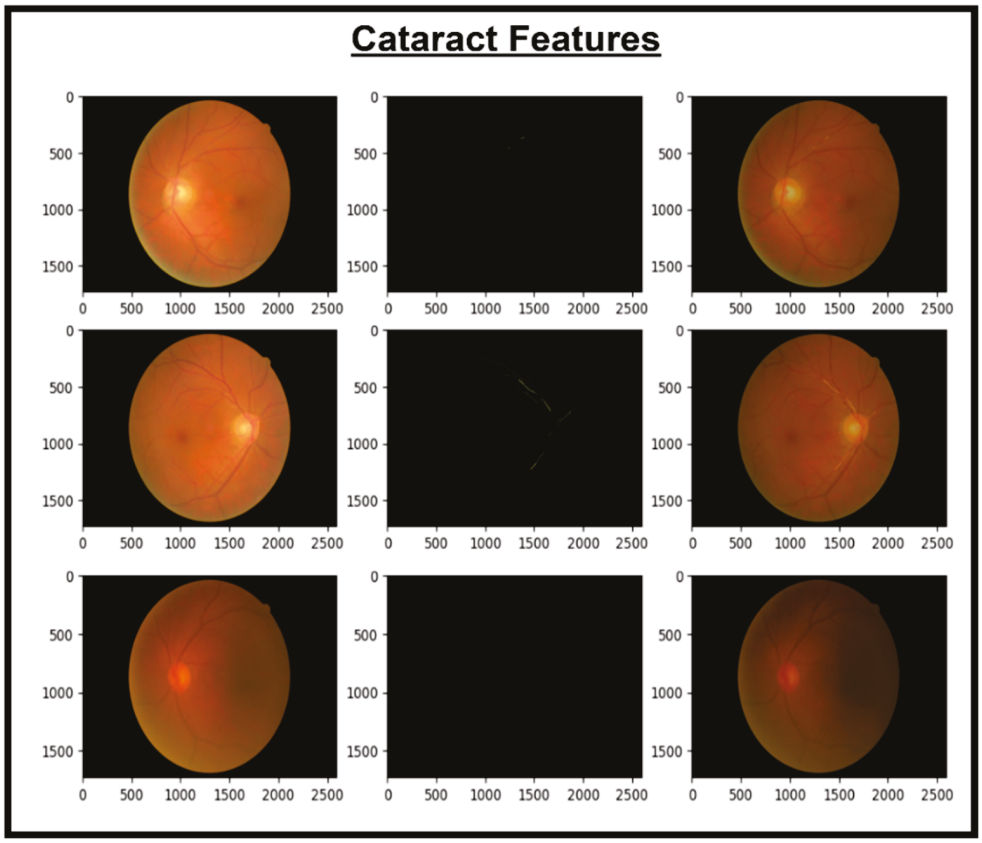


Figure 7. Images and cataract extracted features.

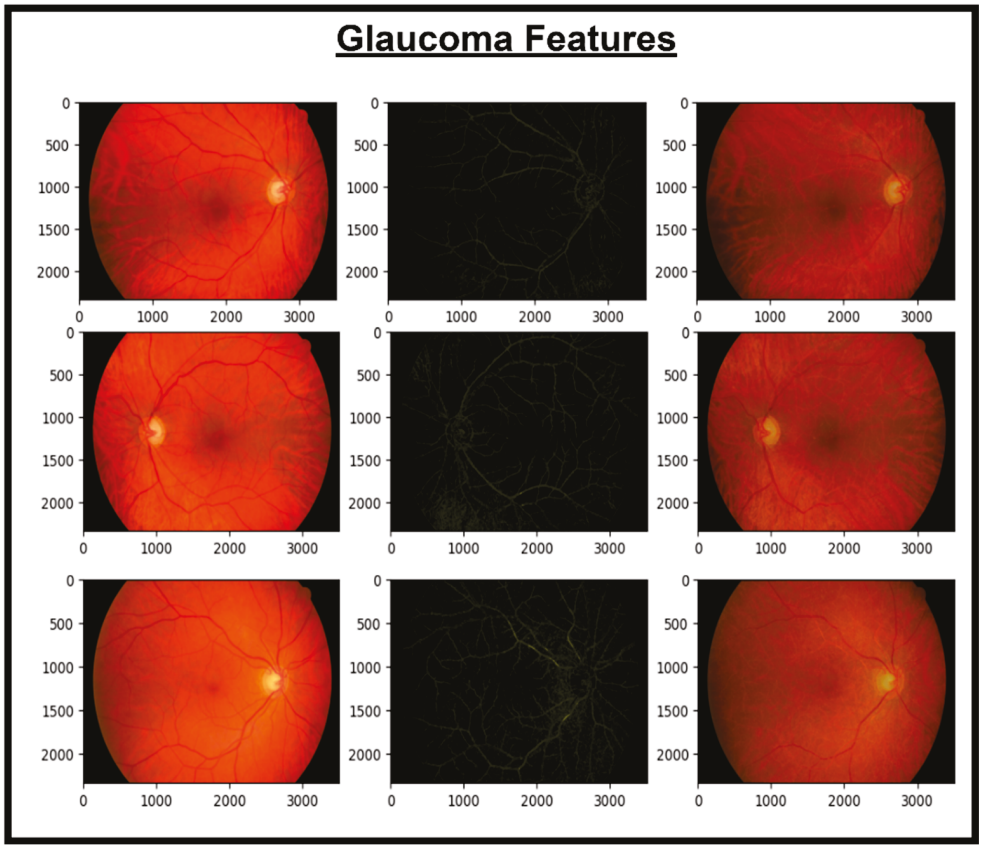


Figure 8. Images and glaucoma extracted features.

5. Recognition of Eye Diseases

In an effort to automate the challenging process of diagnosing eye conditions using fundus photos, our novel fusion system, which we named CAD-EYE-Fusion, cleverly combines the key features of MobileNetV2 and EfficientNetB0. Illustrated in Figure 4, this creative design combines layers like global average pooling, dense, and batch normalization layers while skillfully combining features from both models using a mathematical fusion mechanism. Expertly designed skip connections accelerate network learning. The comprehensive CAD-EYE-Fusion system greatly improves image classification performance for multi-eye diseases by leveraging the integrated feature representation. The architecture boasts multiple dense blocks, constructed with depth-wise convolutional layers, max pooling, ReLU activation, and batch normalization, forming the core of CAD-EYE-Fusion. These dense blocks have skip connections within them that make training and connecting more effective. Comprising a total of three dense blocks, our model ensures consistency in input and output sizes, enabling effective feature learning. An extra layer—which includes layers like dense and batch normalization layers—that customizes the classification procedure provides the final classification outcome while preserving 850 nerve cells for peak performance. As a preprocessing step, batch normalization layers are smoothly integrated into CAD-EYE-Fusion, improving the model's training and convergence. The detailed incorporation of batch normalization is provided in Table 6, emphasizing its significance in the comprehensive fusion system designed for robust eye disease classification from medical images.

$$B = \{X_{1\cdots m}\}, \gamma, \beta \tag{11}$$

$$B = \{y_i = BN_{\gamma * \beta}(X_i)\} \tag{12}$$

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} X_i \tag{13}$$

$$\sigma_B \leftarrow -\frac{1}{m}\sum_{i=1}^{m} X_i - \mu_B \tag{14}$$

$$X_i - \frac{X_i - \mu B}{\sigma_B^2 + \in} \tag{15}$$

Table 6. Notation table.

Techniques for Augmentation	Values
B	batch
X	batch minimum activating value
μ_B	mini-batch mean
σ_B^2	mini-batch variance
ϵ	constant added for numerical stability
β	learning parameter
γ	learning parameter

6. XGBoost Classifier

In this study, the effectiveness of the XGBoost algorithm for the task of identifying features linked to glaucoma, hypertensive retinopathy, diabetic retinopathy, and contrast is investigated. The reason for using XGBoost was its superior performance in handling complex, high-dimensional data, which aligns well with the feature-rich representations extracted from the MobileNetV2 and EfficientNetB0 models. The algorithm, outlined in detail as Table 7 employs a gradient boosting framework with decision trees as base learners. In this work, the XGBoost model was customized through hyperparameters like

the learning rate (η), regularization term (*lambda*), and the number of trees (τ) because it is especially well suited for binary classification jobs. For the feature extraction process in the realm of computer vision, we adopted depth-wise Conv2D, providing a more specialized convolutional operation. This modification was added to improve the model's capacity in extracting complex patterns in retinal images.

Table 7. Proposed XGBoost classifier algorithm.

Step	Description	Input	Output
1	Initialize XGBoost model	-	XGBoost model with hyperparameters (η, λ, τ)
2	Train XGBoost model on normal and abnormal samples	Training data: X = $(x_1, a_1), (x_2, a_2), \ldots, (x_m, a_m)$ Labels: a = 0, 1	Trained model XGBoost
3	Use depth-wise Conv2D instead of Conv2D	Feature map x = $a_1, a_2, a_3, \ldots, a_n$	Modified map feature
4	Build classifier based on XGBoost • Train XGBoost model • Generate ensemble of trees • Training data: X, Labels: a	Trained XGBoost model and ensemble of decision trees	-
5	Allocate class label for testing samples	*Testing data*: x $xtest1, xtest2, xtest3, \ldots, xtestk$ =	Predicted class labels for X_{test}
6	Output identification of normal retinographic samples and hypertensive retinopathy (HR)	Predicted class labels	Recognition result for diabetic retinopathy, hypertensive retinopathy, glaucoma, and contrast samples

During the training phase, decision trees are built iteratively, and the model is updated by updating the computed gradients (gi) and Hessians (hi) in every training sample. Additionally, the t-th tree prediction is gradually introduced to the ensemble throughout the model training phase of the XGBoost model, with the learning rate and weights established by the optimization process serving as a guide. All of the ensemble's trees' combined predictions make up the final prediction for a test sample, resulting in a robust recognition model. Mathematically, an objective function that balances a loss term and a regularization term is optimized by the XGBoost algorithm. The output for a testing sample (Atest) is computed as the sum of the predictions from each tree, weighted by the learning rate and tree weights. By using an ensemble technique, XGBoost is able to identify intricate relationships within the data and produce a dependable identification result for the categorization of diabetic retinopathy, hypertensive retinopathy, glaucoma, and contrast samples.

7. Results

A dataset comprising 65,871 fundus images, encompassing high-resolution normal, diabetic retinopathy, hypertensive retinopathy, glaucoma, and contrast images, was used to train the CAD-EYE system. These images of eyes have been gathered from a number of trustworthy websites. All 65,871 photos were scaled to 700 × 600 pixels in order to facilitate feature extraction and classification tasks. The CAD-EYE model was trained for 100 epochs. The proposed model evaluation metrics were calculated using a statistical analysis. The proposed model achieved an outstanding F1-score of 0.99. These measured metrics were used to compare the model with other models from the literature. The model was trained and developed on a PC equipped with 2 GB Gigabyte NVIDIA GPU 8 cores.

7.1. Experiment 1

In the first experiment implemented to test the proposed DL model, a comparison is made between the proposed model and other deep learning models like DenseNet-169,

InceptionV3, VGG16, and MobileNet. All models were trained for an equal number of epochs. Comparative findings for accuracy, specificity, sensitivity, and F1-score between the CAD-EYE system and the MobileNet, VGG16, VGG19, Xception, InceptionV3, and ResNet models are shown in Table 8. The outcomes show that the CAD-EYE has better performance than other DL models. Figure 9 shows a graph that compares various DL models with the CAD-EYE model.

Table 8. Comparison between different DL models using performance metrics.

Models	Sensitivity (SE)	Specificity (SP)	Accuracy	F1-Score
VGG16	68%	64%	69%	68.8%
VGG19	70.9%	70.8%	71.1%	71%
InceptionV3	73.2%	74.5%	75%	74.7%
ResNet50	83.5%	82%	80.2%	80.8%
Xception	81.7%	80.4%	81.2%	81%
MobileNet	89.5%	88.9%	89.4%	89%
DenseNet-169	89.9%	90%	90%	90%
EfficientNet	88%	79.8%	88%	88%
CAD-EYE	97.3%	97.9%	98%	98%

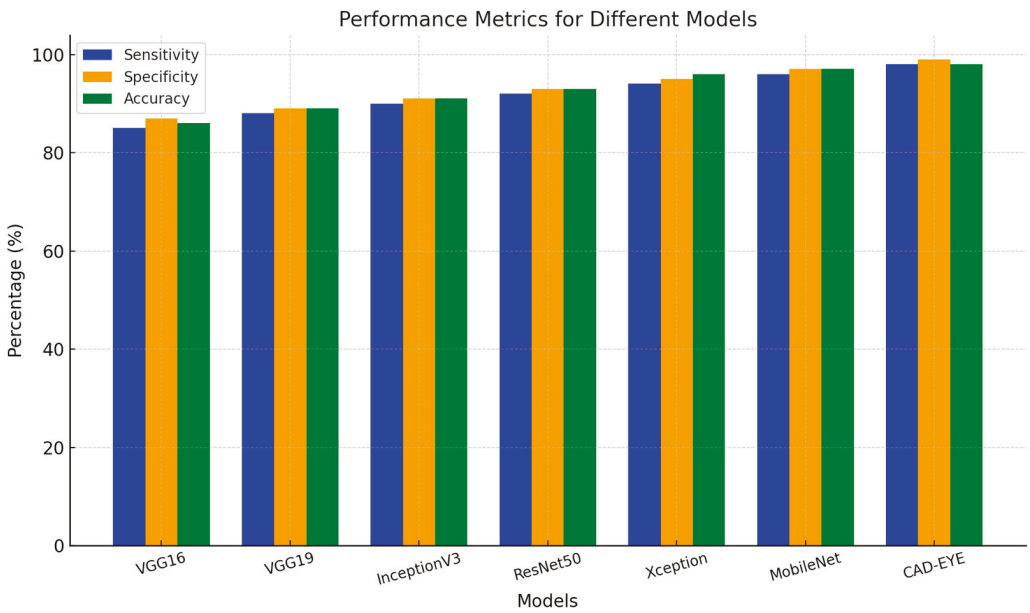


Figure 9. Comparison between some DL models and CAD-EYE.

7.2. Experiment 2

In this experiment, a dataset of 65,871 fundus images obtained from different reputable online sources is used to test the performance of our proposed CAD-EYE system. Figures 10 and 11 show accuracy and loss versus the epoch for the CAD-EYE model using training and validation datasets. The results demonstrate the high efficacy of CAD-EYE in both training and validation scenarios.

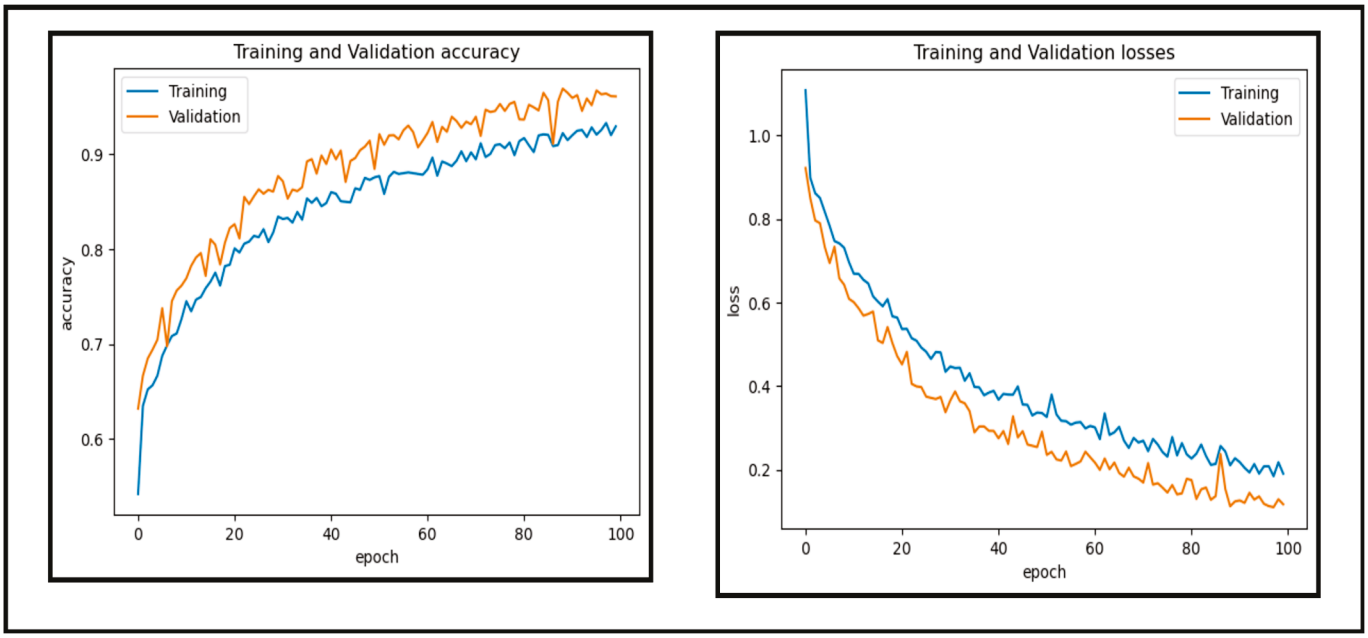


Figure 10. Accuracy and loss versus epoch for training and validation data.

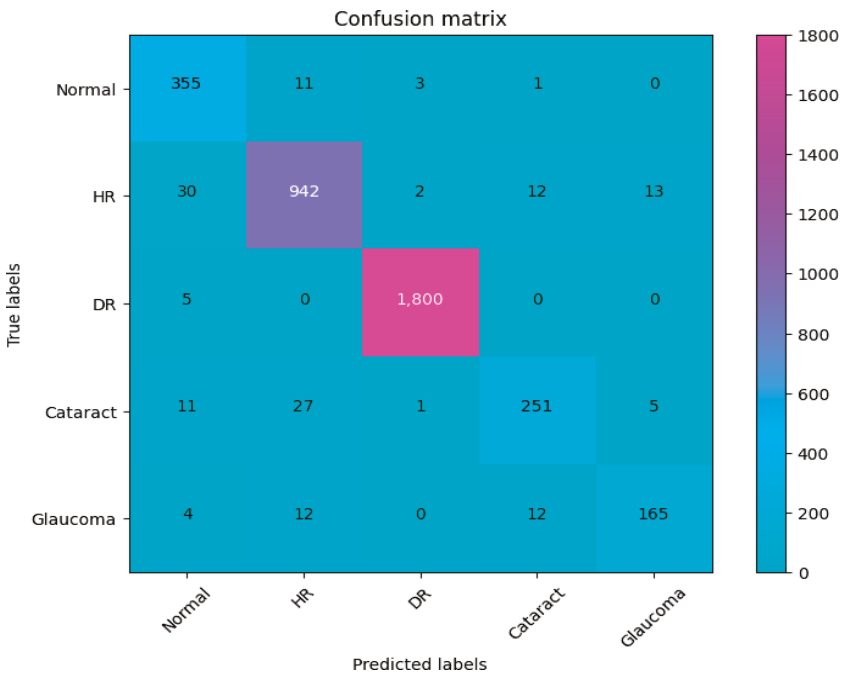


Figure 11. Confusion matrix of CAD-EYE.

7.3. Experiment 3

In this work, we assessed the efficiency of our proposed CAD-EYE technique using the eye disease classification (EDC) dataset [56] that was investigated. First, we looked at the loss function and analyzed how well the model performed using the EDC dataset on the training and validation sets. The training and validation accuracy of the CAD-EYE model trained on this dataset can be seen in Figure 12. Results show how exceptionally successful our model is in both training and validation scenarios. Furthermore, by utilizing the EDC dataset outlined in Table 9, we achieved a remarkable 99% accuracy on the validation and training datasets.

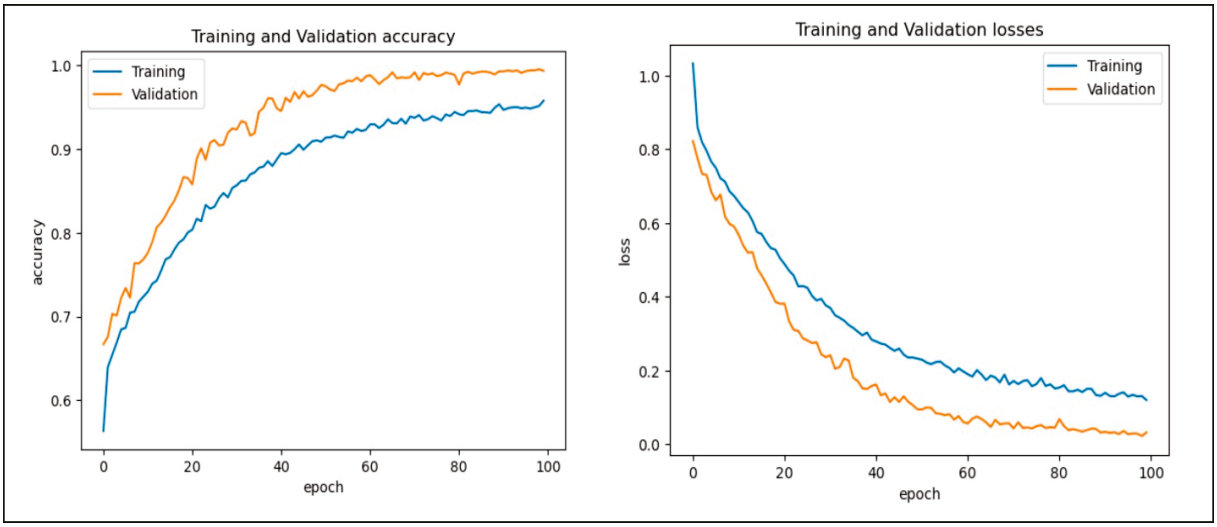


Figure 12. Accuracy and loss versus epoch for training and validation data using EDC dataset.

Table 9. Performance assessment of EDC Dataset.

Model	Dataset	Sensitivity (SE)	Specificity (SP)	F1-Score	Recall	Accuracy
CAD-EYE	EDC	99.50	99.68	99.95	99.98	1.0

8. State-of-the-Art Comparison

Only a few studies investigated using deep learning algorithms to diagnose eye conditions using retinal pictures. Among these, the EDC research [58] stands out for utilizing deep learning with a small dataset to detect normal, cataract, glaucoma, and diabetic retinopathy cases in retinal images. The latest deep learning model addressing normal, cataract, glaucoma, and diabetic retinopathy detection is referred to as EDC [58]. In contrast, the CAD-EYE system we developed demonstrated outstanding outcomes, achieving values of 99.50%, 99.68%, 99.98%, 99.95%, 99.98%, and 1.0 for SE, SP, F1-score, Recall, and ACC, respectively. In EDC, the author of [58] noted that they utilized a very limited set of input fundus images for training, resulting in high precision and accuracy. However, it is crucial to mention that their dataset lacked approval from expert optometrists. In our case, the CAD-EYE system underwent testing and training on a dataset of 65,871 images, which received validation from expert optometrists. Consequently, we achieved a classification accuracy of 98%, representing significant advancement above the state-of-the-art work that is currently available. Table 10 presents a detailed performance comparison between the proposed model CAD-EYE and EDC model.

Table 10. This table reveals that CAD-EYE outperforms EDC [58] with superior performance.

Dataset	Sensitivity (SE)	Specificity (SP)	F1-Score	Recall	Accuracy
EDC Model [58]	98.7	96.3	98.7	98.7	99.4
CAD EYE	99.5	99.68	99.95	99.98	1.0

9. Discussion

The eye is a crucial element for a person's work and activity. There are many illnesses that have an impact on how effectively the eyes work. Effective treatment of these diseases depends on early discovery. This article proposes a system for classifying various eye illnesses. The four different illnesses that the proposed model can identify in eye images are diabetic retinopathy, hypertensive retinopathy, glaucoma, and contrast-related eye diseases.

This work combines different innovative ideas to achieve the goal of the classification of four different eye diseases. This work is the first to propose the use of a fluorescence imaging algorithm in multi-eye disease classifications. This image processing algorithm highlights important areas. The regions that are critical for classification are the focus of this image processing approach. Secondly, the authors suggest combining features extracted from two of the cutting-edge models to create a new feature vector. Also, a dataset is produced by merging several datasets obtained from different internet resources. This step is necessary since the model was trained on classifying four different diseases. The proposed model undergoes three different experiments to prove its superior performance. The first experiment is performed to evaluate the suggested model's performance against other cutting-edge models (VGG16, VGG19, InceptionV3, ResNet, Xception, MobileNet). The presented results show the superior performance of the proposed method in comparison to these models. The results show that CAD-EYE was able to achieve 97% accuracy, which is higher than the best other models by 10%. In reviewing the confusion matrix in Figure 11, we observe that the model achieves higher accuracy for diabetic retinopathy (DR) cases, while normal eye images are occasionally misclassified. This variation in performance may stem from the distinct features associated with DR, which are more easily identifiable by the model compared to the subtle variations in normal eyes. The features of a normal eye may overlap with mild signs of disease or noise, leading to occasional misclassification. The second experiment conducted in this research is testing our proposed model using a dataset created by combining images from different online sources. The dataset contains 65,871 fundus images. The dataset contains high-resolution normal and diabetic retinopathy, hypertensive retinopathy, glaucoma, and contrast eye disorder images. The findings demonstrate that the suggested model can achieve accuracy levels greater than 95%. The third experiment conducted in this research is testing our proposed model against the EDC (Kaggle's eye disease classification) dataset. It is noted that the EDC dataset contains a very limited set of input fundus images for training. However, it is crucial to mention that the EDC dataset lacked approval from expert optometrists. Our collected dataset for the CAD-EYE system consists of 65,871 images, which received validation from expert optometrists. The experiment is necessary to evaluate the model against similar work from the literature. The enhanced performance of the suggested system is demonstrated by a state-of-the-art comparison with a comparable work from the literature. Exceeding published accuracies in the literature, the suggested approach managed to attain better accuracy levels. In summary, the classification strategy for eye illnesses presented in this study is based on the fusion of characteristics and deep feature extraction. To the best of the authors' knowledge, this is the first attempt to develop an automated system for the accurate classification of multi-eye diseases.

10. Conclusions

Millions of individuals worldwide are affected by diabetic retinopathy, hypertensive retinopathy, glaucoma, and contrast-related eye conditions. Preventing the development of these illnesses requires early identification. In this work, a fully automated system for detecting and classifying the four eye diseases is proposed. The model combines innovative preprocessing techniques with deep learning models to obtain the best possible results. The newly introduced Fluorescence Imaging Simulation image preprocessing algorithm is used to enhance the classification performance for these diseases. This novel image processing approach focuses on crucial areas in affected images, aiding in their identification. Additionally, we propose a feature fusion between the features extracted from state-of-the-art models (MobileNet, EfficientNet). The model has been extensively evaluated on four different pre-existing datasets as well as a newly created dataset that the authors assembled from other sources. Our suggested technique performs better than state-of-the-art models in the literature, according to a comparison analysis. An examination of their strengths and limitations proves the effectiveness of our approach over established models. To validate the efficacy of the proposed methodology, additional testing on a

large, diversified dataset comprising a substantial number of possible sickness cases is required. Future research may explore the analysis of new datasets using NASNet or MobileNet, along with other augmentation techniques. This approach is not a decision support system, but rather a prescreening and automated illness detection tool. Future studies should concentrate on reducing the sensitivity to picture quality, in order to offer substantial flexibility and dependability across diverse healthcare settings.

Author Contributions: Conceptualization, M.K. and M.Z.S.; methodology, M.Z.S.; software, M.Z.S.; validation, M.Z.S., A.Y. and N.A.K.; formal analysis, M.Z.S.; investigation, M.K. and M.Z.S.; resources, M.K. and F.A.; data curation, M.Z.S.; writing—original draft preparation, M.Z.S. and M.F.H.; writing—review and editing, A.Y., N.A.K., F.A. and M.F.H.; visualization, M.Z.S.; supervision, M.Z.S., N.A.K. and F.A.; project administration, M.K.; funding acquisition, A.Y. and F.A. 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: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data will be available upon request from the authors for research purposes.

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

References

1. World Health Organization. Blindness and Vision Impairment. 2018. Available online: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment (accessed on 22 September 2022).
2. Sabanayagam, C.; Banu, R.; Chee, M.L.; Lee, R.; Wang, Y.X.; Tan, G.; Jonas, J.B.; Lamoureux, E.L.; Cheng, C.Y.; Klein, B.E.; et al. Incidence and progression of diabetic retinopathy: A systematic review. *Lancet Diabetes Endocrinol.* **2019**, *7*, 140–149. [CrossRef] [PubMed]
3. Orujov, F.; Maskeliūnas, R.; Damaševičius, R.; Wei, W. Fuzzy based image edge detection algorithm for blood vessel detection in retinal images. *Appl. Soft Comput.* **2020**, *94*, 106452. [CrossRef]
4. Cho, N.H.; Shaw, J.E.; Karuranga, S.; Huang, Y.; da Rocha Fernandes, J.D.; Ohlrogge, A.; Malanda, B. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res. Clin. Pract.* **2018**, *138*, 271–281. [CrossRef] [PubMed]
5. Jain, A.; Jalui, A.; Jasani, J.; Lahoti, Y.; Karani, R. Deep learning for detection and severity classification of diabetic retinopathy. In Proceedings of the 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), Chennai, India, 25–26 April 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.
6. Walker, H.K.; Hall, W.D.; Hurst, J.W. *Clinical Methods: The History, Physical, and Laboratory Examinations*; Butterworths: Boston, MA, USA, 1990.
7. Mishra, C.; Tripathy, K. Fundus Camera 2022 [Updated 25 August 2023]. In *StatePearls*; StatePearls Publishing: Treasure Island, FL, USA, 2024.
8. Decencière, E.; Zhang, X.; Cazuguel, G.; Lay, B.; Cochener, B.; Trone, C.; Gain, P.; Ordonez, R.; Massin, P.; Erginay, A.; et al. Feedback on a publicly distributed image database: The Messidor database. *Image Anal. Stereol.* **2014**, *33*, 231–234. [CrossRef]
9. Naveed, K.; Abdullah, F.; Madni, H.A.; Khan, M.A.; Khan, T.M.; Naqvi, S.S. Towards automated eye diagnosis: An improved retinal vessel segmentation framework using ensemble block matching 3D filter. *Diagnostics* **2021**, *11*, 114. [CrossRef]
10. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **2016**, *316*, 2402–2410. [CrossRef]
11. Decenciere, E.; Cazuguel, G.; Zhang, X.; Thibault, G.; Klein, J.C.; Meyer, F.; Marcotegui, B.; Quellec, G.; Lamard, M.; Danno, R.; et al. TeleOphta: Machine learning and image processing methods for teleophthalmology. *Irbm* **2013**, *34*, 196–203. [CrossRef]
12. Wong, T.; Cheung, C.; Larsen, M.; Sharma, S.; Simo, R. Diabetic retinopathy. *Nat. Rev. Dis. Primers* **2016**, *2*, 16012. [CrossRef]
13. Zhang, X.; Saaddine, J.B.; Chou, C.F.; Cotch, M.F.; Cheng, Y.J.; Geiss, L.S.; Gregg, E.W.; Albright, A.L.; Klein, B.E.; Klein, R. Prevalence of diabetic retinopathy in the United States, 2005–2008. *JAMA* **2010**, *304*, 649–656. [CrossRef]
14. Teo, Z.L.; Tham, Y.C.; Yu, M.; Chee, M.L.; Rim, T.H.; Cheung, N.; Bikbov, M.M.; Wang, Y.X.; Tang, Y.; Lu, Y.; et al. Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis. *Ophthalmology* **2021**, *128*, 1580–1591. [CrossRef]
15. Raman, R.; Srinivasan, S.; Virmani, S.; Sivaprasad, S.; Rao, C.; Rajalakshmi, R. Fundus photograph-based deep learning algorithms in detecting diabetic retinopathy. *Eye* **2019**, *33*, 97–109. [CrossRef] [PubMed]

16. Chakrabarti, R.; Harper, C.A.; Keeffe, J.E. Diabetic retinopathy management guidelines. *Expert Rev. Ophthalmol.* **2012**, *7*, 417–439. [CrossRef]
17. Topaloglu, I. Deep learning based convolutional neural network structured new image classification approach for eye disease identification. *Sci. Iran.* **2023**, *30*, 1731–1742. [CrossRef]
18. Choudhary, A.; Ahlawat, S.; Urooj, S.; Pathak, N.; Lay-Ekuakille, A.; Sharma, N. A deep learning-based framework for retinal disease classification. *Healthcare* **2023**, *11*, 212. [CrossRef]
19. Thomas, S.; Hodge, W.; Malvankar-Mehta, M. The cost-effectiveness analysis of teleglaucoma screening device. *PLoS ONE* **2015**, *10*, e0137913. [CrossRef]
20. Harasymowycz, P.; Birt, C.; Gooi, P.; Heckler, L.; Hutnik, C.; Jinapriya, D.; Shuba, L.; Yan, D.; Day, R. Medical management of glaucoma in the 21st century from a Canadian perspective. *J. Ophthalmol.* **2016**, *2016*, 6509809. [CrossRef]
21. Weinreb, R.N.; Leung, C.K.; Crowston, J.G.; Medeiros, F.A.; Friedman, D.S.; Wiggs, J.L.; Martin, K.R. Primary open-angle glaucoma. *Nat. Rev. Dis. Primers* **2016**, *2*, 16067. [CrossRef]
22. Tham, Y.C.; Li, X.; Wong, T.Y.; Quigley, H.A.; Aung, T.; Cheng, C.Y. Global prevalence of glaucoma and projections of glaucoma burden through 2040: A systematic review and meta-analysis. *Ophthalmology* **2014**, *121*, 2081–2090. [CrossRef]
23. Allison, K.; Patel, D.; Alabi, O. Epidemiology of glaucoma: The past, present, and predictions for the future. *Cureus* **2020**, *12*, e11686. [CrossRef]
24. Akter, N.; Fletcher, J.; Perry, S.; Simunovic, M.P.; Briggs, N.; Roy, M. Glaucoma diagnosis using multi-feature analysis and a deep learning technique. *Sci. Rep.* **2022**, *12*, 8064. [CrossRef]
25. Greenfield, D.S.; Weinreb, R.N. Role of optic nerve imaging in glaucoma clinical practice and clinical trials. *Am. J. Ophthalmol.* **2008**, *145*, 598–603. [CrossRef]
26. Michelessi, M.; Lucenteforte, E.; Oddone, F.; Brazzelli, M.; Parravano, M.; Franchi, S.; Ng, S.M.; Virgili, G. Optic nerve head and fibre layer imaging for diagnosing glaucoma. *Cochrane Database Syst. Rev.* **2015**, *11*, CD008803.
27. Antón López, A.; Nolivos, K.; Pazos López, M.; Fatti, G.; Ayala, M.E.; Martínez-Prats, E.; Peral, O.; Poposki, V.; Tsiroukis, E.; Morilla-Grasa, A.; et al. Diagnostic accuracy and detection rate of glaucoma screening with optic disk photos, optical coherence tomography images, and telemedicine. *J. Clin. Med.* **2021**, *11*, 216. [CrossRef] [PubMed]
28. Kanse, S.S.; Yadav, D.M. Retinal fundus image for glaucoma detection: A review and study. *J. Intell. Syst.* **2019**, *28*, 43–56. [CrossRef]
29. Shinde, R. Glaucoma detection in retinal fundus images using U-Net and supervised machine learning algorithms. *Intell.-Based Med.* **2021**, *5*, 100038. [CrossRef]
30. Yalçin, N.; Alver, S.; Uluhatun, N. Classification of retinal images with deep learning for early detection of diabetic retinopathy disease. In Proceedings of the 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2–5 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–4.
31. Chakrabarty, N. A deep learning method for the detection of diabetic retinopathy. In Proceedings of the 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Gorakhpur, India, 2–4 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–5.
32. Ting, D.S.W.; Pasquale, L.R.; Peng, L.; Campbell, J.P.; Lee, A.Y.; Raman, R.; Tan, G.S.W.; Schmetterer, L.; Keane, P.A.; Wong, T.Y. Artificial intelligence and deep learning in ophthalmology. *Br. J. Ophthalmol.* **2019**, *103*, 167–175. [CrossRef]
33. Ramakrishnan, P.; Sivagurunathan, P.; Sathish Kumar, N. Fruit classification based on convolutional neural network. *Int. J. Control Autom* **2019**.
34. Feng, X.; Yang, J.; Laine, A.F.; Angelini, E.D. Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, 11–13 September 2017, Proceedings, Part III 20*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 568–576.
35. Fu, H.; Cheng, J.; Xu, Y.; Zhang, C.; Wong, D.W.K.; Liu, J.; Cao, X. Disc-aware ensemble network for glaucoma screening from fundus image. *IEEE Trans. Med. Imaging* **2018**, *37*, 2493–2501. [CrossRef]
36. Fumero, F.; Alayón, S.; Sanchez, J.L.; Sigut, J.; Gonzalez-Hernandez, M. RIM-ONE: An open retinal image database for optic nerve evaluation. In Proceedings of the 2011 24th International Symposium on Computer-Based Medical Systems (CBMS), Bristol, UK, 27–30 June 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 1–6.
37. Verma, K.; Deep, P.; Ramakrishnan, A. Detection and classification of diabetic retinopathy using retinal images. In Proceedings of the 2011 Annual IEEE India Conference, Hyderabad, India, 16–18 December 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 1–6.
38. Gondal, W.M.; Köhler, J.M.; Grzeszick, R.; Fink, G.A.; Hirsch, M. Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 2069–2073.
39. Bock, R.; Meier, J.; Michelson, G.; Nyúl, L.G.; Hornegger, J. Classifying glaucoma with image-based features from fundus photographs. In *Proceedings of the Pattern Recognition: 29th DAGM Symposium, Heidelberg, Germany, 12–14 September 2007*; Proceedings 29; Springer: Berlin/Heidelberg, Germany, 2007; pp. 355–364.
40. Malik, S.; Kanwal, N.; Asghar, M.N.; Sadiq, M.A.A.; Karamat, I.; Fleury, M. Data driven approach for eye disease classification with machine learning. *Appl. Sci.* **2019**, *9*, 2789. [CrossRef]

41. Abbas, Q. Glaucoma-deep: Detection of glaucoma eye disease on retinal fundus images using deep learning. *Int. J. Adv. Comput. Sci. Appl.* **2017**, *8*, 41–45. [CrossRef]
42. Jain, L.; Murthy, H.S.; Patel, C.; Bansal, D. Retinal eye disease detection using deep learning. In Proceedings of the 2018 Fourteenth International Conference on Information Processing (ICINPRO), Bengaluru, India, 21–23 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6.
43. Metin, B.; Karasulu, B. Derin Öğrenme Modellerini Kullanarak İnsan Retinasının Optik Koherans Tomografi Görüntülerinden Hastalık Tespiti. *Veri Bilimi* **2022**, *5*, 9–19.
44. Sarki, R.; Ahmed, K.; Wang, H.; Zhang, Y.; Wang, K. Convolutional neural network for multi-class classification of diabetic eye disease. *EAI Endorsed Trans. Scalable Inf. Syst.* **2021**, *9*, e5. [CrossRef]
45. Umer, M.J.; Sharif, M.; Raza, M.; Kadry, S. A deep feature fusion and selection-based retinal eye disease detection from oct images. *Expert Syst.* **2023**, *40*, e13232. [CrossRef]
46. Gargeya, R.; Leng, T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology* **2017**, *124*, 962–969. [CrossRef] [PubMed]
47. Sajid, M.Z.; Hamid, M.F.; Youssef, A.; Yasmin, J.; Perumal, G.; Qureshi, I.; Naqi, S.M.; Abbas, Q. DR-NASNet: Automated System to Detect and Classify Diabetic Retinopathy Severity Using Improved Pretrained NASNet Model. *Diagnostics* **2023**, *13*, 2645. [CrossRef] [PubMed]
48. Sajid, M.Z.; Qureshi, I.; Youssef, A.; Khan, N.A. FAS-Incept-HR: A fully automated system based on optimized inception model for hypertensive retinopathy classification. *Multimed. Tools Appl.* **2024**, *83*, 14281–14303. [CrossRef]
49. Sajid, M.Z.; Qureshi, I.; Abbas, Q.; Albathan, M.; Shaheed, K.; Youssef, A.; Ferdous, S.; Hussain, A. Mobile-Hr: An ophthalmologic-based classification system for diagnosis of hypertensive retinopathy using optimized MobileNet architecture. *Diagnostics* **2023**, *13*, 1439. [CrossRef]
50. Hockwin, O. Cataract classification. *Doc. Ophthalmol.* **1995**, *88*, 263–275. [CrossRef]
51. Dong, Y.; Zhang, Q.; Qiao, Z.; Yang, J.J. Classification of cataract fundus image based on deep learning. In Proceedings of the 2017 IEEE International Conference on Imaging Systems and Techniques (IST), Beijing, China, 18–20 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–5.
52. Nakayama, L.F.; Ribeiro, L.Z.; Dychiao, R.G.; Zamora, Y.F.; Regatieri, C.V.; Celi, L.A.; Silva, P.; Sobrin, L.; Belfort, R., Jr. Artificial intelligence in uveitis: A comprehensive review. *Surv. Ophthalmol.* **2023**, *68*, 669–677. [CrossRef]
53. Kaggle. Eyepacs, Aptos, Messidor Diabetic Retinopathy. 2024. Available online: https://www.kaggle.com/datasets/ascanipek/eyepacs-aptos-messidor-diabetic-retinopathy (accessed on 15 August 2024).
54. Eye Disease Dataset. 2019. Kaggle. Available online: https://www.kaggle.com/datasets/kondwani/eye-disease-dataset (accessed on 15 August 2024).
55. Dataset for Different Eye Disease. 2022. Kaggle. Available online: https://www.kaggle.com/datasets/dhirajmwagh1111/dataset-for-different-eye-disease (accessed on 15 August 2024).
56. Eye Diseases Classification. 2022. Kaggle. Available online: https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification (accessed on 15 August 2024).
57. Papers with Code—DIARETDB1 Dataset, n.d. Available online: https://paperswithcode.com/dataset/diaretdb1 (accessed on 15 August 2024).
58. Wahab Sait, A.R. Artificial Intelligence-Driven Eye Disease Classification Model. *Appl. Sci.* **2023**, *13*, 11437. [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.

diagnostics

MDPI

Article

Validation of an Eye-Tracking Algorithm Based on Smartphone Videos: A Pilot Study

Wanzi Su [1], Damon Hoad [2], Leandro Pecchia [3] and Davide Piaggio [1,*]

[1] School of Engineering, University of Warwick, Library Road, Coventry CV4 7AL, UK; wanzi.su@warwick.ac.uk
[2] Warwick Medical School, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK; damon.hoad@swft.nhs.uk
[3] Department of Engineering, University Campus Bio-Medico of Rome, Via Alvaro del Portillo 21, 00128 Rome, Italy; l.pecchia@warwick.ac.uk
* Correspondence: davide.piaggio@warwick.ac.uk

Abstract: Introduction: This study aimed to develop and validate an efficient eye-tracking algorithm suitable for the analysis of images captured in the visible-light spectrum using a smartphone camera. **Methods:** The investigation primarily focused on comparing two algorithms, which were named CHT_TM and CHT_ACM, abbreviated from the core functions: Circular Hough Transform (CHT), Active Contour Models (ACMs), and Template Matching (TM). **Results:** CHT_TM significantly improved the running speed of the CHT_ACM algorithm, with not much difference in the resource consumption, and improved the accuracy on the x axis. CHT_TM achieved a reduction by 79% of the execution time. CHT_TM performed with an average mean percentage error of 0.34% and 0.95% in the x and y direction across the 19 manually validated videos, compared to 0.81% and 0.85% for CHT_ACM. Different conditions, like manually opening the eyelids with a finger versus without a finger, were also compared across four different tasks. **Conclusions:** This study shows that applying TM improves the original eye-tracking algorithm with CHT_ACM. The new algorithm has the potential to help the tracking of eye movement, which can facilitate the early screening and diagnosis of neurodegenerative diseases.

Keywords: eye-tracking; eye movement; neurodegenerative condition; biosignal processing

1. Introduction

Eye-tracking is an increasingly popular methodology across various applications. What captivates numerous researchers is its time-efficiency, cost-effectiveness, reliability, objectiveness, quantifiability, easiness to conduct, non-invasiveness, and accessibility. By leveraging techniques such as shape-based, appearance-based, feature-based, and hybrid methods, eye-tracking technologies capture and analyse the movements of key ocular components, including the pupil, cornea, sclera, iris, and retina [1–3]. These movements rely on the physiological mechanism controlled by the central nervous system, involving the brainstem, cerebellum, basal ganglia, and cerebral cortex [4]. Consequently, eye-tracking measurements provide a valuable window into brain function and cognitive processes.

Eye movements are categorised into four basic types: saccades, smooth pursuit movements, vergence movements, and vestibulo-ocular movements [5]. Even during fixation, the eye was not completely still due to fixational eye movements, which are composed of three main kinds of smaller movements: microsaccades, drift, and tremor [6]. Microsaccades

and drifts have an average size of approximately 6′ arc, with drifts exhibiting an average velocity of about 1′ arc per second [7]. These eye movements can influence each other and, sometimes, occur simultaneously. For example, in Miki and Hirata's study [8], microsaccades were observed in parallel with the ongoing vestibulo-ocular reflex. Among the aforementioned eye features, saccades and fixations are commonly analysed in eye-tracking studies [9], although pursuit and pupil dilation are also frequently investigated [10–12].

A systematic literature review has summarised the eye-tracking methods for enhancing productivity and reading abilities [13], categorising studies into four areas: general eye metrics and tracking, comprehension measurement, attention measurement, and typography and typesetting. Notably, the eye metrics such as fixation duration, saccade velocity, and blink rates have been linked to reading behaviours including speed reading and mind-wandering. Although findings across studies have limited generalizability, the integration of machine learning with eye-tracking technologies has shown promising potential in understanding reading behaviour, attention, and cognitive workload.

1.1. Eye-Tracking in Healthcare

Eye-tracking has demonstrated significant potential in healthcare applications, particularly in telemedicine [6], assistive technologies (e.g., wheelchair control [14]), and the assessment of neurological and vestibular disorders such as acute-onset vertigo [15]. It has been used to evaluate memory in conditions such as Amyotrophic Lateral Sclerosis (ALS) and late stages of Alzheimer's Disease (AD), without relying on intact verbal or motor functions, which may not prevail or be reliable in patients of late stages. By minimising the need for complex instructions or decision-making, eye-tracking provides an accessible means of evaluation [9].

Eye-tracking features have been identified as potential biomarkers for multiple neurodegenerative disorders. For instance, impaired convergence is observed in Parkinson's disease (PD) [16], while atypical saccades characterise preclinical and early-stage Huntington's disease [17,18]. Neurological abnormalities could also be detected in long COVID patients. García Cena et al. proved that long COVID patients were affected by altered latency for centrally directed saccades, impaired memory-guided saccades, and increased latencies for antisaccades [19].

1.2. Challenges and Limitations

Despite these promising applications, there are still concerns and challenges. The experimental constraints and demands of certain tasks can lead to changes in microsaccade behaviour, potentially producing responses that do not occur in natural conditions. For example, Mergenthaler and Engbert observed more microsaccades in the fixation task than the naturalistic scene viewing task, suggesting those responses might be task-specific [20]. However, in real-world settings, human behaviour will involve a range of different tasks (with both fixation and free viewing present), resulting in a combination of different types of eye movements. This highlights the difficulty of replicating experimental conditions outside controlled laboratory settings with the use of special eye-tracking equipment.

Therefore, barriers to the widespread adoption of these examinations in daily life or at home require patients to visit a clinic or laboratory for eye-tracking assessments. Moreover, the expensive price and high-calibration requirements, which need specific training, present further obstacles [6,21–23]. While some methods are accurate and effective, they often require skilled professionals to operate, making the process laborious and time-consuming [24].

To address these limitations, researchers have explored the use of consumer-grade electronic devices like smartphones, video recorders, laptops, and tablets. However, these

approaches introduce new challenges, particularly regarding head stabilisation, which significantly affects tracking accuracy. Solutions such as external head restraints and head-mounted equipment have been proposed to mitigate these issues [6,21,25]. Another solution is to use some indicators (e.g., a sticker) to represent the head movement and then subtract it from the eye movement [22].

1.3. State of the Art

Eye-tracking is sometimes combined with functional magnetic resonance imaging (fMRI) to investigate neurological deficits [26,27]. Compared to fMRI, one of the advantages of automated eye feature evaluation is its ability to simultaneously collect multiple measures during the same session and obtain different information for specific analysis and diagnosis [9]. High-frame-rate cameras further enhance detection capabilities, revealing some covert saccades that are difficult to see with the naked eye.

In order to collect and synthesise the available evidence, in a chapter of 'Eye Movement Research', Hutton [28] introduced the range of eye-tracking technologies currently available. There are two common tests to measure eye movement: video recording (video-oculography or VOG) and electromyography (electro-oculography or EOG). The less common approaches are through limbus reflection, Dual Purkinje Image (DPI), and scleral search coils, but they will not be discussed in detail here.

EOG uses a pair of electrode patches and measures the activity of two cranial nerves, which, respectively, connect the brain with the inner ear and the eye muscles. VOG uses a camera to record the eye movements and calibrate the accuracy. According to Corinna Underwood [29], there is a general consensus that the accuracy of directly measuring eye movements is higher than indirectly measuring movements via eye muscle motion. Hutton further highlights that EOG's primary drawback is its susceptibility to spatial inaccuracies due to drift artefacts over time, caused by impedance changes in the electrodes [28]. While most studies favour VOG over EOG, there is insufficient evidence to conclusively recommend one method over the other. VOG, though more precise, is relatively expensive and requires the eyes to remain open. Meanwhile, using EOG requires careful consideration in practice as certain medications (e.g., sedatives) or medical electrical equipment (e.g., cardiac pacemakers) would affect or interfere with the electromyography function [29].

VOG systems require head stability due to vestibulo-ocular movements that cause involuntary eye motion (to maintain the gaze fixation on a location, the head motion would cause the eye movement in the opposite direction [6]). Solutions to this issue fall into two categories: stationary VOG, where the camera remains fixed and external supports such as chin rests stabilise the head, and mobile VOG, where the camera moves with the head (e.g., in wearable eye-tracking glasses). Mobile VOG is generally better suited for real-world tasks, whereas stationary VOG is typically used in research settings that require specific stimuli. However, the need for mobile VOG to be lightweight and wearable imposes limitations on its specifications. As a result, its data quality, sampling rates, and spatial resolution are generally lower than those of high-specification stationary VOG systems [28].

Infrared-based eye-tracking, known as infrared oculography (IOG), is a common VOG method that tracks corneal reflection and the pupil centre [9,30,31]. Under infrared illumination, the pupil appears as a comparatively dark circle compared to the rest of the eye (e.g., the iris and sclera) [28]. However, this approach assumes that the pupils dilate and constrict symmetrically around their centre, which is not always accurate under varying luminance conditions according to Wildenmann et al. [32]. Consequently, researchers continue to explore alternative methods that do not rely on infrared illumination.

Table 1 provides a comparative analysis of different eye-tracking technologies, including fMRI, VOG, and EOG, highlighting their respective advantages and limitations [6,9,28,29].

Table 1. A comparison between the advantages and disadvantages of different state-of-the-art technologies.

Technologies			Advantages	Disadvantages
fMRI and other technology			For specific purposes	Without multiple eye measures
Eye-Tracking	**EOG**		Cheap, well established, readily available, high temporal resolution, no need for eyes to be open	Low accuracy and safety
	VOG	**Stationary**	High specifications	Need stimuli (likely in research scenarios)
		Mobile	No need for head stability, real-world tasks, lightweight	Low data quality, low sampling rates and poor spatial resolution
		Infrared Cameras (pupil centre)	Not influenced by the environment light levels	Pupil change error (pupils do not dilate or constrict symmetrically around their centre)

1.4. Deep Learning Methods and Hardware Limitations

Deep learning, particularly convolutional neural network (CNN) methods, has become a popular approach in modern eye-tracking research due to its capacity to learn direct mappings from raw facial and ocular images to gaze coordinates. These methods have been greatly accelerated by the availability of large-scale gaze datasets (GazeCapture [33] and MPIIGaze [34]) and the outstanding performance of CNNs in computer vision tasks.

Several CNN architectures have been explored in the literature. Early approaches include different CNN structures such as LeNet-based [35], AlexNet-based [36], VGGNet-based [34], and ResNet18-based [37] and ResNet50-based [38] models. Among notable CNN-based models, the iTracker model by Krafka et al. [33] used facial and eye images to predict gaze with high accuracy and was trained using the GazeCapture dataset. Valliappan et al. [39] demonstrated the feasibility of smartphone-based gaze tracking. Their proposed multi-layer feed-forward CNN achieved a gaze estimation accuracy of 0.6–1°, which is comparable to commercial solutions such as Tobii Pro Glasses.

Despite their effectiveness, these models typically demand a high computational burden, which limits their applicability for real-time deployment on mobile devices. They often require CPUs and GPUs with high cost and power consumption. Due to hardware limitations such as constrained computational power, limited battery life, and network bandwidth, deep learning models are not well suited for low-resource settings and edge computing, particularly on standard smartphone devices.

To address these limitations, several lightweight CNN architectures have been proposed, such as SqueezeNet [40] and ShuffleNet [41]. Notably, MobileNet [42] and MobileNet-V2 [43] employ depthwise convolutions to significantly reduce parameter count and computational cost while maintaining reasonable accuracy. When combined with edge computing platforms such as the Raspberry Pi, Google Coral USB Accelerator, and NVIDIA Jetson TX2, this direction appears promising.

However, these lightweight models are typically only suitable for gaze estimation on specific smartphone models with relatively high hardware specifications, even with support from mobile-optimised frameworks such as TensorFlow Lite (TFLite) and PyTorch Mobile. While this may become increasingly viable with ongoing advancements in smartphone technology and reductions in device cost, such solutions are not yet feasible

in low-resource settings. In these contexts, a traditional processing algorithm embedded within the smartphone may offer a more accessible and practical solution. Ultimately, the choice reflects a trade-off between the required accuracy for the research question and the limitations of the target hardware.

1.5. Smartphone-Based Eye-Tracking

Modern smartphones present an excellent opportunity for accessible and cost-effective eye-tracking. Equipped with high-resolution cameras and increasing on-device processing power, they are well-suited for edge computing, i.e., bringing the computation closer to the devices where data is gathered rather than cloud computing. This allows smartphones to function as both data acquisition and processing units, enabling real-time analysis, reducing latency, and preserving user privacy, which are particularly valuable in medical diagnostics. Their internet connectivity further supports remote deployment and monitoring when needed. Previous research has demonstrated the feasibility of using smartphones as an alternative to commercial-grade equipment for pupillometry [23]. Their affordability, ease of use, and resilience to power supply make them an attractive solution for large-scale, low-resource applications.

One challenge with smartphone-based tracking is its small screen size, which can restrict visual stimulus presentation. Many researchers have addressed this by incorporating external displays. For instance, Azami et al. [44] used an Apple iPad display and an Apple iPhone camera to record eye movements in ataxia and PD participants. After comparing different machine learning methods, they decided to use principal components analysis (PCA) and linear support vector machine (SVM), achieving 78% accuracy in distinguishing individuals with and without oculomotor dysmetria. Similarly, before switching to tablets, Lai et al. [21] initially designed their experiment based on a laptop display and an iPhone 6 camera. The iTracker-face algorithm can estimate mean saccade latency with a precision of less than 10 ms error. However, they did not report gaze accuracy. One exception is the study by Valliappan et al. [39], using a smartphone (Pixel 2 XL) as both a display and camera on healthy subjects.

While there is a wide variety of equipment and algorithms, few studies integrate eye-tracking with edge computing. Gunawardena et al. [45] found only one study (by Otoom et al. [46]) that combined edge computing with eye-tracking. As for the accuracy, Molina-Cantero et al. [47] reviewed 40 studies on visible-light eye-tracking using a low-cost camera, reporting an average visual angle accuracy of 2.69 degrees. Among these, only a subset achieved errors below 2.3 degrees, including the studies by Yang et al. [48], Hammal et al. [49], Liu et al. [50,51], Valenti et al. [52], Jankó and Hajder [53], Wojke et al. [54], Cheng et al. [55], and Jariwala et al. [56].

1.6. Contribution

Given the gap between the considerable potential of smartphones and the lack of edge computing solutions in eye-tracking, the proposed algorithm aims to introduce a lightweight alternative to conventional resource-demanding deep learning neural networks. As highlighted in the literature review, this cross-modal, smartphone-based approach is computationally efficient. Its low requirements make it especially suitable for smartphones and its compatibility with edge computing supports deployment in low-resource settings, where conventional systems are impractical or unavailable.

Although most studies would unintentionally mix the concept of eye-tracking and gaze estimation, it is important to distinguish between these two methods. Eye-tracking involves analysing visual information from images or frames in VOG, focusing on identifying facial landmarks, the eye region, and the pupil centre. In contrast, gaze estimation

further requires stimulus calibration, head orientation detection, and gaze angle calculation. This article aims to develop a stationary VOG application that prioritises eye-tracking over gaze tracking, utilising smartphone cameras as an alternative to infrared-based systems.

This study evaluates the performance of two eye-tracking algorithms—Circular Hough Transform plus Active Contour Models (CHT_ACM) and Circular Hough Transform Template Matching (CHT_TM)—in tracking the iris, particularly in neurodegenerative conditions. The findings will inform the second phase of the study, aimed at developing a novel gaze estimation software that is based on images acquired in the visible-light spectrum and will be validated against a commercial infrared eye tracker.

Modern eye trackers in gaming (e.g., Tobii Eye Tracker 5 [57] or Gazepoint 3 Eye Tracker [58]) achieve an optimistic performance in gaze tracking with unconstrained head movements. However, in line with the United Nations' Sustainable Development Goal 3 (Good Health and Wellbeing) and the principle of frugal innovation, this study aims to develop a cost-effective, smartphone-based eye-tracking solution that eliminates the need for additional accessories, thereby improving accessibility in low-resource settings. This study further contributes to the ongoing development of smartphone-based eye-tracking, paving the way for wider adoption in both clinical and non-clinical settings.

2. Methods

2.1. Experiment Definition

In order to build a dataset to develop and try the proposed algorithm on, an ethical approval application was submitted at the Biomedical and Scientific Research Ethics Committee of the University of Warwick and was granted (Ethical approval number 08/22-23).

This protocol was developed under the supervision of a medical doctor, who was an expert in neurorehabilitation, and was tested on three subjects with different iris colours, to investigate the robustness of the proposed algorithm. The three subjects had brown, green, and gray eyes, respectively. The experiment took place in the same laboratory setting and the environmental illumination was measured by a Dr.meter LX1010B digital lux meter (DrMeter, Union City, CA, USA).

During the experiment, the subject was also recorded in two different conditions, i.e., one with the eyelids naturally maintained open, and one where they were asked to keep them open by using their fingers. This was done to understand the influence of blinks and eyelids' partial closure on the algorithm performance. The experiment stimuli were designed to have an orange circle as the target and a dotted line to show the trajectory (shown in Figure 1). There were four tasks in total: (a) vertical task; (b) horizontal task; (c) fixation task; and (d) circular task. The target would follow the target along the trajectory only once, i.e., one complete round of the circle and back and forth for the linear tasks. The duration of each movement was either 1 or 5 s in order to mimic both slow and fast type of movements (e.g., smooth pursuit vs. nystagmus). The participant was asked to use fingers to open the eyelids, after one session without doing so.

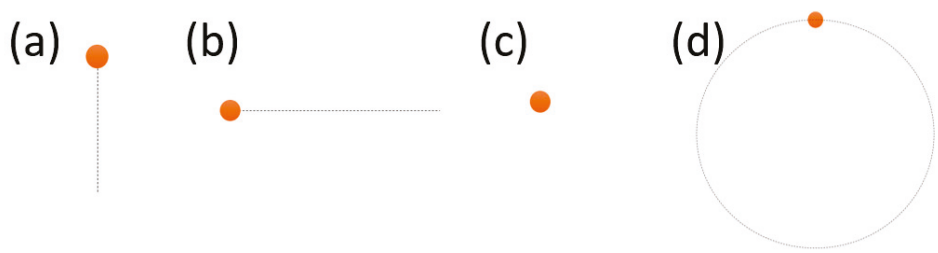


Figure 1. The initial set of the stimuli in the experiment protocol. (**a**) Vertical task; (**b**) horizontal task; (**c**) fixation task; and (**d**) circular task.

The equipment preparation involved merely a smartphone (iPhone 12, Apple Inc., Cupertino, CA, USA, 1920 × 1080 pixels, 30 frames/s), a desktop display (Samsung Electronics Co., Ltd., Suwon, South Korea), and a tripod (PEYOU Ltd., Wan Chai, Hong Kong). The participant was asked to sit in a private room in the author's department with a desktop screen right in front of them and a smartphone slightly shifted aside to avoid blocking the screen. The participant was asked to perform these tests without any other tool including the use of chin rest, which is normally used for aiding them to keep their head still. The participants were asked to follow certain instructions on the screen in front, while the different aforementioned tasks were being shown. In the experiment, the smartphone tracker placed on a tripod was used to track and record the eye movement of the participants. The detailed experimental set-up with the relevant quantifications is displayed in Figure 2.

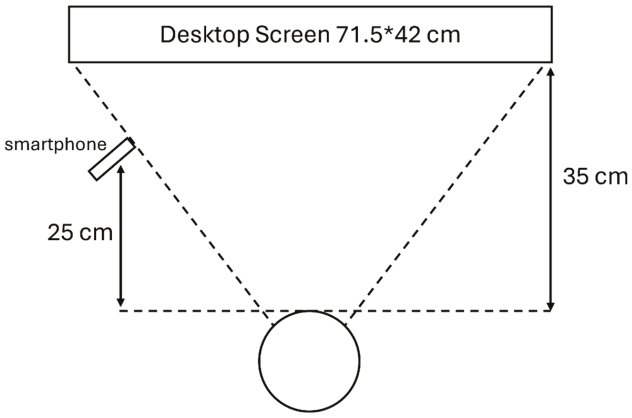


Figure 2. The experimental set-up.

2.2. Algorithm Development

Starting from the acquired videos, which featured the whole face, and after excluding the invalid videos (i.e., those with interfering movements) and manually cropping them to the region of interest (ROI), i.e., the whole eye, the first step in algorithm development entailed the selection of the best pipeline for image pre-processing and iris detection. At this stage, the working platform was Python 3.9.12, and the OpenCV (4.6.0) library was used for most functions. In terms of image pre-processing, the sampling rate for extracting the frames from the video was every 10 frames. Morphological operations [59] (i.e., erosion and dilation) and binarisation were applied. Erosion was used to reduce the influence of the eyelashes on the iris, and dilation was applied to remove the light spots caused by the reflection of light. Consequently, the CHT [60] was applied (minDist: 40; param1: 180; param2: 10; minRadius: 15; maxRadius: 50) in two different cases, one on data pre-processed as above plus ACM, and one on data pre-processed as above plus TM.

CHT is an image analysis technique designed to detect circular shapes within an image, even when the circles are partially occluded or incomplete. It operates by identifying radial symmetry through the accumulation of "votes" in a three-dimensional parameter space defined by the circle's centre coordinates and radius. Each edge pixel in the input image contributes votes in a circular pattern for potential circle centres. When multiple edge pixels support a circle of the same radius and centre, the votes accumulate at those coordinates. The peaks or local maxima in the voting space indicate the presence and location of the centroid. In real-world scenarios, such as eye-tracking, the pupil often appears elliptical due to the angle between the eye and the camera. Consequently, the voting space may produce a cluster of local maxima rather than a single point, as illustrated in Figure 3. This effect can be mitigated by adjusting the voting threshold and defining a minimum distance between detected peaks.

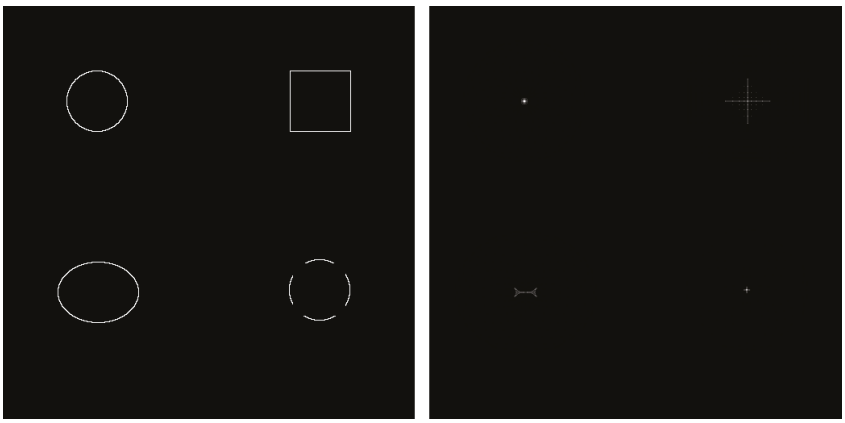


Figure 3. The voting mechanism from the edge pixel in a circular pattern [61].

CHT on its own, despite several adjustments to its parameters to reach optimal performance, was not powerful enough, resulting in too many eligible circle candidates to be the actual iris as output. In order to select the correct circle, binarisation [62] was then applied. This allowed for the removal of other spurious round circles. As mentioned above, ACM was also applied as a pre-processing step for one of the algorithms. ACM [63] works by aligning a deformable model with an image through energy minimisation. This deformable spine, also called a snake, responds to both constraint and image forces. These forces act collaboratively, pulling the snake toward object contours, while internal forces counteract undesired deformations. Specifically, snakes are tailored to address scenarios where an approximate understanding of the boundary's shape is available.

TM [64], conversely, represents an advanced machine vision technique designed to discern portions of an image that correspond to a predefined template. The process involves systematically placing the template over the image at all conceivable positions, calculating a numerical measure of similarity between the template and the currently overlapped image segment each time. Subsequently, the positions yielding the highest similarity measures are identified as potential occurrences of the template. The computation of the similarity measure between the aligned template image and the overlapped segment of the input image relies on the cross-correlation technique, entailing a straightforward summation of pairwise multiplications of corresponding pixel values from the two images.

Our CHT_TM relies on CHT_ACM only for the first frame. In detail, ACM is used in the first frame to fit an approximate boundary initiated with a circular snake inside of the ROI. With this approximate boundary, the image could be further cropped in order to reduce the noise and enhance the performance. CHT provides the potential circle candidates and binarisation selects the best iris circle. After obtaining an intact iris template through CHT_ACM, TM is applied in the following frames to find the position with the highest similarity. The position is translated into the corresponding iris centre and is validated against the manual measurements.

2.3. Algorithm Validation and Error Avoidance

To quantify the speedup, the running speed was calculated by measuring the running time of the same algorithm on the data of 10, 40, 100, and 400 frames on the same video. Moreover, in order to perform an initial validation of the algorithm, manual measurements were taken every 10 frames (same as algorithm development) on the 19 acquired videos from 9 different stimuli. Manually annotating every frame of the videos is not efficient. These stimuli represented various conditions that needed to be analysed, including finger/no finger and quick/slow movement.

Manually annotated iris centres were used as ground truth and comparison for the output of the algorithm. Several measurements of performance including mean absolute error (MAE), mean percentage error (MPE), root mean square error (RMSR), and Pearson Correlation Coefficient (PCC) were calculated by the Euclidean distance between the actual location and ground truth in both x and y directions. For a given set of coordinates on the x axis as an example, x_{actual} is the ground truth, $x_{measure}$ is the output of the algorithm, and n is the number of the frames. Here are the mathematical equations of the metrics:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_{actual,i} - x_{measure,i}\right| \tag{1}$$

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|x_{actual,i} - x_{measure,i}\right|}{x_{actual,i} + EPSILON} * 100\% \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{actual,i} - x_{measure,i}\right)^2} \tag{3}$$

$$Pearson\ r = \frac{\sum_{i=1}^{n}\left(x_{actual,i} - \overline{x}_{actual}\right)\left(x_{measure,i} - \overline{x}_{measure}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{actual,i} - \overline{x}_{actual}\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_{measure,i} - \overline{x}_{measure}\right)^2}} \tag{4}$$

Once Pearson Correlation Coefficient r is calculated, the t-statistic is then computed:

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}} \tag{5}$$

Assuming $t \sim t_{n-2}$, the *p*-value is calculated using the survival function (two-tailed):

$$pval = 2 \cdot sf(|t|) \tag{6}$$

This validation process compared the difference between several conditions including ones with/without fingers (to keep the eyelid open), 1 s/5 s movement, with/without TM, and all types of eye movements (fixation, vertical, horizontal, circle).

The frames in which the subject was blinking were deemed invalid. Moreover, due to the tendency of CHT_TM to match the lower half of the non-intact eye (shown in Figure 4a) resulting in a larger y axis error, a masking technique (shown in Figure 4b) that only calculates the pixels within the central circle was applied. The larger error on y axis (still quite a low error) also represented the technical issue while developing CHT_TM, i.e., that CHT_TM always tended to match the lower half of the non-intact eye (shown in Figure 4a).

(**a**) (**b**)

Figure 4. An example of matching the lower half of non-intact eye. (**a**) Before using masking. (**b**) After using masking.

3. Results

A total of 19 videos were acquired from 3 subjects, resulting in a total of 633 analysed frames. The illumination was natural daylight with an average intensity of 1680 ± 282.49 lux. The first subject performed 9 videos with 288 images in total, and the other two subjects performed 5 videos with 173 images in total.

Figure 5 describes the steps of the algorithm. The valid frames of the videos were cropped into the eye region and converted into grayscale. After that, the first frame was processed by CHT_ACM. An intact iris template was generated, and the iris centre position of the first frame was calculated. With this template, the following frames were processed by CHT_TM, which output the iris centre positions of the frames. The results could be calculated accordingly, and the performance was validated with manual measurements (see Figure 5).

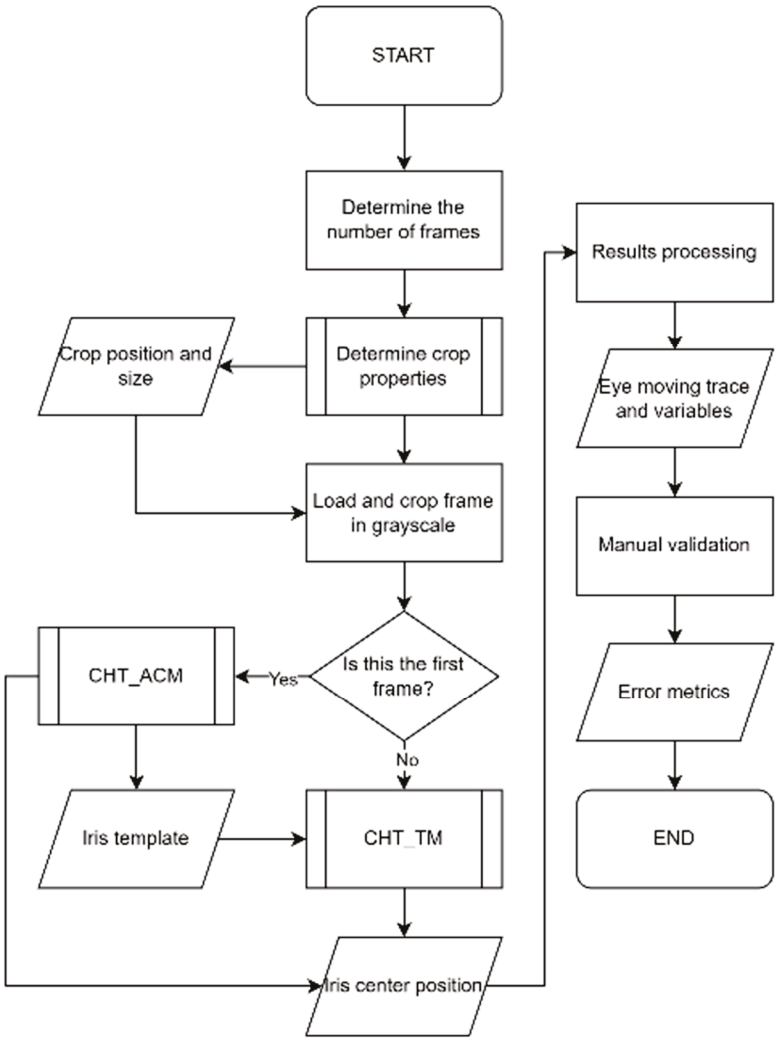


Figure 5. The flowchart describes the steps of the algorithm.

3.1. CHT_ACM Versus CHT_TM

Figure 6 compares the two different algorithms: CHT_ACM and CHT_TM. The CHT_ACM method begins with Active Contour Models to define a region of interest, followed by histogram-based binarisation of the image. It then applies the Circular Hough Transform to detect eligible circles and evaluates them by summing white pixel values in their respective binarised regions. A circle is identified as the iris if its white pixel values are below a 60% threshold. In contrast, the CHT_TM method uses a predefined iris template for Template Matching and identifies locations with a similarity metric above a threshold.

For each candidate, it calculates the pixel value within a circular region and selects the darkest (lowest value) location as the final iris centre.

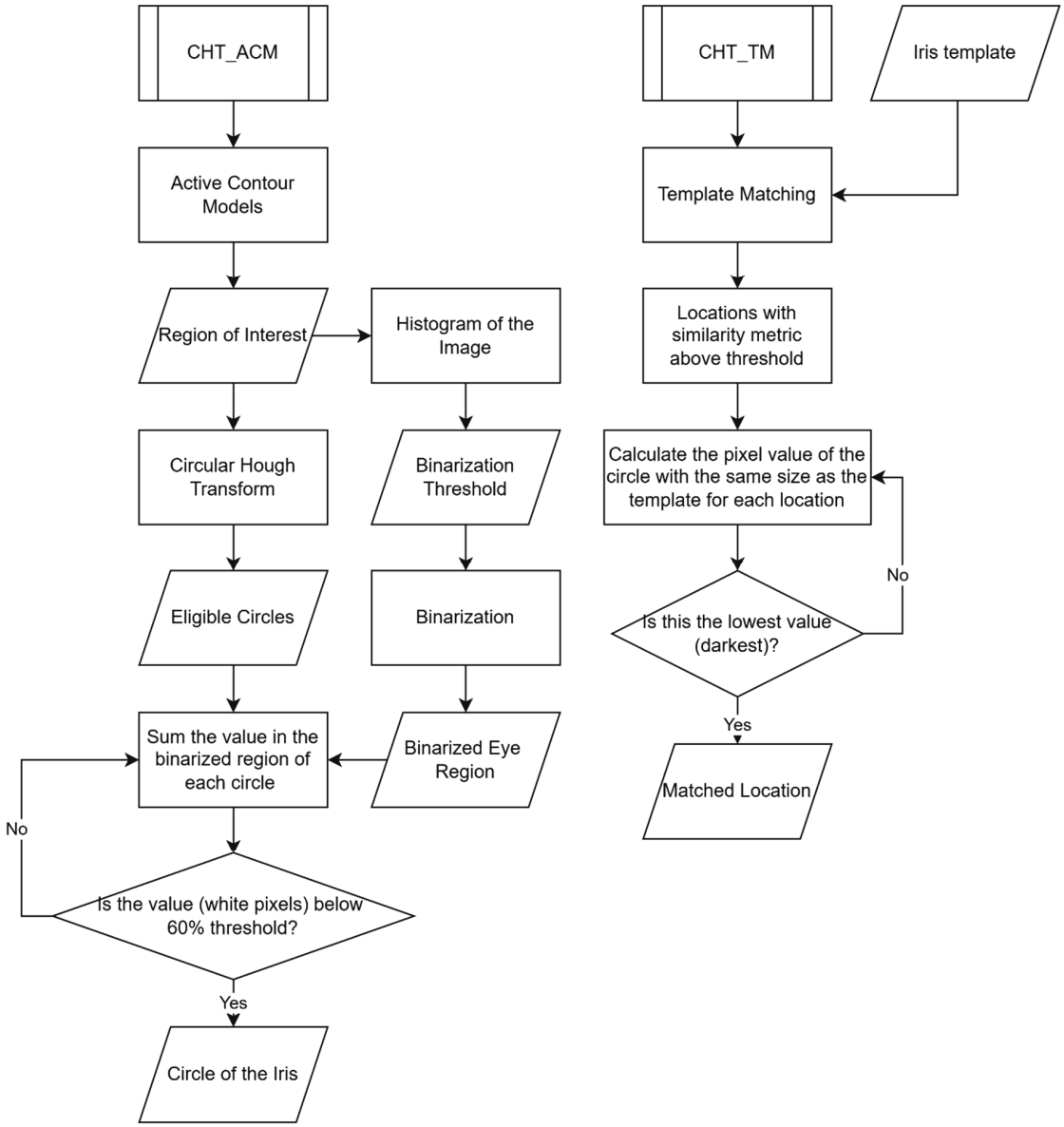


Figure 6. The flowchart comparing the two different algorithms in this study.

The TM function would significantly increase the running speed of the CHT_ACM (as shown in Figure 7). The average running time per frame (seconds) is 1.298 (CHT_ACM) and 0.271 (CHT_TM). CHT_TM saved 79% of the execution time, which means it is around 5 times faster, with a good trade-off between execution time and resource use. Most of the resource consumption is on reading the image so CPU and memory usage is not much different. Therefore, Template Matching greatly improved the running speed of CHT_ACM.

The MAE of CHT_TM was 1.43 pixels and 1.75 pixels in the x and y axis (shown in Table 2). Meanwhile, the MAE of CHT_ACM was 1.77/2.08 for x/y axis, less accurate than CHT_TM. CHT_TM performed with an average MPE of 1.06% and 1.21% in the x and y direction across the 19 videos, compared to 1.29% and 1.42% for CHT_ACM. Moreover, Pearson's Correlation Coefficients for CHT_TM were extremely high (r_x of 0.993, p $value_x$ less than 0.001; r_y of 0.990, p $value_y$ less than 0.001). The conclusion is that Template Matching not only improved the running speed but also the accuracy of CHT_ACM.

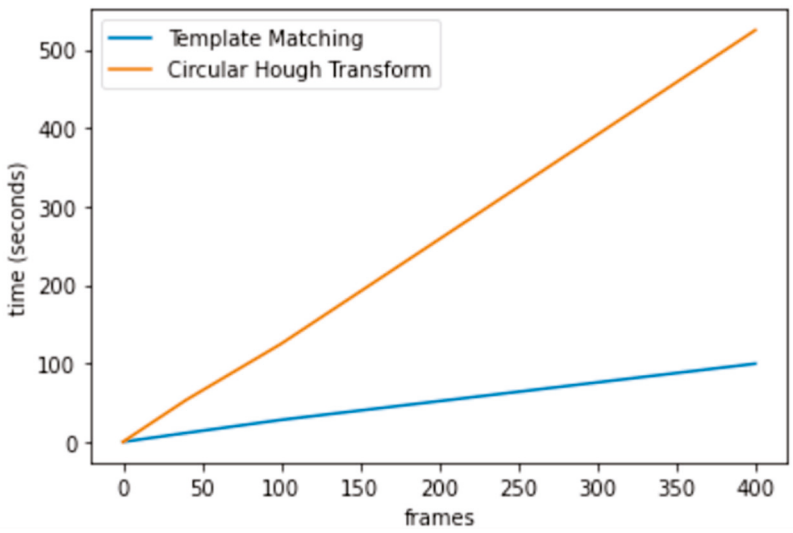


Figure 7. The running speed of CHT_ACM and CHT_TM.

Table 2. Mean average error, mean percentage error, root mean square error, and Pearson Correlation Coefficient calculated in both x and y directions of all 19 videos. * denotes better performance (lower errors).

	CHT_TM	CHT_ACM
MAE for x	1.432 *	1.774
MAE for y	1.752 *	2.079
MPE for x	1.059 *	1.290
MPE for y	1.213 *	1.423
RMSE for x	2.362 *	2.514
RMSE for y	2.554 *	3.069
Pearsons r and pval for x	0.993, <0.0001	0.993, <0.0001
Pearsons r and pval for y	0.990, <0.0001	0.986, <0.0001

3.2. With-Fingers Versus Without-Fingers Condition

In order to investigate the influence of using fingers to open the eyelid, the results of two conditions were analysed, i.e., with fingers or without fingers. In Table 3, it is obvious that the with-finger condition has lower error and higher Pearson Correlation Coefficient value, except one row, namely "x_CHT_TM". This suggests that using fingers can improve the performance of CHT_ACM and CHT_TM on the y axis, but it reduces the accuracy of CHT_TM on the x axis. PCC_p values of Table 3 are all below 0.00001. Therefore, these values are not reported in Table 3.

Table 3. The comparison of with-fingers and without-fingers conditions. * denotes better performance (lower errors) between with-fingers and without-fingers conditions.

No Finger	MAE	MPE	RMSE	PCC_r
x_CHT_TM	0.467 *	0.304 *	0.748 *	0.999
y_CHT_TM	1.697	1.241	2.396	0.984
x_CHT_ACM	1.36	0.904	1.997	0.995
y_CHT_ACM	1.337	0.969	1.876	0.990
Finger	**MAE**	**MPE**	**RMSE**	**PCC_r**
x_CHT_TM	0.541	0.403	0.849	0.999
y_CHT_TM	0.739 *	0.476 *	0.995 *	0.997
x_CHT_ACM	0.901 *	0.653 *	1.252 *	0.998
y_CHT_ACM	1.009 *	0.653 *	1.760 *	0.993

3.3. Comparison Among Tasks

Table 4 presents the performance of the two algorithms CHT_ACM and CHT_TM on the four tasks, i.e., vertical task, horizontal task, circular task, and fixation task. This is used to analyse the characteristics of each algorithm on different tasks, which helps identify the specific improvements of CHT_TM compared to CHT_ACM. Moreover, it provides evidence of what is sacrificed by applying TM. It is worth noticing that CHT_ACM performs better on the x axis in the fixation task but is worse in the horizontal task and vertical task regardless of axis. On the other hand, CHT_TM is more accurate on the y axis but has larger error on the x axis in the circular task and fixation task. PCC_p values of Table 4 are all below 0.00001. Therefore, these values are not reported in Table 4.

Table 4. The comparison of different tasks. * denotes the better performance (lower error) among the four tasks.

Vertical	**MAE**	**MPE**	**RMSE**	**PCC_r**
x_CHT_TM	1.291	0.911	2.175	0.995
y_CHT_TM	1.659	1.153	2.190	0.992
x_CHT_ACM	1.665	1.147 *	2.237 *	0.996
y_CHT_ACM	2.015	1.367	2.786	0.984
Horizontal	**MAE**	**MPE**	**RMSE**	**PCC_r**
x_CHT_TM	1.020 *	0.726 *	1.553 *	0.997
y_CHT_TM	1.672	1.184	2.520	0.964
x_CHT_ACM	2.128	1.516	3.014	0.988
y_CHT_ACM	2.298	1.588	3.489	0.934
Circular	**MAE**	**MPE**	**RMSE**	**PCC_r**
x_CHT_TM	1.755	1.260	2.925	0.985
y_CHT_TM	2.054	1.386	3.039	0.993
x_CHT_ACM	1.747	1.273	2.416	0.990
y_CHT_ACM	2.009	1.347 *	3.086	0.993
Fixation	**MAE**	**MPE**	**RMSE**	**PCC_r**
x_CHT_TM	1.537	1.353	2.182	0.999
y_CHT_TM	1.276 *	0.922 *	1.665 *	0.947
x_CHT_ACM	1.428 *	1.186	2.246	0.996
y_CHT_ACM	1.997 *	1.427	2.685 *	0.850

The best and the worst performing applications for both algorithms are summarised in Table 5. The fixation task is the most accurate on the y axis regardless of CHT_ACM or CHT_TM and the PCC is the lowest. The horizontal task can be well tracked on the x axis by CHT_TM but is the worst for CHT_ACM no matter the axis. The vertical task is best measured on the x axis by CHT_ACM. The circular task has the largest error in general, which means it is the hardest to track, especially by CHT_TM. However, the PCC of the circular task is the highest so the direction is similar between the real output and the manual validation result. PCC_p values of Table 5 are all below 0.00001. Therefore, these values are not reported in Table 5.

Table 5. The best performer/the worst performer among the four tasks for all the metrices. The best performer means the lowest error or the highest Pearson Correlation Coefficient value, and the worst performer is the opposite.

	MAE	**MPE**	**RMSE**	**PCC_r**
x_CHT_TM	Horizontal/Circle	Horizontal/Fixation	Horizontal/Circle	Fixation/Circle
y_CHT_TM	Fixation/Circle	Fixation/Circle	Fixation/Circle	Circle/Fixation

Table 5. *Cont.*

	MAE	MPE	RMSE	PCC_r
x_CHT_ACM	Fixation/Horizontal	Vertical/Horizontal	Vertical/Horizontal	Vertical/Horizontal
y_CHT_ACM	Fixation/Horizontal	Circle/Horizontal	Fixation/Horizontal	Circle/Fixation

3.4. Comparison Among Iris Colours

To understand the robustness of the proposed algorithm, the error metrics from subjects with different iris colours are presented in Table 6. The performance table indicates that there is a slight decrease in the accuracy of the algorithm for subjects with light iris colour but the final error is still pixel-level. Specifically, subject 1 with brown iris colour had the lowest error, especially on x axis, and subject 2 with green iris colour had the highest error. Subject 3 with gray iris colour showed a higher error on x axis than y axis, which was different from the other two subjects. The accuracy of CHT_TM was better than CHT_ACM across all three subjects.

Table 6. The performance of both algorithms on subjects with different iris colours. * denotes the better performance (lower error) among three subjects.

Subject 1	MAE	MPE	RMSE	PCC_r
x_CHT_TM	0.497 *	0.343 *	0.789 *	0.999
y_CHT_TM	1.325 *	0.945 *	1.974	0.992
x_CHT_ACM	1.182 *	0.807 *	1.746 *	0.996
y_CHT_ACM	1.210 *	0.847 *	1.832 *	0.993
Subject 2	**MAE**	**MPE**	**RMSE**	**PCC_r**
x_CHT_TM	2.791	1.985	3.732	0.972
y_CHT_TM	2.814	1.759	3.686	0.973
x_CHT_ACM	2.528	1.802	3.291	0.981
y_CHT_ACM	3.689	2.329	4.695	0.924
Subject 3	**MAE**	**MPE**	**RMSE**	**PCC_r**
x_CHT_TM	1.822	1.493	2.512	0.988
y_CHT_TM	1.415	1.143	1.957 *	0.991
x_CHT_ACM	2.148	1.712	2.823	0.990
y_CHT_ACM	2.037	1.583	2.734	0.980

4. Discussion

This study was aimed at developing and validating an effective eye-tracking algorithm to be used on visible-light images captured by a smartphone camera, in order to unlock more affordable and user-friendly technologies for eye-tracking. Two algorithms, named CHT_TM and CHT_ACM, were compared in terms of performance and computational efficiency. The selection of these algorithms was based on their potential to enhance accuracy and speed without increasing resource consumption.

Comprehensively, CHT_TM demonstrated improved runtime and superior performance in vertical eye movement tracking (y axis), although CHT_ACM outperformed it in horizontal tracking (x axis) in two out of four tasks. Whether it was CHT_ACM or CHT_TM, larger errors were consistently observed along the y axis, as seen in Table 2. This can be attributed to the anatomical reality that the upper and lower regions of the iris are more likely to be covered by the eyelids, especially during upward or downward gaze, or when participants are fatigued and the eyes are half-closed. This introduces inaccuracies in iris centre detection.

From the results comparing with/without-fingers conditions, using the fingers to open the eyelid seems to improve accuracy in most cases. This supports the hypothesis

that using fingers can help the algorithm diminish the error of non-intact iris as the eyelid is no longer covering the iris. While CHT_TM performed better along the x axis even without finger assistance, the benefit of finger usage was more significant on the y axis, where eyelid interference is typically greater on the upper edge or the lower edge. Despite this, the use of fingers was reported as uncomfortable for participants and is therefore not advisable in future studies. Alternative non-invasive strategies or postprocessing solutions are recommended.

Task-wise, the circular task produced the largest errors, suggesting that the primary challenge lay in the instability of the task rather than the algorithm itself. CHT_ACM remains less accurate than CHT_TM. As the mean absolute errors of both CHT_ACM and CHT_TM were on the pixel level, these errors were overall very small, also when taking into account the limits of the system for manual measurements with a precision of 0.5 pixels. Fixation tasks, being the most stable, resulted in the lowest tracking errors for both algorithms, which means there was no drastic head movement during the experiment. CHT_TM was better at tracking the fixation task on the y axis. Interestingly, CHT_TM improved x axis tracking in the horizontal task, likely due to its robustness in recognising elliptical iris shapes during lateral gaze. In contrast, CHT_ACM retained an advantage on the x axis for the fixation task. CHT_TM performed more reliably across tasks, especially in preventing tracking loss.

Comparing subjects with different iris colour, the algorithm showed the best performance in subjects with dark iris colour. Nonetheless, both algorithms produced pixel-level errors across all subjects, with CHT_TM consistently outperforming CHT_ACM. This indicates that the iris colour would influence the accuracy of both algorithms simultaneously, especially on the x axis.

No notable differences were found between the fast and the slow experiment.

Despite minor differences in error rates, it is noteworthy that both algorithms achieved high accuracy, with an average error of just 1.7 pixels (1.2%) across 19 videos. These results underscore the feasibility of both methods for reliable iris centre detection. However, the most significant difference lay in computational performance, with CHT_TM offering faster processing times.

Comparing the proposed method with existing smartphone-based approaches proved challenging due to a lack of validated benchmarks and methodological transparency in the literature. Many studies fail to disclose algorithmic details and rely instead on vague references to platforms like OpenCV or ARKit, thereby hindering reproducibility. In contrast, this study prioritises transparency and reproducibility by making both the data and algorithms publicly available.

Some may question the absence of machine learning or deep learning in this work, especially given their strong performance in image analysis tasks. However, the lack of a suitable public dataset, particularly one containing data from neurodegenerative patients, prevented the use of AI-based models. The mismatch between public eye-tracking datasets and cognitive experiment settings makes it difficult to adopt or adapt open-access datasets, considering the final goal of distinguishing patients and healthy subjects.

Public datasets are typically based on static stimuli and short-duration recordings, and are primarily designed to optimise gaze point accuracy. In contrast, cognitive experiments involve dynamic, goal-directed, task-specific stimuli, capture temporal dependencies in sequential data, and focus on variables linked to cognitive processes such as attention, memory, and response inhibition. Moreover, datasets collected in clinical or cognitive contexts are smaller due to stricter requirements and limited participant availability. These datasets are also subject to privacy and ethical regulations, which restrict data sharing and limit opportunities for large-scale model training. Additionally, variations in experimen-

tal conditions such as viewing distance, screen size, inter-subject variability (patients vs. controls), and ambient lighting hinder experimental reproducibility and limit the transferability of pre-trained models. In fact, data collection from patients would still be inevitable, but replicating the same experiment settings as the used public dataset would remain challenging as explained above.

Additionally, using AI trained on different hardware and image conditions (e.g., infrared cameras) would compromise compatibility with the smartphone-based set-up employed here. For instance, the structure of this algorithm is inspired by and is similar to the one proposed by Zhang et al. [65]. However, they used a CNN to condense the video and their data was collected from a portable infrared video goggle instead of the smartphone intended in this paper. Not only is the video grayscale but also the distance from the camera to the eye is different, making it impossible to use their data in this experiment or to develop the same algorithm based on varied data.

Beyond data limitations, deploying AI models on smartphones presents practical challenges. Deep learning methods typically require powerful processors or graphics processing units (GPUs) designed for computer systems, which is not the best fit for smartphones. In this case, it is necessary to upload the video data to the cloud server and use cloud computing. This reliance introduces new issues such as internet connectivity (not always available in rural areas or within low-resource settings), delayed response times, and potential data privacy risks (identifiable data like face videos). In contrast, a self-contained, built-in algorithm avoids these complications and better serves low-resource environments.

Nonetheless, AI remains a promising avenue for future work. Studies have shown a similar or even better performance with a CNN using the front-facing camera of Pixel 2 XL phone [39] or the RealSense digital camera [66] than a commercial eye tracker. Once a sufficiently large and diverse dataset is collected, including both healthy individuals and patients, AI models could be trained to refine or replace parts of the current algorithm. Such models could automate pre-processing or minimise tracking errors through learned feature extraction.

The future plan for this study is to develop a refined experimental protocol in collaboration with medical professionals, followed by validation against a commercial infrared eye tracker. Then, experiments can be carried out at hospital level on patients affected by neurodegenerative conditions. The ultimate goal is to design a smartphone-compatible eye-tracking toolkit and AI-based system for the early screening of neurodegenerative diseases.

Limitations

One limitation of the current protocol is the visible trace of the target during the task (see Figure 1), which allows participants to predict the target's trajectory. As a result, their eye movements may precede rather than follow the target. Additionally, the absence of a headrest introduces variability due to head movement, which can compromise signal quality. To address this, future studies will explore the feasibility of using a sticker placed in a fixed location as a reference point to track and compensate for head motion, allowing reconstruction of more accurate eye movement data.

Although the study aligns with the principle of frugal innovation and avoids using additional apparatus, a tripod was used as a temporary substitute for a user's hand or arm. One thing worth noticing is the differences in participant height, which can affect the camera angle towards the eye and may contribute to varying errors among subjects.

Another limitation is the small sample size, as this pilot study was primarily intended to demonstrate feasibility. Manual validation was used to assess the algorithm's performance. This method, while effective for small datasets, lacks the efficiency and scalability

of automated validation methods. This limits the generalizability of the results and the potential for large-scale application.

At this stage, comparisons were made between algorithm outputs and manual annotations of actual eye movement centres, rather than estimated gaze points. Each video was relatively short and did not include significant head movement, so the ROI of the eye was manually cropped and fixed at a constant image coordinate. Consequently, all movement was referenced to the same top-left corner of the cropped ROI (coordinate [0, 0]).

5. Conclusions

This paper shows that applying TM improves the original eye-tracking algorithm with CHT_ACM. The improved algorithm shows potential for supporting eye movement tracking in early screening and diagnosis of neurodegenerative diseases. Its integration into a non-invasive, cost-effective telemedicine toolbox could improve diagnostic reliability and accessibility as well as electronic transferability, in a cost-effective way. Such innovations sit perfectly within the current healthcare landscape, which has limited funding and is challenged by health emergencies (e.g., COVID-19 pandemic). Our research contributes to the development of such telemedicine tools by offering a reliable and accessible method for eye-tracking, which can be integrated into telemedicine platforms for remote screening and diagnosis. As highlighted above, the study has several limitations including reliance on a traditional image processing approach rather than AI, visibility of the target trace, and the lack of head movement correction. Future work will aim to refine the experimental protocol with guidance from medical experts, validate the algorithm against commercial benchmarks, and expand testing to patients with neurodegenerative symptoms. The integration of AI techniques will also be explored as a potential enhancement to the system.

Author Contributions: Conceptualisation, W.S., L.P. and D.H.; methodology, W.S. and D.P.; formal analysis, W.S. and D.H.; investigation, W.S.; writing—original draft preparation, W.S.; writing—review and editing, D.H. and D.P.; visualisation, W.S.; supervision, D.P. and L.P.; project administration, W.S. and D.P.; funding acquisition, W.S., L.P. and D.P. All authors have read and agreed to the published version of the manuscript.

Funding: Wanzi Su received support from UKRI Innovate UK grant (grant number 10031483) and the China Scholarship Council (CSC NO. 202308060026).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Biomedical and Scientific Research Ethics Committee of the University of Warwick (Ethical approval number 08/22-23 and granted on 18 January 2023).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets used and/or analysed 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:

AD	Alzheimer's Disease
AI	Artificial Intelligence
ALS	Amyotrophic Lateral Sclerosis
BSREC	Biomedical and Scientific Research Ethics Committee
CHT	Circular Hough Transform
CHT_ACM	CHT plus Active Contour Model
CHT_TM	CHT plus Template Matching

CNN	Convolutional Neural Network
DPI	Dual Purkinje Image
EOG	Electro-Oculography
fMRI	Functional Magnetic Resonance Imaging
IOG	Infrared Oculography
MAE	Mean Absolute Error
MPE	Mean Percentage Error
PCC	Pearson Correlation Coefficient
RMSR	Root Mean Square Error
ROI	Region of Interest
VOG	Video-Oculography

References

1. Mengtao, L.; Fan, L.; Gangyan, X.; Su, H. Leveraging eye-tracking technologies to promote aviation safety—A review of key aspects, challenges, and future perspectives. *Saf. Sci.* **2023**, *168*, 106295. [CrossRef]
2. Hansen, D.W.; Ji, Q. In the Eye of the Beholder: A Survey of Models for Eyes and Gaze. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 478–500. [CrossRef] [PubMed]
3. Majaranta, P.; Bulling, A. Eye Tracking and Eye-Based Human–Computer Interaction. In *Advances in Physiological Computing*; Fairclough, S.H., Gilleade, K., Eds.; Springer: London, UK, 2014; pp. 39–65.
4. Pouget, P. The cortex is in overall control of 'voluntary' eye movement. *Eye* **2015**, *29*, 241–245. [CrossRef]
5. Purves, D.; Augustine, G.J.; Fitzpatrick, D.; Hall, W.C.; LaManlia, A.-S.; Mooney, R.D.; Platt, M.L.; White, L.E. (Eds.) Types of Eye Movements and Their Functions. In *Neuroscience*, 2nd ed.; Sinauer Associates: Sunderland, MA, USA, 2001.
6. Alexander, R.G.; Macknik, S.L.; Martinez-Conde, S. Microsaccades in Applied Environments: Real-World Applications of Fixational Eye Movement Measurements. *J. Eye Mov. Res.* **2020**, *12*, 15. [CrossRef] [PubMed]
7. Mahanama, B.; Jayawardana, Y.; Rengarajan, S.; Jayawardena, G.; Chukoskie, L.; Snider, J.; Jayarathna, S. Eye movement and pupil measures: A review. *Front. Comput. Sci.* **2022**, *3*, 733531. [CrossRef]
8. Miki, S.; Hirata, Y. Microsaccades generated during car driving. In Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 3–7 July 2013.
9. Bueno, A.P.A.; Sato, J.R.; Hornberger, M. Eye tracking–The overlooked method to measure cognition in neurodegeneration? *Neuropsychologia* **2019**, *133*, 107191. [CrossRef]
10. Gooding, D.C.; Miller, M.D.; Kwapil, T.R. Smooth pursuit eye tracking and visual fixation in psychosis-prone individuals. *Psychiatry Res.* **2000**, *93*, 41–54. [CrossRef]
11. Garbutt, S.; Matlin, A.; Hellmuth, J.; Schenk, A.K.; Johnson, J.K.; Rosen, H.; Dean, D.; Kramer, J.; Neuhaus, J.; Miller, B.L. Oculomotor function in frontotemporal lobar degeneration, related disorders and Alzheimer's disease. *Brain* **2008**, *131*, 1268–1281. [CrossRef]
12. Van Gerven, P.W.M.; Paas, F.; Van Merriënboer, J.J.G.; Schmidt, H.G. Memory load and the cognitive pupillary response in aging. *Psychophysiology* **2004**, *41*, 167–174. [CrossRef]
13. Arnold, L.; Aryal, S.; Hong, B.; Nitharsan, M.; Shah, A.; Ahmed, W.; Lilani, Z.; Su, W.; Piaggio, D. A Systematic Literature Review of Eye-Tracking and Machine Learning Methods for Improving Productivity and Reading Abilities. *Appl. Sci.* **2025**, *15*, 3308. [CrossRef]
14. Iturrate, I.; Antelis, J.M.; Kubler, A.; Minguez, J. A noninvasive brain-actuated wheelchair based on a P300 neurophysiological protocol and automated navigation. *IEEE Trans. Robot.* **2009**, *25*, 614–627. [CrossRef]
15. Chang, T.P.; Zee, D.S.; Kheradmand, A. Technological advances in testing the dizzy patient: The bedside examination is still the key to successful diagnosis. In *Dizziness and Vertigo Across the Lifespan*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 9–30.
16. Kang, S.L.; Shaikh, A.G.; Ghasia, F.F. Vergence and strabismus in neurodegenerative disorders. *Front. Neurol.* **2018**, *9*, 299. [CrossRef] [PubMed]
17. Golding, C.; Danchaivijitr, C.; Hodgson, T.; Tabrizi, S.; Kennard, C. Identification of an oculomotor biomarker of preclinical Huntington disease. *Neurology* **2006**, *67*, 485–487. [CrossRef] [PubMed]
18. Hicks, S.L.; Robert, M.P.; Golding, C.V.; Tabrizi, S.J.; Kennard, C. Oculomotor deficits indicate the progression of Huntington's disease. *Prog. Brain Res.* **2008**, *171*, 555–558.
19. García Cena, C.; Costa, M.C.; Saltarén Pazmiño, R.; Santos, C.P.; Gómez-Andrés, D.; Benito-León, J. Eye Movement Alterations in Post-COVID-19 Condition: A Proof-of-Concept Study. *Sensors* **2022**, *22*, 1481. [CrossRef]
20. Mergenthaler, K.; Engbert, R. Microsaccades are different from saccades in scene perception. *Exp. Brain Res.* **2010**, *203*, 753–757. [CrossRef]

21. Lai, H.Y.; Saavedra-Peña, G.; Sodini, C.G.; Sze, V.; Heldt, T. Measuring Saccade Latency Using Smartphone Cameras. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 885–897. [CrossRef]
22. Sangi, M.; Thompson, B.; Turuwhenua, J. An Optokinetic Nystagmus Detection Method for Use With Young Children. *IEEE J. Transl. Eng. Health Med.* **2015**, *3*, 1600110. [CrossRef]
23. Piaggio, D.; Namm, G.; Melillo, P.; Simonelli, F.; Iadanza, E.; Pecchia, L. Pupillometry via smartphone for low-resource settings. *Biocybern. Biomed. Eng.* **2021**, *41*, 891–902. [CrossRef]
24. Ilesanmi, A.E.; Ilesanmi, T.; Gbotoso, G.A. A systematic review of retinal fundus image segmentation and classification methods using convolutional neural networks. *Healthc. Anal.* **2023**, *4*, 100261. [CrossRef]
25. Newman, J.L.; Phillips, J.S.; Cox, S.J.; Fitzgerald, J.; Bath, A. Automatic nystagmus detection and quantification in long-term continuous eye-movement data. *Comput. Biol. Med.* **2019**, *114*, 103448. [CrossRef] [PubMed]
26. Alichniewicz, K.K.; Brunner, F.; Klünemann, H.H.; Greenlee, M.W. Neural correlates of saccadic inhibition in healthy elderly and patients with amnestic mild cognitive impairment. *Front. Psychol.* **2013**, *4*, 467. [CrossRef]
27. Witiuk, K.; Fernandez-Ruiz, J.; McKee, R.; Alahyane, N.; Coe, B.C.; Melanson, M.; Munoz, D.P. Cognitive deterioration and functional compensation in ALS measured with fMRI using an inhibitory task. *J. Neurosci.* **2014**, *34*, 14260–14271. [CrossRef] [PubMed]
28. Hutton, S.B. Eye tracking methodology. In *Eye Movement Research: An Introduction to Its Scientific Foundations and Applications*; Springer: Cham, Switzerland, 2019; pp. 277–308.
29. Underwood, C. Electronystagmography (ENG). Available online: https://www.healthline.com/health/electronystagmography (accessed on 31 January 2024).
30. Cornsweet, T.N.; Crane, H.D. Accurate two-dimensional eye tracker using first and fourth Purkinje images. *JOSA* **1973**, *63*, 921–928. [CrossRef]
31. Guestrin, E.D.; Eizenman, M. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Trans. Biomed. Eng.* **2006**, *53*, 1124–1133. [CrossRef] [PubMed]
32. Wildenmann, U.; Schaeffel, F. Variations of pupil centration and their effects on video eye tracking. *Ophthalmic Physiol. Opt.* **2013**, *33*, 634–641. [CrossRef]
33. Krafka, K.; Khosla, A.; Kellnhofer, P.; Kannan, H.; Bhandarkar, S.; Matusik, W.; Torralba, A. Eye tracking for everyone. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2176–2184.
34. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *41*, 162–175. [CrossRef]
35. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. Appearance-based gaze estimation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4511–4520.
36. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. It's written all over your face: Full-face appearance-based gaze estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 51–60.
37. Kellnhofer, P.; Recasens, A.; Stent, S.; Matusik, W.; Torralba, A. Gaze360: Physically unconstrained gaze estimation in the wild. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6912–6921.
38. Zhang, X.; Park, S.; Beeler, T.; Bradley, D.; Tang, S.; Hilliges, O. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part V 16. Springer: Cham, Switzerland, 2020; pp. 365–381.
39. Valliappan, N.; Dai, N.; Steinberg, E.; He, J.; Rogers, K.; Ramachandran, V.; Xu, P.; Shojaeizadeh, M.; Guo, L.; Kohlhoff, K.; et al. Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nat. Commun.* **2020**, *11*, 4553. [CrossRef]
40. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.
41. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.
42. Howard, A.G. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
43. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.

44. Azami, H.; Chang, Z.; Arnold, S.E.; Sapiro, G.; Gupta, A.S. Detection of Oculomotor Dysmetria From Mobile Phone Video of the Horizontal Saccades Task Using Signal Processing and Machine Learning Approaches. *IEEE Access* **2022**, *10*, 34022–34031. [CrossRef] [PubMed]
45. Gunawardena, N.; Ginige, J.A.; Javadi, B. Eye-tracking Technologies in Mobile Devices Using Edge Computing: A Systematic Review. *ACM Comput. Surv.* **2022**, *55*, 158. [CrossRef]
46. Otoom, M.; Alzubaidi, M.A. Ambient intelligence framework for real-time speech-to-sign translation. *Assist. Technol.* **2018**, *30*, 119–132. [CrossRef]
47. Molina-Cantero, A.J.; Lebrato-Vázquez, C.; Castro-García, J.A.; Merino-Monge, M.; Biscarri-Triviño, F.; Escudero-Fombuena, J.I. A review on visible-light eye-tracking methods based on a low-cost camera. *J. Ambient. Intell. Humaniz. Comput.* **2024**, *15*, 2381–2397. [CrossRef]
48. Yang, B.; Zhang, X.; Li, Z.; Du, S.; Wang, F. An accurate and robust gaze estimation method based on maximum correntropy criterion. *IEEE Access* **2019**, *7*, 23291–23302. [CrossRef]
49. Hammal, Z.; Massot, C.; Bedoya, G.; Caplier, A. Eyes Segmentation Applied to Gaze Direction and Vigilance Estimation. In Proceedings of the International Conference on Pattern Recognition and Image Analysis, Bath, UK, 22–25 August 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 236–246.
50. Liu, Y.; Lee, B.-S.; Rajan, D.; Sluzek, A.; McKeown, M.J. CamType: Assistive text entry using gaze with an off-the-shelf webcam. *Mach. Vis. Appl.* **2019**, *30*, 407–421. [CrossRef]
51. Liu, Y.; Lee, B.S.; Sluzek, A.; Rajan, D.; Mckeown, M. Feasibility Analysis of Eye Typing with a Standard Webcam. In Proceedings of the Computer Vision—ECCV 2016 Workshops, Amsterdam, The Netherlands, 14–16 October 2016; Springer: Cham, Switzerland, 2016; pp. 254–268.
52. Valenti, R.; Staiano, J.; Sebe, N.; Gevers, T. Webcam-Based Visual Gaze Estimation. In Proceedings of the Image Analysis and Processing—ICIAP 2009, Vietri sul Mare, Italy, 8–11 September 2009; Springer: Berlin/Heidelberg, Germany, 2009; pp. 662–671.
53. Jankó, Z.; Hajder, L. Improving human-computer interaction by gaze tracking. In Proceedings of the 2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom), Kosice, Slovakia, 2–5 December 2012; pp. 155–160.
54. Wojke, N.; Hedrich, J.; Droege, D.; Paulus, D. Gaze-estimation for consumer-grade cameras using a Gaussian process latent variable model. *Pattern Recognit. Image Anal.* **2016**, *26*, 248–255. [CrossRef]
55. Cheng, S.; Ping, Q.; Wang, J.; Chen, Y. EasyGaze: Hybrid eye tracking approach for handheld mobile devices. *Virtual Real. Intell. Hardw.* **2022**, *4*, 173–188. [CrossRef]
56. Jariwala, K.; Dalal, U.; Vincent, A. A robust eye gaze estimation using geometric eye features. In Proceedings of the 2016 Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC), Moscow, Russia, 6–8 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 142–147.
57. Tobii. Tobii Eye tracker 5. Available online: https://gaming.tobii.com/product/eye-tracker-5/ (accessed on 18 June 2024).
58. Gazepoint. GP3 Eye Tracker | Hardware Only. Available online: https://www.gazept.com/product/gazepoint-gp3-eye-tracker/?v=79cba1185463 (accessed on 18 June 2024).
59. Batchelor, B.G.; Waltz, F.M. Morphological Image Processing. In *Machine Vision Handbook*; Batchelor, B.G., Ed.; Springer: London, UK, 2012; pp. 801–870.
60. Cherabit, N.; Chelali, F.Z.; Djeradi, A. Circular hough transform for iris localization. *Sci. Technol.* **2012**, *2*, 114–121. [CrossRef]
61. ctrueden. Hough Circle Transform. Available online: https://imagej.net/plugins/hough-circle-transform (accessed on 9 April 2025).
62. Yousefi, J. *Image Binarization Using Otsu Thresholding Algorithm*; University of Guelph: Guelph, ON, Canada, 2011; p. 10.
63. Kass, M.; Witkin, A.; Terzopoulos, D. Snakes: Active contour models. *Int. J. Comput. Vis.* **1988**, *1*, 321–331. [CrossRef]
64. Brunelli, R. *Template Matching Techniques in Computer Vision: Theory and Practice*; Wiley Publishing: Hoboken, NJ, USA, 2009.
65. Zhang, W.; Wu, H.; Liu, Y.; Zheng, S.; Liu, Z.; Li, Y.; Zhao, Y.; Zhu, Z. Deep learning based torsional nystagmus detection for dizziness and vertigo diagnosis. *Biomed. Signal Process. Control* **2021**, *68*, 102616. [CrossRef]
66. Hassan, M.A.; Yin, X.; Zhuang, Y.; Aldridge, C.M.; McMurry, T.; Southerland, A.M.; Rohde, G.K. A Digital Camera-Based Eye Movement Assessment Method for NeuroEye Examination. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 655–665. [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.

diagnostics

MDPI

Article

Integrating Machine Learning and Deep Learning for Predicting Non-Surgical Root Canal Treatment Outcomes Using Two-Dimensional Periapical Radiographs

Catalina Bennasar [1,*], Antonio Nadal-Martínez [2], Sebastiana Arroyo [1], Yolanda Gonzalez-Cid [3], Ángel Arturo López-González [4] and Pedro Juan Tárraga [5]

1 Academia Dental de Mallorca (ADEMA), School of Dentistry, University of Balearic Islands, 07122 Palma de Mallorca, Spain; s.arroyo@eua.edu.es
2 Soft Computing, Image Processing and Aggregation (SCOPIA) Research Group, University of the Balearic Islands (UIB), 07122 Palma de Mallorca, Spain; antonio.nadal@uib.cat
3 Department of Mathematical Sciences and Informatics, University of the Balearic Islands, 07120 Palma de Mallorca, Spain; yolanda.gonzalez@uib.es
4 ADEMA-Health Group, University Institute of Health Sciences of Balearic Islands (IUNICS), 02008 Palma de Mallorca, Spain; angarturo@gmail.com
5 Faculty of Medicine, University of Castilla-La Mancha, 02001 Albacete, Spain; pjtarraga@sescam.jccm.es
* Correspondence: c.bennasar@eua.edu.es; Tel.: +34-626-388-003

Abstract: Background/Objectives: In a previous study, we utilized categorical variables and machine learning (ML) algorithms to predict the success of non-surgical root canal treatments (NSRCTs) in apical periodontitis (AP), classifying the outcome as either success (healed) or failure (not healed). Given the importance of radiographic imaging in diagnosis, the present study evaluates the efficacy of deep learning (DL) in predicting NSRCT outcomes using two-dimensional (2D) periapical radiographs, comparing its performance with ML models. **Methods**: The DL model was trained and validated using leave-one-out cross-validation (LOOCV). Its output was incorporated into the set of categorical variables, and the ML study was reproduced using backward stepwise selection (BSS). The chi-square test was applied to assess the association between this new variable and NSRCT outcomes. Finally, after identifying the best-performing method from the ML study reproduction, statistical comparisons were conducted between this method, clinical professionals, and the image-based model using Fisher's exact test. **Results**: The association study yielded a p-value of 0.000000127, highlighting the predictive capability of 2D radiographs. After incorporating the DL-based predictive variable, the ML algorithm that demonstrated the best performance was logistic regression (LR), differing from the previous study, where random forest (RF) was the top performer. When comparing the deep learning–logistic regression (DL-LR) model with the clinician's prognosis (DP), DL-LR showed superior performance with a statistically significant difference (p-value < 0.05) in sensitivity, NPV, and accuracy. The same trend was observed in the DL vs. DP comparison. However, no statistically significant differences were found in the comparisons of RF vs. DL-LR, RF vs. DL, or DL vs. DL-LR. **Conclusions**: The findings of this study suggest that image-based artificial intelligence models exhibit superior predictive capability compared with those relying exclusively on categorical data. Moreover, they outperform clinician prognosis.

Keywords: machine learning; deep learning; outcome prediction; non-surgical root canal treatment; apical periodontitis

1. Introduction

Over the past decade, ML and DL have emerged as transformative technologies with significant impacts across multiple scientific disciplines. These techniques, as branches of artificial intelligence (AI), have become essential tools for analyzing large volumes of data, identifying complex patterns, and providing innovative solutions to previously intractable problems. Their relevance has been particularly emphasized in health sciences, including medicine, biology, and, more recently, dentistry, where their application holds the potential to optimize diagnostics, personalize treatments, and improve clinical outcomes.

ML can be defined as a subfield of AI that employs algorithms to enable machines to learn patterns and behaviors from data without being explicitly programmed for each specific task (Samuel, 1959 [1]). Within this field, DL represents a significant evolution, utilizing artificial neural networks with multiple layers capable of learning hierarchical data representations. These deep architectures have proven especially effective in tasks involving large datasets and complex features, such as medical image interpretation, disease prediction, and biological signal analysis [2].

The current utility of ML and DL in health sciences is demonstrated in a wide range of applications, including but not limited to medical image analysis, genomic data mining, drug modeling, and clinical outcome prediction. For instance, in radiology, convolutional neural networks (CNNs) have been successfully employed to detect tumor lesions in computed tomography (CT) and magnetic resonance imaging (MRI) scans, achieving accuracy levels comparable to those of human experts [3]. In genomics, DL algorithms have facilitated the decoding of complex gene interactions, accelerating the development of personalized therapies [4].

In the field of dentistry, the impact of ML and DL is beginning to solidify with promising applications. Dentistry, as a health science, has undergone significant digital transformation in recent years, driven by technologies such as cone-beam computed tomography (CBCT), three-dimensional (3D) printing, and CAD/CAM systems. However, the integration of ML and DL in this domain has opened new avenues for diagnosis, treatment design, and disease monitoring. For example, DL algorithms have been effective in identifying dental caries [5], fractures, periodontal diseases [6], and periapical conditions from digital radiographs, enhancing diagnostic accuracy and reducing variability among professionals [7].

Beyond diagnostics, these technologies are beginning to influence treatment planning and execution. In orthodontics, for example, ML is used to predict tooth movement and optimize the placement of brackets or aligners, leading to more effective and personalized treatments [8]. In implantology, ML models assist in predicting dental implant stability over time by considering factors such as bone density, implant location, and patient characteristics [9]. In endodontics, AI supports professionals by detecting periapical lesions, identifying root fractures, analyzing root canal morphology, predicting retreatment needs, and aiding in regenerative pulpal therapy, all of which contribute to improved diagnostics, treatment planning, and patient care [10–16].

The future of ML and DL in health sciences, particularly in dentistry, promises to be even more revolutionary. It is anticipated that the combination of these technologies with advanced sensor systems and data from wearable devices will enable continuous, real-time monitoring of oral health. For example, the integration of DL with intraoral connected devices could facilitate the early detection of diseases such as oral cancer through the analysis of salivary biomarkers or intraoral images [17]. Additionally, the development of explainable AI (XAI) systems could address one of the most pressing current challenges: the need to provide clear and interpretable explanations of algorithm-generated predictions, fostering clinical acceptance and ethical use [18].

Despite these promises, the implementation of ML and DL in clinical practice faces significant challenges that must be addressed to ensure success. These include the need for large volumes of high-quality data for model training, ethical and legal concerns related to data privacy, and the necessity of educating healthcare professionals in the use of these technologies. These challenges highlight the importance of interdisciplinary collaboration involving researchers, technology developers, clinicians, and policymakers [19,20].

2. Study Objectives

To evaluate DL as an additional variable in an ML study for predicting NSRCTs in cases of AP. This study aims to determine the extent to which deep neural networks can predict the outcome of NSRCTs in teeth with apical periodontitis using digital periapical radiographs of confirmed AP diagnoses.

3. Materials and Methods

3.1. Sample Selection

A retrospective study was conducted based on the analysis of clinical records of patients with AP who underwent NSRCTs for the first time (not retreatments). Cases were randomly selected from the database of a private clinic in Mallorca, Spain. Only patients without reported systemic diseases [21] who received treatment for the first time and whose records included the following were included:

A comprehensive medical and dental history with general, facial, and oral inspection reports, as well as dental percussion and palpation examinations.

Results of complementary tests, such as thermal sensitivity testing using an ice pencil and periapical radiography.

A follow-up period of at least nine years, starting six months after treatment, with documented evaluations of lesion recovery, categorizing cases as successful (0: there are no symptoms or indications for further treatments, and the lesion disappears after NSRCT) or failed (1: the failure occurs when either the clinical or radiographic outcome fails).

Radiographs were acquired using an X Mind Unity Acteon Satelec system with a focal point of 0.4 mm, at 70 kV and 7 mA, employing a Carestream 6100 digital system with a resolution of 15 LP/mm. The bisecting angle technique was used with a Rinn XCD (Dentsply) positioner (Dentsply, Charlotte, NC, USA) [22]. Patients with vertical root fractures or teeth without sufficient ferrule structure for subsequent restoration were excluded.

Due to this filtering process, the final number of patients included in this study was reduced to 119. Patient consent was waived due to the inability to identify participants in the database. The Research Ethics Committee of the Balearic Islands (IB4015/19IP) approved this study.

3.2. Intervention Procedure

The 119 patients with confirmed AP, for whom eight preoperative domain variables were observed as per a recommended data collection template (DCT) for endodontic treatment evaluation studies [23–25], underwent standardized endodontic treatment performed by the same endodontist using identical materials and procedures. The following phases were followed:

Local anesthesia administration and rubber dam placement.

Chamber access and pre-enlargement of the coronal third, followed by apical third negotiation.

Working length determination using a Morita apex locator (Morita, Tokyo, Japan) and radiographic confirmation. The working length was always set at the radiographic apex level.

Instrumentation with K3 (SybronEndo, Orange, CA, USA) and Protaper Gold (Dentsply Maillefer, Woodbridge, ON, USA) rotary systems, complemented with manual instruments.

Irrigation with EDTA and 5.25% sodium hypochlorite.

Obturation using the warm vertical condensation technique with AH Plus sealer.

Following treatment completion, cases were radiographically evaluated to rule out overfills or obturation defects.

3.3. Machine Learning and Deep Learning Analysis

To compare DL with ML models, a previous study, "Second Opinion for NSRCT Prognosis Using Machine Learning Models" [26], was utilized, in which logistic regression (LR), random forest (RF), naive Bayes (NB), and k-nearest neighbor (KNN) algorithms were applied. The RF algorithm demonstrated the best performance.

All periapical radiographs, 108 in total since due to geometric distortion or anatomical noise 11 were excluded, used for DL model training were preoperative images obtained prior to NSRCT, labeled based on post-treatment follow-up outcomes (healed: 0/not healed: 1) after a minimum of 9 years (Figures 1 and 2).

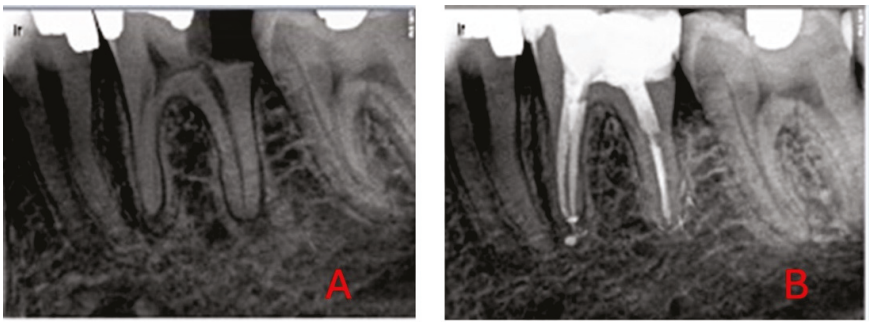


Figure 1. (**A**) AP in tooth 36; (**B**) three-year follow-up: healed.

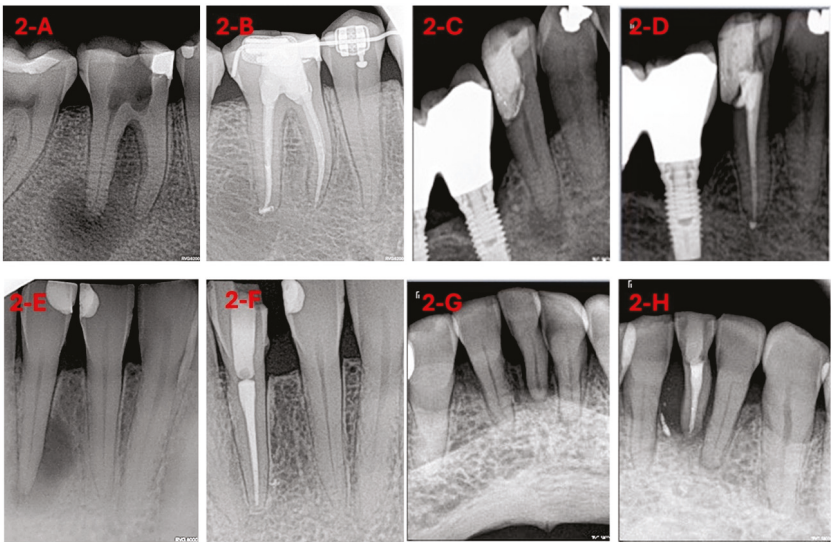


Figure 2. (**A**) AP in tooth 46; (**B**) one-year follow-up: healed; (**C**) AP in tooth 45; (**D**) four-year follow-up: healed; (**E**) AP in tooth 41; (**F**) three-year follow-up: healed; (**G**) AP in tooth 31; (**H**) two-year follow-up: not healed.

The AnotIA software was utilized for the precise segmentation of diagnostic 2D periapical radiographic images of AP, assigning labels to facilitate subsequent analysis (Figure 3). Although the segmented regions produced by AnotIA were not used as direct

input for the DL model (i.e., no semantic segmentation masks were provided to the network), these annotations were employed during model development and testing to ensure that the network's focus aligned with clinically relevant areas. Specifically, the marked regions of apical lesions were used to visually confirm that the model's predictions and activation maps corresponded to diagnostically meaningful locations. This validation strategy contributed to increasing the model's interpretability and may serve as a preliminary step toward incorporating explainable AI (XAI) techniques in future research.

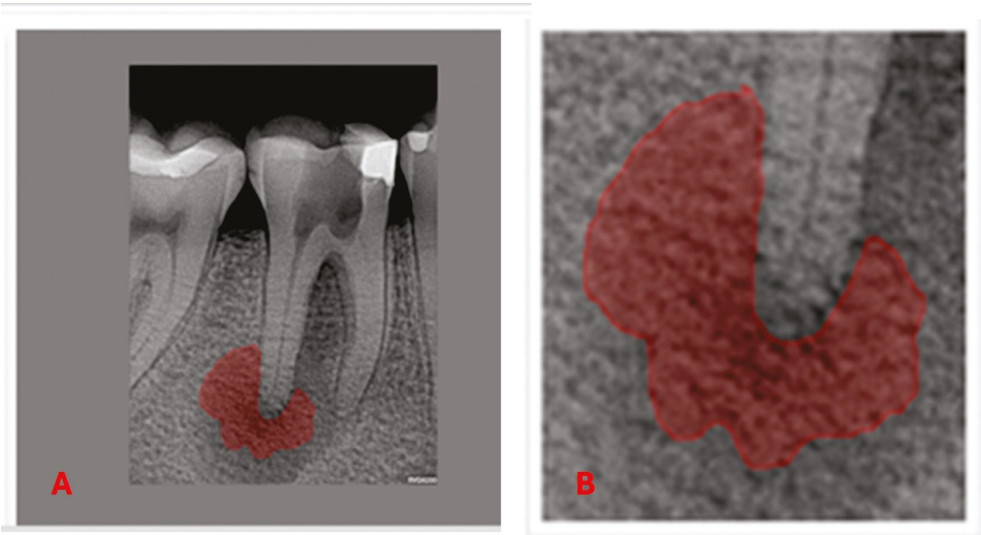


Figure 3. Demarcation of apical periodontitis. (**A**) demarcation of the AP in the vision of the tooth as a whole. (**B**) demarcation of the AP in the apical area of the distal root of the same tooth.

Diagnostic two-dimensional AP images were employed to train a convolutional neural network based on the ResNet-18 architecture, a deep 18-layer network designed for recognizing complex patterns in medical images (Figure 4). This architecture has demonstrated efficacy in various AI applications in dentistry [27–33] due to its residual connections, which facilitate deep network training and mitigate accuracy degradation issues.

The training and validation process of DL follows the same LOOCV scheme used in the evaluation of ML algorithms, including logistic regression (LR), random forest (RF), naive Bayes (NB), and K nearest neighbors (KNN) [26,34]. In this approach, the DL treatment prognosis for each patient is determined by training the model with the images of the remaining patients.

The use of LOOCV is particularly valuable for assessing the performance of artificial intelligence models, as it systematically excludes one data point from the training set, using it as a validation or test instance. Subsequently, a predictive value is generated for the excluded data, and this process is repeated as many times as elements are in the training set. Finally, the predicted values for each excluded data point are compared with the observed values, allowing for a rigorous evaluation of model performance.

In the ML study, once the variable "Prediction by DL" was incorporated into each of the models, the LOOCV scheme was applied again. For variable selection, the Backward Stepwise Selection (BSS) technique was used [34], a commonly employed method for identifying the most relevant features in predictive models.

This study aims to provide evidence regarding the predictive capability of DL in NSRCT prognosis, comparing its performance with conventional ML models and validating its applicability in clinical settings.

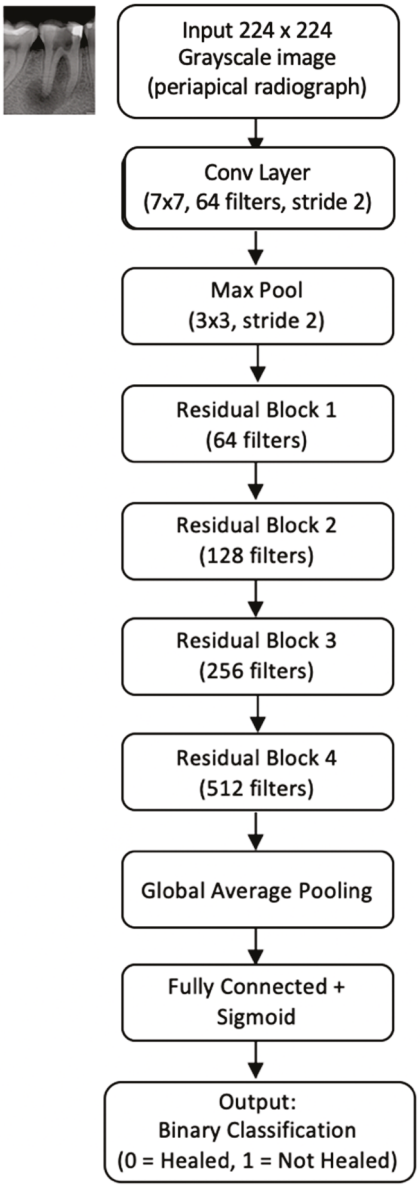


Figure 4. ResNet-18 architecture employed for binary classification of non-surgical root canal treatment (NSRCT) outcomes. The model receives a 224 × 224 grayscale periapical radiograph as input. It processes the image through a series of convolutional layers and residual blocks with increasing filter sizes (64, 128, 256, and 512). A Global Average Pooling layer precedes the Fully Connected Layer, which outputs a binary classification (0: Healed, 1: Not Healed). This architecture enables the model to capture hierarchical features relevant to apical periodontitis prognosis.

3.4. Statistical Analysis

For statistical analysis, we relied on the results obtained in our previous study, where a set of preoperative patient variables, both clinical and demographic, were used as explanatory covariates in various ML models to predict treatment outcomes [26]. In the present study, we included an additional explanatory covariate: the treatment outcome prediction obtained by applying convolutional networks to the diagnostic images of 108 patients, training the networks to forecast the prognosis.

After establishing the performance of DL and the best performing ML model a series of statistical comparisons will be conducted.

For this analysis, Fisher's exact test will be employed, setting a significance level of 0.05. Any result with a p-value below 0.05 will be considered statistically significant. The comparisons to be evaluated are as follows:

- Comparison between the best ML model from the previous study [26], random forest (RF), and the combined DL and best performing ML model.
- Comparison between random forest and DL in general: (RF vs. DL).
- Comparison between the clinical professional's prediction (DP) and the combined DL and best performing ML model.
- Comparison between the clinical professional's prediction (DP) and the DL model by itself: (DP vs. DL).
- Comparison between the combined DL and best performing ML model and DL model.

These comparisons will assess the relative efficacy of different predictive approaches, providing valuable insights into the applicability of AI models in predicting the success of NSRCT.

4. Results

Using the results obtained from the DL study, we applied the chi-square test to detect the association between DL results (DL prediction) and the dentist's outcome, obtaining a p-value of 0.000000127 and an effect size of 0.53 (Table 1) supporting the inclusion of this variable as part of the analysis.

Table 1. Variables associated with the results of the previous ML study, incorporating the DL prediction.

Variable	Levels	p-Value	Effect Size
Age	15–24; 25–34; 35–44; 45–54; 55–64; ≥65	0.0056	0.372
Highest level of education	Primary; Secondary; Post secondary	0.0016	0.33
Arch	Mandible; Maxilla	0.02	0.21
Smoking	No; Every day; Some days; Former	0.046	0.26
Patient co-operation	No; Yes	0.028	0.21
Pain relieved by	None; Cold; Medication	0.003	0.31
Duration of the pain	Sec; Min; Continuous	0.027	0.245
Periapical	Asymptomatic AP; Symptomatic AP; Chronic Apical Abscess; Acute Apical Abscess	0.01	0.31
Estimated prognosis by clinician	Hopeless; Questionable; Fair; Good; Excellent	0.034	0.29
Prediction by DL	Success; Failure	0.000000127	0.53

After replicating the study in [26] with this new variable, the best performance was obtained with an LR model, in which the most influential variables were "DL" (predictions generated by DL networks), "Age", "Smoking", "Level_Education", "Periapical" (periapical condition), and "Prognosis".

The performance of all methods used in this study is presented in Table 2.

Having established the performance of DL and LR, the above-mentioned comparisons were performed.

Table 2. Performance of AI algorithms and the dentist prognosis (DP).

Metric	DP	RF	Logistic Regression (DL-LR)	DL
TP	42	57	57	59
FN	27	12	8	6
FP	21	15	15	18
TN	29	35	28	25
Sensitivity	0.61 (0.48, 0.72)	0.83 (0.72, 0.91)	0.87 (0.77, 0.94)	0.90 (0.80, 0.90)
Specificity	0.58 (0.43, 0.72)	0.7 (0.55, 0.82)	0.65 (0.49, 0.78)	0.58 (0.42, 0.72)
PPV	0.67 (0.54, 0.78)	0.79 (0.68, 0.88)	0.79 (0.67,0.87)	0.76 (0.65, 0.85)
NPV	0.52 (0.38, 0.65)	0.74 (0.6, 0.86)	0.77 (0.60, 0.89)	0.80 (0.62, 0.92)
Accuracy	0.6 (0.5, 0.69)	0.77 (0.69, 0.84)	0.78 (0.69, 0.86)	0.77 (0.68, 0.85)

4.1. Comparison Between Random Forest (RF) and the Deep Learning–Logistic Regression Model (DL-LR)

Overall, DL-LR outperformed the best-performing machine learning model from the previous study [26], random forest, achieving sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy of 0.87, 0.65, 0.79, 0.77, and 0.78, respectively. In comparison, random forest yielded values of 0.83, 0.70, 0.79, 0.74, and 0.77 for the same metrics. However, the differences were not statistically significant, suggesting similar performance between the two models.

4.2. Comparison Between Random Forest and Deep Learning in General (RF vs. DL)

The comparative analysis between the overall DL model and random forest showed no statistically significant differences in their performance.

4.3. Comparison Between the Clinical Professional's Prediction (DP) and Deep Learning–Logistic Regression (DP vs. DL-LR)

In the comparison between DL-LR and DP, DL-LR demonstrated better performance in sensitivity, NPV, and accuracy. Using the true positive (TP) and false negative (FN) values from Table 2, a statistically significant difference was observed in the sensitivity of the logistic regression model with DL compared to the professional's prediction (p-value = 0.00041). However, using the false positive (FP) and true negative (TN) values from the same table, no significant differences were found in specificity or PPV. Nevertheless, significant differences were identified in NPV (p-value = 0.01563) and accuracy (p-value = 0.00253).

4.4. Comparison Between the Clinical Professional's Prediction (DP) and Deep Learning (DL) (DP vs. DL)

Similarly, when comparing the standalone DL model with DP, statistically significant differences were found in sensitivity (p-value = 0.00005), NPV (p-value = 0.0108), and accuracy (p-value = 0.00421), indicating superior performance of DL in these key metrics.

4.5. Comparison Between Deep Learning and the Combined Logistic Regression–Deep Learning Model (DL vs. DL-LR)

Finally, when comparing the individual DL model with the logistic regression model supplemented with categorical variables and the output of the DL model (DL-LR), no statistically significant differences were found in any of the evaluated metrics, suggesting equivalent performance between both models.

4.6. Interpretation of Statistical Comparisons

Based on statistical comparisons: DP vs. DL, RF vs. DL, and DL-LR vs. DL, the following conclusions can be drawn:

Categorical variables have a lower predictive value compared to image-based models.

Sensitivity, NPV, and accuracy metrics show minimal or non-significant differences between RF vs. DL and DL-LR vs. DL models.

High *p*-values (>0.05) in model comparisons indicate no real difference between categorical data-based approaches.

No clear improvements were found in any metric for RF vs. DL and DL-LR vs. DL comparisons.

In contrast, the DP vs. DL comparison showed significant differences:

DL vs. DP exhibited very low *p*-values (0.00005 in sensitivity, 0.00421 in accuracy, 0.0108 in NPV), suggesting that image-based models (likely represented by DL) have superior predictive power compared with dental professionals (DPs).

Similarly, DP vs. AI-based methods showed significant differences:

DL-LR vs. DP displayed very low *p*-values (0.00041 in sensitivity, 0.00253 in accuracy, 0.01563 in NPV), indicating that AI-based methods (DL-LR) have better predictive value than dental professionals (DPs).

5. Discussion

The results obtained in this study, supported by the statistical data collected, highlight the need to compare our AI-based NSRCT prediction for AP with the existing literature to validate our findings scientifically. However, this comparison is challenging due to the limited number of studies dedicated to predicting NSRCT outcomes for apical periodontitis using AI applied to 2D periapical radiographs in endodontics.

AI systems have demonstrated significant advancements in medical imaging, substantially contributing to diagnosis and treatment planning across various specialties. In medicine, convolutional neural networks (CNNs) have been employed for the automatic analysis of pathologies such as breast cancer [35], lung cancer [36,37], and Alzheimer's disease [38]. In dentistry, AI applications have included dental caries detection [39,40], implant classification [40,41], periodontal bone loss quantification [40,42], and cyst evaluation using various types of radiographs, including periapical, panoramic, cephalometric, and CBCT images [43]. In endodontics, AI has been applied to detect apical periodontitis [7] and C-shaped root canals [44].

Although DL applications in medicine are well established [45], studies on disease and treatment outcome prediction in endodontics remain considerably limited [27,35,36]. In this context, Lee et al. (2023) [27] conducted a study predicting endodontic treatment and retreatment outcomes over a three-year period using 598 preoperative periapical radiographs of single-rooted premolars. Utilizing a ResNet-18 CNN model, trained, validated, and tested, their study focused on two main objectives: detecting various clinical features and predicting treatment outcomes. Their findings confirmed the feasibility of DCNN algorithms for feature detection and endodontic prognosis prediction.

Our study shares the objective of evaluating the predictive capability of endodontic treatments using DL with a ResNet-18 architecture, however, our methodology considers all tooth types, not just single-rooted premolars. The selection of single-rooted premolars in Lee et al.'s study [27] was based on the lower anatomical variability of these teeth compared to incisors or molars, which can present heterogeneous periapical conditions [46,47]. Additionally, all cases analyzed in our study exhibited AP, reducing treatment outcome variability. Unlike Lee et al.'s study [27], our research did not include retreatments, which can influence treatment success rates. Furthermore, our study's evaluation period was

extended to nine years, whereas Lee et al. [27] conducted a three-year follow-up. This distinction is relevant, as short-term evaluations may not fully capture the healing process [47].

A key methodological aspect in endodontic treatment evaluation is the use of the periapical index (PAI) score [48]. In Lee et al.'s study [27], only PAI scores 1, 4, and 5 were considered, omitting stage PAI 3, which reflects bone structural changes with minimal demineralization characteristic of apical periodontitis [49]. In our study, we opted to dichotomize the PAI evaluation to avoid ambiguities. Moreover, our study accounted for working length and obturation type, which are critical parameters influencing treatment success rates [50–52].

In a broader context, the literature has explored various AI applications in endodontics. A study employing the AGMB-Transformer model used a dataset of 245 radiographic images of root canal treatments to evaluate its performance in anatomical structure segmentation and outcome classification [53]. Although this study did not focus on treatment prediction, it demonstrated that combining segmentation and classification data significantly improves automated evaluations.

Systematic reviews by Aminoshariae et al. [11], Khanagar et al. [13,16] and Herbst et al. [54] have consolidated knowledge of AI in endodontics, addressing areas such as diagnosis, clinical decision-making, and therapeutic success prediction. However, predicting endodontic treatment outcomes remains an unexplored research gap. Parvathi et al. [10] analyzed AI applications in endodontics, including apical foramen localization, root fracture detection, and retreatment prediction. Campo et al. [55] introduced a case-based reasoning (CBR) system to minimize failed retreatments, however, research addressing NSRCT outcome prediction remains scarce [56].

The use of ResNet-18 architectures in dentistry has proven to be an effective methodology for various applications, including dental caries classification [29], apical periodontitis detection [28], and periodontal disease evaluation [32]. Other studies have employed DL for anatomical structure segmentation [57], predicting inferior alveolar nerve paresthesia after third molar extraction [31], and detecting external root resorptions [30].

Despite advancements in AI applications in endodontics, the current literature presents a shortage of studies focusing on predicting the outcomes of primary endodontic treatments for apical periodontitis. As evidenced by Lee et al. [27] and Yunxiang Li et al. [53], additional studies are imperative. Compared to medicine, where AI has demonstrated significant advancements, efforts in endodontics remain focused on detecting periapical lesions [15,19,28,56,58–60], root morphology analysis [15,19,56,58], and retreatment prediction [55,61], leaving considerable room for future research on NSRCT outcome prediction.

6. Conclusions

The findings of this study suggest that image-based artificial intelligence models (DL) exhibit superior predictive capability compared with those relying solely on categorical data. Significant improvements in DL were observed compared with professional prognosis (DP), whereas, differences among models utilizing categorical data were minimal or statistically insignificant. This finding supports the hypothesis that the information contained in images provides greater richness and discriminatory power in predicting endodontic treatment success compared with categorical data.

These results reinforce the importance of radiographic analysis in evaluating AP and its potential progression, highlighting the critical role of AI models in optimizing clinical diagnoses and therapeutic decision-making. Additionally, further exploration of hybrid models that integrate categorical and imaging data is recommended to enhance predictive accuracy in endodontics.

7. Limitations

Despite the promising findings, this study presents certain limitations that must be considered when interpreting the results. First, the model was developed and validated using a restricted dataset collected from a single institution and obtained using a single radiographic device. This lack of heterogeneity in the sample may affect the generalizability of the results to other populations and clinical settings. Additionally, the limited number of samples available for training could contribute to model overfitting, where the algorithm performs well on the training data but fails to generalize to unseen cases. Although LOOCV was employed to mitigate this risk and maximize data usage, this approach can still yield high variance in performance metrics when applied to small datasets. As a result, the robustness and reliability of the model in broader clinical applications may be limited, underscoring the need for further validation with larger, more diverse cohorts.

Furthermore, the scarcity of previous studies addressing the prediction of the success of NSRCTs for apical periodontitis using artificial intelligence poses a challenge for comparing and validating our findings against the existing literature. The limited availability of specific bibliographic material hinders the direct comparison of our results with other predictive models in endodontics, highlighting the need for further research in this area.

Therefore, we recommend conducting multicenter studies with larger sample sizes and diverse radiographic equipment, as well as integrating complementary clinical data to enhance the applicability of these models in dental practice.

Author Contributions: Conceptualization and methodology, C.B. and A.N.-M.; formal analysis, investigation, data curation, writing—original draft preparation, C.B., A.N.-M., S.A. and Á.A.L.-G.; writing—review and editing, supervision, C.B., A.N.-M., S.A., Y.G.-C., Á.A.L.-G. and P.J.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: This study was conducted in accordance with the Declaration of Helsinki and approved by the Balearic Islands Research Ethics Committee (IB4015/19IP, 2 December 2019).

Informed Consent Statement: Patient consent was waived due to the lack of the possibility of identifying participating patients in the datasets.

Data Availability Statement: Full datasets and R scripts are available upon reasonable request to the corresponding author.

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

References

1. Samuel, A.L. Some Studies in Machine Learning Using the Game of Checkers. *IBM J. Res. Dev.* **1959**, *44*, 206–226. [CrossRef]
2. Hwang, J.-J.; Jung, Y.-H.; Cho, B.-H.; Heo, M.-S. An overview of deep learning in the field of dentistry. *Imaging Sci. Dent.* **2019**, *49*, 1–7. [CrossRef]
3. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
4. Angermueller, C.; Pärnamaa, T.; Parts, L.; Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **2016**, *12*, 878. [CrossRef]
5. Anwar, S.M.; Majid, M.; Qayyum, A.; Awais, M.; Alnowami, M.; Khan, M.K. Medical Image Analysis using Convolutional Neural Networks: A Review. *J. Med. Syst.* **2018**, *42*, 226. [CrossRef]
6. Lin, P.L.; Huang, P.Y.; Huang, P.W. Automatic methods for alveolar bone loss degree measurement in periodontitis periapical radiographs. *Comput. Methods Programs Biomed.* **2017**, *148*, 1–11. [CrossRef] [PubMed]
7. Ekert, T.; Krois, J.; Meinhold, L.; Elhennawy, K.; Emara, R.; Golla, T.; Schwendicke, F. Deep Learning for the Radiographic Detection of Apical Lesions. *J. Endod.* **2019**, *45*, 917–922.e5. [CrossRef]

8. Wang, X.; Cai, B.; Cao, Y.; Zhou, C.; Yang, L.; Liu, R.; Long, X.; Wang, W.; Gao, D.; Bao, B. Objective method for evaluating orthodontic treatment from the lay perspective: An eye-tracking study. *Am. J. Orthod. Dentofac. Orthop.* **2016**, *150*, 601–610. [CrossRef]
9. Alarifi, A.; AlZubi, A.A. Memetic Search Optimization Along with Genetic Scale Recurrent Neural Network for Predictive Rate of Implant Treatment. *J. Med. Syst.* **2018**, *42*, 202. [CrossRef]
10. Gehlot, P.M.; Sudeep, P.; Murali, B.; Mariswamy, A.B. Artificial intelligence in endodontics: A narrative review. *J. Int. Oral Health* **2023**, *15*, 134–141. [CrossRef]
11. Aminoshariae, A.; Kulild, J.; Nagendrababu, V. Artificial Intelligence in Endodontics: Current Applications and Future Directions. *J. Endod.* **2021**, *47*, 1352–1357. [CrossRef]
12. Ourang, S.A.; Sohrabniya, F.; Mohammad-Rahimi, H.; Dianat, O.; Aminoshariae, A.; Nagendrababu, V.; Dummer, P.M.H.; Duncan, H.F.; Nosrat, A. Artificial intelligence in endodontics: Fundamental principles, workflow, and tasks. *Int. Endod. J.* **2024**, *57*, 1546–1565. [CrossRef] [PubMed]
13. Khanagar, S.B.; Alfadley, A.; Alfouzan, K.; Awawdeh, M.; Alaqla, A.; Jamleh, A. Developments and Performance of Artificial Intelligence Models Designed for Application in Endodontics: A Systematic Review. *Diagnostics* **2023**, *13*, 414. [CrossRef] [PubMed]
14. Asiri, A.F.; Altuwalah, A.S. The role of neural artificial intelligence for diagnosis and treatment planning in endodontics: A qualitative review. *Saudi Dent. J.* **2022**, *34*, 270–281. [CrossRef] [PubMed]
15. Karobari, M.I.; Adil, A.H.; Basheer, S.N.; Murugesan, S.; Savadamoorthi, K.S.; Mustafa, M.; Abdulwahed, A.; Almokhatieb, A.A. Evaluation of the Diagnostic and Prognostic Accuracy of Artificial Intelligence in Endodontic Dentistry: A Comprehensive Review of Literature. *Comput. Math. Methods Med.* **2023**, *2023*, 7049360. [CrossRef]
16. Khanagar, S.B.; Al-ehaideb, A.; Maganur, P.C.; Vishwanathaiah, S.; Patil, S.; Baeshen, H.A.; Sarode, S.C.; Bhandi, S. Developments, application, and performance of artificial intelligence in dentistry—A systematic review. *J. Dent. Sci.* **2021**, *16*, 508–522. [CrossRef]
17. Machoy, M.E.; Szyszka-Sommerfeld, L.; Vegh, A.; Gedrange, T.; Woźniak, K. The ways of using machine learning in dentistry. *Adv. Clin. Exp. Med.* **2020**, *29*, 375–384. [CrossRef]
18. Zhang, Y.; Weng, Y.; Lund, J. Applications of Explainable Artificial Intelligence in Diagnosis and Surgery. *Diagnostics* **2022**, *12*, 237. [CrossRef]
19. Pethani, F. Promises and perils of artificial intelligence in dentistry. *Aust. Dent. J.* **2021**, *66*, 124–135. [CrossRef]
20. Schwendicke, F.; Samek, W.; Krois, J. Artificial Intelligence in Dentistry: Chances and Challenges. *J. Dent. Res.* **2020**, *99*, 769–774. [CrossRef]
21. Alasqah, M.; Alotaibi, F.D.; Gufran, K. The Radiographic Assessment of Furcation Area in Maxillary and Mandibular First Molars while Considering the New Classification of Periodontal Disease. *Healthcare* **2022**, *10*, 1464. [CrossRef] [PubMed]
22. Basrani, B. *Endodontic Radiology, Segunda*; Wiley-Blackwell: Hoboken, NJ, USA, 2012; pp. 36–38.
23. Azarpazhooh, A.; Khazaei, S.; Jafarzadeh, H.; Malkhassian, G.; Sgro, A.; Elbarbary, M.; Cardoso, E.; Oren, A.; Kishen, A.; Shah, P.S. A Scoping Review of Four Decades of Outcomes in Nonsurgical Root Canal Treatment, Nonsurgical Retreatment, and Apexification Studies: Part 3—A Proposed Framework for Standardized Data Collection and Reporting of Endodontic Outcome Studies. *J. Endod.* **2022**, *48*, 40–54. [CrossRef]
24. Azarpazhooh, A.; Sgro, A.; Cardoso, E.; Elbarbary, M.; Lighvan, N.L.; Badewy, R.; Malkhassian, G.; Jafarzadeh, H.; Bakhtiar, H.; Khazaei, S.; et al. A Scoping Review of 4 Decades of Outcomes in Nonsurgical Root Canal Treatment, Nonsurgical Retreatment, and Apexification Studies—Part 2: Outcome Measures. *J. Endod.* **2022**, *48*, 29–39. [CrossRef] [PubMed]
25. Azarpazhooh, A.; Cardoso, E.; Sgro, A.; Elbarbary, M.; Lighvan, N.L.; Badewy, R.; Malkhassian, G.; Jafarzadeh, H.; Bakhtiar, H.; Khazaei, S.; et al. A Scoping Review of 4 Decades of Outcomes in Nonsurgical Root Canal Treatment, Nonsurgical Retreatment, and Apexification Studies—Part 1: Process and General Results. *J. Endod.* **2022**, *48*, 15–28. [CrossRef]
26. Bennasar, C.; García, I.; Gonzalez-Cid, Y.; Pérez, F.; Jiménez, J. Second Opinion for Non-Surgical Root Canal Treatment Prognosis Using Machine Learning Models. *Diagnostics* **2023**, *13*, 2742. [CrossRef]
27. Lee, J.; Seo, H.; Choi, Y.J.; Lee, C.; Kim, S.; Lee, Y.S.; Lee, S.; Kim, E. An Endodontic Forecasting Model Based on the Analysis of Preoperative Dental Radiographs: A Pilot Study on an Endodontic Predictive Deep Neural Network. *J. Endod.* **2023**, *49*, 710–719. [CrossRef]
28. Li, S.; Liu, J.; Zhou, Z.; Zhou, Z.; Wu, X.; Li, Y.; Wang, S.; Liao, W.; Ying, S.; Zhao, Z. Artificial intelligence for caries and periapical periodontitis detection. *J. Dent.* **2022**, *122*, 104107. [CrossRef] [PubMed]
29. Panyarak, W.; Wantanajittikul, K.; Suttapak, W.; Charuakkra, A.; Prapayasatok, S. Feasibility of deep learning for dental caries classification in bitewing radiographs based on the ICCMS™ radiographic scoring system. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **2023**, *135*, 272–281. [CrossRef]
30. Mohammad-Rahimi, H.; Dianat, O.; Abbasi, R.; Zahedrozegar, S.; Ashkan, A.; Motamedian, S.R.; Rohban, M.H.; Nosrat, A. Artificial Intelligence for Detection of External Cervical Resorption Using Label-Efficient Self-Supervised Learning Method. *J. Endod.* **2024**, *50*, 144–153.e2. [CrossRef]

31. Kim, B.S.; Yeom, H.G.; Lee, J.H.; Shin, W.S.; Yun, J.P.; Jeong, S.H.; Kang, J.H.; Kim, S.W.; Kim, B.C. Deep Learning-Based Prediction of Paresthesia after Third Molar Extraction: A Preliminary Study. *Diagnostics* **2021**, *11*, 1572. [CrossRef]
32. Vilkomir, K.; Phen, C.; Baldwin, F.; Cole, J.; Herndon, N.; Zhang, W. Classification of mandibular molar furcation involvement in periapical radiographs by deep learning. *Imaging Sci. Dent.* **2024**, *54*, 257–263. [CrossRef]
33. Kim, Y.-H.; Park, J.-B.; Chang, M.-S.; Ryu, J.-J.; Lim, W.H.; Jung, S.-K. Influence of the depth of the convolutional neural networks on an artificial intelligence model for diagnosis of orthognathic surgery. *J. Pers. Med.* **2021**, *11*, 356. [CrossRef] [PubMed]
34. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*, 2nd ed.; Springer: New York, NY, USA, 2021; pp. 229–232.
35. El Adoui, M.; Drisis, S.; Benjelloun, M. Multi-input deep learning architecture for predicting breast tumor response to chemotherapy using quantitative MR images. *Int. J. Comput. Assist. Radiol. Surg.* **2020**, *15*, 1491–1500. [CrossRef] [PubMed]
36. Mukherjee, P.; Zhou, M.; Lee, E.; Schicht, A.; Balagurunathan, Y.; Napel, S.; Gillies, R.; Wong, S.; Thieme, A.; Leung, A.; et al. A shallow convolutional neural network predicts prognosis of lung cancer patients in multi-institutional computed tomography image datasets. *Nat. Mach. Intell.* **2020**, *2*, 274–282. [CrossRef] [PubMed]
37. Xie, H.; Zhang, T.; Song, W.; Wang, S.; Zhu, H.; Zhang, R.; Zhang, W.; Yu, Y.; Zhao, Y. Super-resolution of Pneumocystis carinii pneumonia CT via self-attention GAN. *Comput. Methods Programs Biomed.* **2021**, *212*, 106467. [CrossRef]
38. Ricucci, D.; Siqueira, J.F., Jr. Biofilms and Apical Periodontitis: Study of Prevalence and Association with Clinical and Histopathologic Findings. *J. Endod.* **2010**, *36*, 1277–1288. [CrossRef]
39. Lee, J.-H.; Kim, D.-H.; Jeong, S.-N.; Choi, S.-H. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. *J. Dent.* **2018**, *77*, 106–111. [CrossRef]
40. Shan, T.; Tay, F.R.; Gu, L. Application of Artificial Intelligence in Dentistry. *J. Dent. Res.* **2021**, *100*, 232–244. [CrossRef]
41. Kim, J.-E.; Nam, N.-E.; Shim, J.-S.; Jung, Y.-H.; Cho, B.-H.; Hwang, J.J. Transfer Learning via Deep Neural Networks for Implant Fixture System Classification Using Periapical Radiographs. *J. Clin. Med.* **2020**, *9*, 1117. [CrossRef]
42. Krois, J.; Ekert, T.; Meinhold, L.; Golla, T.; Kharbot, B.; Wittemeier, A.; Dörfer, C.; Schwendicke, F. Deep Learning for the Radiographic Detection of Perio-dontal Bone Loss. *Sci. Rep.* **2019**, *9*, 8495. [CrossRef]
43. Hung, K.; Montalvao, C.; Tanaka, R.; Kawai, T.; Bornstein, M.M. The use and performance of artificial intelligence applications in dental and maxillofacial radiology: A systematic review. *Dentomaxillofacial Radiol.* **2019**, *49*, 20190107. [CrossRef] [PubMed]
44. Yang, S.; Lee, H.; Jang, B.; Kim, K.-D.; Kim, J.; Kim, H.; Park, W. Development and Validation of a Visually Explainable Deep Learning Model for Classification of C-shaped Canals of the Mandibular Second Molars in Periapical and Panoramic Dental Radiographs. *J. Endod.* **2022**, *48*, 914–921. [CrossRef] [PubMed]
45. Umer, F.; Habib, S. Critical Analysis of Artificial Intelligence in Endodontics: A Scoping Review. *J. Endod.* **2021**, *48*, 152–160. [CrossRef] [PubMed]
46. Chugal, N.M.; Clive, J.M.; Spångberg, L.S. A prognostic model for assessment of the outcome of endodontic treatment: Effect of biologic and diagnostic variables. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endodontol.* **2001**, *91*, 342–352. [CrossRef]
47. Friedman, S. Prognosis of initial endodontic therapy. *Endod. Top.* **2002**, *2*, 59–88. [CrossRef]
48. Moidu, N.P.; Sharma, S.; Chawla, A.; Kumar, V.; Logani, A. Deep learning for categorization of endodontic lesion based on radiographic periapical index scoring system. *Clin. Oral Investig.* **2021**, *26*, 651–658. [CrossRef]
49. Jiménez Pinzón, A.; Segura Egea, J.J. Valoración clínica y radiológica del estado periapical: Registros e índices periapicales. *Endodoncia* **2003**, *21*, 220–228.
50. Chugal, N.M.; Clive, J.M.; Spångberg, L.S. Endodontic infection: Some biologic and treatment factors associated with outcome. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endodontol.* **2003**, *96*, 81–90. [CrossRef]
51. Friedman, S.; Abitbol, S.; Lawrence, H. Treatment Outcome in Endodontics: The Toronto Study. Phase 1: Initial Treatment. *J. Endod.* **2003**, *29*, 787–793. [CrossRef]
52. Farzaneh, M.; Abitbol, S.; Lawrence, H.; Friedman, S. Treatment Outcome in Endodontics—The Toronto Study. Phase II: Initial Treatment. *J. Endod.* **2004**, *30*, 302–309. [CrossRef]
53. Li, Y.; Zeng, G.; Zhang, Y.; Wang, J.; Jin, Q.; Sun, L.; Zhang, Q.; Lian, Q.; Qian, G.; Xia, N.; et al. AGMB-Transformer: Anatomy-Guided Multi-Branch Transformer Network for Automated Evaluation of Root Canal Therapy. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 1684–1695. [CrossRef] [PubMed]
54. Herbst, C.S.; Schwendicke, F.; Krois, J.; Herbst, S.R. Association between patient-, tooth- and treatment-level factors and root canal treatment failure: A retrospective longitudinal and machine learning study. *J. Dent.* **2022**, *117*, 103937. [CrossRef]
55. Campo, L.; Aliaga, I.J.; De Paz, J.F.; García, A.E.; Bajo, J.; Villarubia, G.; Corchado, J.M. Retreatment Predictions in Odontology by means of CBR Systems. *Comput. Intell. Neurosci.* **2016**, *2016*, 7485250. [CrossRef] [PubMed]
56. Ramezanzade, S.; Laurentiu, T.; Bakhshandah, A.; Ibragimov, B.; Kvist, T.; EndoReCo, E.; Bjorndal, L. The efficiency of artificial intelligence methods for finding radiographic features in different endodontic treatments—A systematic review. *Acta Odontol. Scand.* **2023**, *81*, 422–435. [PubMed]

57. Sunnetci, K.M.; Kaba, E.; Beyazal Çeliker, F.; Alkan, A. Comparative parotid gland segmentation by using ResNet-18 and MobileNetV2 based DeepLab v3+ architectures from magnetic resonance images. *Concurr. Comput.* **2023**, *35*, e7405. [CrossRef]
58. Boreak, N. Effectiveness of Artifcial Intelligence Applications Designed for Endodontic Diagnosis, Decision-making, and Pre-diction of Prognosis: A Systematic Review. *J. Contemp. Dent. Pract.* **2020**, *21*, 926–934. [CrossRef]
59. Pauwels, R.; Brasil, D.M.; Yamasaki, M.C.; Jacobs, R.; Bosmans, H.; Freitas, D.Q.; Haiter-Neto, F. Artificial intelligence for detection of periapical lesions on intraoral radiographs: Comparison between convolutional neural networks and human observers. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **2021**, *131*, 610–616. [CrossRef]
60. Sadr, S.; Mohammad-Rahimi, H.; Motamedian, S.R.; Zahedrozegar, S.; Motie, P.; Vinayahalingam, S.; Dianat, O.; Nosrat, A. Deep Learning for Detection of Periapical Radiolucent Lesions: A Systematic Review and Meta-analysis of Diagnostic Test Accuracy. *J. Endod.* **2023**, *49*, 248–261.e3. Available online: https://linkinghub.elsevier.com/retrieve/pii/S0099239922008457 (accessed on 17 January 2023). [CrossRef]
61. Sherwood, A.A.; Setzer, F.C.; K, S.D.; Shamili, J.V.; John, C.; Schwendicke, F. A Deep Learning Approach to Segment and Classify C-Shaped Canal Morphologies in Mandibular Second Molars Using Cone-beam Computed Tomography. *J. Endod.* **2021**, *47*, 1907–1916. [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.

diagnostics

MDPI

Article

Comparison of Supervised Machine Learning Models to Logistic Regression Model Using Tooth-Related Factors to Predict the Outcome of Nonsurgical Periodontal Treatment

Ali J. B. Al-Sharqi [1], Mohammed Taha Ahmed Baban [2], Nada K. Imran [1], Sarhang S. Gul [3,4,*] and Ali A. Abdulkareem [1]

[1] Department of Periodontics, College of Dentistry, University of Baghdad, Bab Al Mudam, Baghdad P.O. Box 1417, Iraq; alijaaferalsharki@gmail.com (A.J.B.A.-S.); nada.k.omran@codental.uobaghdad.edu.iq (N.K.I.); ali.abbas@codental.uobaghdad.edu.iq (A.A.A.)
[2] Department of Dental Nursing, Sulaimani Technical Institute, Sulaimani Polytechnic University, Sulaymaniyah P.O. Box 20-236, Iraq; mohammed.baban@spu.edu.iq
[3] Medical Laboratory Department, College of Health and Medical Technology, Sulaimani Polytechnic University, Sulaymaniyah P.O. Box 20-236, Iraq
[4] Department of Periodontics, College of Dentistry, University of Sulaimani, Sulaymaniyah 46001, Iraq
* Correspondence: sarhang.hama@univsul.edu.iq or sarhang.hama@spu.edu.iq

Abstract: Background/Objectives: Conventional logistic regression is widely used in the field of dentistry, specifically for prediction purposes in longitudinal studies. This study aimed to compare the validity of different supervised machine learning (ML) models to the conventional logistic regression (LR) model to predict the outcomes of nonsurgical periodontal treatment (NSPT). **Methods:** Patients diagnosed with periodontitis received full periodontal charting, including bleeding on probing (BoP), probing pocket depth (PPD), and clinical attachment loss (CAL). Furthermore, the tooth type, tooth location, tooth surface, arch type, and gingival phenotype were also collected as site-specific predictors. Later, root surface debridement was provided and treatment outcomes were evaluated after 3 months. Site-specific predictors were used to train five ML models, including random forest (RF), decision tree (DT), support vector classifier (SVC), K-nearest neighbors (KNN), and Gaussian naïve Bayes (GNB), to develop predictive models. **Results:** Site-specific predictors of 1108 examined sites were used, and the overall accuracy prediction of the conventional LR model was 70.4%, with PPD statistically significantly associated with the outcome of NSPT (odds ratio = 0.577, $p = 0.001$). Among the ML models examined, only GNB and SVC showed comparable prediction accuracy (71.0% and 70.4%, respectively) to the LR model, whereas the prediction accuracies of KNN, RF, and DT were 65.0%, 62.0%, and 61.0%, respectively. Similarly, baseline PPD was shown to be the most important featured predictor by both the RF and DT models. **Conclusions**: The evidence suggests that supervised ML models do not outperform the LR model in predicting the outcomes of NSPT. A larger sample size and more predictors of periodontitis are necessary to enhance the accuracy of ML models over the LR model in predicting the outcomes of NSPT.

Keywords: periodontitis; nonsurgical periodontal treatment; tooth-related factors; machine learning models; logistic regression

1. Introduction

There is ample evidence supporting the efficacy of nonsurgical periodontal treatment (NSPT) for treating inflammatory periodontal disease, which is one of the most prevalent

diseases affecting populations worldwide [1]. The term NSPT mainly refers to mechanical methods used to eradicate dysbiotic biofilms, calculus, and other debris coronally and from below the gingival line, performed with either manual or ultrasonic scalers [2]. NSPT is the recommended technique for treating periodontal pockets with shallow to moderate depth and reducing bacterial load from deep periodontal pockets at the early stage before surgical periodontal treatment [2]. Rapid improvement of clinical periodontal parameters, within the 12-week follow-up period after NSPT was observed, with marked reduction in the bacterial load of the subgingival and salivary microbiotas [3]. However, microbiological analysis of periodontal pockets 3 months after NSPT showed that the non-responding sites were dominated by the bacterial species *Porphyromonas gingivalis* and *Treponema denticola* [4]. The current guideline recommends another cycle of NSPT following step 1 and 2 therapy of residual pockets of 4–5 mm; however, for sites with residual pockets of $\geq$6 mm, surgical periodontal treatment is recommended [5].

It is acknowledged that the efficacy of NSPT is highly influenced by both patient- and tooth-related factors, with many variables identified [6–11]. The expected success rate could be more than 80% or less than 60% depending on a range of factors, including tooth type, anatomical landmarks, operator skill, genetics, smoking habits, uncontrolled diabetes, and pathogenicity of bacteria [1,12]. Indeed, identifying these factors in advance before providing NSPT is of paramount importance not only for the clinician evaluating the efficacy of NSPT but also in enabling them to provide patients with clearer information regarding expected outcomes, possible additional cycles of NSPT, or selecting other therapeutic options, such as surgical periodontal therapy in the future [2].

Researchers have examined several tooth-related factors as predictive indicators of NSPT, such as baseline probing pocket depth (PPD), clinical attachment loss (CAL), bleeding on probing (BoP), location of the tooth, number of roots, surface of the tooth, and arch [6]. However, evidence on the impact of tooth-related factors on the efficacy of NSPT is still inconclusive, and further studies are encouraged [13–16].

On the other hand, concern has been raised regarding periodontal data with an inherent hierarchical structure, which might pose difficulties for the statistical analysis. It has been reported that aggregation of site data for an individual patient and its expression as a mean might lead to masking of critical information and overestimation of the standard errors. In contrast, analyzing data at the tooth or site level without considering the interdependence of teeth or sites within the same patient can lead to an underestimation of the standard error. One way to address this issue is by using multilevel modeling—a statistical technique originally developed in educational research, but now widely applied across various areas of health sciences [17,18].

The implication of logistic regression (LR) is that it allows the impact of different tooth-related factors on the outcome of NSPT to be identified. Furthermore, the use of a backward stepwise method in LR helps to exclude redundant parameters without affecting the overall accuracy of the model. Thus, this conventional statistical test is widely used in the field of dentistry, specifically for prediction purposes in longitudinal studies [19,20]. Additionally, predictive models frequently lack externally validation, leaving uncertainty about whether these algorithms can be reliably applied to other diverse populations.

Recently, machine learning (ML) models have gained growing attention, as they rely on developing algorithm and statistical models from data rather than direct programming for statistical analysis. ML models offer significant benefits for analyzing large-scale, complex, and diverse data commonly produced in biomedical research. While their use in dental data analysis is still emerging, they show strong potential for improving the accuracy of diagnostic pattern recognition [21,22]. Furthermore, recent systematic reviews that have examined studies using ML models for periodontal screening, diagnosis, and

prognosis found that supervised ML has potential for use in prognostic models applied in periodontology; however, the current evidence is not conclusive [23,24].

Supervised ML models, such as decision tree (DT), K-nearest neighbors (KNN), random forest (RF), support vector classification (SVC), and Gaussian naïve Bayes (GNB) have been used for diagnostic and prognostic purposes based on the input data [24–26]; however, the reliability and accuracy of these models in comparison to the most widely used LR statistical test have yet to be examined [27]. Therefore, the current study aimed to compare the accuracy of the above supervised ML models to the LR model using the baseline data of several tooth-related factors as a "typical" dataset to predict the outcome of NSPT. It is important to acknowledge that the tooth-related factors used in this study were solely included for the purpose of comparison between supervised ML models and LR rather than actual prediction of NSPT.

2. Materials and Methods

2.1. Study Design

This cohort study included patients with periodontitis seeking periodontal treatment who were recruited consecutively by the Department of Periodontics, College of Dentistry, University of Baghdad between April 2024 and April 2025. Ethical rules declared by the World Medical Association Declaration of Helsinki and the later amendments were followed in the current study. Initially, ethical approval for conducting this study was issued by the Ethics Committee, College of Dentistry, University of Baghdad (Reference No. 663). The purpose of the study, details of the intended treatment plan, and other relevant information were fully declared to each patient before signing a consent form.

2.2. Eligibility Criteria

The inclusion criteria were patients aged $\geq$18 years with no history of systemic diseases, such as diabetes mellitus, that would affect NSPT. The recruited patients needed to be diagnosed with periodontitis, regardless of the stage and grade, with 10 pairs of occluding teeth. The case definition of periodontitis as stated by the American Academy of Periodontology/European Federation of Periodontology was followed in this study. Evidence of interproximal CAL in $\geq$2 non-adjacent teeth, or presence of a detectable CAL $\geq$ 3 mm combined with PPD $\geq$ 4 mm on the buccal/lingual surfaces were the criteria used to diagnose periodontitis [28]. Periodontal pockets included in this study needed to exhibit a PPD between 4 and 6 mm, with no furcation involvement, as indications for NSPT. Current smokers, pregnant women, those receiving recent or ongoing periodontal treatment, and patients not willing to consent were excluded from this cohort.

2.3. Tooth-Related Parameters and Intervention

Periodontal clinical parameters, including BoP [29], PPD, CAL, and gingival phenotype, were recorded for each patient at baseline and after 3 months by the same clinician (A.J.B.A.) at four surfaces per tooth, using the UNC-15 periodontal probe (Medesy, Maniago, Italy). Absence or presence of BoP was recorded as 0 or 1, respectively. Simultaneously, PPD and CAL were measured, representing the linear distances from the gingival margin and cementoenamel junction, respectively, to the base of the sulcus/pocket.

Gingival phenotype was determined using the Colorvue™ Biotype Probe (Hu-Friedy®, Chicago, IL, USA), which is a set of three probes with different color coding, used in a sequential order to define the phenotype [30]. In detail, the white-colored probe was first inserted in the gingival crevice, and if the color was visible through the gingiva, the phenotype was thin. If the transparency of the white color through the gingiva was completely blocked, the same procedure was repeated with the green-colored probe, which

indicated medium phenotype if the color was apparent. If not apparent, the blue tip probe was inserted in a similar manner. Transparency of the blue color through the gingiva indicated a thick phenotype, whereas if the blue color was not visible, the gingival tissue was classified as very thick. These last two thickness possibilities were considered as a thick phenotype group. Furthermore, the tooth-related factors collected were tooth type (single-rooted or multirooted teeth), tooth location (incisors, premolars, and molars), arch type (maxillary and mandibular teeth), and tooth surfaces (mesial, distal, facial, and oral surfaces).

Oral hygiene instructions were provided for each patient, followed first by professional mechanical plaque removal using an ultrasonic scaler (Woodpecker, Ultrasonic Piezoelectric Scaler UDS-A, Guilin, China) and then by polishing the teeth. The patients were instructed to return after one week to receive root surface debridement, using Gracey curettes (Medesy, Maniago, Italy) for all periodontal pockets. Similar types of tooth brush, toothpaste, and interdental brush (Curaprox, Kriens, Switzerland) were provided for each patient. All measurements were performed by the same calibrated examiner, using a UNC-15 periodontal probe (Medesy, Maniago, Italy). The level of consistency for categorical parameters (BoP) was >80% using the kappa coefficient test, and for continuous parameters (PPD and CAL), it was >90%, as indicated by the interclass coefficient test.

2.4. Outcomes and Sample Size

Reduction of the PPD to $\leq$4 mm without BoP [31] after 3 months was set as the primary outcome indicating the success of NSPT. Secondary outcomes included BoP, CAL, gingival phenotype, tooth type, tooth location, arch type, and tooth surfaces. Accordingly, successful/unsuccessful pocket closure was accounted for when calculating the priori sample size. Data from a previous study indicated a successful/unsuccessful ratio of NSPT as 1.6 [6], with 0.05 α-error, and 0.8 power. The estimated sample size was 65 patients, which was rounded to 72 patients to allow for 10% dropout.

2.5. Statistical Analysis

2.5.1. Conventional Statistical Analysis

Mean, standard deviation, frequency, and percentage were used to describe continuous and categorical variables. The Shapiro–Wilk test was used to determine whether the data were parametric or non-parametric. Differences in the PPD and CAL clinical parameters between the baseline and endpoint were determined using paired *t*-tests, while a chi-squared test was performed to detect differences between categorical clinical parameters of BoP, gingival phenotype, tooth type, tooth location, arch type, and tooth surfaces. A binary LR model was used to determine the potential of the tooth-related clinical parameters to predict the outcomes of NSPT. The dependent variable was the successful/unsuccessful NSPT, which was dichotomized into 1 or 0, respectively. For the binary or categorical tooth-related factors (BoP, gingival phenotype, tooth type, tooth location, arch type, and tooth surfaces), the first variables were used as a reference. The level of significance was set at a p value less than 0.05, and all statistical analyses were conducted using GraphPad Prism (GraphPad Prism Software, version 9, San Diego, CA, USA) and SPSS (version 26, IBM, Boston, MA, USA) software.

2.5.2. Machine Learning Models Algorithms

All ML models in this study were developed and evaluated on a Lenovo laptop (Model: 20M3, Lenovo group limited, Beijing, China) equipped with an 8th-generation Intel Core i5 processor (Intel corporation, Chandler, AZ, USA). Five ML models of DT, RF, KNN, SVC, and GNB were applied. The dataset was randomly divided into training (80%) and testing (20%) subsets. Feature importance analysis was conducted to identify the most influential

predictors underlying model decisions. Model performance was assessed using confusion matrices and ROC curves for each model. To ensure robustness, tenfold cross-validation with random shuffling and ten repetitions was employed, and the mean accuracy and standard deviation were calculated. All procedures, including data preprocessing, model training, evaluation, and visualization, were conducted using Python software, version 3.13.5, developed by the Python Software Foundation, Wilmington, DE, USA.

3. Results

3.1. Basic Demographic and Tooth-Related Parameters

After applying the eligibility criteria, 72 periodontitis patients were recruited (out of 274). A total of 5 patients did not commit to the planned appointment, leaving 67 patients for the final analysis (Figure 1).

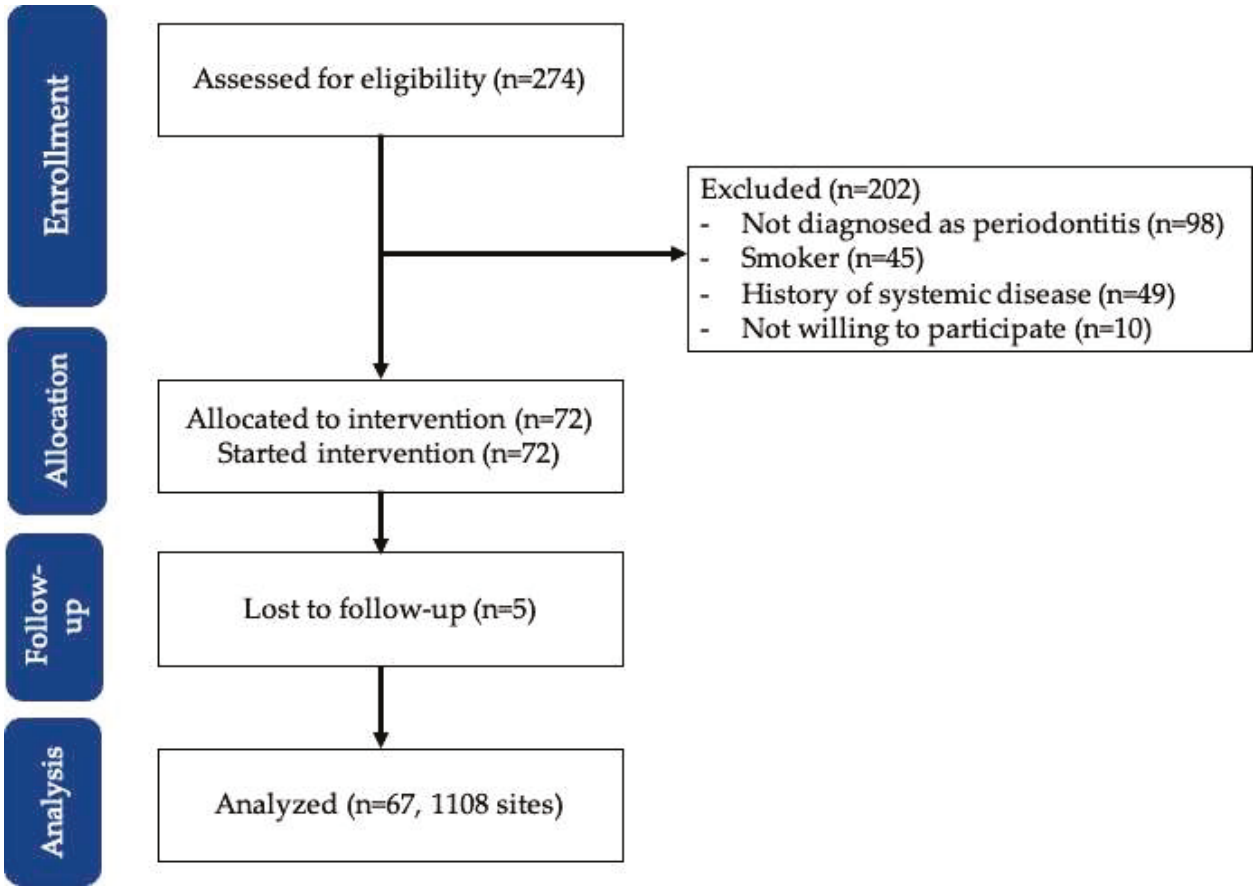


Figure 1. Flow diagram of the study.

The mean age of recruited patients (33 male and 34 female) was 56.72 ± 7.10 years old (Table 1). The clinical periodontal parameters PPD, CAL, and BoP were significantly improved 3 months after NSPT. Out of 1108 sites examined, 781 (70.49%) periodontal pockets exhibited pocket closure as a successful outcome (Table 1). On the other hand, when the baseline tooth-related factors were analyzed versus their outcome (successful or unsuccessful) after 3 months, only baseline PPD and tooth type (single-rooted vs. multirooted teeth) showed statistically significant differences (Table 2).

Table 1. Basic demographic and clinical characteristics of the study population (n = 67).

Variables	Baseline	Endpoint
Age † (years)	56.72 ± 7.10	
Sex ‡		
Male	33, 49.30%	
Female	34, 50.70%	
PPD † (mm)	4.74 ± 0.66	3.83 ± 0.69 *
CAL † (mm)	3.57 ± 0.66	3.19 ± 0.61 *
BoP ‡		
No bleeding	231, 20.85%	880, 79.42% **
Bleeding	877, 79.15%	228, 20.58% **
Gingival phenotype ‡		
Thin	142, 12.82%	
Medium	642, 57.94%	
Thick	324, 29.24%	
Tooth type ‡		
Single-rooted	696, 62.82%	
Multirooted	412, 37.18%	
Tooth location ‡		
Incisors	413, 37.27%	
Premolars	371, 33.49%	
Molars	324, 29.24%	
Arch ‡		
Maxillary	637, 57.49%	
Mandibular	471, 42.51%	
Tooth surface ‡		
Mesial	497, 44.86%	
Distal	487, 43.95%	
Facial	116, 10.47%	
Oral	8, 0.72%	
Site-specific outcomes (n = 1108) ‡		
Successful		781, 70.49%
Unsuccessful		327, 29.51%

BoP: bleeding on probing, CAL: clinical attachment level, PPD: probing pocket depth. † Mean ± SD; ‡ frequency, percent; * significant difference at $p < 0.05$ using paired t-test; ** significant difference at $p < 0.05$ using paired chi-squared test.

Table 2. Comparison of baseline tooth-related factors in successful (n = 782) and unsuccessful (n = 329) sites.

Tooth-Related Factors	Successful	Unsuccessful	p Value
PPD † (mm)	4.65 ± 0.61	4.96 ± 0.71	<0.001 *
CAL † (mm)	3.57 ± 0.66	3.57 ± 0.64	0.9

Table 2. *Cont.*

Tooth-Related Factors	Successful	Unsuccessful	*p* Value
BoP ‡			
No bleeding	621, 56.05%	256, 23.10%	0.57
Bleeding	159, 14.35%	72, 6.50%	
Gingival phenotype ‡			
Thin	100, 9.03%	42, 3.79%	0.93
Medium	450, 40.61%	192, 17.33%	
Thick	231, 20.85%	93, 8.39%	
Tooth type ‡			
Single-rooted	510, 46.03%	186, 16.79%	0.01 **
Multirooted	271, 24.46%	141, 12.72%	
Tooth location ‡			
Incisors	309, 27.80%	105, 9.48%	0.054
Premolars	255, 22.82%	118, 10.65%	
Molars	218, 19.68%	106, 9.57%	
Arch ‡			
Maxillary	436, 39.35%	201, 18.14%	0.11
Mandibular	344, 31.05%	127, 11.46%	
Tooth surface ‡			
Mesial	332, 29.96%	165, 14.89%	0.10
Distal	356, 32.13%	131, 11.83%	
Facial	87, 7.85%	29, 2.62%	
Oral	5, 0.45%	3, 0.27%	

† Mean ± SD; ‡ frequency, percent; * significant difference at $p < 0.05$ using paired *t*-test; ** significant difference at $p < 0.05$ by using chi-squared test.

3.2. Logistic Regression Model

Ability of tooth-related factors to predict the outcomes of NSPT was assessed by a binary LR model (Table 3). The results indicate that only PPD was statistically significantly associated with outcomes of NSPT (odds ratio = 0.577), with a negative estimated coefficient. This suggests that increasing the PPD at baseline was a predictor of unsuccessful treatment outcome after 3 months and vice versa. No other clinical tooth-related factors showed any significant effect on the outcomes of NSPT. Lastly, the overall accuracy of this model was 70.4%.

Table 3. Logistic regression for predictors of site-specific outcome for nonsurgical periodontal treatment.

Predictors	B	SE	*p* Value	Exp (B)	95% CI for EXP (B)
PPD	−0.550	0.103	0.001	0.577	0.471 to 0.706
CAL	0.067	0.102	0.51	1.070	0.875 to 1.307
BoP [a]					
Bleeding	−0.28	0.169	0.098	0.75	0.542 to 1.053
Gingival phenotype [b]					

Table 3. *Cont.*

Predictors	B	SE	*p* Value	Exp (B)	95% CI for EXP (B)
Medium	0.009	0.234	0.971	1.009	0.637 to 1.579
Thick	0.049	0.167	0.772	1.05	0.756 to 1.457
Tooth type [c]					
Multirooted	0.4	0.281	0.155	1.491	0.86 to 2.585
Tooth location [d]					
Premolars	−0.348	0.346	0.314	0.706	0.358 to 1.391
Molars	−0.399	0.277	0.149	0.671	0.39 to 1.154
Arch [e]					
Mandibular	−0.118	0.153	0.44	0.889	0.659 to 1.199
Tooth surface [f]					
Distal	−0.16	0.741	0.983	0.984	0.23 to 4.206
Facial	0.261	0.743	0.725	1.299	0.303 to 5.568
Oral	0.28	0.767	0.715	1.323	0.294 to 5.942
Accuracy of the model	70.4%				

B: estimated coefficient, SE: standard error, Exp (B): odds ratio, CI: confidence interval, PPD: probing pocket depth, CAL: clinical attachment loss. Reference: ([a]) non-bleeding, ([b]) thin, ([c]) single-rooted, ([d]) incisors, ([e]) maxillary, ([f]) mesial.

3.3. Machine Learning Models

Amongst the ML models, SVC and GNB showed the highest AUCs of 0.61 and 0.59, respectively, indicating their ability to differentiate the successful and unsuccessful treatment outcomes of NSPT after 3 months. In contrast, both the RF and KNN models demonstrated an AUC of 0.53, and the lowest AUC (0.5) was shown by DT (Figure 2).

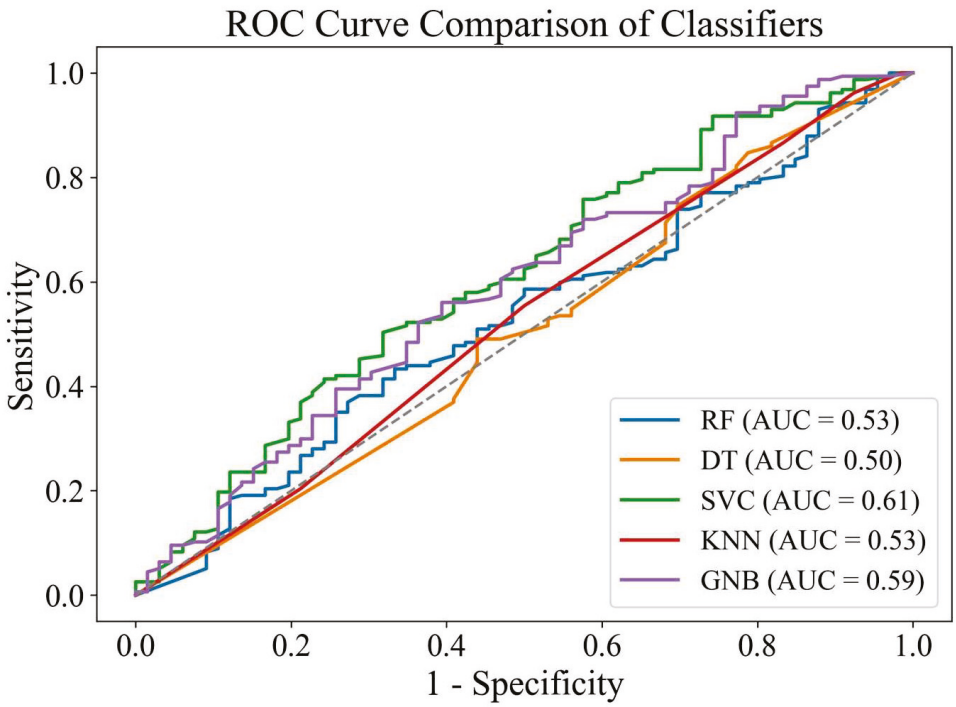


Figure 2. ROC and AUC of tested ML models to predict the outcome of NSPT after 3 months (the reference line is presented as dashed line).

It is apparent that GNB and SVC had the highest testing accuracies of 0.71 and 0.704, respectively, with the lowest training accuracy (GNB = 0.72, SVC = 0.703) (Table 4). Furthermore, the confusion matrix of GNB was able to correctly identify 13 out of 66 unsuccessful sites and 147 out of 157 successful sites. In contrast, the confusion matrix of SVC did not identify any unsuccessful sites and was able to identify all successful sites (Figure 3). RF and DT had the highest training accuracies (0.81 both) with the lowest testing accuracies of 0.62 and 0.61, respectively. The confusion matrix of RF only identified 11 out of 66 unsuccessful sites and 128 out of 167 successful sites, while the DT confusion matrix was able to identify 20 out of 66 unsuccessful sites and 117 out of 167 successful sites. Although SVC and GNB recorded the highest f1-scores of 0.83 and 0.82 for successful sites, the f1-scores of SVC and GNB for the unsuccessful sites were 0 and 0.29, respectively. Among the ML models, DT had the highest f1-score for unsuccessful sites (0.32). The precision, recall, and f1-scores, as well as the training accuracy and the testing accuracy of all the ML models are presented in Table 4. Additionally, the results for the confusion matrices of tested ML models are presented in Figure 3.

Table 4. The precision, recall, f1-score, training accuracy, and testing accuracy of the ML models to predict the outcomes of NSPT.

ML Models	Treatment Outcome	Precision	Recall	F1-Score	Training Accuracy	Testing Accuracy
Random forest	Unsuccessful	0.28	0.17	0.21	0.81	0.62
	Successful	0.7	0.82	0.75		
Decision tree	Unsuccessful	0.33	0.3	0.32	0.81	0.61
	Successful	0.72	0.75	0.73		
Support vector classifier	Unsuccessful	0	0	0	0.703	0.704
	Successful	0.7	1	0.83		
K-nearest neighbors	Unsuccessful	0.34	0.17	0.22	0.76	0.65
	Successful	0.71	0.87	0.78		
Gaussian naïve Bayes	Unsuccessful	0.57	0.2	0.29	0.72	0.71
	Successful	0.73	0.94	0.82		

On the other hand, examination by RF and DT of the importance of tooth-related factors in contributing to the prediction of the outcome of NSPT showed that PPD and CAL are the most important predictors for both models. Next, tooth surfaces and tooth location were identified as the third most important tooth-related factors by the RF and DT models, respectively. Finally, gingival phenotype was shown to be the fourth most important tooth-related factor by both models (Figure 4).

Tenfold cross-validation was used to examine the performance of all tested ML models. Similarly, GNB (0.713 ± 0.037) and SVC (0.7 ± 0.011) demonstrated the highest accuracy, followed by RF (0.667 ± 0.049) and KNN (0.662 ± 0.032), whilst DT showed the lowest accuracy (0.613 ± 0.057) (Table 5).

Table 5. Tenfold cross-validation for the testing accuracy of all the ML models.

Training Set	RF	DT	SCV	KNN	GNB
1	0.642	0.625	0.705	0.651	0.687
2	0.693	0.630	0.702	0.693	0.684
3	0.657	0.621	0.702	0.594	0.720
4	0.639	0.603	0.666	0.657	0.648

Table 5. *Cont.*

Training Set	RF	DT	SCV	KNN	GNB
5	0.621	0.522	0.702	0.657	0.720
6	0.621	0.576	0.702	0.666	0.720
7	0.621	0.540	0.702	0.639	0.720
8	0.702	0.603	0.702	0.648	0.729
9	0.747	0.675	0.702	0.693	0.792
10	0.747	0.729	0.711	0.720	0.747
Mean ± SD	0.667 ± 0.049	0.613 ± 0.057	0.7 ± 0.011	0.662 ± 0.032	0.713 ± 0.037

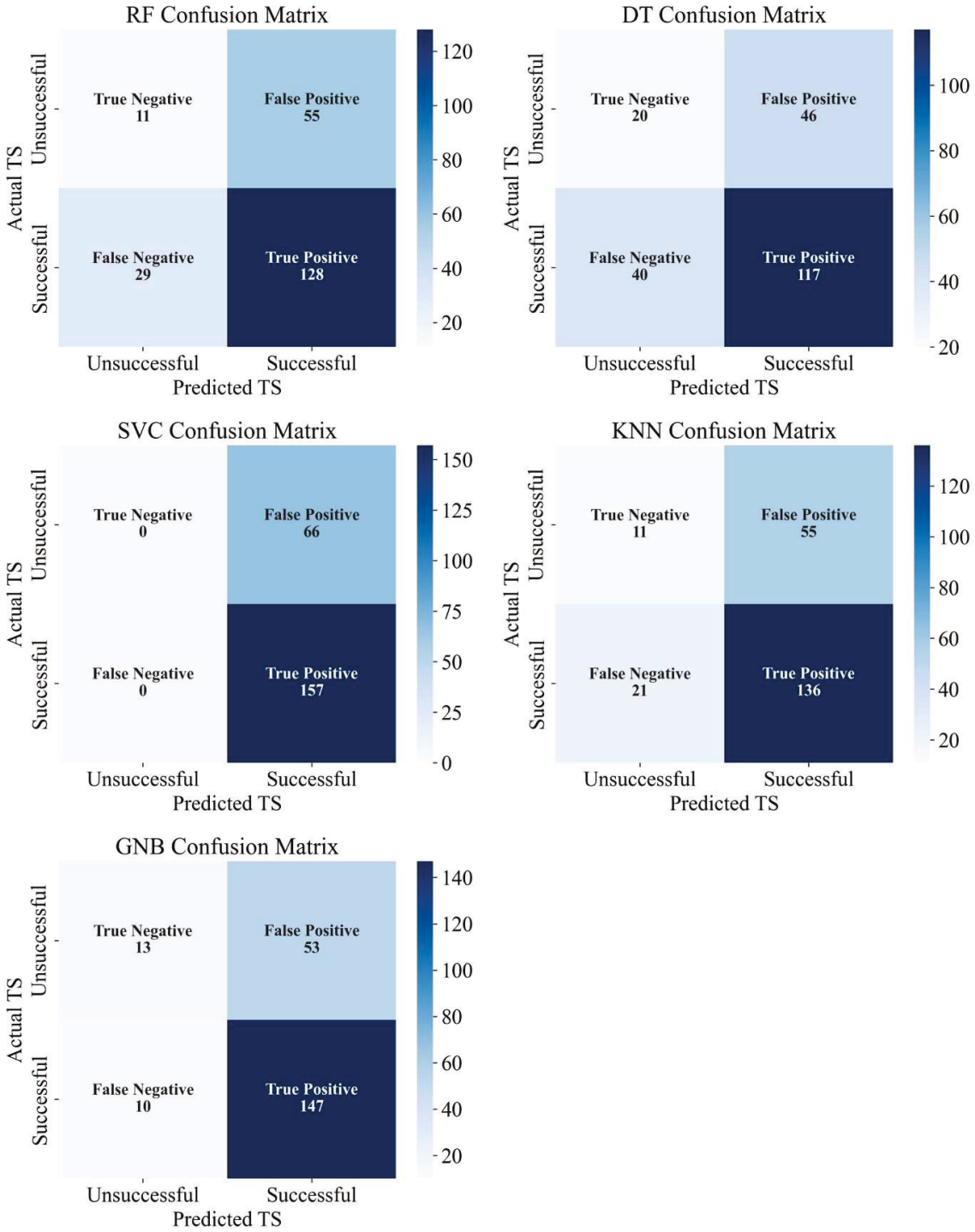


Figure 3. Confusion matrices of examined ML models to predict the outcomes of NSPT.

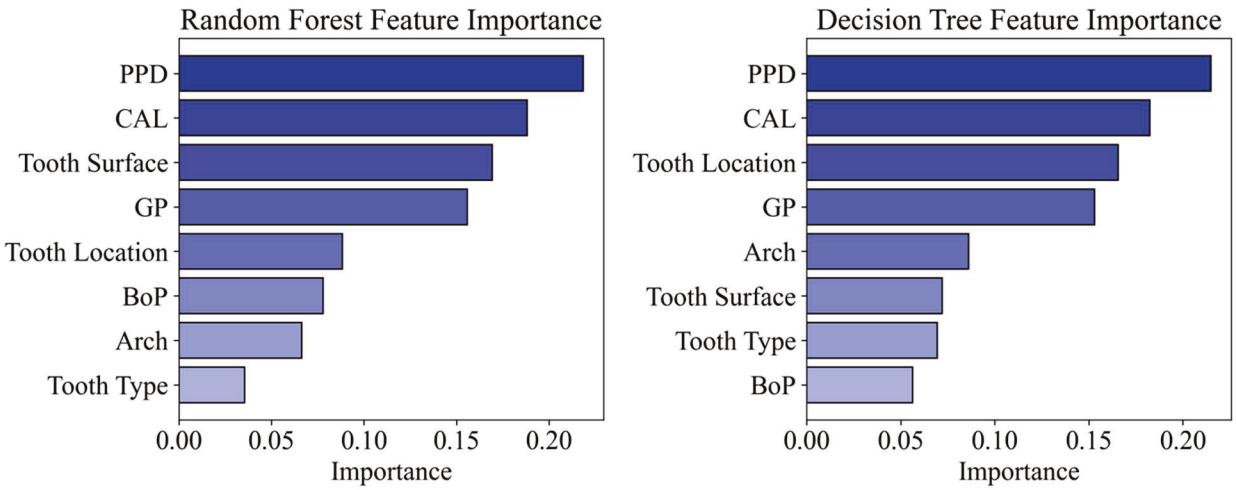


Figure 4. The most important predictors of NSPT as shown by RF and DT models. PPD: probing pocket depth, CAL: clinical attachment loss, GP: gingival phenotype, BoP: bleeding on probing.

3.4. Logistic Regression vs. Machine Learning Models

Comparison between the LR model and ML models can be performed via assessing their overall accuracy and the most important tooth-related factors contributing to their prediction capabilities. It is apparent that the overall accuracy of LR (70.4%, Table 3) was almost equal to that recorded by GNB (71%) and SVC (70.4%), as shown in Table 4 and again following tenfold cross-validation (Table 5). Moreover, in the case of the LR model, among the tooth-related factors, only PPD had a statistically significant association with the outcome of NSPT (Table 3), which is similar to the corresponding results for the RF and DT models (Figure 4).

4. Discussion

Supervised ML models have recently been used for predictive purposes in medicine [32]. The data concerning the use of supervised ML models for prediction purposes in dentistry are inconclusive, as no comparative data were used to evaluate the efficacy of supervised ML models against the traditional risk assessment tools and LR model [24,27]. Consequently, this study was carried out to compare the efficacy of the recent supervised ML models to the most commonly used statistical tests, namely LR with a typical dataset of tooth-related factors. The results of the current study show that the GNB and SVC models have prediction accuracy comparable to that of the LR model regarding outcomes of NSPT. Additionally, similarly to LR, both the RF and DT models identified PPD as the most important factor for prediction of the outcomes of NSPT.

Indeed, a huge number of studies have investigated various biomarkers, such as microbials, host-related markers in saliva, and gingival crevicular fluid for prediction of the outcomes of NSPT [33–35]. Moreover, as treatment outcomes may not only differ between patients but also between different sites of the same tooth within the same patient [36–38], researchers have examined the most convenient clinical and tooth-related factors that might also have an impact on the outcomes of NSPT [6,12,13].

Introducing different clinical, molecular, host-derived, and microbiological biomarkers into dentistry for prediction, diagnosis, and monitoring dental and periodontal/peri-implant diseases has led to increasing interest [39,40]. Among clinical parameters, the severity of calculus accumulation and suppuration were suggested as predictors of tooth survival [41]. Baseline shallow pockets with interproximal location were suggested as predictors of NSPT success [42]. The latter result was consistent with findings of this study indicating shallow baseline PPD has an impact on the outcomes of NSPT. Indeed,

periodontal pockets with shallow or moderate depth (<6 mm) are easier to access and debride than deeper pockets, which are inaccessible to mechanical cleaning and highly infected with pathogenic Gram-negative bacteria [43]. The key factors for success of periodontal therapy are efficient debridement of subgingival dysbiotic microbiota, severity of disease, absence of local/systemic risk factors, compliance to recall visits, and having proper oral hygiene measures [12,44].

The definitions of successful and non-successful NSPT vary according to its application at the patient level and site-specific level. For example, at the patient level, for the baseline PPD of 4–6 mm, a mean PPD reduction of around 1.29 mm was expected [45], whereas, for the site-specific level, as used in the current study, a reduction of PPD to $\leq$4 mm without BoP after 3 months was set as the primary outcome indicating the success of NSPT [31].

This study used the most common tooth-related factors of PPD, CAL, BoP, gingival phenotype, tooth type, arch, tooth location, and tooth surfaces to predict the outcomes of NSPT by the LR model and supervised ML models. Both the conventional LR and ML models demonstrated moderate prediction abilities (around 70%), and with such limited prediction abilities, the clinical implications of the current study might be arguable. However, it is worth mentioning that the purpose of this study was merely to compare the conventional LR statistical test to ML models rather than assessing the actual prediction abilities of the above tooth-related factors regarding NSPT outcomes.

It is important to acknowledge that both the DT and RF models indicated PPD as the most important prediction parameter of outcomes of NSPT, which aligns with the finding for the LR model that only PPD had a statistically significant association with NSPT outcomes. Interestingly, CAL was shown to be the second most important tooth-related factor by both the DT and RF models. However, the LR model did not demonstrate a statistically significant association of CAL with NSPT outcomes. This can be explained by the strong association between PPD and CAL as clinical parameters and, as the prediction role of CAL might be already explained by PPD, CAL would have no significant impact on the overall prediction rate of LR. Similarly, the responses of different gingival phenotypes to biofilm-induced periodontal disease ranged from CAL in the thin, scalloped phenotype to the formation of deep PPD in the thick, flat phenotype [46]. Again, the non-statistically significant association of gingival phenotype to the outcomes of NSPT may already have been explained by PPD.

Indeed, choosing parameters with strong mutual associations hinders the overall prediction ability of LR and, therefore, selecting such parameters for prediction purposes when applying the LR model is highly discouraged. Additionally, it has been reported that the LR model can lead to oversimplification, especially in complex diseases, such as periodontitis, which are characterized by intricate interactions between variables and outcomes [32,47,48]. In contrast, in ML models (RF and DT), such overfitting of the model may not strongly affect the overall prediction result, and consequently, such parameters (CAL and gingival phenotype) are still presented among the most important parameters affecting the prediction accuracy of ML models. On the other hand, the outcomes of NSPT after 3 months have been shown to differ statistically according to the tooth type (single-rooted vs. multirooted teeth). The possible explanation could be that the single-rooted teeth can be cleaned easily due to the absence of possible furcation involvement, which causes difficulties during NSPT [49]. However, the LR model did not show a statistically significant association of tooth type with NSPT outcomes; again, both the DT and RF models ranked tooth type as the least important parameter in predicting the outcomes of NSPT. A possible explanation could be that many algorithms with different mechanisms are used to train DT and RF models rather than a static algorithm (paired *t* test) to compare between successful and unsuccessful sites [50].

These results indicate that ML models have similar or lower capabilities for NSPT prediction 3 months after treatment, based on the above-mentioned tooth-related factors. These results are not commensurate with the results of the current systematic reviews showing that supervised ML methods can demonstrate better performance than traditional statistical tests for periodontal screening, diagnosis, disease progression, and prognosis [23,24]. This can be explained by the fact that the primary outcome of the current study is that NSPT is fundamentally different from periodontal screening and diagnosis. Thus, caution should be used when comparing the same ML models in applications of different prediction models, as well as when different independent parameters are used.

Although internal validation of the models was performed, and GNB and SVC produced comparable results to the LR, the lack of external validation and using data of one population are considered as the main limitations of the current study and hence, generalization of the results should be avoided. In fact, this highlights the need for external validation of the ML models in other heterogeneous populations, using larger datasets to confirm their efficacy in a real-world scenario. Nonetheless, this is the first study to compare the efficacy of ML models to conventional LR in predicting the outcomes of NSPT using tooth-related factors, and it could pave the way for future studies to apply ML models with various predictors, such as patient-related, microbial, and host-related factors, to enhance the overall predictive accuracy of ML models.

5. Conclusions

Available evidence indicates that, in predicting outcomes of NSPT, supervised ML models do not offer substantial advantages over the conventional LR model. Therefore, considering its transparency and interpretability, LR remains the suitable and often preferable modeling approach in clinical settings, particularly in the absence of large, high-dimensional datasets. Certainly, to effectively leverage the capabilities of ML models in predicting the outcomes of NSPT, it is imperative to employ larger sample sizes alongside more complex predictors. Moreover, any predictive models developed should be subjected to rigorous external validation prior to making assertions regarding their generalizability or potential impact on broader populations.

Author Contributions: A.J.B.A.-S. contributed to the conception and design, data acquisition, analysis, interpretation, and writing of the original draft; M.T.A.B. contributed to the analysis, interpretation, and writing—review and editing; N.K.I. contributed to the conception and design, data acquisition, analysis, interpretation, and writing of the original draft; S.S.G. contributed to the conception, supervision, and writing—review and editing; A.A.A., contributed to the conception and design, analysis, and interpretation, and drafted and critically revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee, College of Dentistry, University of Baghdad (Reference No. 663, Date 13 September 2022).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request due to ethical reasons.

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

References

1. Cobb, C.M. Non-surgical pocket therapy: Mechanical. *Ann. Periodontol.* **1996**, *1*, 443–490. [CrossRef] [PubMed]
2. Sanz, M.; Herrera, D.; Kebschull, M.; Chapple, I.; Jepsen, S.; Beglundh, T.; Sculean, A.; Tonetti, M.S. Treatment of stage I-III periodontitis-The EFP S3 level clinical practice guideline. *J. Clin. Periodontol.* **2020**, *47* (Suppl. S22), 4–60. [CrossRef]
3. Belstrøm, D.; Grande, M.A.; Sembler-Møller, M.L.; Kirkby, N.; Cotton, S.L.; Paster, B.J.; Holmstrup, P. Influence of periodontal treatment on subgingival and salivary microbiotas. *J. Periodontol.* **2018**, *89*, 531–539. [CrossRef]
4. Byrne, S.J.; Chang, D.; Adams, G.G.; Butler, C.A.; Reynolds, E.C.; Darby, I.B.; Dashper, S.G. Microbiome profiles of non-responding and responding paired periodontitis sites within the same participants following non-surgical treatment. *J. Oral Microbiol.* **2022**, *14*, 2043595. [CrossRef]
5. Sanz-Sánchez, I.; Montero, E.; Citterio, F.; Romano, F.; Molina, A.; Aimetti, M. Efficacy of access flap procedures compared to subgingival debridement in the treatment of periodontitis. A systematic review and meta-analysis. *J. Clin. Periodontol.* **2020**, *47*, 282–302. [CrossRef]
6. Werner, N.; Heck, K.; Walter, E.; Ern, C.; Bumm, C.V.; Folwaczny, M. Probing pocket depth reduction after non-surgical periodontal therapy: Tooth-related factors. *J. Periodontol.* **2024**, *95*, 29–39. [CrossRef]
7. Axtelius, B.; Söderfeldt, B.; Attström, R. A multilevel analysis of factors affecting pocket probing depth in patients responding differently to periodontal treatment. *J. Clin. Periodontol.* **1999**, *26*, 67–76. [CrossRef]
8. D'Aiuto, F.; Ready, D.; Parkar, M.; Tonetti, M.S. Relative Contribution of Patient-, Tooth-, and Site-Associated Variability on the Clinical Outcomes of Subgingival Debridement. I. Probing Depths. *J. Periodontol.* **2005**, *76*, 398–405. [CrossRef] [PubMed]
9. Hughes, F.J.; Syed, M.; Koshy, B.; Marinho, V.; Bostanci, N.; McKay, I.J.; Curtis, M.A.; Croucher, R.E.; Marcenes, W. Prognostic factors in the treatment of generalized aggressive periodontitis: I. Clinical features and initial outcome. *J. Clin. Periodontol.* **2006**, *33*, 663–670. [CrossRef]
10. Pretzl, B.; Kaltschmitt, J.; Kim, T.-S.; Reitmeir, P.; Eickholz, P. Tooth loss after active periodontal therapy. 2: Tooth-related factors. *J. Clin. Periodontol.* **2008**, *35*, 175–182. [CrossRef] [PubMed]
11. Bunæs, D.F.; Lie, S.A.; Åstrøm, A.N.; Mustafa, K.; Leknes, K.N. Site-specific treatment outcome in smokers following 12 months of supportive periodontal therapy. *J. Clin. Periodontol.* **2016**, *43*, 1086–1093. [CrossRef]
12. Van der Weijden, G.A.F.; Dekkers, G.J.; Slot, D.E. Success of non-surgical periodontal therapy in adult periodontitis patients: A retrospective analysis. *Int. J. Dent. Hyg.* **2019**, *17*, 309–317. [CrossRef]
13. Bumm, C.V.; Schwendicke, F.; Pitchika, V.; Heck, K.; Walter, E.; Ern, C.; Heym, R.; Werner, N.; Folwaczny, M. Effectiveness of nonsurgical re-instrumentation: Tooth-related factors. *J. Periodontol.* **2025**, *96*, 748–759. [CrossRef]
14. Herz, M.M.; Hoffmann, N.; Braun, S.; Lachmann, S.; Bartha, V.; Petsos, H. Periodontal pockets: Predictors for site-related worsening after non-surgical therapy—A long-term retrospective cohort study. *J. Clin. Periodontol.* **2024**, *51*, 680–690. [CrossRef]
15. Liu, H.-J.; Wang, B.; Wang, A.-C.; Zhang, D.-H.; Mao, C.; Li, Q.-H. Prognostic factors affecting the short-term efficacy of non-surgical treatment of chronic periodontitis: A multilevel modeling analysis. *Eur. J. Med. Res.* **2021**, *26*, 50. [CrossRef]
16. Ferrarotti, F.; Baima, G.; Rendinelli, M.; Citterio, F.; Mariani, G.M.; Mussano, F.; Romano, F.; Romandini, M.; Aimetti, M. Pocket closure after repeated subgingival instrumentation: A stress test to the EFP guideline for stage III-IV periodontitis. *Clin. Oral Investig.* **2023**, *27*, 6701–6708. [CrossRef]
17. Goldstein, H.; Browne, W.; Rasbash, J. Multilevel modelling of medical data. *Stat. Med.* **2002**, *21*, 3291–3315. [CrossRef]
18. Leyland, A.H.; Groenewegen, P.P. Multilevel modelling and public health policy. *Scand. J. Public Health* **2003**, *31*, 267–274. [CrossRef]
19. Tomasi, C.; Leyland, A.H.; Wennström, J.L. Factors influencing the outcome of non-surgical periodontal treatment: A multilevel approach. *J. Clin. Periodontol.* **2007**, *34*, 682–690. [CrossRef]
20. Gul, S.S.; Griffiths, G.S.; Stafford, G.P.; Al-Zubidi, M.I.; Rawlinson, A.; Douglas, C.W. Investigation of a novel predictive biomarker profile for the outcome of periodontal treatment. *J. Periodontol.* **2017**, *88*, 1135–1144. [CrossRef]
21. Pitchika, V.; Büttner, M.; Schwendicke, F. Artificial intelligence and personalized diagnostics in periodontology: A narrative review. *Periodontol. 2000* **2024**, *95*, 220–231. [CrossRef] [PubMed]
22. Bashir, N.Z.; Rahman, Z.; Chen, S.L.-S. Systematic comparison of machine learning algorithms to develop and validate predictive models for periodontitis. *J. Clin. Periodontol.* **2022**, *49*, 958–969. [CrossRef]
23. Montero, E.; Sánchez, N.; Sanz-Sánchez, I.; López-Durán, M.; de Albornoz, A.C.; Dietrich, T. Emerging Technologies and Algorithms for Periodontal Screening and Risk of Disease Progression in Non-Dental Settings: A Scoping Review. *J. Clin. Periodontol.* **2025**, *52*, 246–291. [CrossRef]
24. Farina, R.; Simonelli, A.; Trombelli, L.; Ettmayer, J.B.; Schmid, J.L.; Ramseier, C.A. Emerging Applications of Digital Technologies for Periodontal Screening, Diagnosis and Prognosis in the Dental Setting. *J. Clin. Periodontol.* **2025**, *52*, 211–245. [CrossRef]
25. Baban, M.T.A.; Mohammad, D.N. A new approach for sex prediction by evaluating mandibular arch and canine dimensions with machine-learning classifiers and intraoral scanners (a retrospective study). *Sci. Rep.* **2024**, *14*, 27974. [CrossRef]

26. Yan, Y.; Sharma, P.; Suvan, J.; D'Aiuto, F. The Association of Periodontal Inflammation and Systemic Health Indicators: A Machine Learning Approach. *J. Clin. Periodontol.* **2025**, *52*, 1466–1477. [CrossRef]
27. Chow, D.Y.; Tay, J.R.H.; Nascimento, G.G. Systematic Review of Prognosis Models in Predicting Tooth Loss in Periodontitis. *J. Dent. Res.* **2024**, *103*, 596–604. [CrossRef] [PubMed]
28. Tonetti, M.S.; Greenwell, H.; Kornman, K.S. Staging and grading of periodontitis: Framework and proposal of a new classification and case definition. *J. Periodontol.* **2018**, *89*, S159–S172. [CrossRef]
29. Ainamo, J.; Bay, I. Problems and proposals for recording gingivitis and plaque. *Int. Dent. J.* **1975**, *25*, 229–235.
30. Uysal, B.F.; Köse, T.; Gürkan, A. Gingival phenotype classification by visual and probe visibility assessments: Relationship with thickness and probe design. *J. Periodontol.* **2025**, *96*, 309–320. [CrossRef] [PubMed]
31. Loos, B.G.; Needleman, I. Endpoints of active periodontal therapy. *J. Clin. Periodontol.* **2020**, *47*, 61–71. [CrossRef]
32. Chiarito, M.; Luceri, L.; Oliva, A.; Stefanini, G.; Condorelli, G. Artificial Intelligence and Cardiovascular Risk Prediction: All That Glitters is not Gold. *Eur. Cardiol.* **2022**, *17*, e29. [CrossRef]
33. Bostanci, N.; Manoil, D.; Van Holm, W.; Belibasakis, G.N.; Teughels, W. Microbial Markers for Diagnosis and Risk Assessment for Periodontal Diseases: A Systematic Literature Search and Narrative Synthesis. *J. Clin. Periodontol.* **2025**, *52*, 125–154. [CrossRef]
34. Rakic, M.; Calciolari, E.; Grant, M.M.; Radovanovic, S.; Bostanci, N.; Preshaw, P.M. Host Markers of Periodontal Diseases: Meta-Analysis of Diagnostic Accuracy Studies. *J. Clin. Periodontol.* **2025**, *52*, 155–181. [CrossRef]
35. Sha, A.M.; Abdulbaqi, H.R.; Qasim, S.S.B. Microbial and Inflammatory Salivary Biomarkers of Periodontal Diseases. *Kurd. J. Appl. Res.* **2024**, *9*, 113–125. [CrossRef]
36. Badersten, A.; Nilveus, R.; Egelberg, J. Effect of nonsurgical periodontal therapy. *J. Clin. Periodontol.* **1984**, *11*, 63–76. [CrossRef] [PubMed]
37. Serino, G.; Rosling, B.; Ramberg, P.; Socransky, S.S.; Lindhe, J. Initial outcome and long-term effect of surgical and non-surgical treatment of advanced periodontal disease. *J. Clin. Periodontol.* **2001**, *28*, 910–916. [CrossRef] [PubMed]
38. Van der Weijden, G.A.; Timmerman, M.F. A systematic review on the clinical efficacy of subgingival debridement in the treatment of chronic periodontitis. *J. Clin. Periodontol.* **2002**, *29*, 55–71. [CrossRef] [PubMed]
39. Mohammed, H.A.; Abdulkareem, A.A.; Zardawi, F.M.; Gul, S.S. Determination of the Accuracy of Salivary Biomarkers for Periodontal Diagnosis. *Diagnostics* **2022**, *12*, 2485. [CrossRef]
40. Gul, S.S.; Abdulkareem, A.A.; Sha, A.M.; Rawlinson, A. Diagnostic Accuracy of Oral Fluids Biomarker Profile to Determine the Current and Future Status of Periodontal and Peri-Implant Diseases. *Diagnostics* **2020**, *10*, 838. [CrossRef]
41. Baelum, V.; López, R. Defining and predicting outcomes of non-surgical periodontal treatment: A 1-yr follow-up study. *Eur. J. Oral Sci.* **2016**, *124*, 33–44. [CrossRef]
42. Chen, M.H.; Yin, H.J.; Chang, H.H.; Kao, C.T.; Tu, C.C.; Chen, Y.W. Baseline probing depth and interproximal sites predict treatment outcomes of non-surgical periodontal therapy. *J. Dent. Sci.* **2020**, *15*, 50–58. [CrossRef] [PubMed]
43. Mombelli, A. Microbial colonization of the periodontal pocket and its significance for periodontal therapy. *Periodontol. 2000* **2018**, *76*, 85–96. [CrossRef]
44. Abdalla, M.; Peck, M.; Rayner, C.; Kimmie-Dhansay, F.; Jeftha, A. Factors that affected the efficacy of non-surgical periodontal treatment carried out by postgraduate periodontology students. *S. Afr. Dent. J.* **2023**, *78*, 126–129. [CrossRef]
45. Killoy, W.J. The clinical significance of local chemotherapiesAbstract. *J. Clin. Periodontol.* **2002**, *29*, 6–13. [CrossRef]
46. Olsson, M.; Lindhe, J. Periodontal characteristics in individuals with varying form of the upper central incisors. *J. Clin. Periodontol.* **1991**, *18*, 78–82. [CrossRef]
47. Harrell, F.E., Jr.; Lee, K.L.; Mark, D.B. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **1996**, *15*, 361–387. [CrossRef]
48. Shameer, K.; Johnson, K.W.; Glicksberg, B.S.; Dudley, J.T.; Sengupta, P.P. Machine learning in cardiovascular medicine: Are we there yet? *Heart* **2018**, *104*, 1156–1164. [CrossRef]
49. Ehnevid, H.; Jansson, L.E. Effects of Furcation Involvements on Periodontal Status and Healing in Adjacent Proximal Sites. *J. Periodontol.* **2001**, *72*, 871–876. [CrossRef]
50. Faouzi, J.; Colliot, O. Classic Machine Learning Methods. In *Machine Learning for Brain Disorders*; Colliot, O., Ed.; Springer: New York, NY, USA, 2023; pp. 25–75.

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.

diagnostics

MDPI

Review

The Role of ChatGPT in Dermatology Diagnostics

Ziad Khamaysi [1,2,*], Mahdi Awwad [3], Badea Jiryis [1,2], Naji Bathish [4] and Jonathan Shapiro [5]

[1] Department of Dermatology, Rambam Health Care Campus, Haifa 3109601, Israel
[2] Bruce Rappaport Faculty of Medicine, Technion Institute of Technology, Haifa 3525433, Israel
[3] Ophthalmology Unit, Tzafon Medical Center, Tiberias 1528001, Israel
[4] Dermatology Unit, Ziv Medical Center, Safed 13100, Israel; najibathish@gmail.com
[5] Maccabi Healthcare Services, Tel Aviv 6817110, Israel
* Correspondence: z_khamaysi@rambam.health.gov.il; Tel.: +972-53-230-1230

Abstract: Artificial intelligence (AI), especially large language models (LLMs) like ChatGPT, has disrupted different medical disciplines, including dermatology. This review explores the application of ChatGPT in dermatological diagnosis, emphasizing its role in natural language processing (NLP) for clinical data interpretation, differential diagnosis assistance, and patient communication enhancement. ChatGPT can enhance a diagnostic workflow when paired with image analysis tools, such as convolutional neural networks (CNNs), by merging text and image data. While it boasts great capabilities, it still faces some issues, such as its inability to perform any direct image analyses and the risk of inaccurate suggestions. Ethical considerations, including patient data privacy and the responsibilities of the clinician, are discussed. Future perspectives include an integrated multimodal model and AI-assisted framework for diagnosis, which shall improve dermatology practice.

Keywords: ChatGPT; diagnosing; dermatology

1. Introduction

Artificial intelligence (AI) has become an increasingly pivotal force in medicine, offering transformative potential across diagnostics, treatment planning, and patient care [1]. In recent years, AI applications have significantly expanded, encompassing tools like natural language processing (NLP), image analysis, and machine learning, which enable precise analysis and prediction in diverse medical fields [2]. Large language models (LLMs), like ChatGPT, now propel an important frontier of healthcare innovation. While originally designed for use as conversational agents, these are now adapted for other purposes, such as patient communication, clinical decision support, and research assistance [3].

One of the very few emerging areas where substantial promise exists for ChatGPT and other similar LLMs is dermatology diagnostics. As dermatology relies predominantly on seeing and writing descriptively accurate texts, it emerges as a distinctive context within which an AI-driven tool such as ChatGPT can glean a major benefit. This includes facilitating the interpretation and generation of medical texts and diagnostic image descriptions that could assist healthcare professionals by generating descriptions or pinpointing characteristic patterns of various skin conditions [4,5]. In addition, NLP capabilities may allow ChatGPT to automatically mine the open access medical literature to help clinicians stay up to date with complex dermatological cases and relevant guidelines [4].

Electronic medical records (EMRs) are essential for dermatology, yet much of the data remains unstructured, making analysis challenging. Natural language processing (NLP) helps streamline EMR data, enabling automated documentation, improved patient history gathering, differential diagnosis suggestions, and integration with AI image analysis. These

capabilities enhance patient care, especially in teledermatology, where NLP assists in organizing patient data for remote consultations. With models like ChatGPT-4 gaining traction, understanding their role in dermatology can maximize their potential in diagnostics and clinical support [6–9].

The objectives of this review are to examine the diversified applications of ChatGPT in dermatology diagnostics, such as interpreting images, natural language processing for synthesizing and summarizing clinical knowledge, integration with machine learning, and programmatic support for dermatological research. By undertaking these dimensions, we aim to dissect the potential, limitations, and ethical perspectives that must be taken into account for responsible implementation within dermatological practice.

We will be considering ChatGPT since it represents a significant step forward in using AI in the clinical setting, particularly in dermatology, having been designed by OpenAI. It is one of the most exciting commercial products that intersects with the clinical world, especially where physicians do not have the luxury of specializing in AIs. Its NLP abilities allow it to interact with clinical data for undertaking differentiation of diagnoses, interpretations of patient symptoms, and clinical documentation. With the increasing adoption rate among dermatologists, it is imperative to carry out an exhaustive review of the social uses, benefits, and drawbacks of ChatGPT. The review will help to illustrate how ChatGPT can further improve dermatologic care while also highlighting areas where improvement is needed so that clinical applications can be used more effectively and responsibly.

Background and Current State of AI in Dermatology

Over the last few decades, artificial intelligence (AI) has become an indispensable element in the field of dermatology diagnostics. Going from pattern recognition to deep learning systems, the evolution of AI in dermatological studies, both technologically and academically, has become more intricate [4]. Many of these early programs were designed to help clinicians recognize simple patterns in skin lesions through imaging techniques and computerized algorithms primarily to diagnose common skin conditions, such as melanoma and psoriasis [10].

Due to progress in machine learning (ML) and deep learning (DL), AI systems have become far more capable and adaptable [11]. DL models have shown to be especially useful in the analysis of skin lesions, where they learn from large sets of carefully annotated images and find patterns that are difficult for the human eye to detect [12]. For example, convolutional neural networks (CNNs) have attained an accuracy high enough to distinguish between benign and malignant lesions, sometimes performing at a level comparable to that of dermatologists [13,14]. However, traditional AI models have limitations. These include problems of generalizability across populations and lighting conditions and reliance on large, diverse datasets that are often difficult to collect and annotate within dermatology [15,16].

The development of AI has ushered in the era of large language models (LLMs), including OpenAI's ChatGPT, which do indeed mark a change in utility for AI in fields of medicine like dermatology. While LLMs began with the purposes of natural language processing, they have become advanced enough to process and generate text in a manner similar to that of humans [17]. Unlike conventional ML or DL models, which mainly focus on image analysis, LLMs like ChatGPT can analyze text inputs for clinical decision support and facilitate patient–physician communication by explaining diagnostic reasoning or treatment options in lay language [18]. Such versatility makes LLMs an exciting addition to dermatology, especially in areas where image-based AI tools fall short, like patient-specific concerns or contextualizing image findings with patient histories [19].

As an AI algorithm, ChatGPT can contribute to diagnostic workflows beyond the limits of image-only models through analysis of substantial amounts of general medical knowledge such as addressing patient concerns, providing proofs detailing the diagnosis through evidence-based insights [20], and even an interpretation of complicated skin conditions through detailed descriptions given by the patient. For instance, ChatGPT can offer believable differential diagnoses or suggest follow-ups, especially advantageous in a primary-care setting [21], where dermatologists are not readily accessible. However, while this area may be promising, it is imperative to approach the integration of LLMs such as ChatGPT trained on dermatology datasets cautiously, as they are not free from certain limitations, for example, precision-dependent cases or situations requiring critical diagnostic accuracy [22].

2. Image Analysis Using ChatGPT and Its Integration with Diagnostic Tools

In dermatology, image analysis plays a pivotal role in diagnostics, for clinical images, dermoscopic images, histology images, and confocal microscopy images, with the use of instruments like dermoscopy and clinical imaging giving us high-resolution views that not only help with early detection but also facilitate accurate diagnosis of conditions like melanoma and other skin problems [23]. Dermatoscopic images play, in particular, a key role in enabling dermatologists to detect fine changes occurring within skin lesions since such structures are not readily visible to the naked eye. All of that, in turn, practically demonstrates drastic increases in diagnostic accuracy [24]. ChatGPT, being a large language model, is notable for interpreting prompts and formulating self-consistent text descriptions. Thus, processing detailed image-related descriptions for clinicians can assist ChatGPT in providing relevant information for potential diagnostic implications or possibly suggest areas for further examination [25]. Such capability enables dermatologists to use ChatGPT as an auxiliary tool for the generation or refinement of descriptions, notably in cases where immediate computer vision assessment is unavailable.

With descriptive assistance, ChatGPT may help dermatologists generate detailed and relevant descriptions emphasizing clinical signs and, potentially, minimize human errors while improving diagnostic accuracy [26]. ChatGPT can, therefore, serve as an assistance for the clinician in potentially rephrasing observations or suggesting an alternative diagnostic wording to elaborate an uncertain lesion description in a complex case; by adding the input of an expert, increased reliability and diagnosis were achievable [27].

Incorporating computer vision (CV) neural networks and deep machine learning frameworks into ChatGPT might enhance the existing setup of diagnostic workflows in dermatology [28]. The output of CNNs, trained on extensive datasets and influential in the constructive identification of features of skin lesions, could do much better with the input of ChatGPT for its natural processing of language in preparation for interpretive summaries or comments that could provide human interpreters with a little help [28,29]. Thereby, they adopt a symbiotic approach that involves portable knowledge pairing that combines the strengths of both systems, whereby computer vision interrogates the feature identifications and associated descriptions to relate this information to a clinically relevant narrative.

Previous GPT models more exclusively input text data and interpreted visual data with detailed descriptions. Long outputs were often not described properly or lost something vital, especially in areas like dermatology, where visual subtleties have always been in question [30]. With GPT-4's multimodal capabilities, the limitation has been somewhat circumvented as the model can directly analyze an image together with text. But GPT-4 is still behind the achievements of specialized computer vision models for medical imaging and lacks the ability to cross-check its interpretations. This means that it could be a very handy adjunct but is not useful as an independent diagnostic tool.

The concomitant improvement of new developments in dermatologic diagnostic technologies may lie with the combination of off-the-shelf large language models, like ChatGPT, and next-generation computer vision technologies, thus developing a joint exterior approach, wherein LLMs process visual data properly with clinic parameters of such nature [22]. A conceptual overview of such a multimodal system is illustrated in Figure 1.

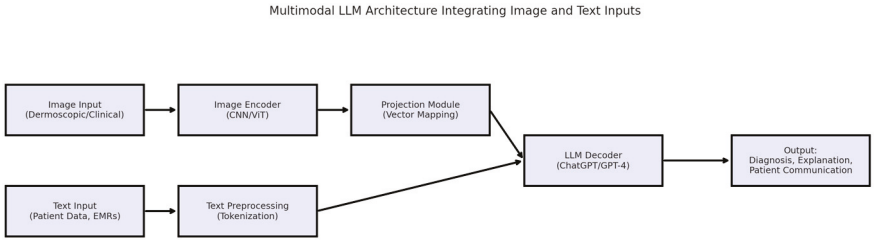


Figure 1. Illustration of a multimodal large language model (LLM) system integrating both image and text data. The image pathway is encoded via convolutional neural networks (CNNs) or vision transformers (ViTs), while text inputs (e.g., patient history) are tokenized and processed by a large language model (e.g., GPT-4). A projection module aligns visual embeddings with the text domain, enabling the model to generate comprehensive outputs such as diagnostic suggestions, patient education, and clinical explanations.

The evolving LLMs, in the process of interacting more directly with visual input, could provide predictive insights along with image analysis, streamlining diagnostics and maximizing early detection of dermatologic conditions. Such advances could lead to systems where LLMs like ChatGPT continuously refine image interpretations made by CNNs, culminating in a hybrid diagnostic model incorporating textual and visual analytics in an integrated approach serving clinical support [10,12].

3. Natural Language Processing (NLP) Capabilities of ChatGPT in Dermatology

NLP has already become a disruptive technology in the healthcare domain, mainly in diagnostics and patient management. NLP can analyze unstructured data in EHRs—patient notes, clinical histories, and diagnostics reports—which helps healthcare providers gain impressions that facilitate better and more accurate decisions in diagnosing and planning. In pathology, NLP can help to extract relevant data from the patient history and allow quick analysis of text-based information, enabling diagnostic workflows while easing the workload of a clinician [6–9].

Another shining example of efficient natural language processing developed by OpenAI in the dermatological field is ChatGPT. The pertinent information that really could guide a clinician's diagnostic decision has now been extracted, analyzed, and presented on a platter by ChatGPT, using the patient anamnesis and clinical notes. To put it plainly, while scrutinizing patient medical histories, ChatGPT can present dermatologic symptoms along the lines of pruritus, erythema, and lesions, in this way allowing better initial assessments. In addition, ChatGPT is able to parse out relevant dermatologic symptoms among free-text inputs, allowing clinicians a more efficient means of garnering information needed in diagnostic metrics, which leads to smoothening the initial triage process in a wide range of clinical and telemedicine settings [6,9,30,31].

Teledermatology, telemedicine for the diagnosis of skin conditions, has gained ground in recent years, ever since newly found demands for remote healthcare solutions arose. ChatGPT's new-wave NLP capabilities constitute a boost to documentation tools, wherein automated summaries of patient interactions and symptom reporting are significant aiding

features. These features permit dermatologists to concentrate on the case, ensuring that clinically necessary medical records are accurate and comprehensive. Further, it helps communication by making sure the responses generated sufficiently cover the questions, clarifications of instructions, and laying down the treatment plan in a manner a patient can understand, especially when referring to a patient seeking dermatoscopy through virtual modalities [8,9,30,32–34].

Recent studies have begun with a glance at the applicability of ChatGPT and its equivalent natural language processing models in dermatology. In one of the studies analyzed, ChatGPT was competent enough to extract disease-specific terminology from dermatological case notes in support of improving the diagnostic work of clinicians [6,7,26]. Another study suggested that ChatGPT and various NLP approaches may further compress and summarize patient records, as well as create documentation errors in other medical methods [26,35,36]. Based on comparisons with traditional dermatological diagnostic means, some studies used OpenAI's ChatGPT and other models of NLP to help non-specialist providers identify common skin diseases, which would allow these patients more access to dermatological services in under-served areas [9,37].

Like any other natural language processing technologies used in the healthcare system, ChatGPT in dermatology presents significant ethical concerns with regard to patient data, confidentiality, and privacy. The model must conform to the legality of patient confidentiality, meaning all data protection standards must be observed to guarantee patient secrecy, such as compliance with HIPAA in the United States. Further, there lies an ethical dilemma pertaining to the accuracy of NLP: in precise words and interpretation, this may lead to improper diagnoses. In a way, therefore, it presents risk in the automated processing of clinical data in dermatology, and hence appropriate data governance mechanisms and oversight from clinicians are necessary to mitigate risks associated with automated data processing [38–40].

4. Machine Learning and Programming Applications in Dermatology Diagnostics

ChatGPT is a powerful tool that researchers use to create and refine ML algorithms [41] oriented toward dermatology diagnostics. These revolve around feature selection, model optimization, and data preprocessing strategies tailored to meet clinical requirements and fit within the data structure [42]. For example, it may stipulate ways to improve performance in convolutional neural networks (CNNs) while analyzing dermatological images, possibly by data augmentation, thereby reducing overfitting [28,43,44]. ChatGPT greatly aids dermatology researchers in developing, optimizing, and debugging their programming skills, so they can perform programming activities more quickly [9,45]. For example, it can help develop Python scripts in training ML models on dermatological image datasets, creating pipelines for data preprocessing on skin lesion image datasets, or solving case problems of integrating libraries like TensorFlow (version 2.19.0) or PyTorch (v2.7.0) into existing workflows. ChatGPT can also provide explanations for coding errors or suggest alternative methods to improve computational efficiency [46,47].

5. Summary of the Literature Findings on Machine Learning and Programming Applications in Dermatology Diagnostics

Recent advances in artificial intelligence widen its applications in dermatology, and ChatGPT emerges as a new tool to aid in diagnosis. This review presents findings from the studies about ChatGPT's utility in dermatological diagnosis, being compared with traditional AI models, and outlines its reception by dermatologists and patients.

ChatGPT has been evaluated for its potential in dermatology diagnostics in a few studies. For example, [Ferreira et al., 2023] found that ChatGPT offers competent evidence of recognizing some symptoms of conditions like eczema and acne, showing an 88% agreement with expert dermatologists [37]. In the same way, [Lam Hoai & Simonart, 2023; Shapiro et al., 2024] evaluated its attempts at giving treatment advice, reporting a moderate to high degree of accuracy in conformity with current clinical guidelines [6,7].

ChatGPT has recently been compared with traditional AI models in dermatology, bringing various strengths and performance to light. One study comparing ChatGPT and Claude 3 Opus examined their respective capabilities with both malignancies and benign birthmarks of 100 dermoscopic images. ChatGPT was more correct in classifying the top three diagnoses than an alternative technology, being slightly superior at 78% to 76%, while Claude 3 Opus, having higher sensitivity and specificity, was more competent than ChatGPT in malignancy discrimination. ChatGPT, with a better breadth of knowledge and contextual considerations, sometimes failed to classify benign conditions as being malignant, a concern that underlines the necessity for better fine-tuning before clinical use [48].

One of the major areas of study that is focused chiefly on is the user experience of ChatGPT. Based on the findings of Goktas and Grzybowski, ChatGPT was found to be conversationally useful in supporting patient education and decision-making [45,49]. In the meantime, [Alanezi, 2024] validated that the patients had high satisfaction with the use of ChatGPT for preliminary consultations, adding that accessibility and a non-threatening interface were the main strengths [50].

In a recent review to assess the performance of ChatGPT for a multitude of dermatological conditions, it was shown to have varying degrees of good and poor accuracy. According to studies, ChatGPT has been cited with high accuracy in issues such as psoriasis and eczema, yet users complained of poor performance with more complex instances of, say, cutaneous neoplasms, otherwise complicated by the fact that some of these neoplasms require nuance for proper identification. For example, ChatGPT fared better in identifying benign lesions by description but had serious trouble correctly diagnosing malignant lesions compared to image-based AI models.

There was a large variance in the performance, necessitating improvement in the size of the dataset or adjustment to the models of improvement so that the complexity and variability of dermatological manifestations may be dealt with [45].

The reliability of ChatGPT in generating accurate differential diagnoses remains a topic of ongoing debate. Some studies highlight its ability to provide consistent and reasonable diagnostic suggestions for common dermatological conditions, demonstrating its potential as a supportive tool in clinical settings [7,9,45]. However, other research highlights occasional inaccuracies, particularly in cases involving rare or atypical presentations [48,49,51]. These discrepancies underline that while ChatGPT can complement clinical efforts by offering differential diagnoses, it should not be relied upon as a standalone diagnostic tool. Experts consistently emphasize its role as an adjunct to enhance, rather than replace, professional judgment.

To date, several studies on the uses of Generative Pre-trained Transformers (GPTs) in dermatology have explored various domains in diagnostics, therapy recommendations, and patient education. More precisely, Table 1 presents a summary of 18 key studies that picture the variety of GPT applications.

The mainstay across several studies (10) illustrates the capacity of GPTs in dermatologic diagnosis, allowing treatment through text, multimodal, or image-based systems. For instance, GPT-4 impressed with extraordinary diagnostic accuracy across several dermatologic conditions in both text and multimodal examples, but does not yet show a

satisfactory level of accuracy in image-based diagnosis, as discussed by Pillai et al. [52]. In comparative studies, such as Liu et al., the strength of GPTs was examined against that of other AI models, showing exceptional areas as well as those needing refinement in melanoma diagnosis [18].

Less research has focused on management and treatment recommendations, but a study by Iqbal et al. indicated that GPT-4 may offer parallel second opinions for dermatology treatments that earned wide approval among dermatologists [53]. Yet it has trouble with more difficult cases, which opens up room for enhancement.

As for educational and examination performance, GPT-4 is notedly more effective than its predecessors, attaining accuracies from 75 to 93% in various dermatological examinations (Elias et al.; Passby et al.) [54,55]. These works show promise in being able to apply GPT-4 to vetting an adjunct to medical education; however, the shortcomings involving high-difficulty questions indicate a need for clinical assessment on its behalf.

Finally, studies on NLP applications have indicated GPT's participation in conducting an analysis of unstructured patient data and building materials of patient education, to cite examples like Shapiro et al., where GPT-4 provided high accuracy amid psoriasis patient records [6], and Lambert et al., where it generated accessible patient education materials tailored to different reading levels [56].

In general, while most studies center on its diagnostic applications, the growing exploration of GPT in treatment and education reflects the growing interest in its multifaceted potential. However, the consideration of further research into the integration of GPT into dermatology is needed due to various challenges such as ethical concerns, limitations in dealing with complex cases, and variable performance. An overview of the published studies on ChatGPT and other LLMs in dermatology is summarized in Table 1.

Table 1. Categorization of studies on GPT in dermatology.

Type	Focus Area	Paper Title/Details	Authors	Year	Summary of Findings	Version of GPT	Image Dataset Source	Diagnosis Compared to	Study Type	Diagnosis Made Based on Images Alone or Metadata	Image no.	Types of Images
Diagnostics	Image Processing	Evaluating the Diagnostic and Treatment Recommendation Capabilities of GPT-4 Vision in Dermatology [52]	Pillai et al.	2024	GPT-4V demonstrated strong diagnostic accuracy for dermatological conditions, especially in text-based scenarios, with 89% accuracy in both text and multimodal setups. However, its image-based diagnosis showed lower performance, highlighting the need for further model development.	GPT-4.0	Publicly available sources: dermnet.nz and dermatlas.org.	Two board-certified dermatologists	N/A	A combination of both images and metadata	54 images.	Images depicting 9 common dermatological conditions, showcasing classic manifestations of these conditions.
		Argentine dermatology and ChatGPT: infrequent use and intermediate stance [57]	Ko et al.	2024	A survey of 257 Argentine dermatologists showed 83.7% were familiar with ChatGPT, but 65.4% had never used it. While 74.9% expressed interest in future use, only 5.4% used it frequently. Most were 'early majority' adopters.	N/A	N/A	N/A	Prospective	N/A	N/A	N/A
		Claude 3 Opus and ChatGPT With GPT-4 in Dermoscopic Image Analysis for Melanoma Diagnosis: Comparative Performance Analysis [48]	Liu et al.	2024	This study compared the diagnostic performance of Claude 3 Opus and ChatGPT for melanoma detection, finding no significant difference in primary diagnosis accuracy but superior malignancy discrimination by Claude 3 Opus. Both models showed potential, but their limitations highlight the need for further development in AI-driven dermatology tools.	GPT-4.0	The International Skin Imaging Collaboration (ISIC) archive.	N/A	N/A	Image	100	Dermoscopic images of melanocytic lesions.
		ChatGPT versus clinician: challenging the diagnostic capabilities of artificial intelligence in dermatology [58]	Stoneham et al.	2024	ChatGPT correctly diagnosed 56% of cases with expert data and 39% with non-specialist data, lower than dermatologists (83%). It always provided a differential diagnosis but did not significantly improve diagnostic accuracy in primary or secondary care.	GPT-4.0	N/A	The diagnosis made by ChatGPT was compared to those made by dermatologists (experts) and nonspecialists.ective	Retrospective	Metadata	N/A	N/A

Table 1. *Cont.*

Type	Focus Area	Paper Title/Details	Authors	Year	Summary of Findings	Version of GPT	Image Dataset Source	Diagnosis Compared to	Study Type	Diagnosis Made Based on Images Alone or Metadata	Image no.	Types of Images
		A Qualitative Analysis of Provider Notes of Atopic Dermatitis-Related Visits Using Natural Language Processing Methods [8]	Pierce et al.	2021	This study analyzed provider notes for 133,025 patients with atopic dermatitis (AD), revealing a focus on symptoms (primarily itch) and treatment, but limited documentation of AD's impact on patients' work or lifestyle. The findings highlight a care gap that requires further investigation.	N/A	N/A	N/A	Retrospective	N/A	N/A	N/A
NLP		Application of a natural language processing artificial intelligence tool in psoriasis: A cross-sectional comparative study on identifying affected areas in patients' data [6]	Shapiro et al.	2024	ChatGPT-4 accurately analyzed unstructured EMR data from psoriasis patients, identifying affected body areas with 92.8% accuracy. It demonstrated high performance in detecting nail and joint involvement, though errors were more common in complex cases.	GPT-4.0	The study does not involve images; it uses unstructured text data from EMRs.	Senior dermatologist	Retrospective	Metadata	N/A	N/A
		Comparing Meta-Analyses with ChatGPT in the Evaluation of the Effectiveness and Tolerance of Systemic Therapies in Moderate-to-Severe Plaque Psoriasis [7]	Lam Hoai et al.	2023	ChatGPT-4 accurately analyzed psoriasis patient data, identifying affected areas with 92.8% accuracy. It performed well in detecting nail and joint involvement, though errors occurred in complex cases.	N/A	The study does not mention the use of an image dataset; it focuses on evaluating textual data and conclusions from meta-analyses.	Experts	Retrospective	The study did not involve image-based diagnosis; it focused on evaluating textual conclusions from meta-analyses and ChatGPT outputs	N/A	N/A

Table 1. *Cont.*

Type	Focus Area	Paper Title/Details	Authors	Year	Summary of Findings	Version of GPT	Image Dataset Source	Diagnosis Compared to	Study Type	Diagnosis Made Based on Images Alone or Metadata	Image no.	Types of Images
Patient Interaction	Use of ChatGPT for Query Handling	Trends in Accuracy and Appropriateness of Alopecia Areata Information Obtained from a Popular Online Large Language Model, ChatGPT [59]	O'Hagan et al.	2023	ChatGPT 4.0 demonstrated higher accuracy (4.53/5) than ChatGPT 3.5 (4.29/5) in addressing patient questions about alopecia areata. Responses were rated highly appropriate for general information and moderately suitable for EHR drafts, indicating potential for patient education and clinical use.	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		Comparing the quality of ChatGPT- and physician-generated responses to patients' dermatology questions in the electronic medical record [9]	Reynolds et al.	2024	This study was evaluating responses to patient questions, physician-generated responses were preferred over ChatGPT's, especially for readability and empathy. However, ChatGPT was seen as useful for drafting initial responses and providing educational information.	GPT-3.5	N/A	The diagnosis was compared to responses from dermatology physicians, as well as nonphysicians (blinded reviewers).	Retrospective	N/A	N/A	N/A
	Generates learning materials	Assessing the Application of Large Language Models in Generating Dermatologic Patient Education Materials According to Reading Level: Qualitative Study [56]	Lambert et al.	2024	LLMs like GPT-4 generate dermatologic patient education materials (PEMs) at specified reading levels, with GPT-4 performing best at the fifth-grade level for both common and rare conditions. PEMs produced by LLMs are generally accurate, easy to read, and understandable for patients, with variable results at the seventh-grade level.	ChatGPT-3.5, GPT-4.0.	N/A	The diagnosis was compared to 2 blinded dermatology resident trainees.	N/A	N/A	N/A	N/A

Table 1. *Cont.*

Type	Focus Area	Paper Title/Details	Authors	Year	Summary of Findings	Version of GPT	Image Dataset Source	Diagnosis Compared to	Study Type	Diagnosis Made Based on Images Alone or Metadata	Image no.	Types of Images
Others	Performing exams	ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the Specialty Certificate Examination in Dermatology [60]	Lewandowski et al.	2024	ChatGPT-4 outperformed ChatGPT-3.5 in dermatology exams, achieving 80–93% accuracy in English and 70–84% in Polish. While effective in clinical decision support, it struggles with high-difficulty questions. Recommended for aiding but not replacing physicians.	ChatGPT-3.5, GPT-4.0.	N/A	The study compared the performance of ChatGPT to that of a dermatologist with 25 years of experience, who reviewed the questions for compliance with current knowledge.	Retrospective	N/A	N/A	N/A
		Performance of ChatGPT on Specialty Certificate Examination in Dermatology multiple-choice questions [55]	Passby et al.	2024	ChatGPT-4 scored 90% on 84 multiple-choice dermatology questions, outperforming ChatGPT-3.5 (63%). This highlights AI's potential in clinical decision-making, with caution regarding complex cases and patient safety.	ChatGPT-3.5 and ChatGPT-4.	N/A	N/A	N/A	N/A	N/A	N/A
		OpenAI's GPT-4 performs to a high degree on board-style dermatology questions [54]	Elias et al.	2024	GPT-4 achieved 75% accuracy on dermatology board-style questions, showing potential as an educational tool but requiring improvements in response depth and completeness for unsupervised learning. Its performance was consistent across subspecialties and question difficulty.	GPT-4.0	N/A	The diagnosis was compared to the correct answers evaluated by two physicians.	Cross-sectional study	N/A	N/A	N/A
		Pediatric dermatologists versus AI bots: Evaluating the medical knowledge and diagnostic capabilities of ChatGPT [61]	Huang et al.	2024	This study compares OpenAI's ChatGPT (versions 3.5 and 4.0) to pediatric dermatologists in answering multiple-choice and case-based questions. Results show that while human clinicians outperformed both AI versions, ChatGPT-4.0 performed comparably in some areas, highlighting AI's potential with clinician oversight.	GPT-3.5 and GPT-4.0	The image dataset was not used. For cases with accompanying images, only text descriptions were included	The diagnosis was compared to pediatric dermatologists	prospective	text descriptions alone	N/A	N/A

Table 1. *Cont.*

Type	Focus Area	Paper Title/Details	Authors	Year	Summary of Findings	Version of GPT	Image Dataset Source	Diagnosis Compared to	Study Type	Diagnosis Made Based on Images Alone or Metadata	Image no.	Types of Images
Others	management	An evaluation of ChatGPT compared with dermatological surgeons' choices of reconstruction for surgical defects after Mohs surgery [62]	Cuellar-Barboza et al.	2024	This study found that while ChatGPT-4 showed slight concordance with dermatologists in reconstructive decision-making for skin cancer surgery, the agreement was lower than that between dermatologists themselves. The findings highlight the variability in AI-driven medical decisions and the importance of certified expertise.	ChatGPT-4.0	N/A	The diagnosis was compared to dermatological surgeons' choices.	Retrospective	N/A	N/A	N/A
		Evaluation of ChatGPT's acne advice [63]	Li et al.	2024	This study assessed ChatGPT's responses to acne-related queries, evaluating accuracy, completeness, and relevance. Dermatologists rated the responses as satisfactory, limited, or problematic, revealing variable quality in the answers.	ChatGPT-3.5	N/A	The diagnosis was compared to two board-certified dermatologists	Retrospective	N/A	N/A	N/A
Machine Learning	Diagnostics	Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4 [64]	Zhou et al.	2024	SkinGPT-4 is an interactive dermatology diagnostic system that combines a vision transformer and Llama-2-13b-chat, trained on 52,929 skin disease images. It autonomously diagnoses skin conditions, analyzes characteristics, and offers treatment recommendations based on real-life evaluations with dermatologists.							
	Treatment	Can large language models provide secondary reliable opinion on treatment options for dermatological diseases? [53]	Iqbal et al.	2024	Proven potential to provide accurate second opinions on dermatological medication recommendations (98.87% approval rate by dermatologists). However, limitations include occasional coding inaccuracies and incomplete data, suggesting the need for domain-specific knowledge integration.							
	Exams	Assessing large language models' accuracy in providing patient support for choroidal melanoma [65]	Anguita et al.	2024	ChatGPT provided the most accurate answers (92%) for medical advice questions about choroidal melanoma compared to Bing AI and DocsGPT. However, inconsistencies highlight the need for fine-tuning and oversight before clinical use.					Assessing large language models' accuracy in providing patient support for choroidal melanoma [65]	Anguita et al.	2024
		Performance of Three Large Language Models on Dermatology Board Examinations [66]	Mirza et al.	2024	GPT-4 outperformed GPT-3.5 and Google Bard in dermatology board-style questions, achieving 81.7% accuracy and passing CORE and APPLIED exams. Challenges included difficulty with higher-order and complex questions.					Performance of Three Large Language Models on Dermatology Board Examinations [66]	Mirza et al.	2024

Table 1. *Cont.*

Type	Focus Area	Paper Title/Details	Authors	Year	Summary of Findings	Version of GPT	Image Dataset Source	Diagnosis Compared to	Study Type	Diagnosis Made Based on Images Alone or Metadata	Image no.	Types of Images
[53,64] Reviews	Review—diagnostics	Assessing the Impact of ChatGPT in Dermatology: A Comprehensive Rapid Review. [67]	Goktas et al.	2024	ChatGPT shows promise in patient education and teledermatology but faces challenges in diagnosing complex cases and raises ethical concerns regarding data privacy and algorithmic bias. Future research should focus on improving its diagnostic accuracy and addressing these issues.							
		ChatGPT and dermatology [68]	D'AGOSTINO et al.	2024	This review explores the potential applications of ChatGPT in dermatology, highlighting its role in clinical practice and patient support. It emphasizes the synergy between AI and dermatology, driving innovation in healthcare delivery.							
	General review on the potential	Potential applications of ChatGPT in dermatology [69]	Kluger	2023	Supports clinical decision-making and treatment planning with high accuracy for common conditions. Facilitates patient education, simplifies medical writing, and integrates into teledermatology platforms for consultations and triage. Limitations include challenges in multilingual settings, image interpretation, and ethical concerns.							
		Ethical considerations for artificial intelligence in dermatology: a scoping review [70]	Gordon et al.	2024	AI applications span mobile apps for skin cancer detection, clinical image analysis, and large language models for diagnostic queries. Ethical concerns include biases, misdiagnosis risks, data privacy, and exacerbation of health disparities in teledermatology. Benefits include improved access, decision-making, and efficiency in clinical practice, but safeguards are necessary for ethical use.							

Table 1. *Cont.*

Type	Focus Area	Paper Title/Details	Authors	Year	Summary of Findings	Version of GPT	Image Dataset Source	Diagnosis Compared to	Study Type	Diagnosis Made Based on Images Alone or Metadata	Image no.	Types of Images
[53,64] Reviews	Analyzing potential	ChatGPT for healthcare providers and patients: Practical implications within dermatology [71]	Jin et al.	2023	Identified five domains of use in dermatology, including automating administrative tasks, enhancing patient education, supporting medical education, aiding clinical research, and improving health literacy. Ethical challenges include risks of "artificial hallucinations," biases, and outdated datasets, necessitating systematic validation and usage guidelines.							
		The Arrival of Artificial Intelligence Large Language Models and Vision-Language Models: A Potential to Possible Change in the Paradigm of Healthcare Delivery in Dermatology [19]	Gupta et al.	2024	The study explores the potential of large language models (LLMs) and vision-language models (VLMs) in dermatology, addressing how AI can improve patient care amidst challenges like workload and staffing shortages. AI technologies, such as ChatGPT and Google Bard, could transform dermatology by integrating text and image inputs.							

Table 2. Volume of articles on ChatGPT vs. other LLMs in dermatology diagnostics.

Model	Volume of Articles	Applications in Dermatology
ChatGPT (OpenAI)	High (50+ articles)	Used for diagnostic support, clinical note analysis, and patient interaction. Achieves 88% accuracy in handling common queries.
Google Bard (PaLM)	Low (2–5 articles)	No specific dermatology focus found; primarily applied to general conversational AI and integration with Google systems.
Claude by Anthropic	Low (1–2 articles)	Limited mentions in healthcare; prioritized for safe, ethical applications in sensitive tasks.
Microsoft Copilot (Powered by GPT-4)	Moderate (5–10 articles)	Integrated in teledermatology workflows via Office Suite for clinical reporting and diagnostic data organization
LLaMA (Meta)	Low (<5 articles)	Primarily used in researchno specific dermatology-related applications identified.
Mistral	None Found	No known dermatology-related applications.
Cohere	None Found	Primarily enterprise knowledge management; no dermatology-specific use cases identified.
Amazon Bedrock	None Found	Focus on general enterprise flexibility; no dermatol no dermatology-specific use cases identified.
xAI (Grok)	None found	Still emerging; no dermatology-specific use cases identified.
IBM Watson Assistant	Low (1–2 articles)	Some use cases in patient engagement and healthcare support, but minimal focus on dermatology.
Jasper AI	None Found	No known applications in dermatology; focused on content creation.
Character.ai	None Found	Used for entertainment; no healthcare or dermatology use cases.
DeepMind Gemini	Emerging	Promising capabilities in diagnostics, but still in development with no dermatology applications yet.
Perplexity.ai	None Found	Focused on information retrieval with no dermatology-specific applications.

Table 2 displays that among the evaluated LMMs, very few have published studies on their use in dermatology. In this regard, ChatGPT had the most number of articles published: over 50 studies on various applications, including diagnostic support, clinical note analysis, and patient interaction, with an 88% accuracy rating for handling common queries. Microsoft Copilot is powered by GPT-4 with 5–10 articles published, mainly about its integration into teledermatology workflows, clinical reporting, and organizing diagnostic data. Google Bard (PaLM) has 2–5 published studies, and Claude by Anthropic has 1–2 published studies; however, these two do not specifically address dermatology, as they have been more generally applied to broader conversational AI tasks. Other models with limited publication on dermatology include IBM Watson Assistant, LLaMA, and emerging models such as DeepMind Gemini. Many models, such as Mistral, Cohere, and Amazon Bedrock, have no known applications in the field of dermatology and have focused on other domains, such as enterprise knowledge management and more general AI versatility. Importantly, the comparatively higher number of studies examining GPT-based models reinforces their established role in dermatology research while raising the opportunity for additional explorative work around their diagnostic potential.

6. Limitations of Using Large Language Models (LLMs) Like ChatGPT in Dermatology Diagnostics

Though LLMs like ChatGPT have shown possibilities in aiding dermatology diagnostics, they still hold some considerable shortcomings. The inability to interpret dermatoscopic or clinical photographs directly is a major limitation; this is fundamentally central to the diagnosis of most skin conditions. Therefore, their incapacity to make independent diagnoses is largely due to their dependence on other tools or their joint incorporation with image-processing systems [72,73]. Nonetheless, the introduction of GPT-4 has partly resolved this limitation. In this regard, GPT-4, by combining with vision humanoid models or external APIs, has the possibility to process and analyze pictures to some degree, rendering dermatological practice more relevant by joining both text and visual diagnostic workflows [74].

Despite these advancements, the risk of "hallucination"—generating incorrect or misleading suggestions—remains a concern, particularly in high-stakes clinical scenarios [75]. Furthermore, dermatologic conditions with similar visual presentations often require nuanced differential diagnosis, a task that AI-based imaging tools may not reliably perform without contextual clinical information [48,58]. Furthermore, these models still require integration with domain-specific knowledge bases to ensure relevance and accuracy, as their general training data often lack the granularity necessary for specialized medical contexts [76,77].

It is worth mentioning that uploading images to GPT may result in some ethical consequences. Before sharing images, patient permission must be obtained, taking dignity and confidentiality into account. Temporary chat mode—the automatic deletion of the

conversation and any uploaded data when the session closes—is preferable in terms of potential risks. This would further avoid abuse and guarantee better privacy protection. Additionally, one such capacity in allowing this temporary and single-session interaction gives evidence of a huge advantage to keeping ethics in place when such AI technologies are used in healthcare [70,78,79].

7. Future Perspectives

The future of integrating LLMs like ChatGPT into dermatology looks bright, with several anticipated developments listed below:

- Better Integration with Diagnostic Tools:

The combination of ChatGPT with dermatoscopes, smartphone applications, and other diagnostic technologies has the prospective of creating an improved workflow of clinicians, facilitating rapid data interpretation and prompt decision-making. For instance, linking ChatGPT with a smartphone app specially for clinical use would allow for real-time, dialogue-based analyses of patient-reported symptoms in conjunction with visual data from dermatoscopic images.

- Multimodal Models:

Future AI systems combining text and image analysis stand to provide a synthesis of a holistic diagnostic approach. These multimodal models could better connect the language-functioning domain of ChatGPT with various imaging technologies in improving diagnostic accuracy for those conditions requiring visual assessment, such as melanomas and psoriasis.

- Personalized Treatment Recommendations:

LLMs could be personalized to incorporate patient-specific data like genetic profiles, medical history, and lifestyle preferences into treatment recommendations. This will enable dermatologists to provide more targeted accompaniment to them.

- AI-Assisted Diagnostic Frameworks:

The development of frameworks involving models like ChatGPT could reform the practice of dermatology following AI-assisted consultations and a clinical support system for diagnosis that would enhance clinical efficiency and precision.

- Future Research Directions:

Longitudinal studies assessing the long-term reliability and outcomes of these technologies are needed for the promise of AI in dermatology to be fully realized. These can include real-world implementation projects and clinical trials aimed at their exploration of efficiency and scalability, ensuring AI models produce unbiased functioning across different populations and settings.

8. Conclusions

The integration of ChatGPT and similar LLMs in dermatology is a major step forward toward assistance in diagnosis and clinical support. While the model offers significant advantages, especially in terms of natural language processing and image analysis, it is fraught with limitations that could affect the diagnostic interpretation of visual laboratory tests and assignments. ChatGPT can also serve as a very helpful assistant for dermatologists to speed up the processing of patient information and differential diagnosis particularly in teledermatology and remote care contexts. Nevertheless, we argue for further research and development to optimize such instruments. A more accurate, self-informing tool would still command the highest approach because reputation is at stake as far as these case

studies are concerned, alongside all the possible benefits and challenges like data privacy and clinical liability. Owing to the continued evolution of AI technology, the future of dermatology will involve a more seamless integration of AI diagnostic tools for always better efficiencies, accuracy, and equitable patient care improvements.

Author Contributions: Conceptualization, Z.K.; writing—original draft preparation, M.A.; writing—review and editing, B.J.; formal analysis, N.B.; supervision, J.S. All authors contributed equally. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: All 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. Alowais, S.A.; Alghamdi, S.S.; Alsuhebany, N.; Alqahtani, T.; Alshaya, A.I.; Almohareb, S.N.; Aldairem, A.; Alrashed, M.; Saleh, K.B.; Badreldin, H.A.; et al. Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Med. Educ.* **2023**, *23*, 689. Available online: https://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-023-04698-z (accessed on 12 November 2024). [CrossRef] [PubMed]
2. Alqahtani, T.; Badreldin, H.A.; Alrashed, M.; Alshaya, A.I.; Alghamdi, S.S.; bin Saleh, K.; Alowais, S.A.; Alshaya, O.A.; Rahman, I.; Al Yami, M.S.; et al. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Res. Soc. Adm. Pharm.* **2023**, *19*, 1236–1242. [CrossRef] [PubMed]
3. De Angelis, L.; Baglivo, F.; Arzilli, G.; Privitera, G.P.; Ferragina, P.; Tozzi, A.E.; Rizzo, C. ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health. *Front. Public Health* **2023**, *11*, 1166120. Available online: https://pubmed.ncbi.nlm.nih.gov/37181697/ (accessed on 12 November 2024).
4. Du-Harpur, X.; Watt, F.M.; Luscombe, N.M.; Lynch, M.D. What is AI? Applications of artificial intelligence to dermatology. *Br. J. Dermatol.* **2020**, *183*, 423–430. Available online: https://pubmed.ncbi.nlm.nih.gov/31960407/ (accessed on 12 November 2024). [CrossRef]
5. Khalifa, M.; Albadawy, M. AI in diagnostic imaging: Revolutionising accuracy and efficiency. *Comput. Methods Programs Biomed. Update* **2024**, *5*, 100146. [CrossRef]
6. Shapiro, J.; Baum, S.; Pavlotzky, F.; Mordehai, Y.B.; Barzilai, A.; Freud, T.; Gershon, R. Application of a natural language processing artificial intelligence tool in psoriasis: A cross-sectional comparative study on identifying affected areas in patients' data. *Clin. Dermatol.* **2024**, *42*, 480–486. [CrossRef]
7. Lam Hoai, X.L.; Simonart, T. Comparing Meta-Analyses with ChatGPT in the Evaluation of the Effectiveness and Tolerance of Systemic Therapies in Moderate-to-Severe Plaque Psoriasis. *J. Clin. Med.* **2023**, *12*, 5410. Available online: https://www.mdpi.com/2077-0383/12/16/5410/htm (accessed on 15 November 2024). [CrossRef] [PubMed]
8. Pierce, E.J.; Boytsov, N.N.; Vasey, J.J.; Sudaria, T.C.; Liu, X.; Lavelle, K.W.; Bogdanov, A.N.; Goldblum, O.M. A Qualitative Analysis of Provider Notes of Atopic Dermatitis-Related Visits Using Natural Language Processing Methods. *Dermatol. Ther.* **2021**, *11*, 1305–1318. Available online: https://link.springer.com/article/10.1007/s13555-021-00553-5 (accessed on 15 November 2024). [CrossRef]
9. Reynolds, K.; Nadelman, D.; Durgin, J.; Ansah-Addo, S.; Cole, D.; Fayne, R.; Harrell, J.; Ratycz, M.; Runge, M.; Shepard-Hayes, A.; et al. Comparing the Quality of ChatGPT-and Physician-Generated Responses to Patients' Dermatologic Questions in the Electronic Medical Record. Available online: https://academic.oup.com/ced/advance-article-abstract/doi/10.1093/ced/llad456/7511330 (accessed on 15 November 2024).
10. Fliorent, R.; Fardman, B.; Podwojniak, A.; Javaid, K.; Tan, I.J.; Ghani, H.; Truong, T.M.; Rao, B.; Heath, C. Artificial intelligence in dermatology: Advancements and challenges in skin of color. *Int. J. Dermatol.* **2024**, *63*, 455–461. Available online: https://onlinelibrary.wiley.com/doi/full/10.1111/ijd.17076 (accessed on 13 November 2024). [CrossRef]
11. Soori, M.; Arezoo, B.; Dastres, R. Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cogn. Robot.* **2023**, *3*, 54–70. [CrossRef]
12. Abhishek, K.; Kawahara, J.; Hamarneh, G. Predicting the clinical management of skin lesions using deep learning. *Sci. Rep.* **2021**, *11*, 7769. Available online: https://www.nature.com/articles/s41598-021-87064-7 (accessed on 13 November 2024).
13. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. Available online: https://www.nature.com/articles/nature21056 (accessed on 13 November 2024). [CrossRef]

14. Voon, W.; Hum, Y.C.; Tee, Y.K.; Yap, W.S.; Salim, M.I.M.; Tan, T.S.; Mokayed, H.; Lai, K.W. Performance analysis of seven Convolutional Neural Networks (CNNs) with transfer learning for Invasive Ductal Carcinoma (IDC) grading in breast histopathological images. *Sci. Rep.* **2022**, *12*, 19200. Available online: https://www.nature.com/articles/s41598-022-21848-3 (accessed on 13 November 2024). [CrossRef] [PubMed]
15. Venkatesh, K.P.; Raza, M.M.; Nickel, G.; Wang, S.; Kvedar, J.C. Deep learning models across the range of skin disease. *NPJ Digit. Med.* **2024**, *7*, 32. Available online: https://www.nature.com/articles/s41746-024-01033-8 (accessed on 13 November 2024). [CrossRef] [PubMed]
16. Ozmen Garibay, O.; Winslow, B.; Andolina, S.; Antona, M.; Bodenschatz, A.; Coursaris, C.; Falco, G.; Fiore, S.M.; Garibay, I.; Grieman, K.; et al. Six Human-Centered Artificial Intelligence Grand Challenges. *Int. J. Hum. Comput. Interact.* **2023**, *39*, 391–437. Available online: https://www.tandfonline.com/doi/abs/10.1080/10447318.2022.2153320 (accessed on 13 November 2024). [CrossRef]
17. Hubert, K.F.; Awa, K.N.; Zabelina, D.L. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Sci. Rep.* **2024**, *14*, 3440. Available online: https://www.nature.com/articles/s41598-024-53303-w (accessed on 13 November 2024). [CrossRef]
18. Liu, J.; Wang, C.; Liu, S. Utility of ChatGPT in Clinical Practice. *J. Med. Internet Res.* **2023**, *25*, e48568. Available online: https://pmc.ncbi.nlm.nih.gov/articles/PMC10365580/ (accessed on 13 November 2024). [CrossRef]
19. Gupta, A.K.; Talukder, M.; Wang, T.; Daneshjou, R.; Piguet, V. The Arrival of Artificial Intelligence Large Language Models and Vision-Language Models: A Potential to Possible Change in the Paradigm of Healthcare Delivery in Dermatology. *J. Investig. Dermatol.* **2024**, *144*, 1186–1188. Available online: https://www.jidonline.org/article/S0022-202X(24)00004-6/fulltext (accessed on 13 November 2024). [CrossRef]
20. Mu, Y.; He, D. The Potential Applications and Challenges of ChatGPT in the Medical Field. *Int. J. Gen. Med.* **2024**, *17*, 817–826. [CrossRef]
21. Chen, J.; Liu, L.; Ruan, S.; Li, M.; Yin, C. Are Different Versions of ChatGPT's Ability Comparable to the Clinical Diagnosis Presented in Case Reports? A Descriptive Study. *J. Multidiscip. Healthc.* **2023**, *16*, 3825–3831. [CrossRef]
22. Clusmann, J.; Kolbinger, F.R.; Muti, H.S.; Carrero, Z.I.; Eckardt, J.N.; Laleh, N.G.; Löffler, C.M.L.; Schwarzkopf, S.-C.; Unger, M.; Veldhuizen, G.P.; et al. The future landscape of large language models in medicine. *Commun. Med.* **2023**, *3*, 141. Available online: https://www.nature.com/articles/s43856-023-00370-1 (accessed on 13 November 2024). [CrossRef]
23. Sonthalia, S.; Yumeen, S.; Kaliyadan, F. Dermoscopy Overview and Extradiagnostic Applications. Available online: https://www.ncbi.nlm.nih.gov/books/NBK537131/ (accessed on 14 November 2024).
24. Wang, Z.; Wang, C.; Peng, L.; Lin, K.; Xue, Y.; Chen, X.; Bao, L.; Liu, C.; Zhang, J.; Xie, Y. Radiomic and deep learning analysis of dermoscopic images for skin lesion pattern decoding. *Sci. Rep.* **2024**, *14*, 19781. Available online: https://www.nature.com/articles/s41598-024-70231-x (accessed on 14 November 2024). [CrossRef]
25. Omar, M.; Ullanat, V.; Loda, M.; Marchionni, L.; Umeton, R. ChatGPT for digital pathology research. *Lancet Digit. Health* **2024**, *6*, e595–e600. Available online: http://www.thelancet.com/article/S2589750024001146/fulltext (accessed on 14 November 2024). [CrossRef] [PubMed]
26. Goktas, P.; Grzybowski, A. Assessing the Impact of ChatGPT in Dermatology: A Comprehensive Rapid Review. *J. Clin. Med.* **2024**, *13*, 5909. Available online: https://pmc.ncbi.nlm.nih.gov/articles/PMC11477344/ (accessed on 14 November 2024). [CrossRef] [PubMed]
27. Gabashvili, I.S. ChatGPT in Dermatology: A Comprehensive Systematic Review. 2023. Available online: http://medrxiv.org/lookup/doi/10.1101/2023.06.11.23291252 (accessed on 14 November 2024).
28. Nazir, A.; Wang, Z. A comprehensive survey of ChatGPT: Advancements, applications, prospects, and challenges. *Meta-Radiology* **2023**, *1*, 100022. [CrossRef]
29. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. Available online: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00444-8 (accessed on 14 November 2024). [CrossRef]
30. Ray, P.P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys. Syst.* **2023**, *3*, 121–154. [CrossRef]
31. Roumeliotis, K.I.; Tselikas, N.D. ChatGPT and Open-AI Models: A Preliminary Review. *Future Internet* **2023**, *15*, 192. Available online: https://www.mdpi.com/1999-5903/15/6/192/htm (accessed on 15 November 2024). [CrossRef]
32. Shapiro, J.; Lyakhovitsky, A. Revolutionizing teledermatology: Exploring the integration of artificial intelligence, including Generative Pre-trained Transformer chatbots for artificial intelligence-driven anamnesis, diagnosis, and treatment plans. *Clin. Dermatol.* **2024**, *42*, 492–497. Available online: http://www.cidjournal.com/article/S0738081X24001044/fulltext (accessed on 15 November 2024). [CrossRef]

33. McKoy, K.; Halpern, S.; Mutyambizi, K. International Teledermatology Review. *Curr. Dermatol. Rep.* **2021**, *10*, 55. Available online: https://pmc.ncbi.nlm.nih.gov/articles/PMC8317676/ (accessed on 15 November 2024). [CrossRef]
34. Campagna, M.; Naka, F.; Lu, J. Teledermatology: An updated overview of clinical applications and reimbursement policies. *Int. J. Womens Dermatol.* **2017**, *3*, 176–179. [CrossRef]
35. Alomari, E.A. Unlocking the Potential: A Comprehensive Systematic Review of ChatGPT in Natural Language Processing Tasks. *Comput. Model. Eng. Sci.* **2024**, *141*, 43–85. [CrossRef]
36. Bansal, G.; Chamola, V.; Hussain, A.; Guizani, M.; Niyato, D. Transforming Conversations with AI—A Comprehensive Study of ChatGPT. *Cognit. Comput.* **2024**, *16*, 2487–2510. Available online: https://link.springer.com/article/10.1007/s12559-023-10236-2 (accessed on 15 November 2024). [CrossRef]
37. Ferreira, A.L.; Chu, B.; Grant-Kels, J.M.; Ogunleye, T.; Lipoff, J.B. Evaluation of ChatGPT Dermatology Responses to Common Patient Queries. *JMIR Dermatol.* **2023**, *6*, e49280. Available online: https://pmc.ncbi.nlm.nih.gov/articles/PMC10692871/ (accessed on 15 November 2024). [CrossRef] [PubMed]
38. Jerfy, A.; Selden, O.; Balkrishnan, R. The Growing Impact of Natural Language Processing in Healthcare and Public Health. *Inq. J. Health Care Organ. Provis. Financ.* **2024**, *61*, 00469580241290095. Available online: https://journals.sagepub.com/doi/full/10.1177/00469580241290095 (accessed on 15 November 2024). [CrossRef] [PubMed]
39. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940. Available online: https://www.nature.com/articles/s41591-023-02448-8 (accessed on 15 November 2024). [CrossRef] [PubMed]
40. Haltaufderheide, J.; Ranisch, R. The ethics of ChatGPT in medicine and healthcare: A systematic review on Large Language Models (LLMs). *NPJ Digit. Med.* **2024**, *7*, 183. Available online: https://www.nature.com/articles/s41746-024-01157-x (accessed on 15 November 2024). [CrossRef]
41. Hariri, W. Unlocking the Potential of ChatGPT: A Comprehensive Exploration of Its Applications, Advantages, Limitations, and Future Directions in Natural Language Processing. Available online: http://arxiv.org/abs/2304.02017 (accessed on 16 November 2024).
42. OpenAI. ChatGPT (Version GPT-4) [Large Language Model]. *OpenAI*. 2024. Available online: https://openai.com/chatgpt (accessed on 14 June 2024).
43. Nayak, I.; Kanagaraj, K.; Shrimali, M.; Kumar, D.; Ghamande, M.V.; Kumar, C.S. Improved Convolutional Neural Network based Approach for Analyzing the ChatGPT and Different Fields Impact. In Proceedings of the 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 11–13 December 2023; pp. 629–635.
44. Gao, Y.; Zhang, L.; Xu, Y. CKG: Improving ABSA with text augmentation using ChatGPT and knowledge-enhanced gated attention graph convolutional networks. *PLoS ONE* **2024**, *19*, e0301508. Available online: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0301508 (accessed on 16 November 2024). [CrossRef]
45. Ahmed, W.; Afzal, M.S.; Khan, M.N. ChatGPT in Dermatology: Evaluating Its Diagnostic Accuracy and Potential Role in Clinical Practice. *Clin. Cosmet. Investig. Dermatol.* **2023**, *16*, 1795–1802.
46. DePalma, K.; Miminoshvili, I.; Henselder, C.; Moss, K.; AlOmar, E.A. Exploring ChatGPT's code refactoring capabilities: An empirical study. *Expert Syst. Appl.* **2024**, *249*, 123602. [CrossRef]
47. Using ChatGPT for Coding Challenges | by Mustakim Masum | Medium. Available online: https://medium.com/@m.m.kiron/using-chatgpt-for-coding-challenges-4cab6ca3bdc7 (accessed on 16 November 2024).
48. Liu, X.; Duan, C.; Kim, M.K.; Zhang, L.; Jee, E.; Maharjan, B.; Huang, Y.; Du, D.; Jiang, X. Claude 3 Opus and ChatGPT With GPT-4 in Dermoscopic Image Analysis for Melanoma Diagnosis: Comparative Performance Analysis. *JMIR Med. Inform.* **2024**, *12*, e59273. Available online: https://pubmed.ncbi.nlm.nih.gov/39106482/ (accessed on 17 November 2024). [CrossRef]
49. Lakdawala, N.; Channa, L.; Gronbeck, C.; Lakdawala, N.; Weston, G.; Sloan, B.; Feng, H. Assessing the Accuracy and Comprehensiveness of ChatGPT in Offering Clinical Guidance for Atopic Dermatitis and Acne Vulgaris. *JMIR Dermatol.* **2023**, *6*, e50409. Available online: https://pmc.ncbi.nlm.nih.gov/articles/PMC10685272/ (accessed on 17 November 2024). [CrossRef]
50. Alanezi, F. Factors influencing patients' engagement with ChatGPT for accessing health-related information. *Crit. Public. Health* **2024**, *34*, 1–20. Available online: https://www.tandfonline.com/doi/abs/10.1080/09581596.2024.2348164 (accessed on 17 November 2024). [CrossRef]
51. Zhang, D.; Zhao, X. Understanding adoption intention of virtual medical consultation systems: Perceptions of ChatGPT and satisfaction with doctors. *Comput. Human Behav.* **2024**, *159*, 108359. [CrossRef]
52. Pillai, A.; Parappally Joseph, S.; Hardin, J. Evaluating the Diagnostic and Treatment Recommendation Capabilities of GPT-4 Vision in Dermatology. *medRxiv* **2024**. [CrossRef]
53. Iqbal, U.; Lee, L.T.J.; Rahmanti, A.R.; Celi, L.A.; Li, Y.C.J. Can large language models provide secondary reliable opinion on treatment options for dermatological diseases? *J. Am. Med. Inform. Assoc.* **2024**, *31*, 1341–1347. Available online: https://pubmed.ncbi.nlm.nih.gov/38578616/ (accessed on 20 November 2024). [CrossRef] [PubMed]

54. Elias, M.L.; Burshtein, J.; Sharon, V.R. OpenAI's GPT-4 performs to a high degree on board-style dermatology questions. *Int. J. Dermatol.* **2024**, *63*, 73–78. Available online: https://pubmed.ncbi.nlm.nih.gov/38131454/ (accessed on 20 November 2024). [CrossRef] [PubMed]
55. Passby, L.; Jenko, N.; Wernham, A. Performance of ChatGPT on Specialty Certificate Examination in Dermatology multiple-choice questions. *Clin. Exp. Dermatol.* **2024**, *49*, 722–727. Available online: https://pubmed.ncbi.nlm.nih.gov/37264670/ (accessed on 20 November 2024). [CrossRef]
56. Lambert, R.; Choo, Z.Y.; Gradwohl, K.; Schroedl, L.; De Luzuriaga, A.R. Assessing the Application of Large Language Models in Generating Dermatologic Patient Education Materials According to Reading Level: Qualitative Study. *JMIR Dermatol.* **2024**, *7*, e55898. Available online: https://pubmed.ncbi.nlm.nih.gov/38754096/ (accessed on 20 November 2024). [CrossRef]
57. Ko, E.A.; Torre, A.C.; Hernandez, B.; Bibiloni, N.; Covián, E.; Salerni, G.; Alonso, C.; Ochoa, A.K.; Mazzuoccolo, L.D. Argentine dermatology and ChatGPT: Infrequent use and intermediate stance. *Clin. Exp. Dermatol.* **2024**, *49*, 734–736. [CrossRef]
58. Stoneham, S.; Livesey, A.; Cooper, H.; Mitchell, C. ChatGPT versus clinician: Challenging the diagnostic capabilities of artificial intelligence in dermatology. *Clin. Exp. Dermatol.* **2024**, *49*, 707–710. Available online: https://pubmed.ncbi.nlm.nih.gov/37979201/ (accessed on 20 November 2024). [CrossRef]
59. O'Hagan, R.; Kim, R.H.; Abittan, B.J.; Caldas, S.; Ungar, J.; Ungar, B. Trends in Accuracy and Appropriateness of Alopecia Areata Information Obtained from a Popular Online Large Language Model, ChatGPT. *Dermatology* **2023**, *239*, 952–957. Available online: https://pubmed.ncbi.nlm.nih.gov/37722370/ (accessed on 20 November 2024). [CrossRef]
60. Lewandowski, M.; Łukowicz, P.; Świetlik, D.; Barańska-Rybak, W. ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the Specialty Certificate Examination in Dermatology. *Clin. Exp. Dermatol.* **2024**, *49*, 686–691. Available online: https://pubmed.ncbi.nlm.nih.gov/37540015/ (accessed on 20 November 2024). [CrossRef] [PubMed]
61. Huang, C.Y.; Zhang, E.; Caussade, M.C.; Brown, T.; Stockton Hogrogian, G.; Yan, A.C. Pediatric dermatologists versus AI bots: Evaluating the medical knowledge and diagnostic capabilities of ChatGPT. *Pediatr. Dermatol.* **2024**, *41*, 831–834. Available online: https://pubmed.ncbi.nlm.nih.gov/38721744/ (accessed on 20 November 2024). [CrossRef] [PubMed]
62. Cuellar-Barboza, A.; Brussolo-Marroquín, E.; Cordero-Martinez, F.C.; Aguilar-Calderon, P.E.; Vazquez-Martinez, O.; Ocampo-Candiani, J. An evaluation of ChatGPT compared with dermatological surgeons' choices of reconstruction for surgical defects after Mohs surgery. *Clin. Exp. Dermatol.* **2024**, *49*, 1367–1371. Available online: https://pubmed.ncbi.nlm.nih.gov/38738492/ (accessed on 20 November 2024). [CrossRef]
63. Li, K.; Hsu, J.T.S.; Li, M.K. Evaluation of ChatGPT's acne advice. *Clin. Exp. Dermatol.* **2024**, *49*, 746–749. [CrossRef] [PubMed]
64. Zhou, J.; He, X.; Sun, L.; Xu, J.; Chen, X.; Chu, Y.; Zhou, L.; Liao, X.; Zhang, B.; Afvari, S.; et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nat. Commun.* **2024**, *15*, 5649. Available online: https://www.nature.com/articles/s41467-024-50043-3 (accessed on 19 November 2024). [CrossRef]
65. Anguita, R.; Downie, C.; Ferro Desideri, L.; Sagoo, M.S. Assessing large language models' accuracy in providing patient support for choroidal melanoma. *Eye* **2024**, *38*, 3113–3117. Available online: https://pubmed.ncbi.nlm.nih.gov/39003430/ (accessed on 20 November 2024). [CrossRef]
66. Mirza, F.N.; Lim, R.K.; Yumeen, S.; Wahood, S.; Zaidat, B.; Shah, A.; Tang, O.Y.; Kawaoka, J.; Seo, S.-J.; DiMarco, C.; et al. Performance of Three Large Language Models on Dermatology Board Examinations. *J. Investig. Darmatol.* **2024**, *144*, 398–400. Available online: https://pubmed.ncbi.nlm.nih.gov/37541614/ (accessed on 20 November 2024). [CrossRef]
67. Diazo, J.M.; Ngo, D.Q.; Wetter, D.A.; Davis, D.M.R.; Tollefson, M.M. Performance of ChatGPT in Dermatology: A Systematic Review. *JAAD Int.* **2024**, *14*, 100137.
68. D'Agostino, M.; Feo, F.; Martora, F.; Genco, L.; Megna, M.; Cacciapuoti, S.; Villani, A.; Potestio, L. ChatGPT and dermatology. *Ital. J. Dermatol. Venereol.* **2024**, *159*, 566–571. Available online: https://pubmed.ncbi.nlm.nih.gov/39039954/ (accessed on 19 November 2024). [CrossRef]
69. Kluger, N. Potential applications of ChatGPT in dermatology. *J. Eur. Acad. Dermatol. Venereol.* **2023**, *37*, e941–e942. Available online: https://onlinelibrary.wiley.com/doi/full/10.1111/jdv.19152 (accessed on 20 November 2024). [CrossRef]
70. Gordon, E.R.; Trager, M.H.; Kontos, D.; Weng, C.; Geskin, L.J.; Dugdale, L.S.; Samie, F.H. Ethical considerations for artificial intelligence in dermatology: A scoping review. *Br. J. Dermatol.* **2024**, *190*, 789–797. Available online: https://pubmed.ncbi.nlm.nih.gov/38330217/ (accessed on 20 November 2024). [CrossRef]
71. Jin, J.Q.; Dobry, A.S. ChatGPT for healthcare providers and patients: Practical implications within dermatology. *J. Am. Acad. Dermatol.* **2023**, *89*, 870–871. Available online: https://pubmed.ncbi.nlm.nih.gov/37315798/ (accessed on 20 November 2024). [CrossRef]
72. Jabour, T.B.F.; Ribeiro, J.P.; Fernandes, A.C.; Honorato, C.M.A.; Queiroz Mdo, C.A.P. ChatGPT: Performance of artificial intelligence in the dermatology specialty certificate examination. *An. Bras. Dermatol.* **2023**, *99*, 277. Available online: https://pmc.ncbi.nlm.nih.gov/articles/PMC10943280/ (accessed on 19 November 2024). [CrossRef]

73. Ullah, E.; Parwani, A.; Baig, M.M.; Singh, R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology—A recent scoping review. *Diagn. Pathol.* **2024**, *19*, 43. Available online: https://diagnosticpathology.biomedcentral.com/articles/10.1186/s13000-024-01464-7 (accessed on 19 November 2024). [CrossRef]
74. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
75. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Toronto, ON, Canada, 3–10 March 2021; pp. 610–623. Available online: https://dl.acm.org/doi/10.1145/3442188.3445922 (accessed on 19 November 2024).
76. Kololgi, S.P.; Lahari, C. Harnessing the Power of Artificial Intelligence in Dermatology: A Comprehensive Commentary. *Indian J. Dermatol.* **2024**, *68*, 678. Available online: https://pmc.ncbi.nlm.nih.gov/articles/PMC10868991/ (accessed on 19 November 2024). [CrossRef]
77. Jeong, H.K.; Park, C.; Henao, R.; Kheterpal, M. Deep Learning in Dermatology: A Systematic Review of Current Approaches, Outcomes, and Limitations. *JID Innov.* **2023**, *3*, 100150. [CrossRef]
78. Busik, V. How artificial intelligence and large language models are revolutionizing dermatology. *Dermatologie* **2024**, *75*, 743–746. Available online: https://pubmed.ncbi.nlm.nih.gov/38900290/ (accessed on 19 November 2024). [CrossRef]
79. Porter, E.; Murphy, M.; O'Connor, C. Chat GPT in dermatology: Progressive or problematic? *J. Eur. Acad. Dermatol. Venereol.* **2023**, *37*, e943–e944. Available online: https://pubmed.ncbi.nlm.nih.gov/37150950/ (accessed on 19 November 2024). [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.

diagnostics

MDPI

Review

Machine Learning-Powered Smart Healthcare Systems in the Era of Big Data: Applications, Diagnostic Insights, Challenges, and Ethical Implications

Sita Rani [1], Raman Kumar [2,3,*], B. S. Panda [4], Rajender Kumar [5,6], Nafaa Farhan Muften [7], Mayada Ahmed Abass [8] and Jasmina Lozanović [9,*]

[1] Department of Computer Science and Engineering, Guru Nanak Dev Engineering College, Ludhiana 141006, Punjab, India; sitasaini80@gmail.com
[2] Department of Mechanical and Production Engineering, Guru Nanak Dev Engineering College, Ludhiana 141006, Punjab, India
[3] Jadara Research Center, Jadara University, Irbid 21110, Jordan
[4] Department of CSE, Raghu Engineering College, Visakhapatnam 531162, Andhra Pradesh, India; panda.bs@raghuenggcollege.in
[5] Department of Mechanical Engineering, Graphic Era (Deemed to be University), Clement Town, Dehradun 248002, Uttarakhand, India; rajender6291@gmail.com
[6] Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura 140401, Punjab, India
[7] Department of Medical Laboratories Technology, Mazaya University College, Nasiriyah 64001, Iraq; nafaaalomari10@gmail.com
[8] College of Pharmacy, Al-Mustaqbal University, Babylon 51001, Iraq; mayada.ahmed@uomus.edu.iq
[9] Department of Engineering, FH Campus Wien, University of Applied Sciences, Favoritenstraße 226, 1100 Vienna, Austria
* Correspondence: sehgal91@yahoo.co.in (R.K.); jasmina.lozanovic@fh-campuswien.ac.at (J.L.)

Abstract: Healthcare data rapidly increases, and patients seek customized, effective healthcare services. Big data and machine learning (ML) enabled smart healthcare systems hold revolutionary potential. Unlike previous reviews that separately address AI or big data, this work synthesizes their convergence through real-world case studies, cross-domain ML applications, and a critical discussion on ethical integration in smart diagnostics. The review focuses on the role of big data analysis and ML towards better diagnosis, improved efficiency of operations, and individualized care for patients. It explores the principal challenges of data heterogeneity, privacy, computational complexity, and advanced methods such as federated learning (FL) and edge computing. Applications in real-world settings, such as disease prediction, medical imaging, drug discovery, and remote monitoring, illustrate how ML methods, such as deep learning (DL) and natural language processing (NLP), enhance clinical decision-making. A comparison of ML models highlights their value in dealing with large and heterogeneous healthcare datasets. In addition, the use of nascent technologies such as wearables and Internet of Medical Things (IoMT) is examined for their role in supporting real-time data-driven delivery of healthcare. The paper emphasizes the pragmatic application of intelligent systems by highlighting case studies that reflect up to 95% diagnostic accuracy and cost savings. The review ends with future directions that seek to develop scalable, ethical, and interpretable AI-powered healthcare systems. It bridges the gap between ML algorithms and smart diagnostics, offering critical perspectives for clinicians, data scientists, and policymakers.

Keywords: smart healthcare systems; smart diagnostics; artificial intelligence; machine learning; big data; electronic health records (EHRs)

1. Introduction

The rapid advancement of digital technologies and the integration of networked biomedical devices have profoundly transformed the healthcare landscape, resulting in a substantial increase in the volume, variety, and complexity of healthcare data [1]. Intelligent health systems that leverage machine learning (ML) and big data analytics (BDA) are poised to revolutionize care delivery by enabling advanced prediction, personalized treatment, and real-time patient monitoring [2]. Despite this potential, several critical challenges remain. These include heterogeneity in data sources, ranging from wearable sensors and genomic data to electronic health records (EHRs), along with growing concerns around data privacy in an increasingly hostile cyber threat environment, the computational burden of real-time analytics, and the interpretability and clinical validity of ML-driven decision support systems [3]. Addressing these challenges necessitates the development of scalable, workflow-compatible, and regulatory-compliant solutions, thereby motivating the focus of this review.

The healthcare sector is already experiencing an exponential growth in data volume, primarily attributed to the proliferation of EHR systems, wearable devices, and Internet of Things (IoT) enabled medical technologies. As of 2023, the global healthcare big data market was valued at approximately USD 67 billion and is projected to exceed USD 540 billion by 2035. Healthcare currently accounts for 30% of global data generation and is expected to grow at a compound annual growth rate (CAGR) of 36% by 2025, outpacing all other sectors. Hospitals alone generate an estimated 50 petabytes of data annually, with individual EHRs and medical imaging contributing roughly 80 megabytes of data per patient per year [4,5]. This magnitude of data highlights the urgent need for sophisticated ML and BDA systems to extract actionable insights effectively.

The emerging landscape of smart healthcare is shaped by a technological synergy that integrates blockchain, IoT, artificial intelligence (AI), and big data to enhance clinical outcomes, service personalization, and overall healthcare delivery. Wearable technologies have narrowed the gap in remote monitoring by providing continuous patient surveillance, enabling early detection, and timely intervention [6]. Predictive analytics are being employed to optimize resource allocation and forecast disease outbreaks, while AI-powered diagnostic tools improve the speed and accuracy of clinical decision-making. Telemedicine has emerged as a cost-effective and accessible mode of care, and blockchain offers a secure and interoperable framework for patient data exchange across institutions [1]. Furthermore, AI-driven automation is revolutionizing hospital operations, while personalized mobile applications that monitor vital signs and suggest lifestyle modifications are empowering patients to take a proactive role in their healthcare.

In the field of ML, algorithmic advancements have significantly enhanced the capacity to process complex healthcare data, detect patterns, and generate predictive models. Unstructured data types, such as medical imaging, genomic sequences, and clinical narratives, are being effectively managed through DL and natural language processing techniques, enabling improved real-time monitoring and operational efficiency [7]. These developments support the evolution of precision medicine and foster the transition towards a proactive and patient-centered healthcare ecosystem.

Previous Reviews and Research Landscape

Some researchers and authors have made significant contributions in healthcare to manage voluminous data, taking advantage of ML and BDA. In [8], the authors explored how the COVID-19 pandemic highlighted the limitations of traditional healthcare and emphasized the need for smart, connected wearables powered by IoT. These devices generate vast health data, posing challenges in management and decision-making. BDA

and ML have gained traction for improving IoTM applications. ML techniques were reviewed for big data analysis in healthcare, highlighting strengths, weaknesses, and research challenges to guide practitioners and policymakers. Semantic perception refers to the capacity of AI and ML systems to interpret and understand the meaning embedded within complex inputs, such as textual data, images, or sensor-generated signals, along with the contextual and relational dynamics among these data components [9]. In the context of healthcare, semantic perception enables systems to extract clinically relevant information from unstructured sources, such as physician notes or radiological images. This capability significantly enhances diagnostic accuracy and supports more informed clinical decision-making. In [10], the authors survey multimodal data-driven approaches in smart healthcare, covering decision-making processes, data fusion, and AI-driven transformations. This work presented semantic perception, entity association, and intelligent decision support, highlighting disease analysis, diagnosis, and privacy protection applications while addressing trends, challenges, and benefits in modern healthcare systems.

In [11], the authors presented ML and DL approaches in healthcare prediction, emphasizing their role in disease diagnosis from clinical data and images. It highlights AI-driven predictive analytics' impact on accuracy, patient outcomes, and industry transformation while addressing challenges and the need for reliable, efficient methods in healthcare predictive analysis. In [12], the authors discussed how advancements in communication technologies and the Internet of Medical Things (IoMT) have revolutionized AI-driven smart healthcare. Traditional AI relies on centralized data processing, posing scalability and privacy challenges. Federated learning (FL) enables decentralized AI training across multiple clients without sharing raw data. This survey explores FL's role in smart healthcare, covering recent developments, key applications, security considerations, and challenges. It also reviews FL-based healthcare projects and future research directions. In [13], the authors emphasized that detecting and classifying drug–drug interactions (DDIs) is crucial for preventing hospital-acquired conditions and improving smart healthcare. Given the vast medical literature, automated extraction using NLP techniques like information extraction (IE) and relationship extraction (RE) is essential. Traditional ML methods have limitations, while DL shows promise. They proposed SEV-DDI, a deep neural network model, enhancing precision in DDI classification and sentiment analysis to assess interaction severity for better clinical decision-making. In [14], the authors discussed how ML and DL have revolutionized healthcare, aiding in disease prediction, drug discovery, and medical imaging. They presented a comprehensive survey on ML and DL in smart healthcare, covering key advancements, applications, and future directions. We analyzed top journals and conferences, explored ML-DL integration, and highlighted research challenges and recommendations, providing a concise yet thorough overview of their impact on the healthcare industry. In [15], the authors discussed the role of BDA and FL in transforming healthcare by improving patient care and efficiency while addressing privacy concerns. FL enables collaborative data analysis without sharing sensitive information. This survey explored FL's role in healthcare, its challenges, and future prospects, including real-time monitoring and predictive modeling. It offers insights into integrating FL with BDA, ensuring security, data quality, and interoperability for researchers, policymakers, and healthcare professionals. In [16], the authors explore the role of mobile health (m-Health) and big data in transforming healthcare. With vast structured and unstructured data generated daily, this paper examined big data issues in m-Health, focusing on communication, sensors, and computing. It explores integrating BDA with m-Health and the role of AI. Balancing innovation with clinical complexities requires stakeholder collaboration, shaping the future of intelligent, connected healthcare systems.

The swift convergence of big data and ML with smart healthcare systems represents a paradigm shift in medical data processing, interpretation, and application to clinical decision-making. As healthcare settings become increasingly digitized and networked, the capability to extract actionable insights from voluminous, disparate data sources becomes paramount. This review is a timely and extensive synthesis of how ML methods revolutionize healthcare delivery, improve diagnostics, facilitate real-time monitoring, enhance treatment personalization, and maximize operational efficiency.

Additionally, although there is a plethora of separate literature on AI and big data, few research studies explore their convergence in applied healthcare settings comprehensively, especially from the perspectives of real-world applications, cutting-edge technologies, and security concerns. By bridging such a gap, the current review synthesizes a cohesive framework that reflects recent developments and addresses limitations, challenges, and avenues for future research. It is hoped to advise researchers, developers, healthcare professionals, and policymakers on leveraging intelligent technologies to construct more scalable, secure, and patient-focused healthcare systems.

2. Methodology

This study adopts a narrative literature review methodology to address the research problem situated within the domains of ML, BDA, and IoT-enabled healthcare systems. Unlike systematic reviews, which follow a predefined protocol with structured data extraction procedures, narrative reviews provide a flexible framework for synthesizing knowledge across diverse sources and interdisciplinary perspectives.

To ensure transparency in the review process, the study outlines the general inclusion criteria that guided the selection of literature. Sources primarily comprised peer-reviewed journal articles indexed in reputable databases, including Scopus, Web of Science, IEEE Xplore, PubMed, and Google Scholar. Priority was given to recent publications (2018–2025), written in English, and directly relevant to smart healthcare applications, particularly those involving the integration of ML and BDA. The selection process was based on a comprehensive literature search prioritizing peer-reviewed journal articles, high-impact conference proceedings, and other scholarly publications published within the last decade, thereby ensuring coverage of the most recent advancements in the field. Studies were included if they addressed smart healthcare systems, ML techniques, and diagnostic applications while also engaging with ethical implications. Particular emphasis was placed on works offering unique insights, detailed case studies, or discussions concerning implementation challenges. This methodological strategy enabled the integration of both established knowledge and emerging perspectives across interdisciplinary domains.

It should be noted that no formal review protocol was registered, nor was a PRISMA flowchart employed, as this work does not constitute a systematic or replicable search. Instead, a broad thematic synthesis is presented, supported by illustrative examples from the literature.

3. Smart Healthcare Systems

Hospitals and other healthcare institutions today have much larger healthcare surveillance systems, and many nations worldwide are now seriously concerned about portable healthcare monitoring systems using emerging technology. In addition to tracking basic vital signs like body temperature and pulse rate, the device provided smart healthcare by measuring hospital room variables, including CO, CO_2, and humidity levels. The application of information and communication technologies in the healthcare sector is referred to as "digital healthcare". From merely keeping electronic patient records and offering patient online portals, electronic health care has advanced to allow for greater comfort and

convenience in managing healthcare [17]. These days, it is known as connected health. The use of intelligent systems that support and enhance biological or medical processes and healthcare is accelerating due to recent advancements in IoT technology. In order to enable efficient patient monitoring even when patients are at home, smart health has developed from connected health, which uses wearable medical devices (such as heart rate monitors, glucose meters, smart watches, smart contact lenses, etc.) and IoT devices (such as implantable or ingestible sensors) in addition to traditional mobile devices (such as smartphones) [18]. By monitoring non-critical patients at home instead of in a hospital, smart healthcare could reduce the demand on medical resources like beds and physicians. It could be employed to facilitate access to healthcare for individuals residing in remote areas or enable older people to stay independently at home for longer. The core of the smart healthcare network is wireless technology, as illustrated in Figure 1. There are various wireless technologies like Wi-Fi, Bluetooth, 6LoWPAN, radio frequency identification (RFID), etc., that play a crucial role in the exchange of information between the numerous physical elements that make up the healthcare network [19].

Figure 1. Wireless technologies: a significant contributor to smart healthcare systems. (Source: created by the authors).

In addition to giving people more control over their health at all times, it may encourage access to healthcare services while reducing the strain on medical systems. It is anticipated that smart health would lower healthcare costs while offering timely care for a range of illnesses. Enhancing the productivity of medical infrastructures and biomedical systems is among contemporary society's most difficult undertakings. These include automated biomedical applications and monitoring of individuals in hospitals, appropriate drug-patient connections, and real-time monitoring of patients' physiological data for the early detection of clinical deterioration [20].

The most promising technologies for smart healthcare systems are UHF, RFID, WSN, and smart mobile. Battery-powered WMSs are becoming more common as low-power sensor, processing, and communication technologies progress. According to Business Insider, more than 56 million wearable sensors were shipped worldwide in 2015, with that figure expected to rise to 123 million by 2018. WSNs are self-organizing ad hoc networks of small, minimal devices (motes) that interact multi-hop to provide monitoring and control capabilities in critical applications such as industrial, defense, residential, automotive, and medical. Most WSN motes are battery-powered computing platforms with analogue and digital sensors. These devices capture, preserve, and transmit physiological data in a quiet, productive, and efficient manner. Wearable sensors and body-area networks

collect physiological signals such as ECG, SpO_2, and blood pressure. Projects like CodeBlue and MobiHealth demonstrate secure data transmission for mobile health monitoring. Integrating sensor and EMR data using ML for structured analysis remains challenging but essential for intelligent, autonomous healthcare systems [21–23].

The growing adoption of a patented, decentralized, scalable, and immutable ledger system, underpinned by blockchain technology, for the secure and verifiable exchange of healthcare data. Blockchain facilitates the creation of tamper-resistant, time-stamped, and access-controlled medical records, which are retrievable solely by authorized stakeholders within smart healthcare systems. This architecture not only reinforces patient privacy but also significantly enhances data interoperability among healthcare institutions.

Furthermore, blockchain-enabled smart contracts offer the potential to automate insurance claims processing, streamline pharmaceutical supply chain management, and enforce patient consent protocols for data sharing. Collectively, these capabilities contribute to the establishment of a secure, transparent, and patient-centric digital health ecosystem. Table 1 summarizes recent advances in AI, IoT, and cybersecurity, showcasing smart healthcare innovations and future integration opportunities.

Table 1. Summary of recent smart healthcare systems research articles.

Focus Area	Proposed Approach	Key Findings	Challenges Addressed	Future Prospects	Study Evaluation	Ref.
Human Activity Recognition (HAR) Using Sensor Data	LSTM-based HAR model Attention and squeeze-and-excitation blocks	99% accuracy in activity recognition; improved feature extraction lower computational complexity	Variability in sensor data; imbalanced Datasets; high computational cost	Enhancing real-world applications of HAR by improving model adaptability and efficiency	Well-defined model; limited to specific dataset; 99% on public HAR dataset	[24]
IoT-based Clinical Decision Support System	Cloud-based C-IoT model with ANN and lightweight encryption for secure health monitoring	91% diagnostic accuracy; health monitoring; enhanced data security	Data security; accuracy	Expanding the model for Society 5.0 by integrating AI-driven secure diagnostics	Small-scale deployment; lacks generalizability; moderate reproducibility	[25]
Intrusion Detection in Smart Healthcare IoT Systems	IADCL framework using feature selection (IRKO), AConBN classifier, and SA-HHO optimization	High accuracy in cyberattack detection using public datasets like CIC-IDS 2017 and 2018	Lack of robust security in resource-constrained IoTM devices	Improving security mechanisms for real-time IoT-based healthcare systems	Benchmark datasets used (CIC-IDS); good methodology; reproducible	[26]
AI-Enabled Smart Acne Diagnosis	Convolutional neural network (CNN)-based cloud-connected IoT device for facial acne severity assessment	AI-driven real-time acne tracking with geographic adaptability, bridging e-healthcare access gaps	Delayed/inaccurate acne diagnosis and lack of remote dermatological consultation	Expanding AI-based dermatology solutions for various skin conditions	Small sample; limited clinical validation; strong conceptual design	[27]

Table 1. *Cont.*

Focus Area	Proposed Approach	Key Findings	Challenges Addressed	Future Prospects	Study Evaluation	Ref.
Privacy-Preserving Authentication in IoMT Healthcare	Blockchain-based double anonymity strategy with cross-hospital authentication	Enhanced privacy, decentralization, and a 23–87% reduction in computational costs	Privacy breaches, untraceability, single point of failure in authentication	Implementing blockchain for broader healthcare interoperability	High potential; lacks empirical multi-hospital deployment proof	[28]
Unconstrained Health Monitoring via BCG	Smart wireless flexible sensing system for heart and respiration monitoring	High-sensitivity flexible sensor capturing subtle physiological signals	Discomfort and movement limitations in traditional long-term monitoring	Developing advanced flexible sensors for more wearable health monitoring applications	Strong sensing performance; tested in controlled setup	[29]
Nursing Professional Values and Job Satisfaction	Survey-based analysis using SEM-PLS modeling	Activism and justice are the key influencers of job satisfaction among Vietnamese nurses	Limited professional development and high patient-to-nurse ratios	Studying nursing satisfaction factors across diverse healthcare environments	Statistical rigor; limited to Vietnam; low generalizability	[30]
Secure IoT Communication in Smart Healthcare	Three-factor lightweight mutual authentication scheme with elliptic curve cryptography	Improved security with minimal computational cost, meeting 15 security criteria	Security vulnerabilities in resource-constrained IoT healthcare environments	Enhancing lightweight cryptographic protocols for broader IoT applications	Good cryptographic model; theoretical validation; real-world deployment pending	[31]

3.1. Smart Healthcare Framework

Innovation is key to improving the quality of life across all age groups. With an aging population and advancements in AI and ML, the IoT offers a powerful means to monitor and deliver personalized healthcare. IoT devices collect and transmit patient data to caregivers and physicians, enabling timely and effective medical interventions [32–34]. With IoT-based healthcare equipment, patients are no longer confined to a single location or dependent on bulky medical apparatus. This innovation allows patients to move freely while being continuously and non-intrusively monitored. This also results in reduced cost monitoring. The following are the categorizations of IoT-based healthcare equipment/devices (shown in Figure 2) along with various aspects of the smart healthcare systems (Figure 3).

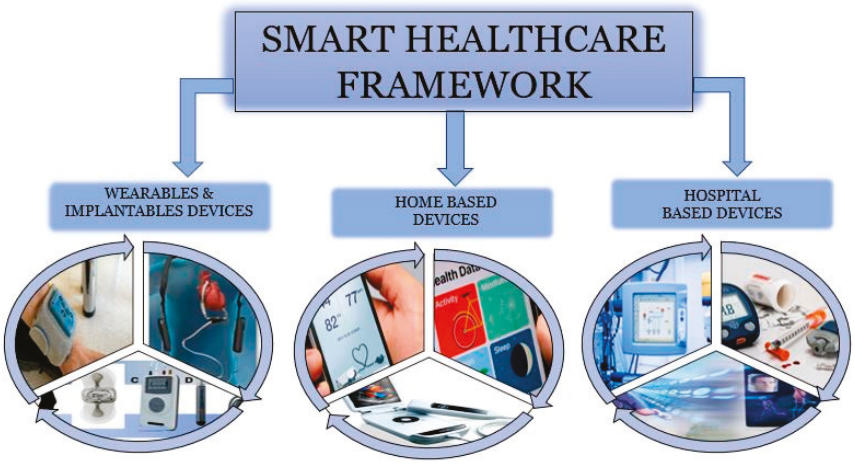


Figure 2. IoT-based smart healthcare equipment. (Source: created by the authors).

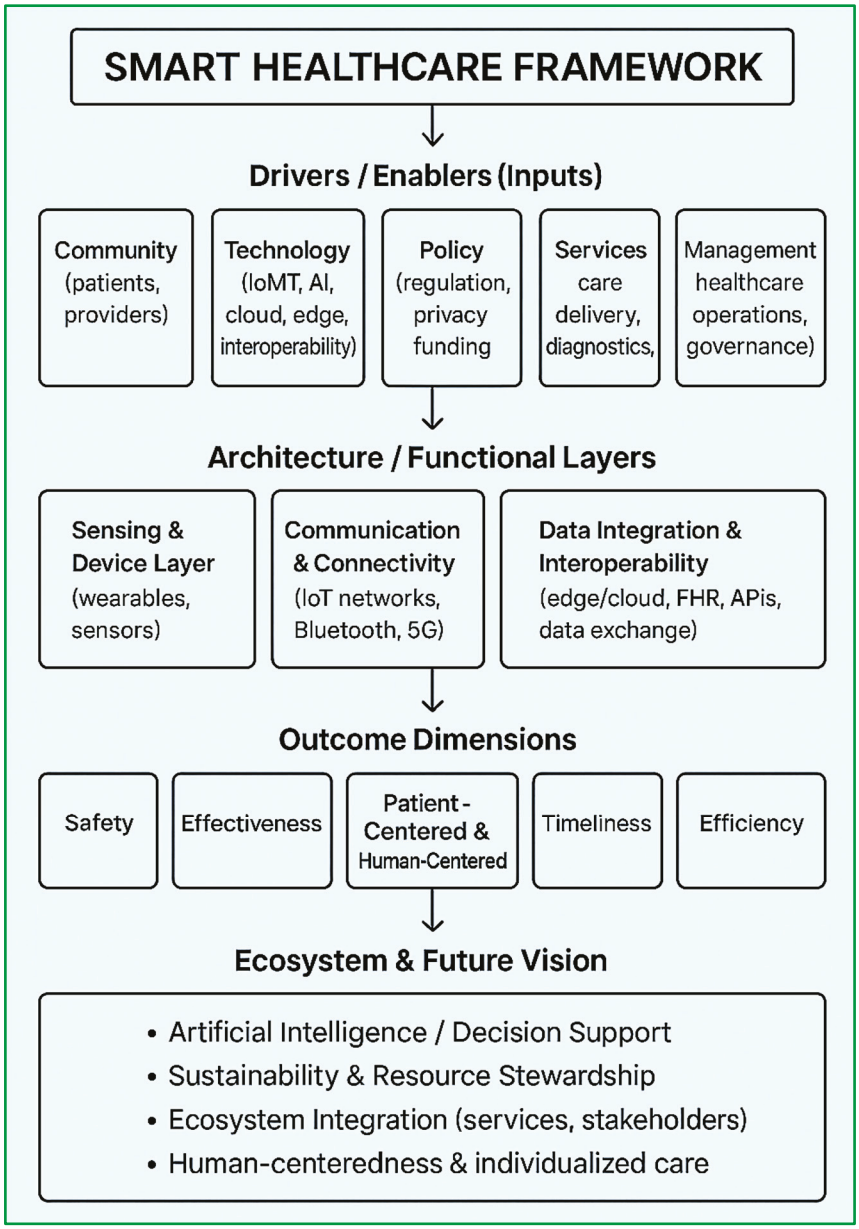


Figure 3. Smart healthcare framework dimensions. (Source: created by the authors).

3.1.1. Wearables IoT Devices

Wearable equipment is any small, wearable gadget that displays information to consumers and allows them to interact with it, whether that be via voice commands or manual input. It is frequently sold as a "casual clothing accessory". Wearable technologies have steadily evolved as additions (e.g., bands, rings), smart clothes, body implants, and body penetrations (e.g., insulin pumps, artificial hearts). The development of multiple quasi-wireless sensors for measuring human physical parameters has expanded the potential for smart sensor network-based medical systems, wherein the sensors are wearable or portable. Innovative tools, such as surgically implanted and Wearable Healthcare Products, have permitted the interpretation and processing of physiologic information from anybody, anywhere, at any time [35]. Wearable technological advances include wireless transceiver modalities such as Bluetooth, Zigbee, infrared, broadcasting identification, Wireless internet, and relatively close communication. Such a platform enables wearables to link to other electronic objects (smartphones) for remote diagnosis and follow-up, resulting in improved quality of treatment. Surgically implanted and Wearable Biomedical Gadgets, which enable quasi-avoidance, accurate intervention, and ongoing medication of physical illnesses, are projected as critical pillars of modern healthcare. Wearable electronics have characteristics that concentrate on lower power consumption, compact size, manageability, and comfort of wear and use. Advances in wearable technology and significant analytics techniques make it possible to capture and modify medical application information in a genuine moment, which can strengthen quality healthcare by lowering injury rates, enhancing doctor–patient conversation, and discovering previously unseen monitoring and sensory features [36]. For example:

- A biker employing a vital sign system, specifically a wristband, is involved in an accident. The body sensor network detects the fall, which transmits an alarm to the city's infrastructure. The technology replies by evaluating road congestion and dispatching ambulances through the most direct route. Furthermore, municipal traffic signals are automatically modified to shorten the time the ambulances take to reach the rider.
- Fitbit Inspire, a wearable fitness band, records all-day activity, relaxation, calorie counting, and breathing rate, among other things, and shows the results on an Android smartphone while storing the cloud to help the user achieve his or her chosen fitness goal. Google Fit, Fitbit Coach, Nike Training Club, Runtastic, and other mobile phone overall fitness applications track wellness and exercise regimens for individual requirements.

Table 2 summarizes recent advancements in smart healthcare, IoT security, and emerging technologies. It presents cutting-edge advancements in smart healthcare, IoT security, wearable technology, and military IoT applications. The research demonstrates how AI, ML, blockchain, edge computing, and next-gen networks are being integrated to improve security, efficiency, and real-time processing. Despite these advancements, challenges such as scalability, privacy risks, computational efficiency, and security vulnerabilities remain. Future research will likely focus on developing more resilient security frameworks, improving AI-driven analytics, and optimizing resource allocation in IoT-based systems. These findings emphasize the growing intersection of AI, IoT, and healthcare, shaping a future where secure, efficient, and intelligent systems will drive better healthcare outcomes, safer military operations, and more immersive Metaverse experiences.

Table 2. Comprehensive summary of recent research in smart healthcare, IoT security, and emerging technologies.

Focus Area	Proposed Approach	Key Findings	Challenges Addressed	Future Prospects	Study Evaluation	Ref.
Big Data and Gene Therapy in Smart Healthcare	Wearables and tracking devices; early health risk prediction BDA for gene therapy	Early risk identification; before genetic data is available; smart data processing for cost-effective healthcare	Integration of multi-source data; affordability of advanced healthcare analytics	Enhancing predictive healthcare models; expanding data-driven gene therapy applications	Small pilot studies; scalability untested	[37]
Security and Privacy in Smart Healthcare Systems	Comprehensive review of SHS security challenges and proposed solutions	Identifies key risks Cyber threats, privacy issues, and attack vulnerabilities in IoMT	Scalability; complexity; securing inter-connected smart healthcare devices	Developing robust security frameworks for safeguarding IoMT-based smart healthcare	Synthesis without empirical validation	[38]
Smart Textiles for Healthcare and IoT Integration	Van der Waals (vdW) force-based 2D functional material integration with textiles	Preserves textile flexibility Adding intelligence; applications (healthcare, human–machine interaction)	Scalability; performance stability Safety concerns in commercial applications	Expansion into fully connected IoT-integrated wearable healthcare systems	Lab prototypes; real-world use pending	[39]
5G Security for IoT and Wearables	Security en-hancements in 5G-AKA au-thentication protocol	Improved user authentication and reduced exposure to security vulnerabilities	Lack of mutual au-thentication; privacy risks in 5G commu-nication	Developing more secure and resilient authentication mechanisms for next-gen networks	Theoretical analysis; real-world deployment needed	[40]
Self-Powered Wearable Sensors	Triboelectric sensors with AI-based LSTM for sign language recognition	Achieved 96.15% recognition accuracy for sign language patterns	Eliminating external power needs while maintaining accuracy	Expanding self-powered sensors for diverse wearable applications	Limited dataset; controlled conditions	[41]
Secure IoT-based Health Monitoring	Cloud and IoT-based secure health monitoring system	Enhanced security and privacy for wearable healthcare technologies	Data privacy concerns and secure remote patient tracking	Improving security standards for cloud-integrated health monitoring	Limited pilot; reproducibil-ity unclear	[42]
Ensemble Learning for IoT Intrusion Detection	Bagging-based IDS integrating DNN and CNN for IoT security	Improved threat detection using the Edge-IIoTset dataset	Balancing IDS accuracy with computa-tional efficiency	Enhancing real-time IoT security using AI-driven ensemble models	Good on benchmark dataset; edge deployment pending	[43]

3.1.2. Home-Based IoT Devices

In-home remote monitoring apps enabled by the IoT are among the primary smartphone health (digital health) apps that allow effective preventive digital health treatments. Mobile-based IoT is a collection of software applications for healthcare solutions housed on digital devices that are based on sensors, computation, connectivity, and data and analytics. Home-based M-health is a mixture of telecommunications, imagery, detection, and human-machine interaction techniques that are used to diagnose and provide follow-up care without interfering with their daily lives. Telemonitoring applications use sensors and a residential hub to connect the patient to doctors and domain experts via cloud-based data offerings. Telemedicine provides the patient with assurance that their ailments (e.g., pulse rate, blood pressure, Blood oxygen saturation levels, and daytime sleepiness) are being tracked, and alarms may be created to notify their healthcare providers in actual environments. A wireless mesh network-based Remote health monitoring application framework incorporates wearable tech to help control various features of mobile computation devices, such as smartphones, smart iPads, and fitness trackers, among others [44,45]. The Open Smartphone Health (mHealth) initiative aims to prevent and treat serious illnesses regularly by leveraging data from medical systems on personal devices, such as smartphones. There seems to be a rapid growth in the availability of mHealth smartphone apps aimed at wireless connectivity and detection of various diseases, assisting patients in better managing their medical problems, and facilitating a successful life. The utilization of mobile gadgets contributes to more effective work performance. S-Health enables the remote monitoring of patients, encourages interaction between experts, household members, and individuals with health issues, and clarifies primary healthcare in three key ways: it provides access to user-oriented and personalized services and insights. Participant voice or visual home patient assistance via smartphones; IoT disaster preparedness; ambient-assisted living technology can deliver everyday situations, patient vital signs, and geospatial data; virtualized computerized medical monitoring, characteristic registration and categorization approaches; sensing smart gloves equipment for handicapped and developmentally disabled persons [46]. Social networking sites such as Facebook, Instagram, YouTube, and Google, which include content applications, journals, and weblogs, were made possible by Web 2.0. It presents a weight reduction app accessible via an Online social network. People discuss their objectives and achievements on the website, which provides anonymized and individualized assistance. Surveys are used to improve the platform's offerings. By 2012, there would have been over 13,000 safety software applications on Apple's iPhone. These programmers allow sufferers to capture and follow their critical physiological responses digitally regularly, resulting in satisfying customer consequences. Unless sufferers mention it, it is challenging for doctors to determine if their patients are taking their medications appropriately during a typical visit. This might have a negative impact on the patient's recuperation. As a result, technologies for monitoring patients' medication at home have evolved. iHomeHealth-IoT was a smart medicine box and smart pharmaceutical packaging. The smart box notifies the patient when it is time to take their medication [47]. The box also sends a text message to the doctor. For the aged patients, home is at the centre of the concept. It empowers patients via illness prevention, treatment, and consciousness of chronic medical conditions [48]. For example:

- Glycaemic monitoring apps supervise the patient's glucose level. They can undertake a variety of actions relying just on the app, such as measuring the information in the device and/or the infrastructure as a service, displaying on connected phones, notifying caregivers, advising the general practitioner, and tracking to the health insurer, among other things.

- Electrocardiogram and cardiovascular monitoring applications use ECG, heart rate readings, and basic pattern recognition to follow the heart's electrical activity. Depending on the app's circumstances, they can forecast some fundamental problems, such as arrhythmias and cardiac ischemia, and notify people, clinicians, and caretakers.

Smart supervision, quick access to actual figures, and the transmission of on-demand judgments have laid the groundwork for a reliable integrated IoTM system that can guarantee more correct diagnoses, credible proof of treatment, fewer hospitalizations, and the best possible resource utilization regarding time, energy, and cost savings.

3.1.3. Hospital-Based IoT Devices

Signaling is a technique employed in practically every element of surgery center care, from the most basic digital thermometers to complicated ultrasonically operating rooms. A sensor is a component that turns a physical measurement into a signal that an observer or equipment can analyze. There seems to be an interest in developing sophisticated medical equipment (e.g., heart rate devices, fructose meters, temperature monitoring, muscle mass scales, etc.). The LifeShirt Equipment is an example of a multi-sensor monitoring method for gathering, evaluating, and presenting a surgeon's health data. It collects continuing patient data concerning data screenshots, which are typically acquired during the patient's frequent appointments to the clinic or hospital [49,50].

The Smart-Health framework is intended to aid in the process of customizing and striving to improve the patient care prototype in a variety of ways, including the following:

- Concentrating on sensor technologies for health checks and appraisal techniques in the home and neighborhood atmospheres to minimize direct stress on healthcare environments and convert it to a digital dissemination of knowledge.
- Transforming the pharmaceutical procedure from a reaction to a proactive risk management strategy can dramatically reduce hospitalization costs for acute occurrences.
- Enhancing the personalized recommendations of the health and care system so that private citizens can supervise and recognize their potential risk characterization, preventive medicine intervention, and diagnosis, allowing patients to stay independent while being cared for, which has a strong positive influence on their psychological makeup and, as a result, their health status.
- Facilitating improved clinical maintenance and permitting the health service to prioritize sick people in the greatest need properly.
- Assisting self-care clinical techniques to observe health status and other varied metrics, in which these data are exchanged with a physician to conduct a diagnosis, either in person or by teleconsultation. Likewise, for mild conditions such as influenza, diagnosis can sometimes be computerized.
- Maximizing moment-in-time testing by lowering diagnostic time by eliminating the need to transfer specimens elsewhere to be analyzed. Computerized monitoring using blood pressure cuffs and electronic thermometers, for example, can assist the doctor in reviewing a medical history as statistics have been taken.

Table 3 summarizes a comparison of all three types, i.e., wearable, home-based, and hospital-stationed smart devices, from ML and big data aspects.

Table 3. Wearable, home-based, and hospital-based healthcare devices from the perspectives of ML and big data.

Aspect	Wearable Devices	Home-Based Devices	Hospital-Based Devices
Data Collection	Continuous real-time monitoring (e.g., heart rate, steps, SpO_2)	Periodic monitoring (e.g., BP monitors, glucose meters)	High-frequency; high-resolution medical data (e.g., MRI, CT scans, ECG)
Data Volume	High but fragmented	Moderate to high	Very high (large datasets per patient)
Data Variety	Limited (mainly physiological data)	Moderate (includes environmental and physiological data)	Extensive (medical imaging, lab tests, patient history)
Data Velocity	Real-time streaming	Near real-time to scheduled readings	Batch processing and real-time for critical care
Privacy and Security	Risk of data breaches via cloud or mobile apps	Moderate security concerns (home network vulnerabilities)	High-security standards (hospital IT infrastructure)
ML Applications	Activity recognition; anomaly detection; predictive analytics	Disease monitoring; early warning systems	Diagnosis; precision medicine; treatment planning
ML Model Complexity	Lightweight models (on-device processing)	Moderate complexity (edge/cloud computing)	High complexity (DL AI-driven diagnostics)
Big Data Challenges	Data fragmentation, interoperability issues	Data inconsistency; integration challenges	High computational demands; need for scalable storage
Computational Resources	Low (wearables have limited processing power)	Moderate (some devices leverage cloud/edge computing)	High (dedicated servers, GPUs, Cloud computing)
Integration with Healthcare Systems	Limited (mostly user-driven insights)	Moderate (telemedicine and EHR integration)	Full integration with EHRs and clinical workflows

4. Machine Learning and Big Data Applications in Smart Healthcare Diagnostics

4.1. Overview of Machine Learning

ML, a practical subset of AI, enables systems to learn from data and make predictions or decisions without the need for explicit programming. These algorithms identify patterns within historical datasets and generalize this learned knowledge to unseen data, rendering them particularly suitable for healthcare contexts, which frequently involve complex, high-dimensional information [51].

ML methodologies can be broadly classified into three primary categories: supervised learning, unsupervised learning, and reinforcement learning, as illustrated in Figure 4.

Supervised learning involves training models on labeled datasets. This category encompasses two major subtypes [52]:

Classification, which aims to predict categorical labels (e.g., presence or absence of disease) [53], commonly employs algorithms such as decision trees, support vector machines (SVMs), and logistic regression.

Regression, which is used for predicting continuous outcomes (e.g., blood glucose levels), utilizes techniques like linear regression and neural networks.

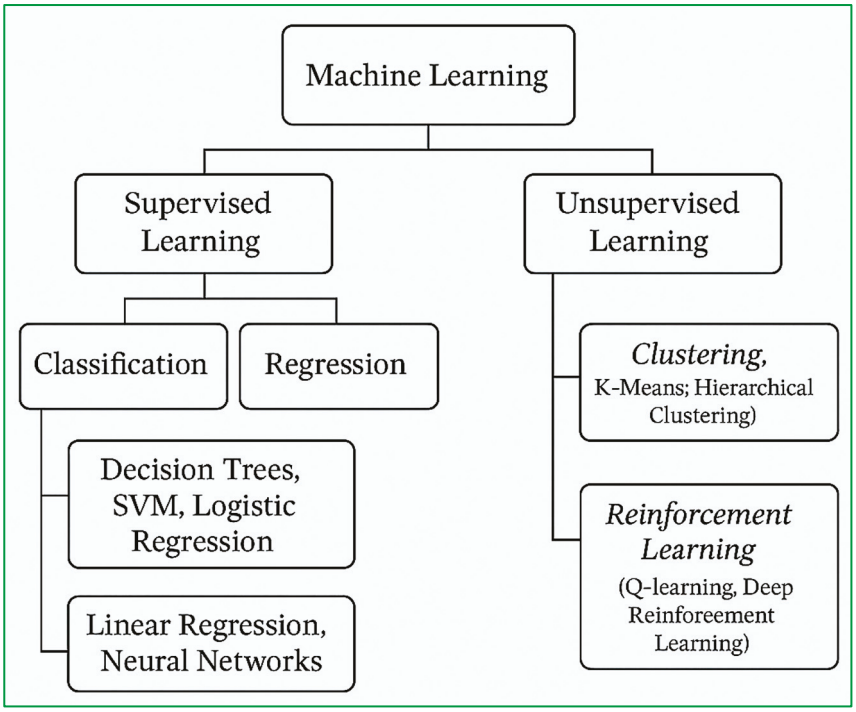


Figure 4. Taxonomy of machine learning methods. (Source: created by the authors).

Unsupervised learning deals with unlabeled data, aiming to uncover hidden structures or intrinsic patterns. Clustering methods such as K-means and hierarchical clustering group similar data points based on inherent similarities without prior knowledge of class labels [54].

Reinforcement learning refers to the process of learning optimal actions within a dynamic environment to maximize cumulative rewards or minimize penalties. Algorithms such as Q-learning and deep reinforcement learning are increasingly being investigated for their potential in adaptive, real-time healthcare decision-making [55]. Figure 5 depicts the taxonomy of machine learning methods categorized by learning approach, with representative algorithms for each.

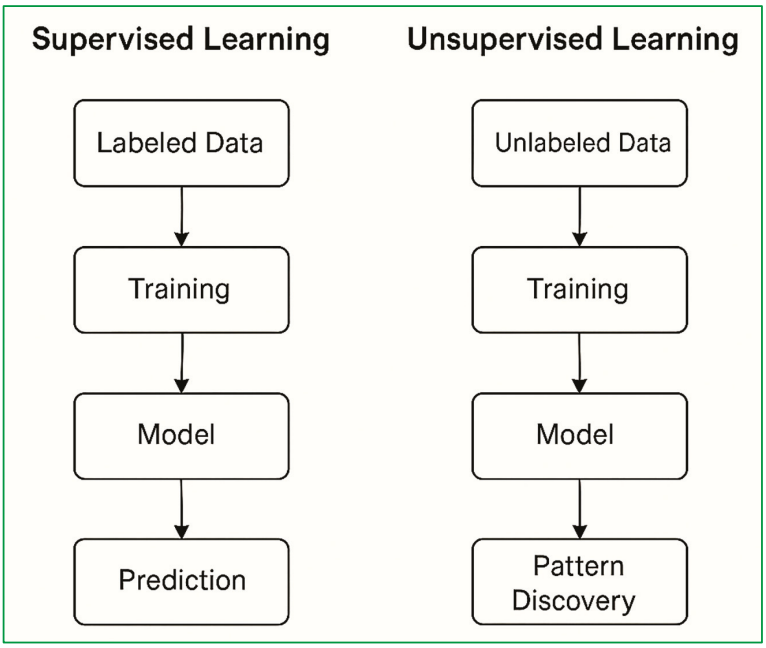


Figure 5. Workflow of supervised vs. unsupervised machine learning. (Source: created by the authors).

Supervised and unsupervised learning also diverge significantly in terms of workflow, as outlined in Figure 5. In supervised learning, the pipeline follows a clear progression: labeled data, model training, and prediction generation [54].

In unsupervised learning, the flow is: unlabeled data, model training, and pattern discovery.

The distinction in workflow between these two paradigms, as visualized in Figure 5, highlights their respective data requirements and end-goals. This foundational understanding of ML methodologies serves as a conceptual precursor for the application of these approaches in smart healthcare systems [56].

4.2. Big Data in Smart Healthcare

The advent of big data in healthcare, powered by technologies such as IoT, cloud computing, and data mining, has enabled real-time patient monitoring, predictive diagnostics, and personalized care delivery. These innovations are transforming traditional healthcare into a proactive, data-driven ecosystem. The main challenge facing healthcare systems is finding ways to detect, collect, analyze, and manage data to assist patients in understanding new diseases and medications, predict outcomes early, and make decisions in real time. The healthcare business generates a substantial amount of data, including patient records, drug information, experimental results, consultations, and real-time data that is both simultaneous and immediate. This data must be processed and handled. Therefore, big data encompasses a range of methods, tools, and strategies for data collection (structured or unstructured), information extraction, organization, analysis, management, and, most importantly, utilization for future purposes. Automating medical data changes the dimensionality of data while also increasing the size of data and the importance of data analytics. In a digitalized healthcare context, healthcare data refers to the vast data that is harder to manage with traditional database management systems, including health centers, fitness trackers, and social platforms. Big data assists in digitizing and integrating information to be user-friendly and easily retrieved. Nevertheless, big data may help us discover hidden patterns, correlations, and cross-correlations in the data it accumulates. BDA is a tool for analyzing a massive amount of digital data associated with patient healthcare and well-being that is exceedingly varied and challenging to evaluate using standard software or hardware [57].

Healthcare BDA has especially come to the forefront as a viable technique for addressing difficulties in a variety of healthcare domains (Figure 6). The primary goal of governments throughout the world is to enhance healthcare facilities while lessening medical expenditures. Big data offers doctors, biochemists, and patient care specialists a fantastic chance to make computational decisions that will eventually strengthen patient care. Big data in healthcare has the potential to transform the medical industry by allowing for early illness identification utilizing appropriate analytical tools and methodologies, as well as by combining and interpreting health-related data in a precise way.

The development and implementation of ML technologies facilitate the efficient utilization of Big Data in Healthcare. Techniques based on ML are sometimes used alternately with methodologies based on AI. The introduction of artificial intelligence and data analytics, coupled with the increasing interconnection opportunities of medical devices, i.e., the IoT, has created a significant opportunity for microbiologists to predict and allocate accurate and effective resources to the appropriate patients. Patient information includes recorded signals such as ECGs, pictures, and videos. Healthcare professionals have only just begun to transfer such healthcare data into EHRs. Attempts are being made to digitize patient records from pre-EHR period entries to support standardization by converting static photos into machine-readable text [58].

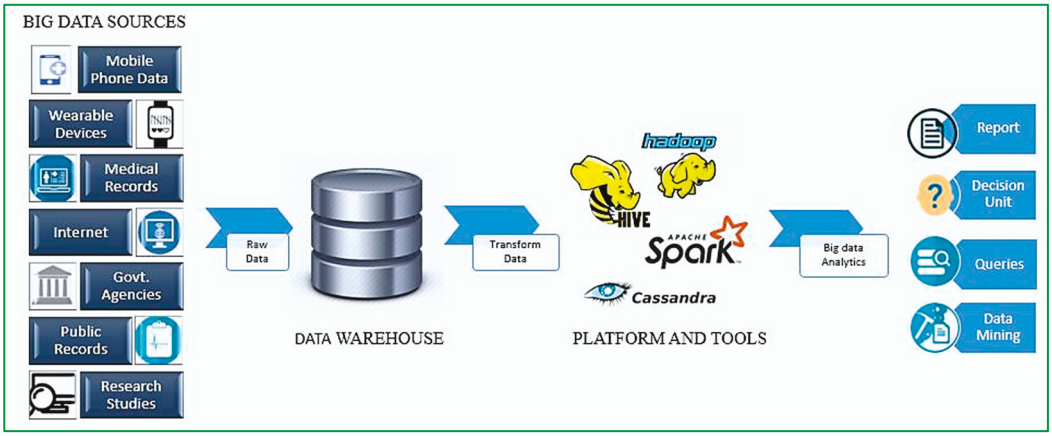


Figure 6. Data analytics framework for big health data. (Source: created by the authors).

4.3. Convergence of Machine Learning and Big Data

The convergence of ML and BDA in healthcare has resulted in tremendous improvements in diagnostic performance, disease forecasting, patient tracking, and decision-making. This convergence allows intelligent healthcare systems to transcend conventional diagnostic processes by tapping into enormous amounts of structured and unstructured data, such as EHRs, sensor readings, medical images, genomic data, and operational hospital data. The capability to analyze such data in virtual real time accelerates early detection, risk stratification, customized treatment planning, and operational effectiveness. As illustrated in Figure 7, the diagnostic insight pipeline combines multimodal health information, from wearables and EHRs to genomic data and medical images, into a machine learning pipeline that facilitates predictive diagnostics, risk stratification, and individualized decision-making. A feedback mechanism allows for iterative model refinement through clinical outcomes, increasing trust, accuracy, and interpretability.

Recent work has illustrated how various ML methods, covering ensemble classifiers and support vector machines to DL models and federated learning, can be efficiently used in multiple clinical fields [59,60]. These are cardiology, ophthalmology, endocrinology, nephrology, hepatology, maternal and geriatric care, intensive care, and hospital-wide operations. Big data sources such as real-world evidence, wearable sensors, IoT-enabled devices, and cross-institutional EHRs offer the size and complexity to enable these models to execute solid diagnostic tasks. Yet, with these opportunities come challenges in data privacy, interpretability of algorithms, generalizability of models, and clinical integration.

Table 4 presents some of the most important recent contributions in the area, highlighting the range of diagnostic applications enabled by ML and big data. The examples vary from predictive models for diabetes complications based on XGBoost to liver fibrosis detection through hybrid classifiers, and smart hospital transformation programs utilizing ERP-integrated AI systems. Notably, several studies point to privacy-preserving designs and non-invasive diagnostic substitutes, respectively, that are essential in scalable clinical rollout.

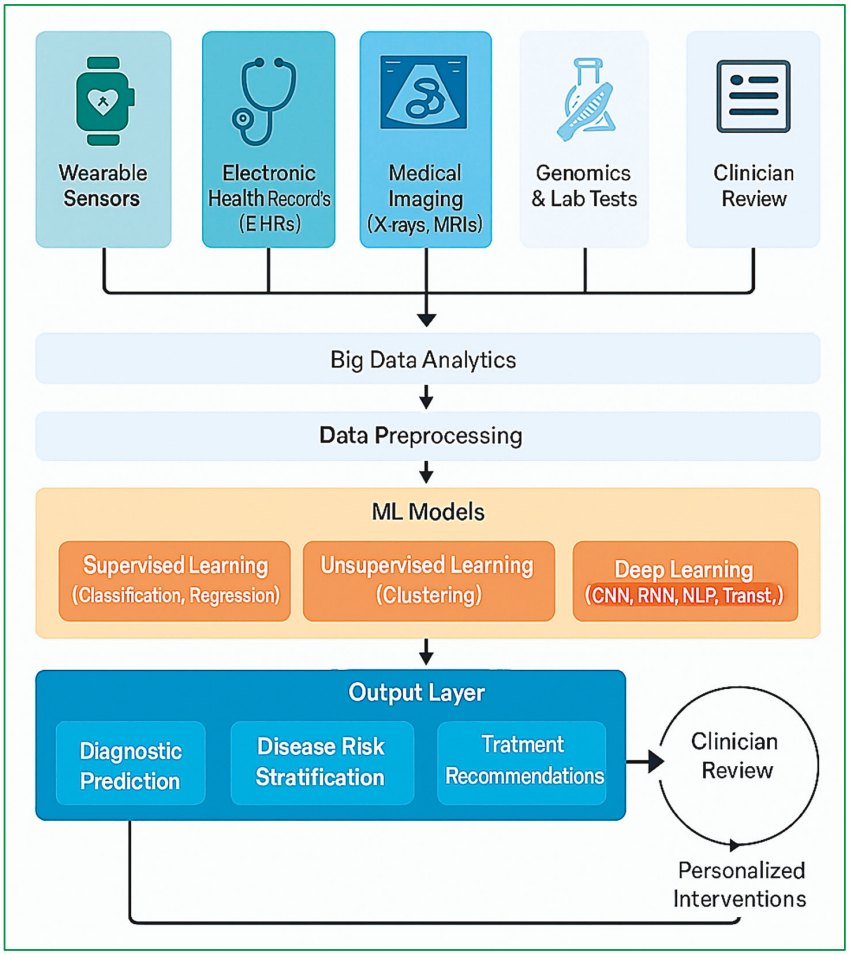


Figure 7. Machine learning diagnostic insight pipeline: from data acquisition to clinical decisions. (Source: created by the authors).

Table 4. Summary of key studies on ML and big data applications in smart healthcare diagnostics.

Study Focus/ Design	Application Domain/ Dataset Size	ML Technique/ Model	Key Contribution/ Reproducibility	Diagnostic Relevance	Ref.
Patient similarity evaluation framework	General healthcare/large-scale patient data	Adaptive semi-supervised recursive tree partitioning (ART)	Efficient patient similarity indexing and retrieval	Supports prognosis, risk assessment, and comparative effectiveness	[61]
AI in operational delivery improvement	Cardiovascular care/healthcare systems	General AI platforms and ML algorithms	Integration of AI in cardiac care operations	Supports diagnosis and risk stratification	[62]
Complaint prediction in diagnostic systems	In-vitro diagnostics/QC data over 90 days	Decision trees, adaptive boosting	Prediction of customer complaints using QC data	Indirect diagnostic support through QC performance monitoring	[63]
Emotion-aware postpartum depression detection	Maternal health/biomedical and sociodemographic data	Ensemble classifiers	Predicts postpartum depression risk	Enables early detection and intervention	[64]

Table 4. *Cont.*

Study Focus/ Design	Application Domain/ Dataset Size	ML Technique/ Model	Key Contribution/ Reproducibility	Diagnostic Relevance	Ref.
Automated retinal image labeling	Ophthalmology/5000 SEED + public datasets	DL + rule-based classifier	RetiSort for high-accuracy retinal photo sorting	Aids diagnostic preprocessing	[65]
DL in IoT healthcare	Healthcare IoT/44 SLR papers	DL frameworks	Review of DL in healthcare IoT	General healthcare diagnostics	[66]
Fall detection from real-life data	Elderly care/400 days of real-life data	ML models	High-sensitivity fall detection system	Automatic fall detection and alert	[67]
ML in intensive care systems	ICU/general data	ML algorithms	Review of ML clinical decision support	Clinical decision support in ICUs	[68]
Prediction of diabetes complications	Diabetes/147,664 patients	XGBoost	Predicts short and long-term complications	Improves prognosis and care quality	[69]
Non-invasive liver fibrosis diagnosis	Hepatitis C/SLBs data	EMD + ANN-J48	Hybrid intelligent classifier	Non-invasive diagnostic tool	[70]
Federated learning in healthcare	Healthcare/multi-center EHRs	Federated learning	Systematic study of FL in healthcare	Privacy-preserving diagnosis modeling	[71]
Elderly activity tracking	Elderly care/sensor data	HDCO + deep ensemble learning	High-accuracy activity recognition	Behavior monitoring for healthcare	[72]
Predictive maintenance of medical devices	Healthcare equipment/SLR data	ML + big data	Review of predictive maintenance techniques	Ensures device reliability	[73]
IDH prediction in dialysis	Dialysis/patient records	DL	Survey on ML for IDH	Preemptive diagnosis and prevention	[74]
HF phenotyping via AI	Heart failure/VA EHRs (20,000 pts)	NLP + ML (SVM)	Efficient HF identification using EHR	Improved diagnosis in HF registries	[75]
Smart hospital optimization	Hospital care/internal data	NLP + ML + BI	Optimized care, diagnostics, and cost	Enhances diagnostic processes	[76]
AI in rare disease transplant support	Rare diseases/Polish national databases	Big data + AI tools	Policy analysis and recommendations	Supports early diagnosis and treatment	[77]
ML in women's health	Women's health/review-based	ML + big data	Perspective on opportunities and biases	Personalized predictive healthcare	[78]
ML for endometriosis detection	Endometriosis/Lucy app (10,000 participants)	Machine learning	Real-world data analysis for earlier detection	Early diagnosis and personalized recommendations	[79]
AI in functional urology	Functional urology/UDS datasets	AI systems	Proposal and discussion on AI use	Improves diagnosis and personalized therapy	[80]

5. Machine Learning for Voluminous Healthcare Data: Real-World Case Studies

Smart healthcare systems increasingly rely on ML techniques, which broadly fall into three primary categories: supervised, unsupervised, and reinforcement learning. These techniques enable the analysis of large-scale structured and unstructured healthcare data, facilitating real-time decision-making, anomaly detection, and the delivery of personalized care.

Supervised learning models are predominantly used for tasks such as disease classification and outcome prediction [81,82]. Unsupervised methods are effective for clustering patients based on symptomatology or behavioral patterns, while reinforcement learning is particularly suited for optimizing treatment pathways in dynamic clinical environments.

The implementation of these methods is particularly crucial in domains such as remote patient monitoring, diagnostic decision support, and clinical workflow automation. ML offers an indispensable toolkit for advanced analytics, empowering healthcare stakeholders to extract insights from vast and heterogeneous datasets, insights that are unattainable through conventional statistical methods alone.

Applications of ML in healthcare span various data-intensive tasks, including predictive analytics, anomaly detection, and clinical decision support, as illustrated in Figure 8. A comprehensive overview of ML applications and their associated Big Data challenges, from both clinical and operational perspectives, is provided in Table 5.

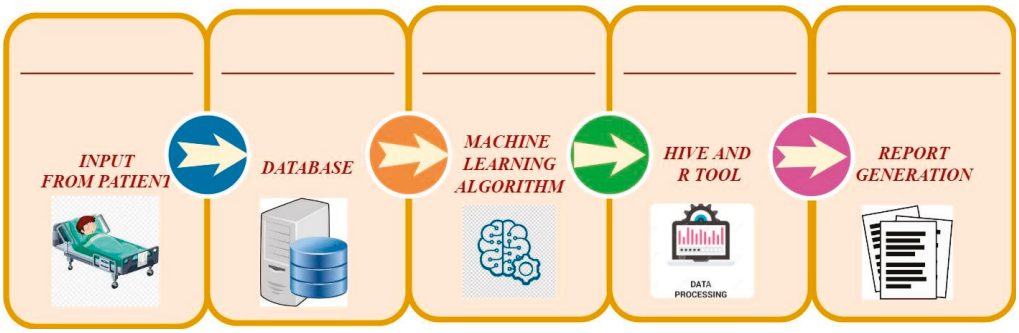


Figure 8. ML for data-intensive healthcare applications. (Source: created by the authors).

Table 5. ML in various healthcare applications and associated data challenges.

Healthcare Application	Role of Machine Learning	Big Data Challenges
Disease Diagnosis	Detects diseases like cancer, diabetes, and neurological disorders using medical imaging and patient data.	Data heterogeneity, interoperability, and privacy concerns.
Predictive Analytics	Forecasts disease outbreaks, patient deterioration, and risk factors.	High-dimensional data processing and real-time analytics.
Medical Imaging	Enhances anomaly detection in X-rays, MRIs, and CT scans using DL.	Large file sizes, data annotation, and storage limitations.
Drug Discovery and Development	Predicts molecular interactions, accelerates drug discovery, and optimizes clinical trials.	Data silos, computational complexity, and regulatory constraints.
Personalized Medicine	Tailors treatments based on genetics, lifestyle, and medical history.	Data integration from diverse sources, ethical concerns.
Remote Patient Monitoring	Analyzes wearable and home-based device data for early intervention.	Real-time processing, connectivity issues, and security risks.

Table 5. *Cont.*

Healthcare Application	Role of Machine Learning	Big Data Challenges
EHR Management	Automates data extraction, summarization, and decision support.	Data inconsistency, duplication, and access control.
Clinical Decision Support	Assists in treatment recommendations using predictive models.	Data accuracy, bias in training data, and model interpretability.
Healthcare Chatbots and Virtual Assistants	Provides symptom checking, appointment scheduling, and medical advice.	NLP limitations, contextual understanding, and data privacy.
Surgical Assistance	Aids robotic surgeries with real-time guidance and precision enhancement.	Sensor data processing and integration with surgical workflows.
Mental Health Analysis	Detects depression, anxiety, and mood disorders using speech and text analysis.	Subjectivity in diagnosis, patient privacy, and data bias.
Fraud Detection and Security	Identifies fraudulent claims and cyber threats in healthcare data.	Anomaly detection in massive datasets, adversarial attacks.

5.1. Real World Case Studies

Diagnostics, treatment planning, and patient care have all significantly improved due to the fundamental transformation of medical data utilization brought about by integrating BDA and ML into healthcare systems. By utilizing big datasets and sophisticated algorithms, healthcare providers can find previously undiscovered patterns and insights, which makes it possible to provide more individualized and efficient treatment. This section presents a collection of real-life case studies demonstrating the revolutionary impacts of big data and ML across various healthcare settings, highlighting the practical applications and benefits of these two technologies.

5.1.1. UK Biobank: Advancing Global Medical Research

The UK Biobank, a central biological database and research tool, includes five hundred thousand people's anonymized genetic and health data. It was established to enhance the prevention, diagnosis, and treatment of various serious and life-threatening diseases. This resource is accessible worldwide to authorized researchers conducting essential studies on prevalent illnesses. The use of BDA and ML methods has allowed scientists to identify patterns and relationships that deepen our comprehension of complex diseases [83].

The UK Biobank prioritizes data security and privacy, ensuring that participant information remains safeguarded while enabling transformative research. By offering a robust data repository, it supports research initiatives that aim to improve health outcomes and foster personalized medicine developments. The Biobank is continually adapting, integrating new types of data, and broadening its scope to support diverse health research endeavors [84].

5.1.2. NVIDIA and GE HealthCare: Transforming Diagnostic Imaging

NVIDIA and GE HealthCare have partnered to develop AI-powered X-ray and ultrasound devices to promote innovation in autonomous imaging. The partnership will enhance diagnostic imaging capabilities by applying NVIDIA's AI technologies, enabling more accurate and effective diagnoses. Clinical operations will be made simpler and patient outcomes will be improved by integrating AI into medical imaging processes. This partnership serves as an example of how BDA and ML may be used to revolutionize healthcare delivery. Healthcare providers can provide faster and more accurate diagnoses because of AI, which will eventually improve patient care. The collaboration also emphasizes the

value of cross-industry collaborations in spurring innovation and enhancing healthcare systems [85].

5.1.3. Healthcare Platform Utilizing Big Data Analytics (BDA)

A medical platform that combines big data analysis and ML was able to detect diseases with more than 95% accuracy and lower costs by 90% in comparison with conventional hospital procedures. The platform's architecture utilizes distributed computing and sophisticated algorithms for efficiently processing large volumes of healthcare information. Through this method, real-time analysis is facilitated and clinical decision support is offered, which improves patient outcomes. Such platforms can improve access to quality healthcare services by providing correct diagnostics at one-sixth the cost. The adaptability and scalability of the system make it easy to implement the system in many different healthcare settings, leading to improvements in patient care on a large scale [86].

5.1.4. Big Data in Oncology Drug Development

Integrating BDA in oncology drug development has shown immense potential in advancing precision medicine. By analyzing large-scale biological experiments, clinical trials, and medical records, researchers can identify biomarkers and genetic profiles associated with specific cancer types [87].

This information enhances patient outcomes and treatment efficacy by facilitating the development of tailored medicines. Big data also expedites the drug development process by making it possible to identify prospective medication candidates and forecast their success probabilities. This strategy has decreased the time and expense of introducing novel cancer medicines to the market. Furthermore, the insights gained from big data analyses support the personalization of cancer care, ensuring that patients receive therapies tailored to their unique genetic makeup [88].

5.1.5. Oncora Medical: Streamlining Oncology Workflows

Oncora Medical integrates big data, automation, and ML to streamline workflows for oncologists to provide faster and more individualized treatment to cancer patients. Its software aggregates information from multiple clinical systems, aligns it to standardized ontologies, and extracts key information automatically to generate a real-time, research-grade data store. This complete data management aids in data-driven clinical decisions and improves treatment planning [89].

Through AI-driven tools, Oncora Medical's product line makes it easier to document and analyze oncology cases, enhancing the speed and precision of cancer treatment. The Cancer Registry Assistant, for instance, makes it easier for healthcare providers to streamline quality control for precise data gathering and reporting. Such innovations are a perfect example of the potential that big data and ML can have in revolutionizing oncology practice [90].

5.1.6. IQVIA's NLP Data Factory for Population Health

IQVIA's NLP Data Factory is transforming the way healthcare organizations manage unstructured information [91]. By aggregating significant insights from clinical notes, lab results, and other unstructured sources, the platform helps healthcare payers enhance population risk stratification [92]. This innovation has made it possible to more effectively identify high-risk patients, enabling timely interventions and improved health outcomes.

NLP Data Factory analyzes millions of records in an hour, making the data analysis process far less time- and resource-intensive. By converting unstructured data into actionable intelligence, IQVIA's solution enables smarter decision-making and better patient care [93].

5.1.7. Digital Health Platform in Colombia

A Colombian healthcare company has developed a digital health platform integrating high-level analytics with operational, clinical, and business data sources. The goal of the integration is to improve population health management decision-making. By combining various data sources, the platform offers an overall overview of patient populations, allowing for more efficient health interventions.

Its utilization of big data and ML methods enables real-time analysis and predictive modeling. It allows proactive healthcare approaches, including early disease detection and resource allocation. The Colombian case study shows the potential for digital health innovations to augment healthcare delivery, especially in resource-poor contexts [94].

5.1.8. BigQuery ML for Diabetes Prediction

Scientists have used Google's BigQuery ML to create a diabetes prediction model, showing the platform's potential to make healthcare analytics more accessible. By allowing ML models to be created through ordinary SQL queries, BigQuery ML reduces the threshold for healthcare experts who lack advanced programming experience. This opens up ML tools for broader use in clinical environments.

The diabetes risk prediction model performed excellently, demonstrating BigQuery ML's effectiveness in processing big data. Integration into current healthcare workflows enables real-time risk calculation and tailored patient care. The case study emphasizes the value of accessible ML capabilities in promoting data-driven healthcare solutions [95].

5.1.9. HealthEdge: Predicting Type 2 Diabetes

HealthEdge is an intelligent health platform that utilizes ML within an integrated cloud, edge, and IoT computing system to forecast Type 2 Diabetes. Through real-time examination of medical sensor and device data, the platform can evaluate unique risk factors. Early detection and intervention are significant in managing chronic conditions such as diabetes.

The architecture of the framework supports efficient data processing at the edge, minimizing latency and conserving bandwidth. ML models are trained in the cloud and pushed to edge devices, enabling scalable and real-time healthcare solutions. HealthEdge demonstrates how the convergence of cutting-edge technologies can improve disease prediction and patient outcomes [96].

5.1.10. AI Predicting 10-Year Heart Disease Risk

Researchers have developed a DL model based on a single chest X-ray that can predict a patient's 10-year chance of dying from a heart attack or stroke. Over 147,000 chest X-rays from more than 40,000 persons were used to train the algorithm, which identified patterns associated with cardiovascular risk. The noninvasive method offers a possible instrument for heart disease prevention and early diagnosis.

The predictions by the AI model were comparable to conventional risk assessment tools that involve blood work and blood pressure readings. This technique may simplify risk stratification and facilitate timely intervention using already available imaging information. The research brings out the value of AI in making diagnostic accuracy possible and improving care for patients in cardiology [97].

The diabetes ML predictive model leverages key patient characteristics, specifically age, body mass index, blood glucose levels, family medical history, and lifestyle patterns, to identify individuals at elevated risk of developing diabetes at an early stage. This predictive capability enables clinicians to implement timely, targeted interventions aimed at preventing or delaying the onset of diabetes-related complications. Such interventions may

include personalized nutrition plans, tailored exercise regimens, and continuous glucose monitoring programs, all of which contribute to reducing the incidence and progression of diabetes.

Table 6 compares ten real-world case studies showcasing BDA and ML in smart healthcare systems. It summarizes essential elements, including the participating organizations, application domains, technologies employed, and the results obtained, demonstrating these technologies' various applications and advantages in improving healthcare delivery.

Table 6. Comparative analysis of big data and ML applications in smart healthcare systems.

Case Study	Application Area	Technologies Used	Key Outcomes
UK Biobank	Medical Research	BDA ML	Enhanced disease understanding through analysis of extensive health data from 500,000 participants.
NVIDIA and GE HealthCare	Diagnostic Imaging	AI ML	Improved diagnostic accuracy and efficiency in X-ray and ultrasound imaging.
Healthcare Platform	Disease Detection	BDA ML	Achieved over 95% accuracy in disease detection with a 90% cost reduction compared to traditional methods.
Oncology Drug Development	Drug Development	BDA	Streamlined drug discovery processes and personalized cancer treatments.
Oncora Medical	Oncology Workflows	BDAML	Enhanced treatment planning and decision-making in oncology through integrated data analysis.
IQVIA's NLP Data Factory	Population Health	NLP BDA	Improved population risk stratification by extracting insights from unstructured health data.
Digital Health Platform in Colombia	Population Health Management	BDA ML	Enhanced decision-making and proactive healthcare interventions in resource-constrained settings.
BigQuery ML for Diabetes Prediction, Google Cloud	Diabetes Prediction	BigQuery ML SQL-based ML	Simplified development of predictive models for diabetes risk assessment.
HealthEdge	Type 2 Diabetes Prediction	ML IoT Edge and Cloud Computing	Enabled real-time diabetes risk prediction using integrated IoT-edge-cloud systems.
AI Predicting 10-Year Heart Disease Risk	Cardiovascular Risk Assessment	DL AI	Predicted 10-year risk of heart disease using single chest X-rays, aiding early intervention strategies.

The real-life case studies demonstrate the revolutionary effect of BDA and ML in different healthcare fields. These applications, from disease prediction and diagnostic imaging to population health management, show how data-driven technology improves decision-making, lowers expenses, and personalizes patient care. One takeaway is the scalability and flexibility of these technologies to both developed and resource-poor environments, affirming their global applicability. In addition, partnerships between health organizations and tech giants (e.g., NVIDIA, Google Cloud) are speeding up innovation and simplifying clinical workflows, setting the stage for future digital health ecosystems.

The case studies highlighted BDA and ML's revolutionary potential in various aspects of healthcare. These technologies are transforming the healthcare landscape, from enhancing diagnostic imaging and optimizing oncology workflows to improving drug development and population health management. Healthcare professionals can deliver more personalized, efficient, and effective care by utilizing massive datasets and complex algorithms. As BDA and ML continue to integrate, they promise further innovations that will enhance patient outcomes and revolutionize healthcare delivery.

5.2. ML-Driven Innovations in Disease Diagnosis and Early Detection

ML has revolutionized disease diagnosis by rapidly and accurately detecting intricate health states from large and diverse datasets. These developments cover various diagnostic modalities, including medical images, genomic sequences, clinical text, wearable sensor signals, and electronic health records. Unlike conventional rule-based diagnostic systems, ML models, particularly DL architectures, can automatically learn patterns and features from data without explicit programming [98]. This makes them well-suited to handle non-linear, high-dimensional healthcare data.

Figure 9 depicts the integrated ML–big data framework for smart diagnostics. It reflects the end-to-end process from data acquisition through wearable sensors and EHRs, big data analytics, and machine learning models to final clinical decisions. The framework has a feedback loop mechanism, allowing continuous model updating and personalized treatment planning, thus improving the accuracy and responsiveness of healthcare systems.

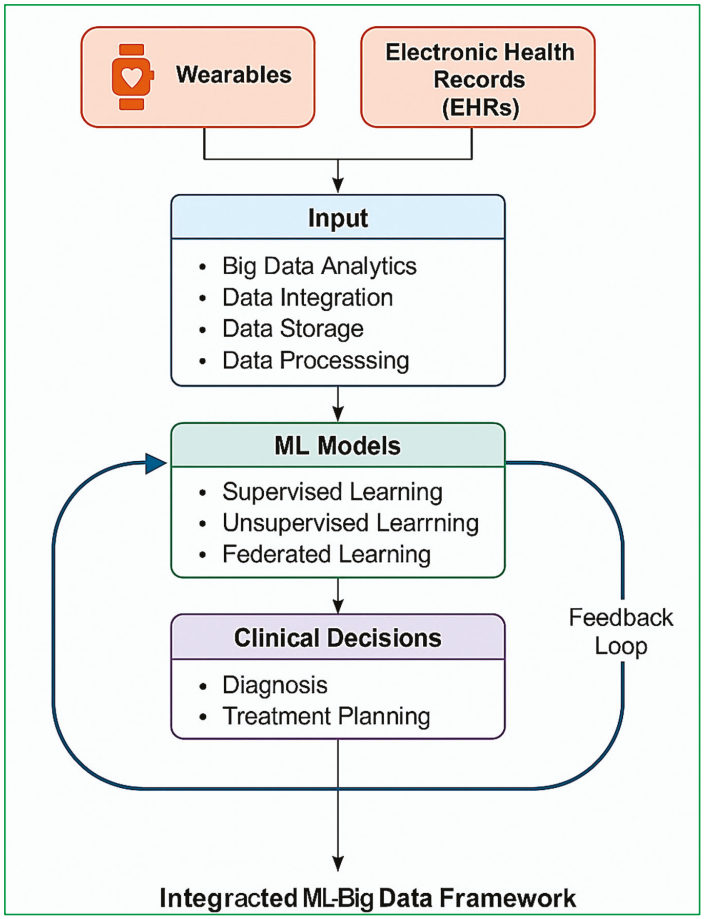


Figure 9. Integrated ML–big data framework for smart diagnostics: from acquisition to decision support. (Source: created by the authors).

5.2.1. Medical Imaging and Radiology

CNNs have transformed radiological diagnosis by achieving high accuracy in detecting conditions [99]. For example, a CNN model achieved 95.19% accuracy in detecting pneumonia from chest X-rays, demonstrating its suitability for real-world clinical use [100]. In another example, a deep CNN ensemble produced an accuracy of 93.91% in diagnosing pneumonia, demonstrating the effectiveness of state-of-the-art ML methods in medical imaging [101].

5.2.2. Natural Language Processing in Clinical Text

Natural language processing (NLP) has been a central innovation in extracting unstructured clinical notes and pathology reports to detect diseases at an early stage. Models based on the Transformer, such as ClinicalBERT and BioBERT, have gained traction in handling clinical documentation, facilitating diagnostic cues extraction, and supporting [76]. NLP develops automated coding and documentation procedures. Such models enable early detection of conditions like Alzheimer's by analyzing electronic health records and clinical notes [102].

5.2.3. Large Language Models (LLMs) in Healthcare

LLMs represent an advanced class of DL architectures trained on extensive text corpora, enabling them to understand and generate human-like language. Prominent examples include BERT, GPT, and domain-specific variants such as ClinicalBERT, all of which have demonstrated exceptional performance in tasks such as text classification, question answering, and summarization [103].

In the healthcare domain, LLMs hold particular promise in extracting meaningful clinical insights from unstructured textual data sources, including physician notes, discharge summaries, and pathology reports. By leveraging their capacity to identify linguistic patterns and contextual relationships, LLMs enhance clinical decision support systems through anomaly detection, pattern recognition, and synthesis of relevant patient information. This, in turn, contributes to faster and more accurate diagnostics, informed treatment planning, and improved risk stratification [104].

5.2.4. Wearable Devices and Continuous Monitoring

Wearable technology is now essential to continuous health monitoring, giving instant physiological data that can be processed using ML to enable early disease detection. For instance, smartwatches have been used to identify early symptoms of Parkinson's disease by tracking fine movement changes in movement patterns, potentially allowing for earlier interventions. Furthermore, AI-enabled wearable sensors have been created to track health markers in real-time and provide actionable data-driven insights that enable early disease detection and personalized interventions [105].

5.2.5. Neurodegenerative Diseases and Early Detection

ML methods have been up-and-coming in the early diagnosis of neurodegenerative disorders like Alzheimer's [106]. Experiments have proved that ML models can forecast the conversion from mild cognitive impairment to Alzheimer's disease dementia by evaluating neuroimaging data and other biomarkers [107]. These predictive functions allow for early interventions and improved management of disease progression.

5.2.6. Integration and Future Prospects

The incorporation of ML across different diagnostic fields is transforming the early detection of disease. It allows for proactive and tailored interventions, minimizes diagnostic delay, and improves outcomes in various areas of medicine. Ongoing research is needed

to enhance model interpretability, resolve data bias, and guarantee safe deployment in real-world clinical settings [108].

6. Major Challenges Faced in Processing Huge Healthcare Data

6.1. Data Challenges

Big data challenges are frequently conceptualized through the lens of the "five Vs": Volume, Velocity, Variety, Veracity, and Volatility, each representing a critical dimension of complexity in managing large-scale data environments.

- Volume: The sheer quantity of data generated by organizations, often reaching hundreds of terabytes or even petabytes, stems from routine business operations and regulatory obligations. In the healthcare domain, this data proliferation quickly leads to saturation, wherein the proportion of actionable or relevant information diminishes. This phenomenon gives rise to what is termed the "blind zone": a segment of data characterized by unknown or unexamined facts that may be either inconsequential or critically informative.
- Variety: The exponential growth of sensor technologies, IoT devices, and diverse communication platforms has resulted in a highly heterogeneous data landscape. Data today spans structured formats (e.g., relational databases), semi-structured formats (e.g., XML or JSON with identifiable markers but lacking a rigid schema), and unstructured formats (e.g., textual data, multimedia, and user-generated content from websites and social media). This diversity presents formidable challenges for integration, storage, and analysis across systems.
- Velocity: The velocity dimension refers to the rapid rate at which data is generated, transmitted, and analyzed. Modern data streams require real-time or near-real-time processing due to their ephemeral value. In contexts such as healthcare, where clinical decisions are time-sensitive, the inability to process high-speed data flows in transit can result in missed opportunities for timely interventions.
- Veracity: Veracity pertains to the reliability, accuracy, and quality of data. Issues such as misinformation, incomplete records, and noise complicate the extraction of meaningful insights. Given the scale of big data, ensuring data veracity becomes a significant hurdle, particularly in fields like healthcare, where analytical precision is paramount. Compounding this issue is the shortage of highly skilled data scientists, professionals adept in data mining, transformation, interpretation, and innovation, whose expertise is often prohibitively expensive and challenging to retain.
- Volatility: Volatility describes the degree to which data can be deemed reliable over time and the duration for which it remains relevant or valid within a system. In an era increasingly reliant on data-driven insights, understanding the temporal sensitivity of data, when it becomes obsolete, is essential for maintaining the integrity and usefulness of information used in decision-making processes.

Collectively, these five dimensions emphasize the multifaceted challenges inherent in the effective utilization of big data, particularly in domains where precision, trustworthiness, and timeliness are critical [109,110].

6.2. Process Challenges

Data provenance refers to the documentation of the origin and transformation of data, enabling the tracking of its evolution and informing decisions regarding its suitability for future processing. In the context of big data, datasets must be horizontally prepared through standardized formatting and consistent extraction processes to facilitate analysis. However, maintaining such a data pipeline can be labor-intensive and complex. Therefore, effective data integration, encompassing the consolidation of disparate systems, operations, and

databases, is critical for ensuring analytical productivity. To support this, mechanisms for aggregation, automation, and system compatibility must be in place, thereby contributing to the development of a robust and reliable analytical framework Similar challenges regarding algorithmic deployment, optimization strategies, and integration across complex data environments have been observed in adjacent domains like manufacturing, where ML and DL have faced barriers related to model generalization and real-time inference [111].

6.3. Management Challenges

Driven by the rapid proliferation of data vulnerabilities, additional big data challenges have emerged, notably in the areas of connectivity, cost, data privacy, and regulatory compliance. These concerns are particularly pronounced in domains such as healthcare, where the protection of personal information is paramount. The improper use or integration of personal data, especially when aggregated from multiple disparate sources, raises serious ethical and legal concerns.

For example, the adoption of cloud technologies to facilitate the sharing of healthcare data introduces potential risks. While cloud-based platforms can enable the pooling of patient data across multiple providers to enhance diagnostic capabilities and preventive care, they also introduce significant vulnerabilities. In contexts where pharmaceutical data security is limited, the reliance on third-party cloud infrastructure compounds the risk of data breaches and unauthorized access.

The increasing reliance on big data analytics has led many enterprises to construct expansive computing environments capable of storing, integrating, and analyzing petabytes of information. These centralized data repositories present attractive targets for malicious actors, including organized cybercriminal groups, due to the potential for high-value data exfiltration, often described as the "crown jewels" of an organization.

A further operational challenge lies in selecting appropriate technological frameworks. Enterprises must evaluate whether open-source platforms such as Hadoop are sufficient for their needs, or whether commercial alternatives, including Cassandra, Cloudera, Hortonworks, and MapR, offer better scalability, security, and support for enterprise-grade big data management. Balancing performance, cost-effectiveness, and compliance with privacy regulations remains a complex but critical endeavor in the modern big data landscape [112].

7. Existing ML-Based Big Data Solutions in Managing Healthcare Data

The advent of ML has dramatically enhanced healthcare activities and services, giving rise to the new discipline of "smart healthcare". ML-based algorithms are frequently more precise than trained practitioners, particularly in pathology and imaging. The most common use of ML in the healthcare field is sickness diagnosis, such as remote healthcare, diagnosis and treatment, healthcare monitoring, treatment adherence, and body sensors. Intelligent software is predicted to aid radiologists and clinicians in assessing patients in the coming years, and ML will transform medical knowledge and education [113]. Medical healthcare has developed as an intriguing application field for ML models, with these models already achieving human-level performance in clinical pathology, radiography, ophthalmology, and dermatology. The primary applications of healthcare that can benefit from ML approaches are as follows (Figure 10).

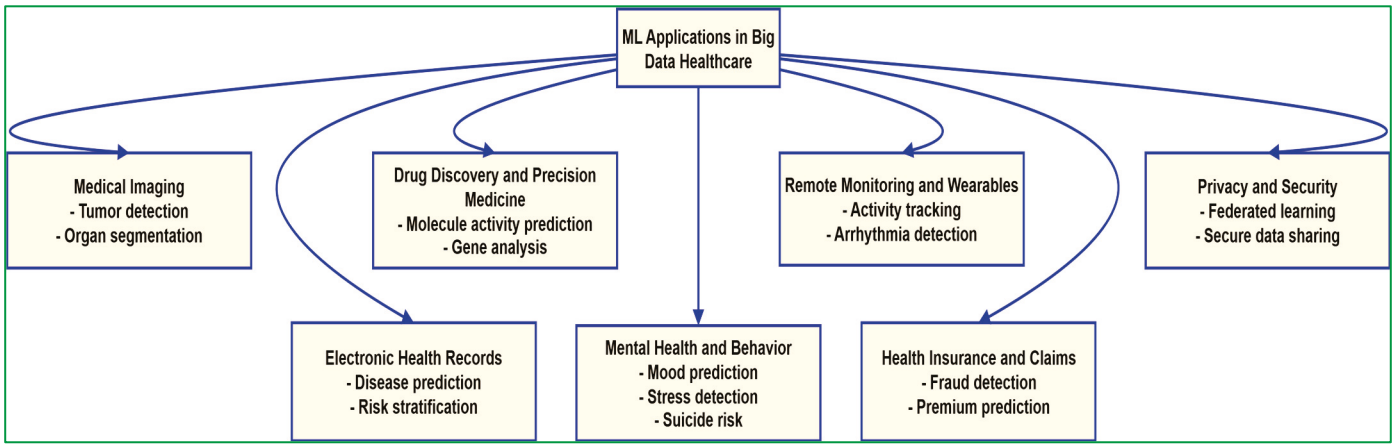


Figure 10. ML applications in voluminous healthcare domains.

7.1. Diagnosis and Treatment

In smart healthcare, various illness detection technologies are connected to computerized decision-making. The categorization system may someday take the role of medical professionals in identifying illnesses if it performs well in terms of overall reliability and testing time. The most notable and outstanding product in clinical decision support systems is IBM's Watson10, an artificial cognitive system that provides the best answer by thoroughly analyzing all medical and literature data. Watson can answer health-related questions in everyday and medical applications in the healthcare industry. Scientific papers, patents, medication and disease terminologies, drug testing, electronic health records, laboratory and radiology data, genomic data, claims data, and online social data are just a few medical resources it can scan and assess. Physicians may use clinical decision support systems to offer expert guidance based on algorithms to improve diagnostic performance, reduce the number of erroneous diagnoses and misinterpretations, and allow patients to get prompt and efficient treatment.

Among the most complex non-communicable diseases to treat at this time are dementia and Alzheimer's disease. Alzheimer's disease diagnosis, dementia diagnosis, Alzheimer's disease-related area identification, mild cognitive impairment patient conversion to Alzheimer's disease prediction, and Alzheimer's disease-related gene discovery are among the uses. Fourteen distinct machine-learning algorithms were used. The patient's health and illness status are more correctly defined using smart diagnosis, which aids in developing a customized treatment plan, and specialists have endorsed the software [114,115].

7.2. Medical Imaging

These days, medical imaging is automated. Several limitations need to be resolved before they can be utilized effectively in healthcare. When combined with radiological imaging, machine learning approaches help identify these sickness symptoms in people. "Medical imaging" refers to various techniques to provide visual depictions of the human body's interior for research, diagnosis, and treatment. This eliminates the need for the conventional clinical standard of surgical therapy. There is an increased risk of infections, strokes, and other complications when any portion of the human body is surgically accessed. ML in medical imaging includes positron emission tomography (PET) scans, ultrasound imaging, computerized axial tomography scans, and magnetic resonance imaging. It has been exposed that the established automated models may be utilized to identify various chest-related disorders, such as TB and pneumonia [116,117].

The most prevalent use of contemporary ML technologies in healthcare is computer-automated detection (CAD), especially in identifying lesions seen in mammograms, brain

scans, and other body scans. Breast imaging has shown the most success with CAD. Computer-aided detection (CADe), in which the machine alerts the radiologists to prospective lesions, and computer-aided diagnosis (CADx), in which the computer estimates the chance that a lesion is cancerous, are both examples of CAD.

The outcome of these imaging methods is a collection or sequence of pictures that a radiologist must interpret and diagnose. ML approaches quickly improve their ability to anticipate and locate pictures that may signal a disease condition or primary concern.

7.3. Drug Discovery and Development

ML techniques offer tools to improve discovery and decision-making for well-defined topics with high-quality data. ML can lower medication development and research failure rates and enhance data-driven decision-making. The most widely used method in drug development is to design drugs (small molecules, enzymes, antibodies, or more recent modalities like short RNAs or cell therapies) that alter a biological target's functioning to impact the disease state. ML holds a lot of promise in the study of fundamental biological concepts to find therapeutic opportunities through alternative modalities or new targets [2]. The investigation of genetic variability in splicing signals is one facet. Drug development and discovery pipelines are drawn out, intricate, and dependent on several variables [116,118,119].

Limitations of ML in Late-Stage Drug Development

ML has demonstrated substantial promise in the early phases of drug discovery, particularly in tasks such as target identification, molecular docking, and virtual screening. However, its efficacy diminishes in the later stages of drug development. A key limitation lies in the paucity of data on clinical failures, drug specificity, and long-term toxicity. This data scarcity significantly impairs the predictive performance of ML models.

Notably, the majority of drug candidates that advance to Phase II or Phase III clinical trials ultimately fail, frequently due to unforeseen adverse effects or inadequate therapeutic efficacy in human populations. ML models are constrained by limited access to training data representing these infrequent yet clinically significant outcomes, as only a small fraction of compounds progress to these advanced stages.

This limitation emphasizes the necessity for enhanced data-sharing mechanisms, the incorporation of real-world evidence, and the development of hybrid modeling approaches that integrate mechanistic insights with data-driven ML techniques. Such strategies are essential to improving decision-making processes during the most resource-intensive and uncertain phases of drug development [120].

7.4. Natural Language Processing of Medical Records

NLP is used in medical documents. Many medical professionals concur that the drive toward EMR in many countries is slow, complex, and sometimes completely mismanaged. As a result, patients may occasionally receive worse overall care. One of the most significant problems in many hospitals and clinics is the amount of tangible medical data and paperwork. Handwritten notes, inconsistent formatting, and incomplete or dispersed data have hindered EMR implementation.

Since language is the primary medium for conveying complex information, doctors' notes and annotated medical records include essential information about the population and the health of specific patients. Analysis is difficult due to the language's unpredictability and inconsistency, and the extraction of higher-level information into relevant subcategories. ML is producing encouraging results when conducting such complex analyses [121].

7.5. ML Applications in Prognosis

In clinical practice, prognosis is the process of forecasting the predicted progression of a disease. It also includes determining if symptoms and indicators linked with a given disease will worsen, improve, or remain constant over time, as well as deciding probable associated health issues, complications, capacity to perform everyday tasks, and the possibility of survival. As in a clinical context, multi-modal patient data, such as phenotypic, genomic, proteomic, pathology test findings, and medical pictures, are collected, which may help ML models with illness prognosis, diagnosis, and therapy. ML approaches are employed in various applications, from identifying and categorizing tumors using X-ray and CRT images to detecting cancers using proteomic and genomic (microarray) tests. Scientists have a substantial bias toward using ML to identify outcomes or hazards linked with breast (24%) and prostate (20%) cancer. However, ML technologies seem to effectively forecast results or risks in almost a dozen types of cancer. This shows that machine-learning approaches can be used to forecast and prognosticate cancer in a broad range of situations [122].

The study revealed many similarities in the sorts of ML algorithms employed, the types of training data incorporated, the kinds of outcome predictions made, the types of cancers analysed, and the overall efficacy of these systems in diagnosing cancer risk or outcomes.

7.6. ML for Medical Time Analysis

One of the duties of clinical processes is to analyze clinical time-series data. Clinical time-series modeling has been used to assess mean arterial blood pressure (ABP) and intracranial pressure (ICP), two important markers of cerebrovascular autoregulation (CA) in TBI patients, to predict death rates in TBI patients, and to predict intervention strategies in intensive care units (ICUs) using CNN and LSTM. By integrating patient records with multivariate and time-series measurement data, attention models were used in recent studies to track ICU forecasting activities (including diagnosis, estimation, and prediction, among other things) [123].

7.7. Prediction of Future Illness Symptoms

Predicting the signs of illnesses that might strike humanity in the future and cause global deaths is vital. ML uses deep neural networks and long-short-term memory learning techniques to forecast the signs and spread of infectious diseases, including malaria, scarlet fever, and chickenpox. Machine-learning algorithms based on information gathered from clinical reports, physician notes, and wearable body sensors may be used to predict patients' future medical circumstances. The machine learning tool predicts and identifies illness causes based on human genetics.

The project to identify relevant information regarding how genetics can influence human health is currently at the forefront of much scientific knowledge. These are next-generation sequencing (NGS) techniques, the explosion of genetic information, and large databases containing population-scale genetic data. Preventive medicine can benefit from knowing how to picture complex diseases and how an individual's genetic composition can augment or reduce their risk. This could give doctors more information about how to alter a patient's course of treatment to minimize the risk of more complicated disorders. The objectives include detecting human activity, tracking health, predicting illness, diagnosing illnesses, and advancing smart medicine. However, the most popular use of machine learning is the development of artificial intelligence infrastructure to make computers smarter, reducing the need for direct patient interaction. Additionally, to enhance current digital technologies so that patients may be recognized, detected, forecasted, and so forth

at home. To restrict the worldwide effect of various diseases, machine learning techniques are essential for modernizing our healthcare and other technological systems [123].

8. Ethical Considerations in AI-Driven Smart Healthcare

Smart healthcare systems stand to gain significantly from the integration of ML and big data, leveraging these technologies to enhance diagnostic accuracy, personalize treatment pathways, and improve operational efficiency. However, this technological reconfiguration introduces a range of ethical and societal considerations that must be addressed to ensure responsible and equitable implementation.

- Data Privacy and Security: A primary ethical concern involves the privacy and security of patient data. Smart healthcare platforms routinely collect vast volumes of sensitive information from sources such as IoT devices deployed in operating theatres, EHRs, and genomic databases. Safeguarding this data against breaches, misuse, or unauthorized access is not only a legal mandate but also an ethical imperative. In alignment with principles of fair data sharing, emerging technologies, such as blockchain, differential privacy, and secure federated learning, are being increasingly adopted to protect patient data while still enabling its utility for clinical insight generation and ML model development.
- Consent and Transparency: A further ethical challenge arises from issues of consent and transparency. The deployment of black-box ML systems often impedes comprehension of decision-making processes by both patients and clinicians. Explainability must, therefore, be prioritized in the design of ethical AI, particularly in healthcare settings where model outputs may inform critical, high-stakes decisions. Patients should be adequately informed about the use of their data, the risks involved, and their rights regarding contesting or overriding automated decisions.
- Algorithmic Accountability: Equally crucial is the matter of algorithmic accountability. As AI systems become integral to decision-making in diagnosis, treatment planning, and resource allocation, mechanisms for tracing errors, identifying points of failure, and assigning accountability must be clearly established. Health systems must avoid uncritical adoption of these tools in the absence of robust governance frameworks, domain expertise, and risk mitigation strategies to prevent unintended harm.

Beyond these established concerns, deeper ethical dilemmas, often insufficiently addressed, are pivotal for equitable AI integration in healthcare:

- Economic Disparity and Access Limitations: AI-powered healthcare tools often entail high implementation and maintenance costs, disproportionately benefiting wealthier individuals and institutions. This economic imbalance risks widening existing disparities in care quality and health outcomes, particularly in under-resourced settings where financial barriers limit access to advanced diagnostics and AI-supported clinical tools.
- Bias in AI Performance Evaluation: Many AI solutions are developed, validated, and promoted by the same entities, raising concerns regarding inflated performance claims and commercial bias. Clinical robustness should be ensured through independent third-party audits, stringent regulatory oversight, and the use of open, transparent benchmarking methodologies.
- Biases in Historical Healthcare Data: The datasets used to train ML models often reflect historical inequities in healthcare delivery. Consequently, models trained on such data risk perpetuating discriminatory outcomes across race, gender, or socioeconomic status. Ensuring fairness requires rigorous auditing of training datasets,

inclusive representation during model design, and continuous model updates to align predictions with equitable health outcomes.

Ultimately, the ethical integration of AI in healthcare necessitates a confluence of technical rigor, transparency, accountability, and inclusivity. This multifaceted approach is essential to the realization of intelligent, trustworthy, and compassionate healthcare systems that serve diverse populations equitably [124–126].

9. Future Research Direction

Despite the rapid advancement of ML, BDA, and IoMT technologies in healthcare, several critical challenges remain before these revolutions can be safely, effectively, and scalably deployed in real-world clinical environments. Future research must be problem-driven, addressing specific gaps while fostering responsible innovation. This review highlights key trends and limitations within current AI research, identifying privacy and security, scalability and infrastructure, and explainability and clinical integration as dominant themes warranting further scholarly attention.

- Privacy and Security

As the healthcare sector becomes increasingly data-centric, ensuring the secure storage and transmission of sensitive patient information is paramount. Emerging approaches such as locality-sensitive hashing and differential privacy aim to enhance model utility while minimizing direct reliance on identifiable patient data. Future research should prioritize the development of privacy-preserving ML techniques, notably federated learning and differential privacy, which enable distributed model training without compromising data confidentiality. These methods significantly reduce the risk of data breaches while maintaining robust model performance. Moreover, blockchain technology presents a promising avenue for constructing decentralized, tamper-proof, and auditable infrastructures that can facilitate secure and transparent data sharing across institutions and jurisdictions. Such innovations are vital for fostering trust and achieving regulatory compliance in multi-stakeholder healthcare collaborations.

- Scalability and Infrastructure

The exponential increase in health-related data generated by wearable devices, remote monitoring sensors, and EHR systems necessitates the development of scalable and efficient computational infrastructures. ML models must operate with low latency and be capable of delivering real-time analytics in resource-constrained environments, particularly at the network edge. Future research should explore the design of cloud-native and hybrid architectures that can dynamically scale while maintaining data integrity and availability. The establishment of interoperable systems and streamlined data pipelines will be critical to maximizing the utility of big data in smart healthcare ecosystems.

- Explainability and Clinical Integration

The successful implementation of AI in clinical practice hinges on the interpretability of ML models and their ability to provide clinically meaningful explanations. Future work should promote the adoption of explainable AI (XAI) techniques that offer transparent, trustworthy, and actionable insights for clinical end-users. These tools must be rigorously evaluated within real-world healthcare settings to support safe and ethical decision-making. Furthermore, seamless integration into clinical workflows, including embedding AI functionalities within EHR platforms, diagnostic systems, and telemedicine solutions, is essential to ensure usability and clinician acceptance. Research should also focus on the development of standardized evaluation protocols and real-world validation

frameworks that assess model robustness, fairness, and generalizability across diverse patient populations.

Addressing these future research priorities is critical to unlocking the transformative potential of ML-powered smart healthcare systems. By ensuring data security, scalable infrastructure, and model transparency, and by enabling clinical usability, future innovations can foster the development of responsible, equitable, and patient-centered AI that meaningfully improves health outcomes across varied clinical settings [127,128].

10. Conclusions

Big data and ML in smart healthcare systems offer a transformative opportunity to enhance the efficiency, accuracy, and accessibility of medical treatment. The importance of machine learning techniques in addressing critical healthcare problems such as disease detection, personalized medication, and operational optimization is highlighted in this review. Because BDA and ML work well together, healthcare systems can process massive amounts of data, uncover hidden patterns, and support evidence-based decision-making. Despite these advancements, the industry still confronts significant challenges, such as scalability, interpretability of ML models, and data privacy. The ethical and fair use of these technologies is crucial to establishing broad adoption and building stakeholder trust. Future research should focus on developing robust frameworks for privacy-preserving analytics, enhancing model explainability, and promoting interdisciplinary collaboration to address healthcare data's unique complexity. By addressing these problems and creating more proactive, patient-centered, and efficient healthcare systems, big data and machine learning have the potential to alter the way healthcare is delivered fundamentally. It will take continual innovation and legislative and regulatory backing to realize these technologies' potential fully.

Disease prediction, medical imaging, drug development, individualized treatment, EHR management, and remote patient monitoring are just a few ML-driven applications highlighted in the paper. Significant real-world results include:

- Diagnostic accuracy of 95%+ is achieved by ML-integrated platforms, comparable to conventional manual-inspection methods [86].
- Big data-enabled AI diagnostic systems have achieved cost reductions of up to 20%, primarily by minimizing unnecessary testing and streamlining decision workflows [82].
- AI-powered imaging systems developed by NVIDIA and GE HealthCare have been successfully deployed for automated X-ray and ultrasound assessments, improving diagnostic efficiency and expanding access globally [86].
- Oncora Medical is advancing the development of ML algorithms and BDA to establish a standardized and automated solution for oncology treatment planning. The platform aggregates patient-specific clinical data, including tumor genetic profiles, medical histories, imaging results, and prior treatment responses, to assist clinicians in efficiently generating personalized therapy recommendations. This automated approach significantly reduces the interval between diagnosis and the initiation of treatment, thereby enhancing patient outcomes through the provision of timely, precise, and individualized cancer care.
- The UK Biobank uses ML to advance population health research through massive genomic datasets.

These case studies illustrate the tangible advantages of ML and big data in different clinical, research, and operational areas.

Despite apparent progress, the review calls out a range of ongoing challenges:

- Heterogeneity and interoperability of data across systems.
- Privacy and security risks in cloud and IoT-enabled environments.

- Computational hardness in handling high-dimensional, real-time data.
- Lack of black-box ML models' explainability hinders clinical trust and uptake.
- Scalability issues in scaling ML models across large, distributed healthcare systems.

These constraints impair the seamless integration of intelligent healthcare technologies and require targeted technological and policy interventions.

To facilitate meaningful advancements in the field, future research must provide greater specificity regarding the clinical features under investigation and elucidate the precise mechanisms by which ML and BDA are translated into learning models. It is essential that studies clearly define the patient populations to which predictive algorithms are applicable and distinguish cases where learning models are successfully applied following predictive insight generation from those where no such application occurs.

Subsequent reviews and empirical investigations should adopt a more focused and clinically grounded methodology to reduce redundancy. Specifically, the inclusion of concise, contextually relevant examples, both generalizable and domain-specific, should be prioritized over extensive but superficial commentary. Such an approach is likely to enhance the clarity, accessibility, and practical relevance of ML and big data research for healthcare practitioners, policymakers, and other stakeholders involved in clinical decision-making.

Limitations of the Review

Although this review offers an extensive overview of the convergence between big data and ML in intelligent healthcare systems, it has several limitations. First, the research is primarily based on a narrative review approach, which, while expansive and exploratory, lacks the statistical sophistication of systematic meta-analyses. Second, literature selection was restricted to English-language publications and potentially excluded relevant non-English language studies. Third, while attempting to address varied applications, some niche use cases and emerging technologies may have been underrepresented due to space and scope limitations. Moreover, the review does not conduct a quantitative comparison of the performance of particular ML models but instead provides conceptual and thematic insights. Lastly, the rapid development of healthcare technologies suggests that newer developments after 2025 may not be included; therefore, ongoing updates are necessary to maintain the findings' relevance.

Author Contributions: S.R.: Conceptualization, Methodology, Investigation, Writing—Original Draft, Visualization; R.K. (Raman Kumar): Supervision, Project Administration, Writing—Review and Editing; B.S.P.: Data Curation, Validation, Software Implementation; R.K. (Rajender Kumar): Visualization, Writing—Review and Editing, Investigation; N.F.M.: Literature Review, Data Collection, Writing—Supporting Drafts; M.A.A.: Ethics Consideration, Validation, Interpretation of Data; J.L.: Supervision, Critical Review, Final Approval of the Manuscript. 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 generated or analyzed during this study are included in this published article.

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

References

1. Xu, G.; Fan, X.; Xu, S.; Cao, Y.; Chen, X.; Shang, T.; Yu, S. Anonymity-Enhanced Sequential Multi-Signer Ring Signature for Secure Medical Data Sharing in IoMT. *IEEE Trans. Inf. Forensics Secur.* **2025**, *20*, 5647–5662. [CrossRef]
2. Kandeel, M. Revolutionizing Healthcare: Harnessing the Power of Artificial Intelligence for Enhanced Diagnostics, Treatment and Drug Discovery. *Int. J. Pharmacol.* **2024**, *20*, 1–10. [CrossRef]
3. Xiao, X.; Li, Y.; Wu, Q.; Liu, X.; Cao, X.; Li, M.; Dai, X. Development and validation of a novel predictive model for dementia risk in middle-aged and elderly depression individuals: A large and longitudinal machine learning cohort study. *Alzheimer's Res. Ther.* **2025**, *17*, 103. [CrossRef]
4. Research and Markets. *Global Big Data in Healthcare Market Trends and Forecasts Report 2024: A $540 Billion Industry by 2035, from $67 Billion in 2023*; GlobeNewswire: El Segundo, CA, USA, 2024.
5. Greene, L. How Healthcare Data Technology is Leveraged by Leaders. *Arcadia Solutions*. 6 September 2023. Available online: https://arcadia.io/resources/healthcare-data-technology (accessed on 10 May 2025).
6. He, W.; Zhu, J.; Feng, Y.; Liang, F.; You, K.; Chai, H.; Wang, W. Neuromorphic-enabled video-activated cell sorting. *Nat. Commun.* **2024**, *15*, 10792. [CrossRef]
7. Zhang, G.; Song, C.; Yin, M.; Liu, L.; Zhang, Y.; Li, Y.; Li, C. TRAPT: A multi-stage fused deep learning framework for predicting transcriptional regulators based on large-scale epigenomic data. *Nat. Commun.* **2025**, *16*, 3611. [CrossRef] [PubMed]
8. Li, W.; Chai, Y.; Khan, F.; Jan, S.R.U.; Verma, S.; Menon, V.G.; Kavita, f.; Li, X. A comprehensive survey on machine learning-based big data analytics for IoT-enabled smart healthcare system. *Mob. Netw. Appl.* **2021**, *26*, 234–252. [CrossRef] [PubMed]
9. Taha, K. Big Data Analytics in IoT, social media, NLP, and information security: Trends, challenges, and applications. *J. Big Data* **2025**, *12*, 150. [CrossRef]
10. Cai, Q.; Wang, H.; Li, Z.; Liu, X. A survey on multimodal data-driven smart healthcare systems: Approaches and applications. *IEEE Access* **2019**, *7*, 133583–133599. [CrossRef]
11. Badawy, M.; Ramadan, N.; Hefny, H.A. Healthcare predictive analytics using machine learning and deep learning techniques: A survey. *J. Electr. Syst. Inf. Technol.* **2023**, *10*, 40. [CrossRef]
12. Nguyen, D.C.; Pham, Q.-V.; Pathirana, P.N.; Ding, M.; Seneviratne, A.; Lin, Z.; Dobre, O.; Hwang, W.-J. Federated learning for smart healthcare: A survey. *ACM Comput. Surv. (CSUR)* **2022**, *55*, 1–37. [CrossRef]
13. Salman, M.; Munawar, H.S.; Latif, K.; Akram, M.W.; Khan, S.I.; Ullah, F. Big data management in drug–drug interaction: A modern deep learning approach for smart healthcare. *Big Data Cogn. Comput.* **2022**, *6*, 30. [CrossRef]
14. Rahman, A.; Debnath, T.; Kundu, D.; Khan, M.S.I.; Aishi, A.A.; Sazzad, S.; Sayduzzaman, M.; Band, S.S. Machine learning and deep learning-based approach in smart healthcare: Recent advances, applications, challenges and opportunities. *AIMS Public Health* **2024**, *11*, 58. [CrossRef] [PubMed]
15. Babar, M.; Qureshi, B.; Koubaa, A. Review on Federated Learning for digital transformation in healthcare through big data analytics. *Future Gener. Comput. Syst.* **2024**, *160*, 14–28. [CrossRef]
16. Istepanian, R.S.; Al-Anzi, T. m-Health 2.0: New perspectives on mobile health, machine learning and big data analytics. *Methods* **2018**, *151*, 34–40. [CrossRef] [PubMed]
17. Selna, A.; Othman, Z.; Tham, J.; Yoosuf, A.K. Challenges to using electronic health records to enhance patient safety, in a Small Island Developing State (SIDS) context. *Rec. Manag. J.* **2022**, *32*, 249–259. [CrossRef]
18. Hu, F.; Yang, H.; Qiu, L.; Wei, S.; Hu, H.; Zhou, H. Spatial structure and organization of the medical device industry urban network in China: Evidence from Specialized, Refined, Distinctive, and Innovative firms. *Front. Public Health* **2025**, *13*, 1518327. [CrossRef]
19. Gardašević, G.; Katzis, K.; Bajić, D.; Berbakov, L. Emerging wireless sensor networks and Internet of Things technologies—Foundations of smart healthcare. *Sensors* **2020**, *20*, 3619. [CrossRef]
20. Baker, S.B.; Xiang, W.; Atkinson, I. Internet of things for smart healthcare: Technologies, challenges, and opportunities. *IEEE Access* **2017**, *5*, 26521–26544. [CrossRef]
21. Balasundaram, A.; Routray, S.; Prabu, A.; Krishnan, P.; Malla, P.P.; Maiti, M. Internet of things (IoT)-based smart healthcare system for efficient diagnostics of health parameters of patients in emergency care. *IEEE Internet Things J.* **2023**, *10*, 18563–18570. [CrossRef]
22. Ijaz, M.; Li, G.; Wang, H.; El-Sherbeeny, A.M.; Moro Awelisah, Y.; Lin, L.; Koubaa, A.; Noor, A. Intelligent fog-enabled smart healthcare system for wearable physiological parameter detection. *Electronics* **2020**, *9*, 2015. [CrossRef]
23. Sikdar, S.; Guha, S. Advancements of healthcare technologies: Paradigm towards smart healthcare systems. *Recent Trends Image Signal Process. Comput. Vis.* **2020**, *1124*, 113–132.
24. Khan, M.; Hossni, Y. A comparative analysis of LSTM models aided with attention and squeeze and excitation blocks for activity recognition. *Sci. Rep.* **2025**, *15*, 3858. [CrossRef] [PubMed]
25. Qi, K. Advancing hospital healthcare: Achieving IoT-based secure health monitoring through multilayer machine learning. *J. Big Data* **2025**, *12*, 1. [CrossRef]

26. Maruthupandi, J.; Sivakumar, S.; Dhevi, B.L.; Prasanna, S.; Priya, R.K.; Selvarajan, S. An intelligent attention based deep convoluted learning (IADCL) model for smart healthcare security. *Sci. Rep.* **2025**, *15*, 1363. [CrossRef]
27. Khalid, U.; Chen, L.; Khan, A.A.; Chen, B.; Mehmood, F.; Yasir, M. A smart facial acne disease monitoring for automate severity assessment using AI-enabled cloud-based internet of things. *Discov. Comput.* **2025**, *28*, 12. [CrossRef]
28. Xie, Q.; Ding, Z. Provably secure and lightweight blockchain based cross hospital authentication scheme for IoMT-based healthcare. *Sci. Rep.* **2025**, *15*, 6461. [CrossRef]
29. Zhou, Z.; Jin, Y.; Fu, J.; Si, S.; Liu, M.; Hu, Y.; Gan, J.; Deng, Y.; Li, R.; Yang, J. Smart wireless flexible sensing system for unconstrained monitoring of ballistocardiogram and respiration. *npj Flex. Electron.* **2025**, *9*, 15. [CrossRef]
30. Pham, T.P.T.; Huang, M.C. Exploring the relationship between nursing professional values and job satisfaction in Vietnam. *Nurs. Pract. Today* **2025**, *12*, 55–64. [CrossRef]
31. Magara, T.; Zhou, Y. EMAKAS: An efficient three-factor mutual authentication and key-agreement scheme for IoT environment. *Cyber. Secur. Appl.* **2025**, *3*, 100066. [CrossRef]
32. Hossain, M.K.; Thakur, V. A performance management framework for smart health-care supply chain based on industry 4.0 technologies. *J. Glob. Oper. Strateg. Sourc.* **2024**, *18*, 285–306. [CrossRef]
33. Mishra, G. A Comprehensive Review of Smart Healthcare Systems: Architecture, Applications, Challenges, and Future Directions. *Int. J. Innov. Res. Technol. Sci.* **2024**, *12*, 210–218.
34. Zhao, X.; Ge, B. An indicator framework for assessing the readiness of hospital in the smart healthcare transformation. *IOP Conf. Ser. Earth Environ. Sci.* **2022**, *1101*, 072010. [CrossRef]
35. Al-Sanjary, O.I.; Vasuthevan, S.; Omer, H.K.; Mohammed, M.N.; Abdullah, M.I. An Intelligent Recycling Bin Using Wireless Sensor Network Technology. In Proceedings of the 2019 IEEE International Conference on Automatic Control and Intelligent Systems, I2CACIS, Selangor, Malaysia, 29 June 2019; pp. 30–33. [CrossRef]
36. Cui, Q.; Ding, Z.; Chen, F. Hybrid Directed Hypergraph Learning and Forecasting of Skeleton-Based Human Poses. *Cyborg Bionic Syst.* **2024**, *5*, 0093. [CrossRef] [PubMed]
37. Taware, R.D.; Deshmukh, A.; Singh, C.; Rathod, N. Magnitude of data science & big data in fitness care. *AIP Conf. Proc.* **2025**, *3162*, 020020.
38. Kumar, A.; Gupta, R.; Kumar, S.; Dutta, K.; Rani, M. Securing IoMT-based healthcare system: Issues, challenges, and solutions. In *Artificial Intelligence and Cybersecurity in Healthcare Cyber Physical Systems*; Agrawal, R., Rathore, P.S., Devarajan, G.G., Divivedi, R.R., Eds.; Scrivener Publishing LLC: Beverly, MA, USA, 2025; pp. 17–56.
39. Du, M.; Li, Z.; Bian, L.; Randriamahazaka, H.; Chen, W. Two-dimensional materials van der Waals assembly enabling scalable smart textiles. *Mater. Sci. Eng. R Rep.* **2025**, *163*, 100915. [CrossRef]
40. Ibrahim, Y.; Abdel-Malek, M.A.; Azab, M.; Rizk, M.R. Privacy-preserved mutually-trusted 5G communications in presence of pervasive attacks. *Internet Things* **2025**, *30*, 101491. [CrossRef]
41. Wang, W.; Bo, X.; Li, W.; Eldaly, A.B.M.; Wang, L.; Li, W.J.; Chan, L.L.H.; Daoud, W.A. Triboelectric Bending Sensors for AI-Enabled Sign Language Recognition. *Adv. Sci.* **2025**, *12*, 2408384. [CrossRef]
42. Lakshmi, S.G.; Lalitha, C.N.; Lavanya, S.; Vijaya, L.S.; Sri Lakshmi, P.S. IOT-enabled cloud solutions for reliable health monitoring. In *Emerging Trends in Computer Science and Its Application*; CRC Press: Boca Raton, FL, USA, 2025; pp. 278–282.
43. Selem, M.; Jemili, F.; Korbaa, O. Deep learning for intrusion detection in IoT networks. *Peer-Peer Netw. Appl.* **2025**, *18*, 22. [CrossRef]
44. Rezaee, K.; Yang, X.; Khosravi, M.R.; Zhang, R.; Lin, W.; Jeon, G. Fusion-based learning for stress recognition in smart home: An IoMT framework. *Build. Environ.* **2022**, *216*, 108988. [CrossRef]
45. Qu, Q.; Sun, H.; Chen, Y. Smart healthcare at home in the era of IoMT. In *Internet of Things—New Insights*; IntechOpen: London, UK, 2023.
46. Komalasari, R. Ambient Assisted Living (AAL) Systems to Help Older People. In *Exploring Future Opportunities of Brain-Inspired Artificial Intelligence*; IGI Global: Hershey, PA, USA, 2023; pp. 84–99.
47. Bedón-Molina, J.; Lopez, M.J.; Derpich, I.S. A home-based smart health model. *Adv. Mech. Eng.* **2020**, *12*, 1687814020935282. [CrossRef]
48. Cheng, C.K.; Bakar, H.A.; Gollasch, M.; Huang, Y. Perivascular Adipose Tissue: The Sixth Man of the Cardiovascular System. *Cardiovasc. Drugs Ther.* **2018**, *32*, 481–502. [CrossRef]
49. Cáceres, C.; Rosário, J.M.; Amaya, D. Proposal of a smart hospital based on Internet of Things (IoT) concept. In Proceedings of the Processing and Analysis of Biomedical Information: First International SIPAIM Workshop, SaMBa 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018; Revised Selected Papers 1 2019; pp. 93–104.
50. Kang, S.; Baek, H.; Jung, E.; Hwang, H.; Yoo, S. Survey on the demand for adoption of Internet of Things (IoT)-based services in hospitals: Investigation of nurses' perception in a tertiary university hospital. *Appl. Nurs. Res.* **2019**, *47*, 18–23. [CrossRef] [PubMed]

51. Zhang, N.; Wang, Y.; Zhang, H.; Fang, H.; Li, X.; Li, Z.; Gong, Z. Application of interpretable machine learning algorithms to predict macroangiopathy risk in Chinese patients with type 2 diabetes mellitus. *Sci. Rep.* **2025**, *15*, 16393. [CrossRef] [PubMed]
52. Li, N.; Ou, J.; He, H.; He, J.; Zhang, L.; Peng, Z.; Jiang, N. Exploration of a machine learning approach for diagnosing sarcopenia among Chinese community-dwelling older adults using sEMG-based data. *J. Neuroeng. Rehabil.* **2024**, *21*, 69. [CrossRef] [PubMed]
53. Abdullah, M.I.; Hao, S.K.; Abdullah, I.; Faizah, S. Parkinson's Disease Symptom Detection using Hybrid Feature Extraction and Classification Model. In Proceedings of the 2023 IEEE 14th Control and System Graduate Research Colloquium, ICSGRC 2023—Conference Proceeding, Shah Alam, Malaysia, 5 August 2023; pp. 93–98.
54. Ding, Z.; Zhang, L.; Zhang, Y.; Yang, J.; Luo, Y.; Ge, M.; Chen, C. A Supervised Explainable Machine Learning Model for Perioperative Neurocognitive Disorder in Liver-Transplantation Patients and External Validation on the Medical Information Mart for Intensive Care IV Database: Retrospective Study. *J. Med. Internet Res.* **2025**, *27*, e55046. [CrossRef]
55. Liang, L.; Liang, X.; Yu, X.; Xiang, W. Bioinformatic Analyses and Integrated Machine Learning to Predict Prognosis and Therapeutic Response Based on E3 Ligase-Related Genes in Colon Cancer. *J. Cancer* **2024**, *15*, 5376–5395. [CrossRef]
56. Kumar, R.; Garg, S.; Kaur, R.; Johar, M.G.M.; Singh, S.; Menon, S.V.; Kumar, P.; Hadi, A.M.; Hasson, S.A.; Lozanović, J. A comprehensive review of machine learning for heart disease prediction: Challenges, trends, ethical considerations, and future directions. *Front. Artif. Intell.* **2025**, *8*, 1583459. [CrossRef]
57. Coskuner-Weber, O.; Alpsoy, S.; Yolcu, O.; Teber, E.; de Marco, A.; Shumka, S. Metagenomics studies in aquaculture systems: Big data analysis, bioinformatics, machine learning and quantum computing. *Comput. Biol. Chem.* **2025**, *118*, 108444. [CrossRef]
58. Sharma, P.; Sharma, P.; Sharma, K.; Varma, V.; Patel, V.; Sarvaiya, J.; Tavethia, J.; Mehta, S.; Bhadania, A.; Patel, I.; et al. Revolutionizing Utility of Big Data Analytics in Personalized Cardiovascular Healthcare. *Bioengineering* **2025**, *12*, 463. [CrossRef]
59. Liu, Z.; Si, L.; Shi, S.; Li, J.; Zhu, J.; Lee, W.H.; Wang, G. Classification of Three Anesthesia Stages Based on Near-Infrared Spectroscopy Signals. *IEEE J. Biomed. Health Informat.* **2024**, *28*, 5270–5279. [CrossRef]
60. Mahawar, K.; Rattan, P.; Jalamneh, A.; Ab Yajid, M.S.; Abdeljaber, O.; Kumar, R.; Lasisi, A.; Ammarullah, M.I. Employing artificial bee and ant colony optimization in machine learning techniques as a cognitive neuroscience tool. *Sci. Rep.* **2025**, *15*, 10172. [CrossRef]
61. Wang, F. Adaptive semi-supervised recursive tree partitioning: The ART towards large scale patient indexing in personalized healthcare. *J. Biomed. Inform.* **2015**, *55*, 41–54. [CrossRef]
62. Tsay, D.; Patterson, C. From Machine Learning to Artificial Intelligence Applications in Cardiac Care: Real-World Examples in Improving Imaging and Patient Access. *Circulation* **2018**, *138*, 2569–2575. [CrossRef]
63. Aris-Brosou, S.; Kim, J.; Li, L.; Liu, H. Predicting the reasons of customer complaints: A first step toward anticipating quality issues of in vitro diagnostics assays with machine learning. *JMIR Med. Inform.* **2018**, *6*, e34. [CrossRef] [PubMed]
64. Moreira, M.W.L.; Rodrigues, J.J.P.C.; Kumar, N.; Saleem, K.; Illin, I.V. Postpartum depression prediction through pregnancy data analysis for emotion-aware smart systems. *Inf. Fusion* **2019**, *47*, 23–31. [CrossRef]
65. Rim, T.H.; Soh, Z.D.; Tham, Y.C.; Yang, H.H.S.; Lee, G.; Kim, Y.; Nusinovici, S.; Ting, D.S.W.; Wong, T.Y.; Cheng, C.Y. Deep Learning for Automated Sorting of Retinal Photographs. *Ophthalmol. Retin.* **2020**, *4*, 793–800. [CrossRef] [PubMed]
66. Bolhasani, H.; Mohseni, M.; Rahmani, A.M. Deep learning applications for IoT in health care: A systematic review. *Inform. Med. Unlocked* **2021**, *23*, 100550. [CrossRef]
67. Saleh, M.; Abbas, M.; Prud'Homm, J.; Somme, D.; Le Bouquin Jeannes, R. A Reliable Fall Detection System Based on Analyzing the Physical Activities of Older Adults Living in Long-Term Care Facilities. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2021**, *29*, 2587–2594. [CrossRef]
68. Montomoli, J.; Hilty, M.P.; Ince, C. Artificial intelligence in intensive care: Moving towards clinical decision support systems. *Minerva Anestesiol.* **2022**, *88*, 1066–1072. [CrossRef]
69. Nicolucci, A.; Romeo, L.; Bernardini, M.; Vespasiani, M.; Rossi, M.C.; Petrelli, M.; Ceriello, A.; Di Bartolo, P.; Frontoni, E.; Vespasiani, G. Prediction of complications of type 2 Diabetes: A Machine learning approach. *Diabetes Res. Clin. Pract.* **2022**, *190*, 110013. [CrossRef]
70. Fatema, N.; Malik, H.; Ahmad, W. Data driven intelligent model for quality management in healthcare. *J. Intell. Fuzzy Syst.* **2022**, *42*, 1155–1169. [CrossRef]
71. Singh, G.; Violi, V.; Fisichella, M. Federated Learning to Safeguard Patients Data: A Medical Image Retrieval Case. *Big Data Cogn. Comput.* **2023**, *7*, 18. [CrossRef]
72. Abidi, M.H.; Umer, U.; Mian, S.H.; Al-Ahmari, A. Big Data-Based Smart Health Monitoring System: Using Deep Ensemble Learning. *IEEE Access* **2023**, *11*, 114880–114903. [CrossRef]
73. Manchadi, O.; Ben-Bouazza, F.E.; Jioudi, B. Predictive Maintenance in Healthcare System: A Survey. *IEEE Access* **2023**, *11*, 61313–61330. [CrossRef]
74. Alqahtani, S.; Luo, S.; Alanazi, M.; Shaukat, K.; Alsubaie, M.G.; Amer, M. Machine Learning for Predicting Intradialytic Hypotension: A Survey Review. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*, 282–293. [CrossRef]

75. Shao, Y.; Zhang, S.; Raman, V.K.; Patel, S.S.; Cheng, Y.; Parulkar, A.; Lam, P.H.; Moore, H.; Sheriff, H.M.; Fonarow, G.C.; et al. Artificial intelligence approaches for phenotyping heart failure in U.S. Veterans Health Administration electronic health record. *ESC Heart Fail.* **2024**, *11*, 3155–3166. [CrossRef]
76. Lin, X.; Duan, G.; Huang, J.; Zhou, Q.; Huang, H.; Xiao, J.; Xu, Z.; Shen, H.; Zhuo, H. Construction of A Smart Hospital Innovation Platform Using the Internet + Technology. *Altern. Ther. Health Med.* **2024**, *30*, 495–505.
77. Bartoszewicz, M.; Prokop, P.; Kosieradzki, M.; Fiedor, P. Are Current Educational and Therapeutic Programs, Directed at Rare Disease Transplant Candidates and Recipients, Sufficient to Support Them on the Path From Diagnosis to Life After Allogenic Transplantation?—Recommendations for Member State Policymakers. *Transplant. Proc.* **2024**, *56*, 907–909. [CrossRef]
78. Joshi, A. Big data and AI for gender equality in health: Bias is a big challenge. *Front. Big Data* **2024**, *7*, 1436019. [CrossRef]
79. Balogh, D.B.; Hudelist, G.; Bļizņuks, D.; Raghothama, J.; Becker, C.M.; Horace, R.; Krentel, H.; Horne, A.W.; Bourdel, N.; Marki, G.; et al. FEMaLe: The use of machine learning for early diagnosis of endometriosis based on patient self-reported data—Study protocol of a multicenter trial. *PLoS ONE* **2024**, *19*, e0300186. [CrossRef]
80. Finazzi Agrò, E.; Rosato, E.; Kheir, G.B.; Rademakers, K.; Averbeck, M.A.; Tarcan, T.; Hashim, H.; Gammie, A.; Sinha, S.; Song, Q.X.; et al. How Can We Show That Artificial Intelligence May Improve Our Assessment and Management of Lower Urinary Tract Dysfunctions?—ICI-RS 2024. *Neurourol. Urodyn.* **2025**, *44*, 616–621. [CrossRef]
81. Yajie, L.; Johar, M.G.M.; Hajamydeen, A.I. Poultry disease early detection methods using deep learning technology. *Indones. J. Electr. Eng. Comput. Sci.* **2023**, *32*, 1712–1723. [CrossRef]
82. Dissanayake, K.; Johar, M.G.M. Comparative study on heart disease prediction using feature selection techniques on classification algorithms. *Appl. Comput. Intell. Soft Comput.* **2021**, *2021*, 1–17. [CrossRef]
83. UK Biobank. *About Us—UK Biobank*; UK Biobank: Stockport, UK, 2025.
84. Collins, R. UK Biobank is safely sharing health data to drive medical research. *The Guardian*, 18 April 2025.
85. Ciborowski, J. NVIDIA and GE HealthCare Collaborate to Advance the Development of Autonomous Diagnostic Imaging With Physical AI. *NVIDIA Newsroom*, 18 March 2025.
86. Islam, M.D.S.; Liu, D.; Wang, K.; Zhou, P.; Yu, L.; Wu, D. A Case Study of HealthCare Platform using Big Data Analytics and Machine Learning. In Proceedings of the 2019 3rd High Performance Computing and Cluster Technologies Conference, Guangzhou, China, 22–24 June 2019; pp. 139–146.
87. Wan, A.E.; Baber Khan, M.S.; Xian Teo, B.S.; Khan, J.; Abdullah, I.; Kaleemullah, M.; Asmani, F.; Suofeiya, M.; Al Dhalli, S.; Kasim, Z.; et al. Screening of antioxidant and antibacterial activity of methanolic extract of ipomoea aquatica leaf and stem against bacteria causes skin infection. *Int. J. Med. Toxicol. Leg. Med.* **2020**, *23*, 169–178. [CrossRef]
88. Qian, T.; Zhu, S.; Hoshida, Y. Use of big data in drug development for precision medicine: An update. *Expert. Rev. Precis. Med. Drug Dev.* **2019**, *4*, 189–200. [CrossRef] [PubMed]
89. Jain, A.K.; Aneja, S.; Fuller, C.D.; Dicker, A.P.; Chung, C.; Kim, E.; Kirby, J.S.; Quon, H.; Lam, C.J.K.; Louv, W.C.; et al. Provider Engagement in Radiation Oncology Data Science: Workshop Report. *JCO Clin. Cancer Inf.* **2020**, *4*, 700–710. [CrossRef] [PubMed]
90. Oncora Medical. Oncology AI Solutions. 2025. Available online: https://www.oncora.ai/ (accessed on 7 May 2025).
91. IQVIA Inc. IQVIA NLP Data Factory: Enterprise NLP Made Easy [Video]. 2021. Available online: https://www.iqvia.com/library/videos/iqvia-nlp-data-factory (accessed on 25 July 2025).
92. IQVIA Inc. Big Data Analytics for Population Health [Case Study]. 7 March 2023. Available online: https://www.iqvia.com/library/case-studies/nlp-big-data-analytics-for-population-health (accessed on 9 May 2025).
93. IQVIA Inc. IQVIA Natural Language Processing (NLP) Risk Adjustment Solution [Fact Sheet]. 3 January 2023. Available online: https://www.iqvia.com/library/fact-sheets/nlp-risk-adjustment-solution (accessed on 15 May 2025).
94. López-Martínez, F.; Núñez-Valdez, E.R.; García-Díaz, V.; Bursac, Z. A Case Study for a Big Data and Machine Learning Platform to Improve Medical Decision Support in Population Health Management. *Algorithms* **2020**, *13*, 102. [CrossRef]
95. Salari, M.; Rahmani, B. Machine Learning for Everyone: Simplifying Healthcare Analytics with BigQuery ML. *arXiv* **2025**, arXiv:2502.07026. [CrossRef]
96. Hennebelle, A.; Materwala, H.; Ismail, L. HealthEdge: A Machine Learning-Based Smart Healthcare Framework for Prediction of Type 2 Diabetes in an Integrated IoT, Edge, and Cloud Computing System. *Procedia Comput. Sci.* **2023**, *220*, 331–338. [CrossRef]
97. Radiological Society of North America (RSNA). AI Predicts Heart Disease Risk Using Single X-Ray [Press Release]. 29 November 2022. Available online: https://press.rsna.org/timssnet/media/pressreleases/14_pr_target.cfm?id=2388 (accessed on 14 May 2025).
98. Wang, Y.; Shen, Y.; Li, J.; Wang, T.; Peng, J.; Shang, X. Enhanced RNA secondary structure prediction through integrative deep learning and structural context analysis. *Nucleic Acids Res.* **2025**, *53*, gkaf533. [CrossRef]
99. Li, H.; Wang, Z.; Guan, Z.; Miao, J.; Li, W.; Yu, P.; Molina Jimenez, C. UCFNNet: Ulcerative colitis evaluation based on fine-grained lesion learner and noise suppression gating. *Comput. Methods Programs Biomed.* **2024**, *247*, 108080. [CrossRef] [PubMed]
100. An, Q.; Chen, W.; Shao, W. A Deep Convolutional Neural Network for Pneumonia Detection in X-ray Images with Attention Ensemble. *Diagnostics* **2024**, *14*, 390. [CrossRef]

101. Mabrouk, A.; Díaz Redondo, R.P.; Dahou, A.; Abd Elaziz, M.; Kayed, M. Pneumonia Detection on Chest X-ray Images Using Ensemble of Deep Convolutional Neural Networks. *Appl. Sci.* **2022**, *12*, 6448. [CrossRef]
102. Loukou, E. Fitness Trackers Could Spot Earliest Signs of Alzheimer's BEFORE Symptoms Start. *The Sun*, 19 March 2024. Available online: https://www.thesun.co.uk/health/26783341/fitness-trackers-signs-of-alzheimers-before-symptoms-start (accessed on 11 May 2025).
103. Iqbal, U.; Tanweer, A.; Rahmanti, A.R.; Greenfield, D.; Lee, L.T.J.; Li, Y.C.J. Impact of large language model (ChatGPT) in healthcare: An umbrella review and evidence synthesis. *J. Biomed. Sci.* **2025**, *32*, 45. [CrossRef] [PubMed]
104. Akbasli, I.T.; Birbilen, A.Z.; Teksam, O. Leveraging large language models to mimic domain expert labeling in unstructured text-based electronic healthcare records in non-english languages. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 154. [CrossRef] [PubMed]
105. Khan, B.; Talha Khalid, R.; Ul Wara, K.; Hasan Masrur, M.; Khan, S.; Khan, W.U.; Amara, U.; Abdullah, S. Reshaping the healthcare world by AI-integrated wearable sensors following COVID-19. *Chem. Eng. J.* **2025**, *505*, 159478. [CrossRef]
106. Yousefi, M.; Akhbari, M.; Mohamadi, Z.; Karami, S.; Dasoomi, H.; Atabi, A.; Sarkeshikian, S.A.; Abdoullahi Dehaki, M.; Bayati, H.; Mashayekhi, N.; et al. Machine learning based algorithms for virtual early detection and screening of neurodegenerative and neurocognitive disorders: A systematic-review. *Front. Neurol.* **2024**, *15*, 1413071. [CrossRef]
107. Singh, S.G.; Das, D.; Barman, U.; Saikia, M.J. Early Alzheimer's Disease Detection: A Review of Machine Learning Techniques for Forecasting Transition from Mild Cognitive Impairment. *Diagnostics* **2024**, *14*, 1759. [CrossRef]
108. Wang, K.; Tan, B.; Wang, X.; Qiu, S.; Zhang, Q.; Wang, S.; Yen, Y.T.; Jing, N.; Liu, C.; Chen, X.; et al. Machine learning-assisted point-of-care diagnostics for cardiovascular healthcare. *Bioeng. Transl. Med.* **2025**, *10*, e70002. [CrossRef]
109. Bhattacherjee, A.; Badhan, A.K. Convergence of Data Analytics, Big Data, and Machine Learning: Applications, Challenges, and Future Direction. In *Data Analytics and Machine Learning: Navigating the Big Data Landscape*; Springer Nature: Singapore, 2024; Volume 145, pp. 317–334.
110. McComb, M.; Bies, R.; Ramanathan, M. Machine learning in pharmacometrics: Opportunities and challenges. *Br. J. Clin. Pharmacol.* **2022**, *88*, 1482–1499. [CrossRef]
111. Kaur, R.; Kumar, R.; Aggarwal, H. Systematic Review of Artificial Intelligence, Machine Learning, and Deep Learning in Machining Operations: Advancements, Challenges, and Future Directions. *Arch. Comput. Methods Eng.* **2025**, 1–54. [CrossRef]
112. Tsang, G.; Xie, X.; Zhou, S.M. Harnessing the Power of Machine Learning in Dementia Informatics Research: Issues, Opportunities, and Challenges. *IEEE Rev. Biomed. Eng.* **2020**, *13*, 113–129. [CrossRef] [PubMed]
113. Zhang, Y.; Liu, C.; Chen, X.; Zhang, Y.; Li, Y.; Hu, X. Effects of Web-Based Acceptance and Commitment Therapy on Health-Related Outcomes Among Patients With Lung Cancer: A Feasibility Randomized Controlled Trial. *Psycho-Oncol.* **2024**, *33*, e70045. [CrossRef] [PubMed]
114. Swanson, K.; Wu, E.; Zhang, A.; Alizadeh, A.A.; Zou, J. From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell* **2023**, *186*, 1772–1791. [CrossRef] [PubMed]
115. Alafif, T.; Tehame, A.M.; Bajaba, S.; Barnawi, A.; Zia, S. Machine and deep learning towards COVID-19 diagnosis and treatment: Survey, challenges, and future directions. *Int. J. Environ. Res. Public. Health* **2021**, *18*, 1117. [CrossRef] [PubMed]
116. Giger, M.L. Machine learning in medical imaging. *J. Am. Coll. Radiol.* **2018**, *15*, 512–520. [CrossRef]
117. Kumar, A.; Bi, L.; Kim, J.; Feng, D.D. Machine learning in medical imaging. In *Biomedical Information Technology*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 167–196.
118. Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477. [CrossRef]
119. Mak, K.-K.; Wong, Y.-H.; Pichika, M.R. Artificial intelligence in drug discovery and development. In *Drug Discovery and Evaluation: Safety and Pharmacokinetic Assays*; Hock, F.J., Pugsley, M.K., Eds.; Springer: Cham, Switzerland, 2024; pp. 1461–1498. [CrossRef]
120. Al Meslamani, A.Z.; Sobrino, I.; de la Fuente, J. Machine learning in infectious diseases: Potential applications and limitations. *Ann. Med.* **2024**, *56*, 2362869. [CrossRef]
121. Badal, V.D.; Depp, C.A. Natural language processing of medical records: New understanding of suicide ideation by dementia subtypes: Commentary on "Suicidal ideation in dementia: Associations with neuropsychiatric symptoms and subtype diagnosis" by Naismith et al. *Int. Psychogeriatr.* **2022**, *34*, 319–321. [CrossRef]
122. Kourou, K.; Exarchos, K.P.; Papaloukas, C.; Sakaloglou, P.; Exarchos, T.; Fotiadis, D.I. Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 5546–5555. [CrossRef]
123. Rahmani, A.M.; Yousefpoor, E.; Yousefpoor, M.S.; Mehmood, Z.; Haider, A.; Hosseinzadeh, M.; Ali Naqvi, R. Machine learning (ML) in medicine: Review, applications, and challenges. *Mathematics* **2021**, *9*, 2970. [CrossRef]
124. Afrifa-Yamoah, E.; Adua, E.; Peprah-Yamoah, E.; Anto, E.O.; Opoku-Yamoah, V.; Acheampong, E.; Macartney, M.J.; Hashmi, R. Pathways to chronic disease detection and prediction: Mapping the potential of machine learning to the pathophysiological processes while navigating ethical challenges. *Chronic Dis. Transl. Med.* **2025**, *11*, 1–21. [CrossRef] [PubMed]

125. Red, G.V.; Palaoag, T.D. Ethical and Privacy Issues in the Use of Machine Learning for Personalized Care for Elderly Patients. *Int. J. Eng. Trends Technol.* **2024**, *72*, 32–42. [CrossRef]
126. Kostick-Quenet, K.; Estep, J.; Blumenthal-Barby, J.S. Ethical Concerns for Remote Computer Perception in Cardiology: New Stages for Digital Health Technologies, Artificial Intelligence, and Machine Learning. *Circ. Cardiovasc. Qual. Outcomes* **2024**, *17*, E010717. [CrossRef]
127. Hajlaoui, R.; Dhahri, S.; Mahfoudhi, S.; Moulahi, T.; Alotibi, G. Protecting machine learning systems using blockchain: Solutions, challenges and future prospects. *Multimed. Tools Appl.* **2025**, *20*, 22755–22782. [CrossRef]
128. Geerthik, S.; Senthil, G.A.; Illakiya, G.; Naveen Raj, S.; Arun Kumar, B. Future Navigation Demand Trends: Accurate Ride Request Forecasting Optimization Using Machine Learning. In Proceedings of the 2024 3rd International Conference on Smart Technologies and Systems for Next Generation Computing, ICSTSN 2024, Villupuram, India, 18–19 July 2024.

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

Diagnostics Editorial Office
E-mail: diagnostics@mdpi.com
www.mdpi.com/journal/diagnostics

Disclaimer/Publisher's Note: The title and front matter of this reprint are at the discretion of the Guest Editor. The publisher is not responsible for their content or any associated concerns. The statements, opinions and data contained in all individual articles are solely those of the individual Editor and contributors and not of MDPI. MDPI disclaims responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-8638-8