diagnostics

Special Issue Reprint

Artificial Intelligence in Clinical Medical Imaging

2nd Edition

Edited by
Daniele Giansanti

mdpi.com/journal/diagnostics

MDPI

Artificial Intelligence in Clinical Medical Imaging: 2nd Edition

Artificial Intelligence in Clinical Medical Imaging: 2nd Edition

Guest Editor

Daniele Giansanti

MDPI

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

Guest Editor
Daniele Giansanti
Centre Tisp
Istituto Superiore di Sanità
Rome
Italy

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/KN912L0L8Y.

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 978-3-7258-8433-9 (Hbk)
ISBN 978-3-7258-8434-6 (PDF)
https://doi.org/10.3390/books978-3-7258-8434-6

© 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**

Daniele Giansanti
Revolutionizing Medical Imaging: The Transformative Role of Artificial Intelligence in Diagnostics and Treatment
Reprinted from: *Diagnostics* **2025**, *15*, 1557, https://doi.org/10.3390/diagnostics15121557 . . . . **1**

Maria Chiara Brunese, Aldo Rocca, Antonella Santone, Mario Cesarelli, Luca Brunese and Francesco Mercaldo
Explainable and Robust Deep Learning for Liver Segmentation Through U-Net Network
Reprinted from: *Diagnostics* **2025**, *15*, 878, https://doi.org/10.3390/diagnostics15070878 . . . . . **9**

Nakarin Inmutto, Suwalee Pojchamarnwiputh and Wittanee Na Chiangmai
Multiphase Computed Tomography Scan Findings for Artificial Intelligence Training in the Differentiation of Hepatocellular Carcinoma and Intrahepatic Cholangiocarcinoma Based on Interobserver Agreement of Expert Abdominal Radiologists
Reprinted from: *Diagnostics* **2025**, *15*, 821, https://doi.org/10.3390/diagnostics15070821 . . . . . **45**

Sevda Gül, Gökçen Cetinel, Bekir Murat Aydin, Devrim Akgün and Rabia Öztaş Kara
YOLOSAMIC: A Hybrid Approach to Skin Cancer Segmentation with the Segment Anything Model and YOLOv8
Reprinted from: *Diagnostics* **2025**, *15*, 479, https://doi.org/10.3390/diagnostics15040479 . . . . . **56**

Eliodoro Faiella, Matteo Pileri, Raffaele Ragone, Anna Maria De Nicola, Bruno Beomonte Zobel, Rosario Francesco Grasso and Domiziana Santucci
Promising Results About the Possibility to Identify Prostate Cancer Patients Employing a Random Forest Classifier: A Preliminary Study Preoperative Patients Selection
Reprinted from: *Diagnostics* **2025**, *15*, 421, https://doi.org/10.3390/diagnostics15040421 . . . . . **82**

Ramazan Berkay Peker and Celal Oguz Kurtoglu
Evaluation of the Performance of a YOLOv10-Based Deep Learning Model for Tooth Detection and Numbering on Panoramic Radiographs of Patients in the Mixed Dentition Period
Reprinted from: *Diagnostics* **2025**, *15*, 405, https://doi.org/10.3390/diagnostics15040405 . . . . . **94**

Hung-Hsien Liu, Chun-Bi Chang, Yi-Sa Chen, Chang-Fu Kuo, Chun-Yu Lin, Cheng-Yu Ma and Li-Jen Wang
Automated Detection and Differentiation of Stanford Type A and Type B Aortic Dissections in CTA Scans Using Deep Learning
Reprinted from: *Diagnostics* **2025**, *15*, 12, https://doi.org/10.3390/diagnostics15010012 . . . . . **111**

Lingeer Wu, Yijun Ling, Ling Lan, Kai He, Chunhua Yu, Zhuhuang Zhou and Le Shen
Automatic Segmentation of the Left Ventricle in Apical Four-Chamber View on Transesophageal Echocardiography Based on UNeXt Deep Neural Network
Reprinted from: *Diagnostics* **2024**, *14*, 2766, https://doi.org/10.3390/diagnostics14232766 . . . . **122**

Yousra Hadhoud, Tahar Mekhaznia, Akram Bennour, Mohamed Amroune, Neesrin Ali Kurdi, Abdulaziz Hadi Aborujilah and Mohammed Al-Sarem
From Binary to Multi-Class Classification: A Two-Step Hybrid CNN-ViT Model for Chest Disease Classification Based on X-Ray Images
Reprinted from: *Diagnostics* **2024**, *14*, 2754, https://doi.org/10.3390/diagnostics14232754 . . . . **137**

Marilena Ianculescu, Corina Petean, Virginia Sandulescu, Adriana Alexandru and Ana-Mihaela Vasilevschi
Early Detection of Parkinson's Disease Using AI Techniques and Image Analysis
Reprinted from: *Diagnostics* **2024**, *14*, 2615, https://doi.org/10.3390/diagnostics14232615 . . . . **153**

Raffaele Pellegrino, Alessandro Federico and Antonietta Gerarda Gravina
Conversational LLM Chatbot ChatGPT-4 for Colonoscopy Boston Bowel Preparation Scoring: An Artificial Intelligence-to-Head Concordance Analysis
Reprinted from: *Diagnostics* **2024**, *14*, 2537, https://doi.org/10.3390/diagnostics14222537 . . . . **172**

Miguel-Angel Gil-Rios, Ivan Cruz-Aceves, Arturo Hernandez-Aguirre, Martha-Alicia Hernandez-Gonzalez and Sergio-Eduardo Solorio-Meza
Improving Automatic Coronary Stenosis Classification Using a Hybrid Metaheuristic with Diversity Control
Reprinted from: *Diagnostics* **2024**, *14*, 2372, https://doi.org/10.3390/diagnostics14212372 . . . . **181**

Yuan Tian, Ruiyang Gao, Xinran Shi, Jiaxin Lang, Yang Xue, Chunrong Wang, et al.
U-Net and Its Variants Based Automatic Tracking of Radial Artery in Ultrasonic Short-Axis Views: A Pilot Study
Reprinted from: *Diagnostics* **2024**, *14*, 2358, https://doi.org/10.3390/diagnostics14212358 . . . . **200**

Barbara Marquez, Zachary T. Wooten, Ramon M. Salazar, Christine B. Peterson, David T. Fuentes, T. J. Whitaker, et al.
Analyzing the Relationship between Dose and Geometric Agreement Metrics for Auto-Contouring in Head and Neck Normal Tissues
Reprinted from: *Diagnostics* **2024**, *14*, 1632, https://doi.org/10.3390/diagnostics14151632 . . . . **214**

Luca Urso, Corrado Cittanti, Luigi Manco, Naima Ortolan, Francesca Borgia, Antonio Malorgio, et al.
ML Models Built Using Clinical Parameters and Radiomic Features Extracted from ^{18}F-Choline PET/CT for the Prediction of Biochemical Recurrence after Metastasis-Directed Therapy in Patients with Oligometastatic Prostate Cancer
Reprinted from: *Diagnostics* **2024**, *14*, 1264, https://doi.org/10.3390/diagnostics14121264 . . . . **227**

Ye Rin Kim, Yu Sung Yoon and Jang Gyu Cha
Opportunistic Screening for Acute Vertebral Fractures on a Routine Abdominal or Chest Computed Tomography Scans Using an Automated Deep Learning Model
Reprinted from: *Diagnostics* **2024**, *14*, 781, https://doi.org/10.3390/diagnostics14070781 . . . . . **238**

Andrea Lastrucci, Nicola Iosca, Yannick Wandael, Angelo Barra, Graziano Lepri, Nevio Forini, et al.
AI and Interventional Radiology: A Narrative Review of Reviews on Opportunities, Challenges, and Future Directions
Reprinted from: *Diagnostics* **2025**, *15*, 893, https://doi.org/10.3390/diagnostics15070893 . . . . . **249**

Michele Mercurio, Federica Denami, Dimitra Melissaridou, Katia Corona, Simone Cerciello, Domenico Laganà, et al.
Deep Learning Models to Detect Anterior Cruciate Ligament Injury on MRI: A Comprehensive Review
Reprinted from: *Diagnostics* **2025**, *15*, 776, https://doi.org/10.3390/diagnostics15060776 . . . . . **292**

Mario Romeo, Marcello Dallio, Carmine Napolitano, Claudio Basile, Fiammetta Di Nardo, Paolo Vaia, Patrizia Iodice and Alessandro Federico
Clinical Applications of Artificial Intelligence (AI) in Human Cancer: Is It Time to Update the Diagnostic and Predictive Models in Managing Hepatocellular Carcinoma (HCC)?
Reprinted from: *Diagnostics* **2025**, *15*, 252, https://doi.org/10.3390/diagnostics15030252 . . . . . **308**

Andrea Lastrucci, Yannick Wandael, Renzo Ricci, Giovanni Maccioni and Daniele Giansanti
The Integration of Deep Learning in Radiotherapy: Exploring Challenges, Opportunities, and Future Directions through an Umbrella Review
Reprinted from: *Diagnostics* **2024**, *14*, 939, https://doi.org/10.3390/diagnostics14090939 . . . . . **330**

Puteri Norliza Megat Ramli, Azimatun Noor Aizuddin, Norfazilah Ahmad, Zuhanis Abdul Hamid and Khairil Idham Ismail
A Systematic Review: The Role of Artificial Intelligence in Lung Cancer Screening in Detecting Lung Nodules on Chest X-Rays
Reprinted from: *Diagnostics* **2025**, *15*, 246, https://doi.org/10.3390/diagnostics15030246 . . . . . **357**

About the Editor

Daniele Giansanti

Daniele Giansanti is a professor at Sapienza University and the Catholic University in Rome, Italy. His research interests span biomedical engineering, medical physics, and e-Health, with a focus on developing wearable medical devices, advancing telemedicine and mHealth applications, and studying the acceptance of robotics and artificial intelligence in healthcare. He currently leads research as Research Director at the Italian National Institute of Health (ISS), where his projects include technology assessment for telerehabilitation and domiciliary monitoring, with emphasis on the design, evaluation, and integration of wearable and digital health technologies for remote patient care, including sensor-based systems, mobile health solutions, and AI-enabled assistive devices. His research also explores advanced data-driven approaches in healthcare, including artificial intelligence applications in medical imaging, with particular attention to radiology and digital pathology, as well as their role in supporting clinical decision-making and personalized care. In parallel, he investigates the ethical, legal, and cybersecurity aspects of digital health, focusing on the safe and responsible integration of artificial intelligence, robotics, and connected technologies in clinical and home-care environments. Dr. Giansanti has authored over 200 Scopus-indexed publications and serves as a board editor and reviewer for several scientific journals.

Preface

The field of medical imaging is undergoing a profound transformation, driven by the rapid integration of artificial intelligence (AI) into clinical workflows and research methodologies. AI technologies, particularly those based on machine learning and deep learning, are demonstrating great potential to enhance the precision, efficiency, and personalization of diagnostic and therapeutic processes. These innovations are shifting medical imaging from traditional manual interpretation toward data-driven, automated, and assistive systems that support clinical decision-making at unprecedented levels.

AI applications in medical imaging span a wide range of areas, including image classification, segmentation, feature extraction, disease progression prediction, multimodal image integration, and treatment planning. These tools are increasingly being integrated into clinical practice, supporting healthcare professionals in tailoring interventions and improving patient outcomes. At the same time, AI adoption raises important considerations regarding data privacy, security, standardization, and ethical use of patient information. Addressing these issues is essential to ensure that technological progress aligns with the core values of medical care.

This Special Issue provides a scientific forum for international scholars to share and exchange ideas on AI applications in medical imaging. It includes 20 contributions plus the concluding editorial, encompassing original research and review articles that collectively highlight the transformative impact of AI across various medical specialties. Research articles in this Special Issue demonstrate advancements in diagnostic accuracy, workflow efficiency, and personalized patient care, showcasing AI's role in oncology, cardiology, dermatology, radiotherapy, neurology, and dental imaging. Complementary review articles offer a broader perspective on current trends, challenges, and future directions, covering topics such as interventional radiology, musculoskeletal imaging, oncology, radiotherapy, and screening practices.

I would like to sincerely thank all the authors, reviewers, and contributors whose outstanding work made the reprint of this Special Issue possible. I also extend my gratitude to the editorial team for their support and dedication, and to Managing Editor Dennis Zhu for his exceptional support and continued guidance. Their combined efforts have been essential in bringing this Special Issue to fruition.

Daniele Giansanti
Guest Editor

diagnostics

MDPI

Editorial

Revolutionizing Medical Imaging: The Transformative Role of Artificial Intelligence in Diagnostics and Treatment

Daniele Giansanti

Centre Tisp, Istituto Superiore di Sanità, 00161 Rome, Italy; daniele.giansanti@iss.it; Tel.: +39-06-49902701

The field of medical imaging is undergoing a significant transformation, driven by the rapid integration of artificial intelligence (AI) into clinical workflows and research methodologies. AI technologies, particularly those based on machine learning and deep learning, have demonstrated immense potential to enhance the precision, efficiency, and personalization of diagnostic and therapeutic processes [1]. These advancements are contributing to a paradigm shift from the traditional, often manual interpretation of medical images to data-driven, automated, and assistive systems that can support clinical decision-making at unprecedented levels [2].

The spectrum of AI applications in medical imaging is growing. Among the most impactful applications are deep learning algorithms designed for image classification and analysis, techniques for robust and accurate image segmentation, and feature extraction methods that can reveal subtle patterns often invisible to the human eye [3]; these innovations have significantly improved the performance of computer-aided diagnosis systems, enabling the earlier and more reliable detection of a wide range of pathologies.

Moreover, AI plays a central role in the prediction of disease progression [4] using longitudinal imaging data, the registration and fusion of multimodal images for more comprehensive assessments, and the development of image-based tools for treatment planning and response evaluation; these tools are increasingly integrated into clinical environments, supporting radiologists, oncologists, and other specialists in tailoring patient-specific interventions [2–4].

Alongside these technical advancements, the integration of AI into medical imaging raises important questions about data governance, including issues of data privacy, security, and the ethical use of patient information [5]. Standardizing imaging protocols with AI is also crucial to ensure reproducibility and interoperability across institutions and platforms [6]. Furthermore, the social and ethical implications of adopting AI in healthcare settings requires careful consideration [7], especially in terms of transparency, accountability, and its potential impact on clinical roles and patient trust [8].

Overall, AI is reshaping the landscape of medical imaging, offering powerful tools to support clinicians and improve patient outcomes. At the same time, it calls for a thoughtful and collaborative approach to ensure that technological progress aligns with the fundamental values of medical care.

With the aim of exploring the evolving landscape of research in this field, and following the success of the first edition of this Special Issue (available online at https://www.mdpi.com/journal/diagnostics/special_issues/3FXN9682V0, accessed on 17 April 2025).

We), we decided to launch a second Special Issue intended to serve as a scientific forum for international scholars to meet and exchange ideas (available online at https://www.mdpi.com/journal/diagnostics/special_issues/KN912L0L8Y, accessed on 17 April 2025).

The Special Issue, in addition to this summarizing editorial, includes 20 contributions (contributions 1–20), 15 research articles (contributions 1–15), four reviews (contributions 16–19), and a systematic review (contribution 20).

Outcomes of the scientific articles

The analyses conducted in the scientific articles reveal how the integration of artificial intelligence (AI) in medical imaging and diagnostics is revolutionizing healthcare, with a wide array of applications across different specialties. These studies collectively underscore the significant advancements in AI-driven methods for disease detection, prediction, and treatment planning, showcasing how AI is reshaping clinical workflows and enhancing diagnostic accuracy.

In the field of oncology, AI models are improving diagnostic precision. For example, Brunese et al. (contribution 1) presented a U-Net-based deep learning model for liver segmentation in CT and MRI scans, focusing on its explainability and robustness in clinical settings, particularly for hepatocellular carcinoma cases. This method allows for the clearer delineation of liver structures, which is crucial for treatment planning and monitoring. Similarly, Inmutto et al. (contribution 2) explored AI's ability to differentiate between hepatocellular carcinoma and intrahepatic cholangiocarcinoma using multiphase CT scans, highlighting the model's performance and interobserver agreement among expert radiologists. The enhanced diagnostic capabilities provided by these AI tools have the potential to significantly improve patient outcomes in liver cancer treatment.

In dermatology, AI applications are making early skin cancer detection more accurate. Gül et al. (contribution 3) introduced a hybrid model combining YOLOv8 and SAM which automates skin lesion detection and segmentation. This approach significantly improves the speed and accuracy of identifying skin cancers such as melanoma, facilitating early intervention and better prognosis for patients.

Regarding prostate cancer diagnostics, Faiella et al. (contribution 4) demonstrated the use of a Random Forest model for predicting lymph node involvement in prostate cancer using mp-MRI data and radiomic features. This method aids in preoperative planning, helping clinicians make informed decisions about treatment strategies. Similarly, Urso et al. (contribution 14) utilized machine learning (ML) models to predict biochemical recurrence in prostate cancer patients following metastasis-directed therapy, based on radiomic features from 18F-choline PET/CT scans. This predictive capability is valuable in tailoring personalized treatment plans and improving patient outcomes.

AI is also transforming the field of cardiology, where it is being applied to enhance diagnostic efficiency. Liu et al. (contribution 6) presented a deep learning model for detecting aortic dissections in CTA scans, specifically Stanford type A and B dissections, a critical condition requiring rapid diagnosis. The model improves diagnostic speed and accuracy, reducing the risk of delayed treatment. In the context of echocardiography, Wu et al. (contribution 7) proposed a UNeXt-based segmentation algorithm for automatic delineation of the left ventricle in transesophageal echocardiography (TEE) images. This tool enhances the precision of cardiac assessments, contributing to improving the diagnosis and management of cardiovascular conditions.

Contributing significantly to cardiac and pulmonary research, Hadhoud et al. (contribution 8) developed a powerful hybrid CNN–ViT framework for the multi-class classification of chest diseases, including tuberculosis and pneumonia, demonstrating that combining convolutional and transformer-based architectures significantly boosts diagnostic accuracy in thoracic imaging.

Dental radiology is another area where AI is having a significant impact. Peker et al. (contribution 5) introduced a YOLOv10-based system for automatic tooth detection and numbering in pediatric panoramic X-rays, offering an efficient solution for pediatric dental

care. This AI-driven approach supports better patient management by reducing human error in dental imaging.

Furthermore, AI's potential in neurological diagnostics is becoming evident. Ianculescu et al. (contribution 9) explored image processing techniques for early Parkinson's disease detection through hand-drawn-spiral analysis, integrating AI to identify subtle changes that may otherwise go unnoticed. This early-detection capability is crucial for improving long-term patient outcomes by enabling earlier intervention and treatment.

AI's role extends beyond disease detection into broader clinical applications. For instance, Pellegrino et al. (contribution 10) examined the interaction between ChatGPT-4 and human experts for evaluating Boston Bowel Preparation (BP) Scale scores during colonoscopy, highlighting AI's growing ability to assist in complex diagnostic assessments. Moreover, Gil-Rios et al. (contribution 11) developed a hybrid metaheuristic algorithm for improving the classification of coronary stenosis from radiological images, further demonstrating AI's capability in optimizing diagnostic workflows and aiding in precise decision-making.

Additionally, Tian et al. (contribution 12) presented U-Net variants for radial artery tracking in ultrasound images, a tool aimed at supporting catheterization procedures by enhancing the accuracy of artery visualization. This kind of AI application directly contributes to improving procedural accuracy in interventions.

AI's utility in radiotherapy is also notable. Marquez et al. (contribution 13) analyzed the correlation between dose and geometric metrics in head and neck auto-contouring, underscoring how AI can aid in creating more reliable and effective radiotherapy treatment plans. This work further illustrates AI's role in optimizing treatment outcomes in oncology.

Lastly, AI is enabling opportunistic screening in other areas, such as vertebral fracture detection. Kim et al. (contribution 15) proposed a deep learning model for screening acute vertebral fractures in routine abdominal and chest CT scans, allowing for opportunistic diagnosis during standard imaging procedures, which could lead to the earlier identification of fractures that might otherwise be missed.

Overall, these studies exemplify how AI is enhancing diagnostic accuracy, accelerating clinical workflows, and providing invaluable support in decision-making across diverse medical specialties. As AI technologies continue to evolve, they will likely become an integral part of clinical practice, improving patient outcomes and transforming healthcare delivery on a global scale.

Table 1 provides a concise overview of the research articles featured in this Special Issue.

Table 1. Overview of scientific articles featured in this Special Issue.

Contribution.	Authors	Clinical Area	AI Application
1	Brunese et al.	Liver (Oncology)	U-Net for liver segmentation in CT/MRI; explainability for HCC diagnosis.
2	Inmutto et al.	Liver (Oncology)	Differentiation between HCC and ICC using multiphase CT with AI model.
3	Gül et al.	Skin (Dermatology)	YOLOv8 + SAM hybrid model for skin lesion detection and segmentation.

Table 1. *Cont.*

Contribution.	Authors	Clinical Area	AI Application
4	Faiella et al.	Prostate (Oncology)	Random Forest on mp-MRI and radiomics to predict lymph node involvement.
5	Peker et al.	Dental	YOLOv10 system for automatic tooth detection in pediatric panoramic X-rays.
6	Liu et al.	Heart (Cardiology)	Deep learning to detect Stanford A/B aortic dissections in CTA.
7	Wu et al.	Heart (Cardiology)	UNeXt segmentation of left ventricle in TEE images.
8	Hadhoud et al.	Thorax (Heart and Lungs)	CNN–ViT for classification of chest diseases (e.g., TB and pneumonia).
9	Ianculescu et al.	Neurology	AI for early Parkinson's disease detection from hand-drawn-spiral analysis.
10	Pellegrino et al.	Gastroenterology	Interaction of ChatGPT-4 and experts on Boston Bowel Preparation Scale scoring.
11	Gil-Rios et al.	Cardiology (Coronary)	Metaheuristic hybrid model for classifying coronary stenosis in radiology.
12	Tian et al.	Vascular Intervention	U-Net variants for radial artery tracking in ultrasound.
13	Marquez et al.	Radiotherapy (Oncology)	Correlation of dose/geometric metrics for auto-contouring in head and neck cancer.
14	Urso et al.	Prostate (Oncology)	ML model predicting recurrence in prostate cancer post-therapy using PET/CT radiomics.
15	Kim et al.	Bone/Spine	Deep learning for opportunistic screening of vertebral fractures in CT scans.

Outcomes of the reviews

The review studies highlight that he integration of artificial intelligence (AI) with clinical and medical imaging is proving transformative, especially in areas such as interventional radiology, musculoskeletal imaging, cancer diagnostics, and radiotherapy.

As demonstrated by Lastrucci et al. (contribution 16) AI's application in interventional radiology is rapidly evolving, exhibiting immense potential to revolutionize healthcare by enhancing clinical outcomes and reducing procedural risks. Their review discusses key themes, opportunities, challenges, and future directions, positioning AI as a significant driver of innovation in this field.

Similarly, Mercurio et al. (contribution 17) highlight the promising role of AI in detecting anterior cruciate ligament (ACL) injuries by applying deep learning models applied MRI scans. Their comprehensive review emphasizes the potential of these models to improve diagnostic accuracy and efficiency, highlighting their potential as a new dimension in musculoskeletal imaging.

In the field of oncology, particularly in liver cancer diagnostics, Romeo et al. (contribution 18) explore how AI is shaping the future management of hepatocellular carcinoma (HCC). Their work underscores AI's growing role in refining diagnostic and predictive models, supporting earlier intervention and more tailored treatment strategies, which are essential given the increasing incidence of HCC.

Lastrucci et al. (contribution 19) further investigate the integration of deep learning in radiotherapy, exploring how this technology is addressing challenges in treatment planning and delivery. Their review emphasizes the acceleration in progress driven by the COVID-19 pandemic, offering valuable insights into the ongoing development and potential of deep learning to optimize radiotherapy procedures.

Lastly, Ramli et al. (contribution 20), in their systematic review, examine AI's role in lung cancer screening, focusing on its application in detecting lung nodules in chest X-rays. They illustrate AI's significant contribution to enhancing early lung cancer detection, improving patient prognosis through quicker and more accurate diagnoses.

Overall, these articles underscore AI's vast potential to improve diagnostic and treatment processes across various medical fields. From interventional radiology to oncology, musculoskeletal imaging, and radiotherapy, AI is reshaping clinical practices, offering new opportunities for more efficient, accurate, and personalized care.

Table 2 provides a concise overview of the review articles featured in this Special Issue.

Table 2. Overview of the review studies featured in this Special Issue.

Contribution.	Authors	Clinical Area	AI Application
16	Lastrucci et al.	Interventional Radiology	Overview of AI integration with interventional radiology, identifying opportunities and challenges.
17	Mercurio et al.	Musculoskeletal (Orthopedics)	Deep learning models for detecting anterior cruciate ligament injuries on MRI.
18	Romeo et al.	Oncology (Hepatology)	AI for managing hepatocellular carcinoma (HCC) through updated diagnostic and predictive models.
19	Lastrucci et al.	Radiotherapy (Oncology)	Deep learning in radiotherapy, exploring its impact on treatment planning and delivery.
20	Ramli et al.	Pulmonology (Lung Cancer)	AI for detecting lung nodules in chest X-rays for early lung cancer screening.

Conclusions and future directions

The integration of artificial intelligence (AI) in medical imaging and diagnostics, as evidenced by the studies (contribution 1–15) and review articles (contribution 16–20)

featured in this Special Issue, has been shown to have transformative potential across multiple medical disciplines. These studies collectively underscore AI's capacity to enhance diagnostic accuracy, streamline clinical workflows, and facilitate personalized treatment strategies, highlighting its role as a driving force in the evolution of healthcare. The findings from the presented research underline the capacity of AI to significantly improve clinical decision-making. For instance, AI-driven models are showing promise in areas like early disease detection, as seen in the diagnostic tools for liver cancer, skin cancer, and prostate cancer. These advancements not only enable more accurate diagnoses but also facilitate faster intervention, which is key to improving patient outcomes. Moreover, AI's role in enhancing procedural accuracy—whether in radiotherapy or cardiology—highlights its potential to optimize treatment planning and delivery, ultimately contributing to better patient care.

In the future, several key directions will shape AI's continued integration into healthcare. One important focus will be the refinement of AI algorithms to improve their accuracy, reliability, and explainability. This will be essential to foster trust among clinicians and patients alike, ensuring that AI is seen as a supportive tool rather than a replacement for human expertise. Enhancing AI's transparency in decision-making processes will be critical for ensuring clinical acceptance and integration. Furthermore, the future of AI in healthcare will depend on the expansion of its capabilities in predictive analytics. Leveraging AI for the early detection and prediction of disease progression can lead to more proactive interventions, potentially reducing the burden of chronic conditions and improving long-term outcomes for patients. As AI models become more sophisticated, they will increasingly support clinicians in making more informed, data-driven decisions that are tailored to individual patient needs. Another important avenue for future exploration will be the continued integration of AI across multi-disciplinary fields. As AI tools become more advanced, their ability to combine data from various imaging modalities and clinical sources will create a more comprehensive picture of patient health. This multi-modal approach can improve diagnostic accuracy, support complex decision-making, and enhance patient outcomes. AI is rapidly becoming an indispensable tool in modern healthcare, offering remarkable potential to improve diagnostic precision, accelerate clinical workflows, and facilitate personalized care. As research continues to advance, AI's role in healthcare will grow, shaping the future of medical practice and patient care in profound and lasting ways.

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

List of Contributions:

1. Brunese, M.C.; Rocca, A.; Santone, A.; Cesarelli, M.; Brunese, L.; Mercaldo, F. Explainable and Robust Deep Learning for Liver Segmentation Through U-Net Network. *Diagnostics* **2025**, *15*, 878. https://doi.org/10.3390/diagnostics15070878.
2. Inmutto, N.; Pojchamarnwiputh, S.; Na Chiangmai, W. Multiphase Computed Tomography Scan Findings for Artificial Intelligence Training in the Differentiation of Hepatocellular Carcinoma and Intrahepatic Cholangiocarcinoma Based on Interobserver Agreement of Expert Abdominal Radiologists. *Diagnostics* **2025**, *15*, 821. https://doi.org/10.3390/diagnostics15070821.
3. Gül, S.; Cetinel, G.; Aydin, B.M.; Akgün, D.; Öztaş Kara, R. YOLOSAMIC: A Hybrid Approach to Skin Cancer Segmentation with the Segment Anything Model and YOLOv8. *Diagnostics* **2025**, *15*, 479. https://doi.org/10.3390/diagnostics15040479.
4. Faiella, E.; Pileri, M.; Ragone, R.; De Nicola, A.M.; Beomonte Zobel, B.; Grasso, R.F.; Santucci, D. Promising Results About the Possibility to Identify Prostate Cancer Patients Employing a Random Forest Classifier: A Preliminary Study Preoperative Patients Selection. *Diagnostics* **2025**, *15*, 421. https://doi.org/10.3390/diagnostics15040421.

5. Peker, R.B.; Kurtoglu, C.O. Evaluation of the Performance of a YOLOv10-Based Deep Learning Model for Tooth Detection and Numbering on Panoramic Radiographs of Patients in the Mixed Dentition Period. *Diagnostics* **2025**, *15*, 405. https://doi.org/10.3390/diagnostics15040405.
6. Liu, H.-H.; Chang, C.-B.; Chen, Y.-S.; Kuo, C.-F.; Lin, C.-Y.; Ma, C.-Y.; Wang, L.-J. Automated Detection and Differentiation of Stanford Type A and Type B Aortic Dissections in CTA Scans Using Deep Learning. *Diagnostics* **2025**, *15*, 12. https://doi.org/10.3390/diagnostics15010012.
7. Wu, L.; Ling, Y.; Lan, L.; He, K.; Yu, C.; Zhou, Z.; Shen, L. Automatic Segmentation of the Left Ventricle in Apical Four-Chamber View on Transesophageal Echocardiography Based on UNeXt Deep Neural Network. *Diagnostics* **2024**, *14*, 2766. https://doi.org/10.3390/diagnostics14232766.
8. Hadhoud, Y.; Mekhaznia, T.; Bennour, A.; Amroune, M.; Kurdi, N.A.; Aborujilah, A.H.; Al-Sarem, M. From Binary to Multi-Class Classification: A Two-Step Hybrid CNN-ViT Model for Chest Disease Classification Based on X-Ray Images. *Diagnostics* **2024**, *14*, 2754. https://doi.org/10.3390/diagnostics14232754.
9. Ianculescu, M.; Petean, C.; Sandulescu, V.; Alexandru, A.; Vasilevschi, A.-M. Early Detection of Parkinson's Disease Using AI Techniques and Image Analysis. *Diagnostics* **2024**, *14*, 2615. https://doi.org/10.3390/diagnostics14232615.
10. Pellegrino, R.; Federico, A.; Gravina, A.G. Conversational LLM Chatbot ChatGPT-4 for Colonoscopy Boston Bowel Preparation Scoring: An Artificial Intelligence-to-Head Concordance Analysis. *Diagnostics* **2024**, *14*, 2537. https://doi.org/10.3390/diagnostics14222537.
11. Gil-Rios, M.-A.; Cruz-Aceves, I.; Hernandez-Aguirre, A.; Hernandez-Gonzalez, M.-A.; Solorio-Meza, S.-E. Improving Automatic Coronary Stenosis Classification Using a Hybrid Metaheuristic with Diversity Control. *Diagnostics* **2024**, *14*, 2372. https://doi.org/10.3390/diagnostics14212372.
12. Tian, Y.; Gao, R.; Shi, X.; Lang, J.; Xue, Y.; Wang, C.; Zhang, Y.; Shen, L.; Yu, C.; Zhou, Z. U-Net and Its Variants Based Automatic Tracking of Radial Artery in Ultrasonic Short-Axis Views: A Pilot Study. *Diagnostics* **2024**, *14*, 2358. https://doi.org/10.3390/diagnostics14212358.
13. Marquez, B.; Wooten, Z.T.; Salazar, R.M.; Peterson, C.B.; Fuentes, D.T.; Whitaker, T.J.; Jhingran, A.; Pollard-Larkin, J.; Prajapati, S.; Beadle, B.; et al. Analyzing the Relationship between Dose and Geometric Agreement Metrics for Auto-Contouring in Head and Neck Normal Tissues. *Diagnostics* **2024**, *14*, 1632. https://doi.org/10.3390/diagnostics14151632.
14. Urso, L.; Cittanti, C.; Manco, L.; Ortolan, N.; Borgia, F.; Malorgio, A.; Scribano, G.; Mastella, E.; Guidoboni, M.; Stefanelli, A.; et al. ML Models Built Using Clinical Parameters and Radiomic Features Extracted from 18F-Choline PET/CT for the Prediction of Biochemical Recurrence after Metastasis-Directed Therapy in Patients with Oligometastatic Prostate Cancer. *Diagnostics* **2024**, *14*, 1264. https://doi.org/10.3390/diagnostics14121264.
15. Kim, Y.R.; Yoon, Y.S.; Cha, J.G. Opportunistic Screening for Acute Vertebral Fractures on a Routine Abdominal or Chest Computed Tomography Scans Using an Automated Deep Learning Model. *Diagnostics* **2024**, *14*, 781. https://doi.org/10.3390/diagnostics14070781.
16. Lastrucci, A.; Iosca, N.; Wandael, Y.; Barra, A.; Lepri, G.; Forini, N.; Ricci, R.; Miele, V.; Giansanti, D. AI and Interventional Radiology: A Narrative Review of Reviews on Opportunities, Challenges, and Future Directions. *Diagnostics* **2025**, *15*, 893. https://doi.org/10.3390/diagnostics15070893.
17. Mercurio, M.; Denami, F.; Melissaridou, D.; Corona, K.; Cerciello, S.; Laganà, D.; Gasparini, G., on behalf of the IORS; Minici, R. Deep Learning Models to Detect Anterior Cruciate Ligament Injury on MRI: A Comprehensive Review. *Diagnostics* **2025**, *15*, 776. https://doi.org/10.3390/diagnostics15060776.
18. Romeo, M.; Dallio, M.; Napolitano, C.; Basile, C.; Di Nardo, F.; Vaia, P.; Iodice, P.; Federico, A. Clinical Applications of Artificial Intelligence (AI) in Human Cancer: Is It Time to Update the Diagnostic and Predictive Models in Managing Hepatocellular Carcinoma (HCC)? *Diagnostics* **2025**, *15*, 252. https://doi.org/10.3390/diagnostics15030252.
19. Lastrucci, A.; Wandael, Y.; Ricci, R.; Maccioni, G.; Giansanti, D. The Integration of Deep Learning in Radiotherapy: Exploring Challenges, Opportunities, and Future Directions through an Umbrella Review. *Diagnostics* **2024**, *14*, 939. https://doi.org/10.3390/diagnostics14090939.

20. Megat Ramli, P.N.; Aizuddin, A.N.; Ahmad, N.; Abdul Hamid, Z.; Ismail, K.I. A Systematic Review: The Role of Artificial Intelligence in Lung Cancer Screening in Detecting Lung Nodules on Chest X-Rays. *Diagnostics* **2025**, *15*, 246. https://doi.org/10.3390/diagnostics15030246.

References

1. Artificial Intelligence in Clinical Medical Imaging. Available online: https://mdpi-res.com/bookfiles/book/9149/Artificial_Intelligence_in_Clinical_Medical_Imaging.pdf?v=1744852066 (accessed on 11 June 2025).
2. Pinto-Coelho, L. How Artificial Intelligence Is Shaping Medical Imaging Technology: A Survey of Innovations and Applications. *Bioengineering* **2023**, *10*, 1435. [CrossRef] [PubMed] [PubMed Central]
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. Bajwa, J.; Munir, U.; Nori, A.; Williams, B. Artificial intelligence in healthcare: Transforming the practice of medicine. *Future Healthc. J.* **2021**, *8*, e188–e194. [CrossRef] [PubMed] [PubMed Central]
5. Herington, J.; McCradden, M.D.; Creel, K.; Boellaard, R.; Jones, E.C.; Jha, A.K.; Rahmim, A.; Scott, P.J.H.; Sunderland, J.J.; Wahl, R.L.; et al. Ethical Considerations for Artificial Intelligence in Medical Imaging: Data Collection, Development, and Evaluation. *J. Nucl. Med.* **2023**, *64*, 1848–1854. [CrossRef] [PubMed] [PubMed Central]
6. Hosny, A.; Parmar, C.; Quackenbush, J.; Schwartz, L.H.; Aerts, H.J.W.L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **2018**, *18*, 500–510. [CrossRef] [PubMed] [PubMed Central]
7. Farhud, D.D.; Zokaei, S. Ethical Issues of Artificial Intelligence in Medicine and Healthcare. *Iran. J. Public Health* **2021**, *50*, i–v. [CrossRef] [PubMed] [PubMed Central]
8. Habli, I.; Lawton, T.; Porter, Z. Artificial intelligence in health care: Accountability and safety. *Bull. World Health Organ.* **2020**, *98*, 251–256. [CrossRef] [PubMed] [PubMed Central]

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 and Robust Deep Learning for Liver Segmentation Through U-Net Network

Maria Chiara Brunese [1], Aldo Rocca [1], Antonella Santone [1], Mario Cesarelli [2], Luca Brunese [1] and Francesco Mercaldo [1,*]

[1] Department of Medicine and Health Sciences "Vincenzo Tiberio", University of Molise, 86100 Campobasso, Italy; antonella.santone@unimol.it (A.S.)
[2] Department of Engineering, University of Sannio, 82100 Benevento, Italy; mcesarelli@unisannio.it
* Correspondence: francesco.mercaldo@unimol.it

Abstract: Background/Objectives: Clinical imaging techniques, such as magnetic resonance imaging and computed tomography, play a vital role in supporting clinicians by aiding disease diagnosis and facilitating the planning of appropriate interventions. This is particularly important in malignant conditions like hepatocellular carcinoma, where accurate image segmentation, delineating the liver and tumor, is a critical initial step in optimizing diagnosis, staging, and treatment planning, including interventions like transplantation, surgical resection, radiotherapy, portal vein embolization, and other procedures. Therefore, effective segmentation methods can significantly influence both diagnostic accuracy and treatment outcomes. **Method:** In this paper, we propose a deep learning-based approach aimed at accurately segmenting the liver in medical images, thus addressing a critical need in hepatic disease diagnosis and treatment planning. We consider a U-Net architecture with residual connections to capture fine-grained anatomical details. We also take into account the prediction explainability, aiming to highlight, in the image under analysis, the areas that are symptomatic for a certain segmentation. In detail, by exploiting the U-Net architecture, two different models are trained with two annotated datasets of computed tomography medical images, resulting in four different experiments. **Results:** We consider two different datasets to improve robustness and generalization across diverse patient populations and imaging conditions. Experimental results demonstrate that the proposed method obtains interesting performances, with an accuracy ranging from 0.81 to 0.93. **Conclusions:** We thus show that the proposed method can provide a reliable and efficient solution for automated liver segmentation, promising significant advancements in clinical workflows and precision medicine.

Keywords: liver; U-Net; segmentation; biomedical images; bioimages; deep learning; explainability; robustness

1. Introduction

The liver, a vital organ responsible for numerous essential functions such as detoxification, protein synthesis, and the production of biochemicals necessary for digestion, plays a critical role in maintaining overall health [1,2]. This is the reason why accurate liver segmentation, the process of delineating the liver and its structures within medical imaging [3,4], has become increasingly important in medical diagnostics and treatment planning. This advanced imaging technique allows for precise identification and analysis of liver anatomy and pathology, aiding in the diagnosis of liver diseases, assessment of tumor growth, and planning of surgical interventions [5,6]. As medical imaging technology

continues to evolve, liver segmentation is crucial, offering enhanced precision in disease monitoring, facilitating targeted therapies, and ultimately improving patient outcomes [7,8]. Understanding the significance of liver segmentation underscores its impact on the quality and effectiveness of healthcare, particularly in the management of hepatic conditions [9].

As a matter of fact, liver segmentation can be helpful in several medical scenarios, for instance, accurate segmentation helps in identifying and characterizing liver tumors, essential for diagnosing liver cancer [3,10]. Moreover, detailed segmentation assists surgeons in planning the removal of liver tumors or damaged tissue while preserving as much healthy liver as possible. Furthermore, precise liver segmentation ensures that radiation is accurately targeted to cancerous tissues, minimizing damage to healthy liver tissue [11,12]. Another context where segmentation is helpful is due to their ability to monitor changes in liver size and shape, helping to track the progression of chronic liver diseases. It also allows for the assessment of how well a patient is responding to treatments such as chemotherapy or antiviral therapy [13]. From the perspective of the development of computational models, segmented liver images are used to create models that can predict disease progression and treatment outcomes and can also support the development of new therapies and diagnostic tools by providing detailed anatomical and pathological data. Segmentation enables precise volume measurements of the liver and its lesions, which are crucial for assessing the severity of diseases and planning treatments, and by integrating with other imaging modalities, segmentation helps in evaluating liver function and perfusion [14–16].

Overall, liver segmentation enhances the accuracy and effectiveness of medical diagnostics, treatment planning, and disease monitoring, significantly improving patient care and outcomes [17–19].

Recently, computer vision (CV) has been emerging as a subfield of artificial intelligence that deals with the interpretation and understanding of visual information from the real world. In particular, segmentation is a fundamental task within computer vision, involving the partitioning of an image into meaningful regions or objects [17,20,21].

As a matter of fact, CV can significantly enhance liver segmentation by automating the detection and delineation of liver boundaries in medical images, such as computed tomography (CT). As a matter of fact, it can reduce the variability and errors associated with manual segmentation, improving accuracy and efficiency in clinical workflows. Through the use of deep learning models, CV algorithms can learn to identify complex patterns and structures in liver tissues, enabling precise segmentation even in cases with challenging anatomical variations. Thus, CV allows for faster and more consistent segmentation, facilitating better diagnosis, treatment planning, and monitoring of liver diseases [22–24].

Starting from these considerations, in this paper, we propose a method aimed at performing automatic liver segmentation by exploiting CV, with particular regard to deep learning. The proposed method also takes into account explainability, with the aim of highlighting the areas on the medical images that are related to a certain prediction. From the model point of view, we exploit the U-Net model, a type of convolutional neural network primarily designed for biomedical image segmentation. It uses an encoder–decoder architecture, where the encoder captures spatial features and the decoder reconstructs the image, allowing precise localization. A feature of U-Net is its skip connections, which help retain spatial information by transferring feature maps from the encoder to corresponding layers in the decoder. In order to take into account the explainability (i.e., a way to enable human users to understand and trust the results and outputs of a machine learning model), we resort to Gradient-weighted Class Activation Mapping (i.e., Grad-CAM) [17,25–27]. Grad-CAM is a visualization technique used to interpret the predictions of convolutional neural networks. It basically works by highlighting the important regions in an input image that influence the decision of the model. It works by computing the gradients of

a target class score with respect to the feature maps in the final convolutional layer and using these gradients to produce a heatmap of class-discriminative regions. This method is particularly useful for understanding model behavior in tasks like image classification, object detection, and medical imaging [17,20,28,29].

With the aim of demonstrating the effectiveness of the proposed method, we resort to two different datasets, both of them freely available for research purposes; in particular, we provide, in the experimental analysis section, four different experiments, the first one by exploiting the model built with the first dataset and evaluated with (not used in the training) images with the first dataset, the second one with the same model of the first experiment that is evaluated with images belonging to the second dataset. The third experiment considers the second dataset for model building and (not used in the training) model testing images, while the last experiment is related to the model of the third experiment evaluated with images belonging to the first dataset. The idea of these four experiments (in particular, the second and the fourth ones, where images belonging from different datasets are exploited for training and testing) is to evaluate the model robustness with different images, acquired from different machinery, with different conditions and from different human operators, exploited for training and testing purposes.

This study aims to investigate the effectiveness, robustness, and interpretability of deep learning-based liver segmentation. Specifically, we address the following research questions:

RQ1: How effectively can a deep learning-based U-Net model with residual connections improve the accuracy of liver segmentation in CT images compared to traditional methods?

RQ2: How does the integration of Grad-CAM enhance the explainability of segmentation predictions, making them more interpretable for clinical applications?

RQ3: Can the proposed model generalize well across different datasets acquired from diverse imaging conditions, scanner types, and patient populations?

Below, we itemize the main contributions of the proposed method:

- Development of an Explainable Liver Segmentation Model: We propose a U-Net-based deep learning model incorporating residual connections to improve feature propagation and segmentation accuracy.
- Integration of Explainability with Grad-CAM: Unlike conventional black-box deep learning models, our approach enhances transparency by employing Grad-CAM to visualize and interpret segmentation decisions.
- Robustness Across Multiple Datasets: The model is trained and evaluated on two different publicly available CT liver segmentation datasets, ensuring improved generalization across imaging conditions, scanner types, and patient populations.
- Comprehensive Experimental Validation: We conduct extensive experiments across four different scenarios, evaluating the model's segmentation accuracy, Dice coefficient, and Intersection over Union (IoU) to assess its robustness and effectiveness.
- Potential for Clinical Application: By improving segmentation accuracy while offering explainability, the proposed method enhances trust in AI-driven liver segmentation, making it more suitable for real-world clinical workflows, including preoperative planning and disease monitoring.

The paper proceeds as follows. In Section 2, we provide preliminary notions related to the U-Net architecture considered by the proposed method, with the aim of making the paper self-contained; Section 3 depicts in detail the proposed method for explainable automatic liver segmentation; the experimental analysis aimed to show the performances, and the robustness of the proposed method is presented in Section 4. An overview of the

state-of-the-art research in the context of liver segmentation is shown in Section 5, and finally, the conclusion and future research plans are presented in Section 6.

2. Background

In this section, we provide preliminary notions about the deep learning model we considered for liver segmentation, with the aim of making the paper self-contained.

Deep learning offers a variety of architectures for medical image segmentation, with Convolutional Neural Networks (CNNs) being the foundation of most. Among these, U-Net stands out as a specialized architecture tailored for biomedical imaging tasks, including segmentation. Its name derives from its distinctive "U" shape, which reflects its symmetric encoder–decoder design. This architecture is particularly well-suited for tasks that demand accurate localization and identification of specific image regions.

The U-Net architecture consists of two primary components: the contracting path, also known as the encoder, and the expansive path, referred to as the decoder. These paths are connected by a bottleneck, forming the central bridge. Together, these components enable the model to learn hierarchical feature representations while retaining spatial details essential for segmentation.

The contracting path is responsible for extracting high-level features and capturing the broader context of the input image. This path comprises several stages, each consisting of two convolutional layers followed by a ReLU activation function and a max-pooling operation. The convolutional layers use small 3×3 kernels to extract features, while the max-pooling layers reduce the spatial dimensions by half, thereby increasing the receptive field and enabling the network to learn more abstract features. This process is repeated multiple times, progressively reducing the spatial dimensions and deepening the feature representations.

At the center of the U-Net architecture lies the bottleneck, which serves as a connection between the encoder and the decoder. This segment includes two 3×3 convolutional layers with ReLU activations. Unlike the contracting path, the bottleneck does not employ max-pooling layers, preserving the spatial resolution of the feature maps. This design ensures that the most critical features are retained for reconstruction.

The expansive path, or decoder, is tasked with reconstructing the segmentation map by progressively upsampling the feature maps. Each stage of the decoder begins with a transposed convolution, also known as up-convolution, which doubles the spatial dimensions of the feature maps. The upsampled features are then concatenated with the corresponding features from the contracting path through skip connections. These skip connections play a pivotal role in preserving spatial information and finer details that might have been lost during downsampling. Following concatenation, two 3×3 convolutional layers with ReLU activations are applied to refine the features and prepare them for the next upsampling stage.

The final layer of the U-Net architecture is a 1×1 convolutional layer that reduces the feature map to the desired number of output classes. For binary segmentation tasks, this layer typically outputs two classes, representing the segmented region and the background. The resulting map is further processed with a sigmoid activation function to produce the final segmentation output.

The U-Net architecture's design, combining hierarchical feature extraction with spatial detail retention, makes it an ideal choice for medical image segmentation. Its ability to deliver accurate and precise segmentation maps has made it a cornerstone in biomedical imaging applications.

This architecture allows U-Net to effectively capture both global context and fine details, making it highly effective for various segmentation tasks, particularly in medical imaging.

In the following, we motivate the reason why we resort to this specific model for liver segmentation, by discussing the several advantages offered by U-Net over traditional segmentation methods, particularly in the context of biomedical imaging.

First, U-Net supports end-to-end learning, training directly on raw image data and corresponding segmentation masks. This enables it to automatically learn features and representations from the data, optimizing for the specific task and reducing the need for manual feature engineering. In contrast, traditional methods often rely on manual feature extraction, which can be time-consuming and prone to human error.

Second, U-Net efficiently handles small datasets by utilizing data augmentation techniques like rotations, shifts, and flips. It leverages symmetry and skip connections, making the most out of limited training data. On the other hand, traditional methods are less capable of handling small datasets without overfitting and often require extensive manual tuning and augmentation strategies.

Third, U-Net excels at precision and accuracy, particularly in localizing and segmenting structures of varying sizes. The use of skip connections allows it to retain fine-grained spatial information, leading to more precise segmentations. Traditional methods may struggle with small or complex structures and often rely on heuristic or rule-based approaches, which can be less precise.

Fourth, U-Net offers flexibility and adaptability, making it versatile enough to be applied to a wide range of segmentation tasks, beyond medical imaging, such as satellite imagery and natural scene segmentation. It is also easy to modify and extend with additional layers or modules to suit specific needs. In contrast, traditional methods are often task-specific, requiring significant redesign and retraining for different segmentation tasks, and are less flexible in adapting to new types of data or changing requirements.

Fifth, U-Net demonstrates efficiency in inference, capable of real-time or near-real-time processing, which is ideal for applications like intraoperative guidance. It is well suited for GPU acceleration, significantly speeding up processing times. Traditional methods tend to be slower, especially when complex pre-processing and feature extraction steps are involved, and may not be optimized for parallel processing or hardware acceleration.

Sixth, U-Net effectively integrates contextual information, capturing both local and global context, which is essential for understanding complex structures in medical images. By combining high-level semantic information with low-level details through its expansive path and skip connections, it achieves a deeper understanding of the data. Traditional methods, however, often lack the ability to integrate multi-scale contextual information, requiring separate modules or steps to handle different scales, thereby increasing complexity.

Seventh, U-Net reduces the annotation effort by providing high-quality initial segmentations that can be fine-tuned by experts. This minimizes the time and effort needed for creating large annotated datasets. In comparison, traditional methods typically demand extensive manual annotation and correction, which can be labor-intensive and time-consuming, with segmentations often requiring significant post-processing and refinement.

Lastly, U-Net demonstrates strong generalization capabilities, performing well across different datasets and imaging modalities. Thanks to its deep learning foundations, it can handle variations in image quality, noise, and artifacts. In contrast, traditional methods are often sensitive to variations in imaging conditions and may require separate tuning for different datasets, with generalization capabilities limited without extensive customization and parameter adjustment.

3. Materials and Methods

In this section, we describe the proposed method for explainable liver segmentation, whose workflow is shown in Figure 1. As introduced in the previous section for liver segmentation, we resort to the U-Net model, enhanced with explainability through Grad-CAM. We also take into account model robustness by considering two different datasets in the experimental analysis.

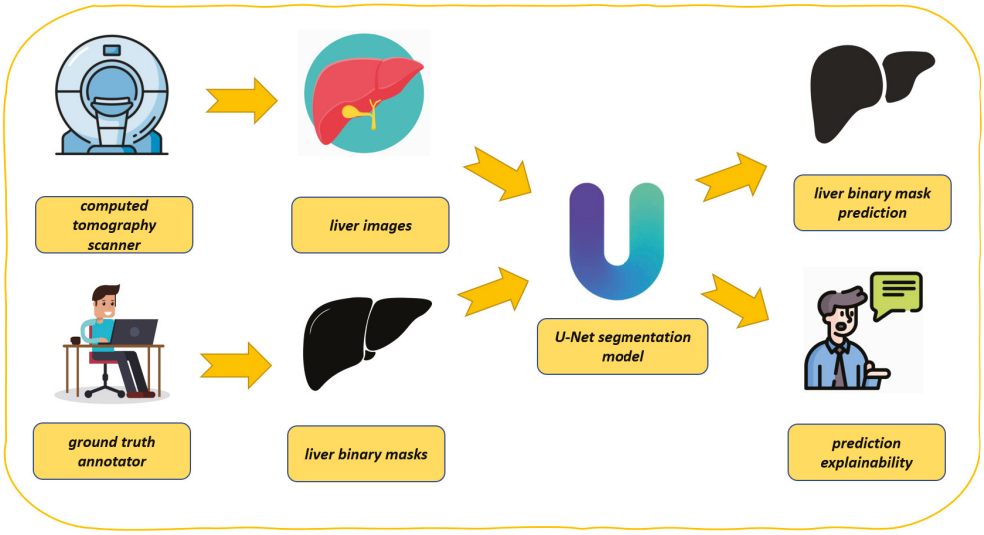


Figure 1. The workflow related to the proposed method for explainable liver segmentation.

Figure 1 presents the proposed workflow for automated liver segmentation from CT images, utilizing a U-Net deep learning model alongside explainability techniques provided by Grad-CAM.

As shown in Figure 1, the proposed method is composed of several steps, explained in detail below.

- The process begins with the acquisition of medical imaging data (i.e., computer tomography scanner in Figure 1). CT scanners capture detailed cross-sectional images of the abdomen, which include the liver. These images are essential for constructing accurate 3D representations of the liver. The CT scanner indicates the source of the initial input data.
- The CT scans, shown as liver images, are extracted from the abdominal region (i.e., liver images in Figure 1). These images serve as the raw input for the segmentation task. Each image is typically a 2D slice of the liver, which is part of a larger 3D dataset.
- Before training the model, the CT images must be annotated to create a reference standard (i.e., ground truth annotator in Figure 1). Expert radiologists or medical professionals manually outline the liver boundaries in each image, generating accurate segmentation masks. These annotations are represented as liver binary masks in the figure, where pixels corresponding to the liver are marked as 1 (or white) and all other pixels as 0 (or black). The annotator is symbolized by a person working at a computer, highlighting the manual effort involved in creating these ground truth labels.
- These masks (i.e., liver binary masks in Figure 1) are critical for supervised learning, as they provide the U-Net model with examples of correct segmentations, against which its predictions can be compared during training.
- The U-Net model, depicted at the center of the figure, is the core of the proposed method (i.e., U-net segmentation model in Figure 1). U-Net is a CNN specifically designed for biomedical image segmentation tasks. It features a unique architecture

with a contracting path (encoder) that captures context and a symmetric expanding path (decoder) that enables precise localization. The liver images and corresponding binary masks are fed into the U-Net model during training. The model learns to map input images to their respective segmentation masks, effectively learning to identify and delineate the liver from the surrounding tissues.

- After training, the U-Net model is capable of segmenting the liver in unseen CT images (i.e., liver binary mask prediction in Figure 1). Given a new image, the model outputs a predicted binary mask, which highlights the liver region. This predicted mask is intended to replicate the accuracy of the ground truth annotations, allowing for automated liver segmentation. The figure shows this as a black-and-white liver shape, emphasizing the binary nature of the output.
- To ensure that the model predictions are reliable and understandable (i.e., prediction explainability in Figure 1), the proposed method incorporates explainability through Grad-CAM. In a nutshell, Grad-CAM generates a heatmap that overlays on the input image, showing which regions the U-Net model focused on while making its prediction. This allows practitioners to verify that the model is basing its decisions on relevant anatomical features, such as the liver's edges, rather than irrelevant background areas. The explainability aspect is depicted by an individual explaining or discussing the model's predictions, symbolizing the interpretability and validation process.

As shown in Figure 1, the proposed method represents a process where CT images are annotated by medical domain experts to create ground truth masks. These masks are used to train a U-Net model that learns to segment the liver automatically. The model's predictions are subsequently explained and validated using Grad-CAM, ensuring that the segmentation is not only accurate but also interpretable, with the aim of boosting the adoption of deep learning in the real-world medical domain.

The use of U-Net ensures high accuracy in segmentation, while Grad-CAM addresses the need for transparency in computational intelligence-driven medical applications. This integration of segmentation and explainability forms a robust approach to liver segmentation, suitable for clinical use where both performance and interpretability are critical.

The use of Grad-CAM for explainability in segmentation tasks might not be immediately apparent, as it is traditionally employed to visualize class-relevant activations in classification networks. Nevertheless, its application in segmentation tasks offers several practical advantages and serves distinct purposes.

Firstly, Grad-CAM provides a means to visualize model attention, helping to identify the regions of an image the segmentation model focuses on when making decisions. This is especially valuable for verifying whether the model is leveraging the correct image features to segment the regions of interest. Importantly, while Grad-CAM does not enhance model performance, it offers a visual representation of the areas that the model considers significant during analysis.

Secondly, Grad-CAM contributes to debugging and interpretability, which are essential in critical domains such as medical segmentation or autonomous driving. By highlighting the specific parts of the image influencing predictions, it enables researchers and engineers to pinpoint and address potential errors in the model's behavior.

Additionally, Grad-CAM serves as a tool for model validation, ensuring that the activated regions align with expectations. For instance, in an organ segmentation task like the liver segmentation addressed in this paper, Grad-CAM can confirm whether the model focuses on the appropriate organs during the segmentation process.

Finally, it aids in feature analysis by revealing the features the model is learning. This insight can guide further refinement of the model or the input data, ultimately contributing to improved overall performance.

To integrate Grad-CAM into the proposed segmentation task, the process involves selecting an intermediate convolutional layer within the U-Net architecture. The activations of this layer are computed, along with the gradients of the loss function with respect to these activations. These gradients are then used to weight the activations, which are subsequently combined to produce a heatmap. Finally, this heatmap is visualized in conjunction with the segmentation prediction to provide insights into the model's focus areas.

Using Grad-CAM in segmentation tasks proves to be a valuable strategy for improving the interpretability, debugging, and validation of segmentation models. Although originally developed for classification, adapting Grad-CAM to segmentation contexts offers meaningful insights into the model's decision-making process and highlights potential areas for further refinement.

To evaluate the performance of the proposed method for explainable liver segmentation, we resort to Accuracy (by comparing the predicted segmentation mask with the ground truth mask and calculating the proportion of correctly classified pixels), Dice and IoU metrics.

The model was developed from authors by exploiting the Python programming language (3.9.19 version).

To evaluate the effectiveness of the proposed method, two different datasets are exploited, both of them retrieved from the Kaggle website and freely available for research purposes (and thus for replication purposes). The first one is the CT liver dataset https://www.kaggle.com/datasets/zxcv2022/digital-medical-images-for--download-resource, accessed on 27 February 2025 (i.e., the DI dataset), while the second one is the Segmentation of Liver https://www.kaggle.com/datasets/priyamsaha17/segmentation-of-liver dataset, accessed on 27 February 2025 (i.e., the DII dataset). The DI dataset is composed of 116 CT liver images with the related binary mask; in particular, 96 CT images are considered for training and the remaining 20 as the testing dataset. The DII dataset is composed of 282 CT liver images with the related binary mask, and in this case as well, 20 CT liver images are exploited as the testing dataset and the remaining 262 CT liver images are considered as the training dataset.

4. Results

In this section we describe the results of the experimental analysis. First of all, we describe in detail the exploited datasets.

4.1. Image Resolution and Format

- The images in both datasets are provided in grayscale with a resolution of 512×512 pixels, ensuring a high level of anatomical detail necessary for accurate segmentation.
- All images are stored in DICOM (Digital Imaging and Communications in Medicine) format, which preserves essential metadata such as acquisition parameters and patient positioning.

4.2. Number of Images and Data Distribution

- DI Dataset: Contains 116 CT liver scans with corresponding binary segmentation masks. The dataset is split into 96 images for training and 20 for testing to maintain a balanced evaluation.
- DII Dataset: Includes 282 CT liver images, where 262 are used for training and 20 for testing to assess the model's generalizability.

4.3. Patient Diversity and Imaging Conditions

- The datasets consist of CT scans from multiple patients, captured using different CT scanners under varied acquisition settings (contrast-enhanced and non-contrast images).
- This diversity ensures the dataset reflects real-world variability in patient anatomy, liver pathology, scanner resolution, and image noise levels, which is crucial for evaluating model robustness.

4.4. Preprocessing

To ensure consistency across both datasets and enhance the training process, the following preprocessing steps were applied:

4.4.1. Resizing and Normalization

- Since the original image sizes varied slightly across samples, all images were resized to 256 $\times$ 256 pixels to optimize computational efficiency while preserving critical anatomical structures.
- Min-max normalization was applied to scale pixel intensity values between 0 and 1, reducing the impact of intensity variations due to contrast agents or different scanning protocols.

4.4.2. Augmentation for Robustness

Data augmentation techniques were applied to prevent overfitting and improve generalization:

- Rotation ($\pm 20^{\circ}$) to account for variations in patient positioning.
- Horizontal and vertical flipping to enhance symmetry recognition.
- Random cropping and zooming ($\pm 10\%$) to simulate variations in scan resolutions.
- Gaussian noise injection to mimic real-world imaging noise.

4.4.3. Segmentation Mask Refinement

- The provided ground truth masks were binarized (liver = 1, background = 0) to standardize segmentation outputs across datasets.
- To correct minor inconsistencies in annotations, morphological operations (erosion and dilation) were applied, improving mask clarity and boundary precision.

By incorporating these preprocessing techniques, we ensure that the deep learning model is trained on a standardized, diverse, and well-balanced dataset, facilitating robust and generalizable liver segmentation.

Figure 2 shows four examples of CT liver images related to the DI dataset.

Figure 2 presents four axial slices from a CT scan, likely of the abdomen. The scans are in grayscale, adhering to standard medical imaging practices. Each image slice is square with dimensions of 250 $\times$ 250 pixels, potentially corresponding to a specific magnification level or scale within the imaging modality. These slices represent axial (cross-sectional) views of the abdomen, showcasing various abdominal organs, including the liver, kidneys, stomach, intestines, and possibly sections of the spine and abdominal musculature.

In the top left image, sections of the liver (on the left side) and the kidney (on the right side) are visible, with part of the spinal column centrally located. The top right image depicts a slightly lower section of the abdomen, where the liver occupies a larger portion of the view, and a section of the bowel is observable. The bottom left slice reveals a broader section of the abdominal cavity, with distinct bowel loops that may show gas-filled intestines or contrast material. The bottom right image features the liver more prominently, extending across the section, and includes parts of the gastrointestinal system, possibly the colon.

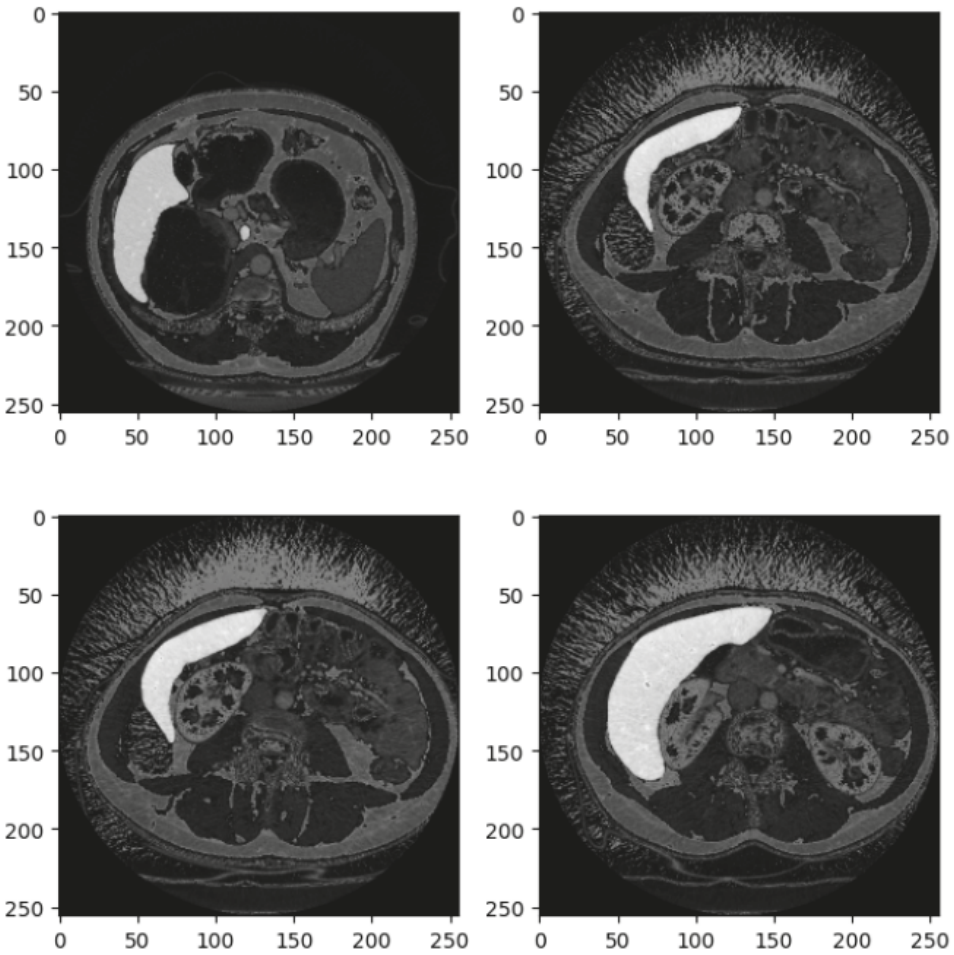


Figure 2. Four examples of CT images with the related mask obtained from the DI dataset.

The images appear to be contrast-enhanced, as indicated by the bright white regions representing areas of high contrast uptake, such as blood vessels or perfused organs. The overall quality of the images is high, with minimal noise, suggesting that they were acquired at an adequate resolution suitable for diagnostic or research purposes. The grayscale intensity is consistent across all slices, indicating uniform windowing and leveling, which is essential for evaluating the radiodensity of the tissues depicted.

Figure 3 shows four related examples of CT liver images belonging to the DII dataset.

Figure 3, similar to Figure 2, consists of four axial slices from a CT scan, each displaying different cross-sections of the abdominal region. These images are part of the DII dataset, and their detailed description is as follows:

The set is arranged in a 2×2 grid of square images, each with a resolution of 250×250 pixels, consistent with the format in Figure 2. Like the previous set, these CT slices are taken from the axial plane, providing horizontal cross-sections of the abdomen at varying anatomical levels.

In the top left slice, portions of the kidneys are visible, with the left kidney prominently displayed. Sections of the gastrointestinal tract are also identifiable, and the spine is centrally located, providing a clear view of the vertebral body. The top right slice captures part of the liver on the left side, the spleen on the right side, and some sections of the bowel. The stomach may also be visible, depending on the level of the scan. The bottom left slice offers a clearer view of the left kidney, with the vertebra still centrally located. This slice also provides a good view of the posterior abdominal wall and adjacent structures. In

the bottom right slice, both kidneys are distinctly visible, and the spine remains centrally located. The intestines and surrounding tissues are also identifiable in this view.

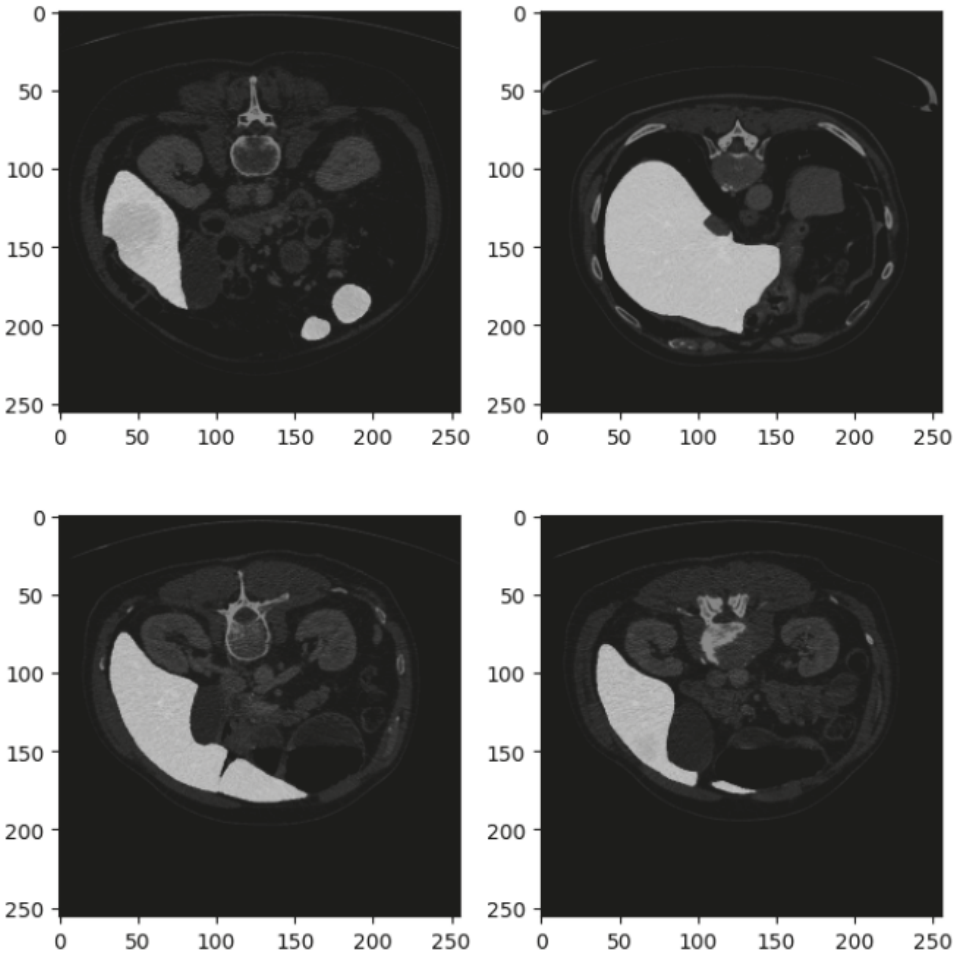


Figure 3. Four examples of CT images with the related mask obtained from the DII dataset.

These images are likely contrast-enhanced, as evidenced by the bright areas, particularly in the liver, kidneys, and gastrointestinal tract, indicating high contrast uptake in these organs and their vasculature. The soft tissue structures are well differentiated due to the contrast, while bone structures, such as the vertebrae, are clearly visible. The grayscale intensities across the images show excellent contrast resolution, which allows for clear differentiation between muscle, fat, organs, and bone.

These slices provide valuable insight into the spatial relationships of major abdominal organs, with the kidneys and liver being especially prominent. This suggests that the images may be focused on assessing the function or structure of these organs. The visibility of the vertebrae in all slices also enables an assessment of the spine in the abdominal context. The images are likely used for diagnostic purposes to assess conditions such as kidney abnormalities (e.g., hydronephrosis or masses), liver conditions (e.g., hepatomegaly or lesions), or bowel issues (e.g., blockages or inflammation).

These slices provide a detailed look at key abdominal organs and would be useful for clinical diagnostics or anatomical study, given the high contrast and clear visualization of both soft tissues and bone structures.

The learning rate is a critical hyperparameter that determines how quickly the model updates its weights during training. In our proposed U-Net with residual connections, we set the initial learning rate to 0.001, with a learning rate decay strategy to ensure stable convergence. Specifically, the initial learning rate is set to 0.001 and is reduced by a factor of

0.1 every 100 epochs. The Adam optimizer is chosen due to its adaptive learning capability and its effectiveness in handling sparse gradients.

To provide a comprehensive view of the training setup, in Table 1, we list all the key hyperparameters used in the training process.

Table 1. Hyperparameters of the proposed U-Net model.

Hyperparameter	Value
Model Architecture	U-Net with residual connections
Batch Size	16
Input Image Size	256 × 256 pixels
Optimizer	Adam
Initial Learning Rate	0.001
Learning Rate Decay	0.1 reduction every 100 epochs
Number of Epochs	500
Loss Function	Dice Loss + Cross-Entropy Loss
Activation Function	ReLU (Hidden Layers)/Sigmoid (Output Layer)s
Dropout Rate	0.3
Weight Initialization	He Normal
Augmentation Methods	Rotation (±20°), Flipping, Cropping, Gaussian Noise

The selection of these hyperparameters is based on best practices in deep learning for biomedical image segmentation. The Adam optimizer is particularly effective as it combines the benefits of momentum and adaptive learning rates, making it well suited for training on medical image datasets. The Dice Loss + Cross-Entropy Loss combination ensures that the model optimizes shape constraints while also improving pixel-wise classification accuracy. To mitigate overfitting, a dropout rate of 0.3 is applied, which is particularly important given the relatively small dataset size. Data augmentation techniques, including rotation, flipping, cropping, and Gaussian noise, enhance the model's ability to generalize across diverse imaging conditions and scanner variations.

Thus, we build two different models by exploiting the same U-Net architecture; the performance of each model is evaluated by using the testing dataset of the same dataset exploited for training and with the testing dataset related to the other model, resulting in four different experiments:

- Experiment I: In this experiment, we train a model with the training dataset belonging to the DI dataset and we evaluate the model trained with the testing dataset related to the DI dataset.
- Experiment II: In this experiment, we consider the model trained in the Experiment I that is evaluated with the DII dataset.
- Experiment III: In this experiment, we train a model with the training dataset belonging to the DII dataset and we evaluate the model trained with the testing dataset related to the DII dataset.
- Experiment IV: In this experiment, we consider the model trained in Experiment III that is evaluated with the DI dataset.

We consider training and testing datasets related to different datasets in experiments III and IV with the aim of showing the generalization capability of the proposed method and thus the model robustness.

For both models, we set the number of epochs equal to 500. The experiment parameters were set in an empirical way.

Figure 4 shows the segmentation capabilities every 25 epochs (for the first 100 epochs) of the U-Net model trained with the DI dataset.

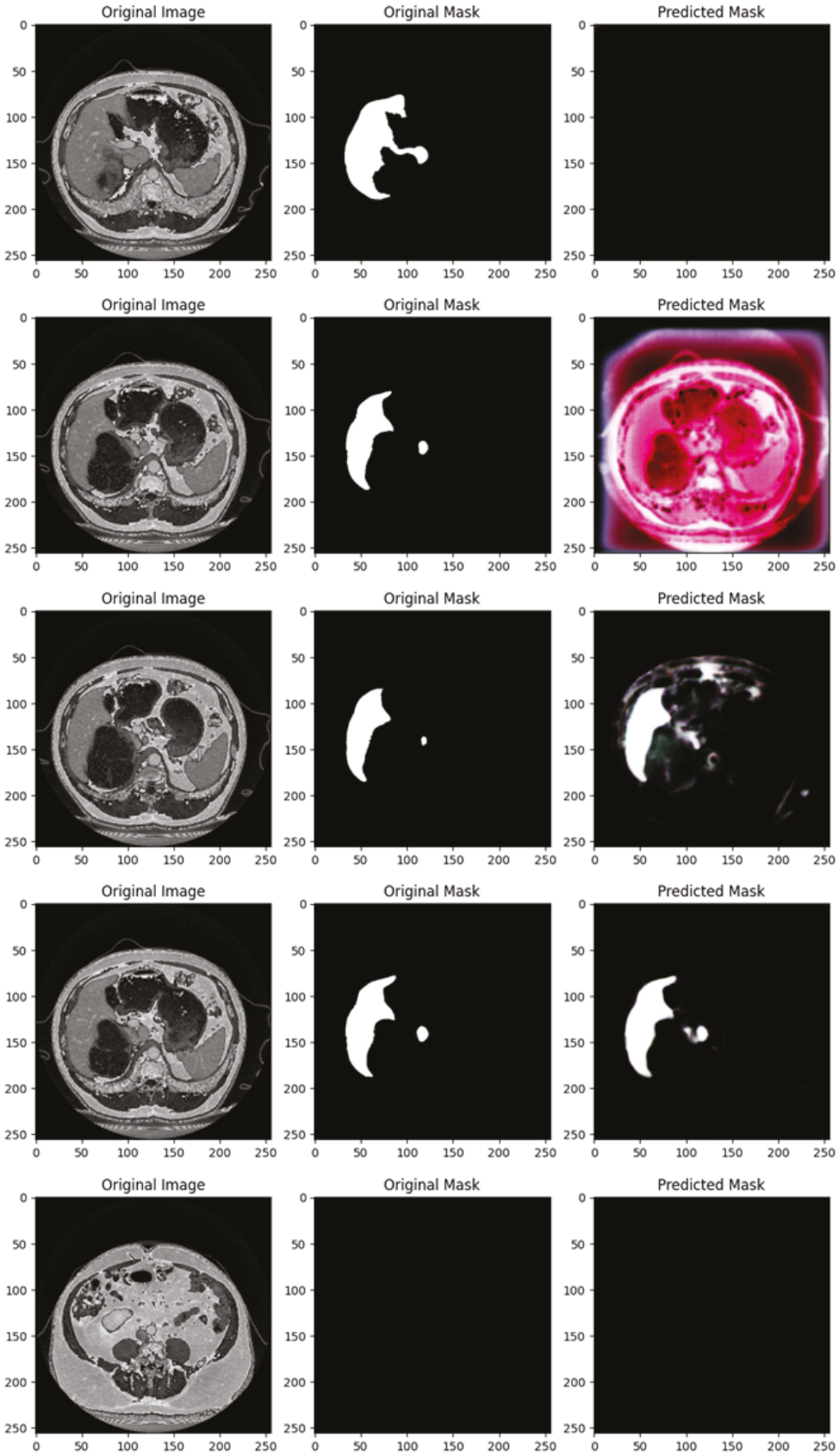


Figure 4. An example of segmentation mask in the U-Net trained, respectively, for 1, 25, 50, 75 and 100 epochs with the DI dataset.

Figure 4 shows, for different epochs, a set of images with the related segmentation mask predicted at a certain epoch. In particular, it shows the original images, original masks (i.e., the ground truth) and the predicted masks generated by a U-Net network trained for different numbers of epochs (1, 25, 50, 75, and 100 from top to bottom). The aim of the U-Net network is likely to segment a specific region of interest in the abdominal CT scans, i.e., the liver, based on the shape of the masks.

In the following, we provide a detailed analysis of the several masks generated for each epoch considered in Figure 4.

At 1 epoch, the original abdominal CT scan clearly shows several internal structures. The original mask marks the liver region in white, indicating its precise location and shape. However, the predicted mask is empty, suggesting that the U-Net model has not yet learned to recognize and segment the liver region. This is expected, as early stages of training often result in poor or no predictions.

At 25 epochs, the model begins to show some improvement. The original image is another abdominal CT slice, displaying a similar view. The liver region is again marked in white in the original mask, now with a small additional structure. The predicted mask, although incomplete and somewhat noisy, shows a partial segmentation of the liver. The areas predicted by the model are highlighted in pink, indicating that while there is some correspondence with the original mask, the model is still in the process of learning to isolate the liver region correctly.

By 50 epochs, the model's performance significantly improves. The original image is another abdominal slice, and the liver, along with a small adjacent structure, is clearly demarcated in the original mask. The predicted mask now captures a larger portion of the liver, though there is still some noise around the region, signaling that the model is not yet fully refined. However, it has made substantial progress in recognizing and outlining the liver.

At 75 epochs, the model's segmentation is much more accurate. The original image remains consistent with the others in this set. The liver region is again highlighted with a small additional white structure in the original mask. The predicted mask now closely approximates the shape of the liver shown in the original mask, with minimal noise. The model has become much better at distinguishing the target region, though a small amount of excess segmentation remains, indicating that further refinement is possible.

At 100 epochs, a new abdominal slice is introduced. However, this slice does not have an original mask, possibly indicating a different region or a non-target slice. The predicted mask is empty, similar to the prediction after 1 epoch. Since the original mask is missing, this suggests that the model is either not designed to segment this specific structure or has successfully learned not to apply a mask where no liver structure is expected.

As shown from Figure 4, as the number of epochs increases, the U-Net network demonstrates clear improvement in segmenting the liver. At 1 epoch, the model has no predictive capability. By 25 epochs, it starts to recognize relevant areas but is still noisy. From 50 epochs onward, the predicted masks align more closely with the original masks, with improved segmentation accuracy and reduced noise. The model at 100 epochs demonstrates the ability to avoid unnecessary predictions in slices where no liver is present, indicating the model has generalized better.

Training the U-Net model over more epochs results in significant improvements in segmentation quality, with the model learning to more accurately capture the target region (liver) and minimize noise. By 75 epochs, the predictions are highly accurate, and by 100 epochs, the model shows the ability to generalize to slices where no segmentation is required, which is a sign of good model training.

Figure 5 shows the segmentation capabilities every 25 epochs (for the first 100 epochs) of the U-Net model trained with the DII dataset.

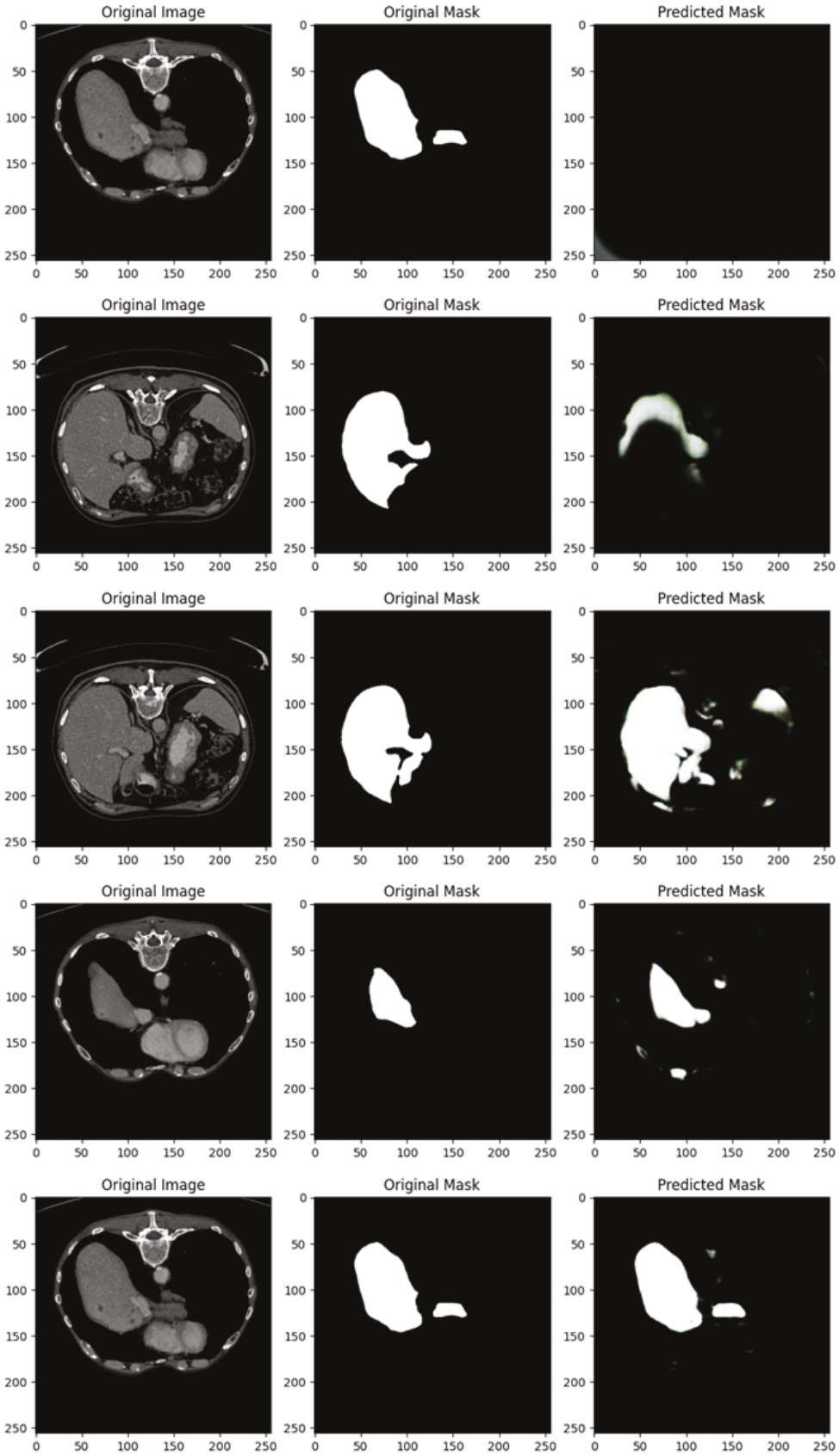


Figure 5. An example of segmentation mask in the U-Net trained, respectively, for 1, 25, 50, 75 and 100 epochs with the DII dataset.

Figure 5 shows, similarly to Figure 4, for different epochs, a set of images with the related segmentation mask predicted at a certain epoch. In particular, Figure 5 shows the original images, original masks (i.e., the ground truth) and the predicted masks generated by a U-Net network trained for different numbers of epochs (1, 25, 50, 75, and 100 from top to bottom).

We provide a detailed analysis for the several masks generated for each epoch considered in Figure 5:

At 1 epoch, the CT scan shows an upper abdominal slice, and the original mask clearly marks a white region, which appears to be a segment of an organ, likely the spleen or liver. However, the predicted mask is empty, suggesting that the U-Net model, at this early stage of training, has not yet learned to recognize and predict the target region. This is typical for early-stage training models and aligns with the expected behavior of the model at this point.

At 25 epochs, the model begins to show some improvement. The original image is another abdominal CT slice with visible internal organs, and the target organ (likely the liver or spleen) is again demarcated in white. The predicted mask at this stage is partial and somewhat noisy, lacking clear boundaries. While the shape begins to resemble the organ, the prediction is still underdeveloped, which is characteristic of intermediate-stage training.

By 50 epochs, the model's performance has significantly improved. The original image is another CT scan slice showing a similar anatomical view, and the organ of interest is clearly outlined in white. The predicted mask captures a much larger portion of the target region, and the segmentation is more refined. Despite this progress, some blurred edges and noise remain, indicating that the model is still refining its boundaries, which is expected as it continues to learn.

At 75 epochs, the model achieves a high level of accuracy. The original image is another abdominal CT slice, and the target organ is clearly marked with a smaller, localized region. The predicted mask closely resembles the original mask, with distinct boundaries and minimal noise compared to earlier stages. At this point, the U-Net model has learned to accurately isolate and segment the region of interest with greater precision.

At 100 epochs, the model demonstrates very high accuracy. The original image is an upper abdominal CT scan, and the original mask again marks the target organ with clear boundaries. The predicted mask is almost identical to the original, with sharp boundaries and very little noise or extraneous segmentation. This shows that by 100 epochs, the U-Net model is able to predict the target region with high fidelity and precision.

As shown from Figure 5, as the number of epochs increases, the U-Net network shows clear improvement in segmenting the liver organ. At 1 epoch, the model fails to make any predictions. By 25 epochs, it begins to recognize some parts of the organ, though the segmentation is noisy and incomplete. From 50 epochs onward, the segmentation becomes progressively more accurate, with the U-Net model capturing more of the organ's shape and reducing background noise. At 75 and 100 epochs, the predicted masks closely match the original masks, showing the model's capability to accurately segment the target region after more extensive training.

With increased training, the model becomes progressively better at recognizing and segmenting the target region. By 100 epochs, the model reaches a stage where the predicted masks are highly accurate, closely mirroring the original masks with minimal noise and strong boundary definition.

Figures 6 and 7 show the trend of the loss relating to the U-Net model, respectively, trained with the DI and the DII datasets.

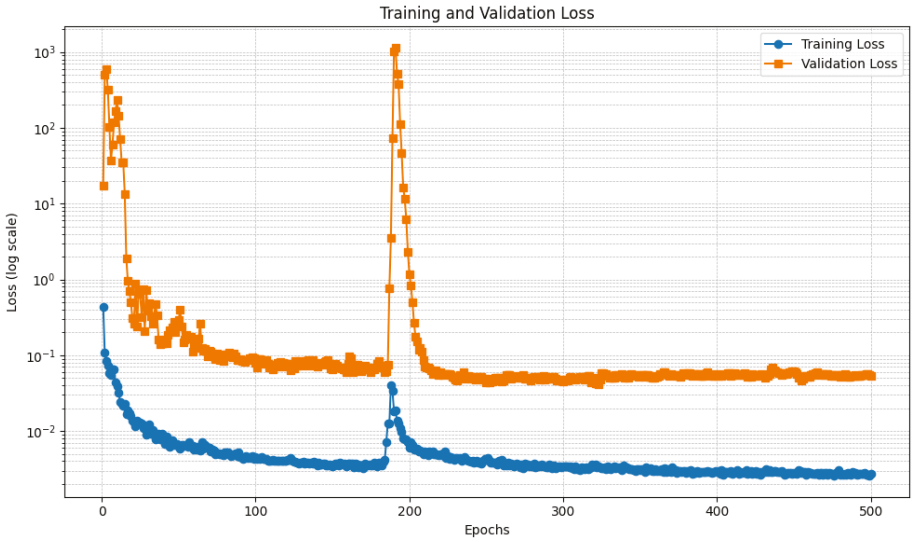


Figure 6. The training and the validation loss trend for the U-Net model trained with the DI dataset.

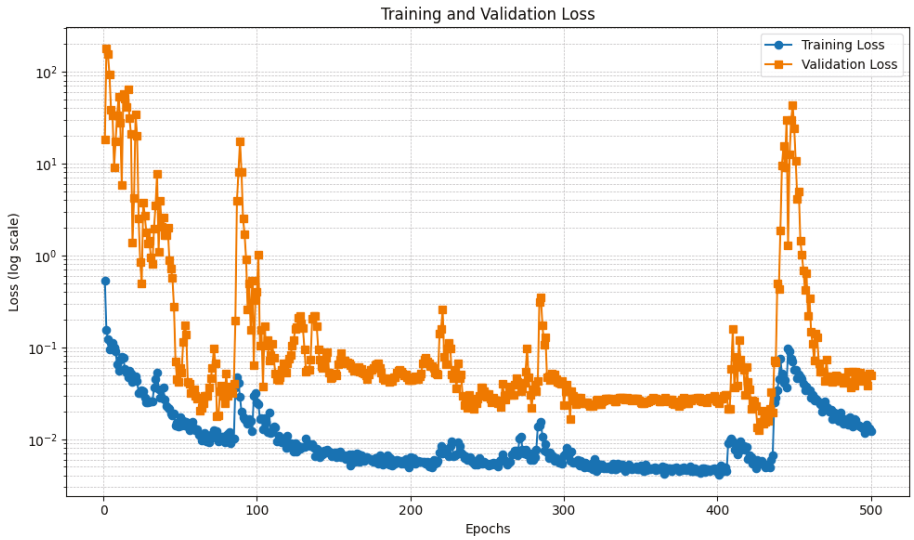


Figure 7. The training and the validation loss trend for the U-Net model trained with the DII dataset.

In detail, Figures 6 and 7 present two plots showing the training loss and validation loss over 500 epochs for the U-Net models each trained on different datasets, with the y-axis in logarithmic scale to better understand the loss trend. Both plots represent how well the models are learning and generalizing during training, but they exhibit significant differences in behavior.

Regarding Figure 6, this plot depicts the training and validation loss trends over 500 epochs, with the y-axis using a logarithmic scale to accommodate the wide range of loss values. The training loss, represented by the blue curve, demonstrates a steady decrease as the training progresses, stabilizing at a low value after approximately 200 epochs. This indicates that the model is effectively learning from the training data. In contrast, the validation loss, represented by the orange curve, starts with significant fluctuations during the initial epochs. This instability is likely due to high variance in the validation data or the model's early-stage adjustments. A notable spike is observed around epoch 200, after which the validation loss stabilizes. This spike could be linked to a sudden change in the training dynamics, such as a learning rate adjustment or a major update in the model's parameters. By the end of the training process, both the training and validation losses converge to relatively low values. However, the final validation loss remains slightly higher than the training loss, suggesting a degree of overfitting. This gap might indicate

the need for additional regularization techniques, such as dropout, early stopping, or data augmentation, to improve the model's generalization. Overall, the plot shows that the model performs well in minimizing both training and validation loss, except for the transient instability around epoch 200. Further analysis of this spike and the training dynamics in that region could provide insights into improving model performance.

With regard to Figure 7, this plot illustrates the training and validation loss trends over 500 epochs with a logarithmic y-axis to visualize the wide range of loss values effectively. The training loss, shown in blue, demonstrates a consistent downward trend, stabilizing at a low value after approximately 100 epochs. This indicates that the model is learning effectively from the training data, with minimal fluctuations observed throughout the process. In contrast, the validation loss, represented by the orange curve, exhibits considerable variability across the epochs. During the initial phase, the validation loss starts high and fluctuates significantly, stabilizing only after approximately 100 epochs. However, there are multiple noticeable spikes throughout the training process, particularly around epochs 100, 300, and 400. These spikes suggest disruptions in training dynamics, possibly due to changes in the learning rate, issues with batch selection, or sensitivity to certain patterns in the validation set. Despite the fluctuations, the overall trend shows that the validation loss stabilizes to relatively low values in later epochs, indicating reasonable generalization. However, the final validation loss remains consistently higher than the training loss, which could point to slight overfitting. Addressing this may involve incorporating regularization methods, adjusting the learning rate schedule, or increasing the validation data's representativeness. Overall, while the model shows effective learning and generalization trends, the recurring spikes in validation loss require further investigation. Identifying the root cause of these spikes could help improve model stability and overall performance.

Relating to the differences between the two U-Net models, the first plot indicates more stable validation performance, with only a single notable spike in the validation loss before stabilization. This suggests that the training process was smoother and potentially better tuned, possibly due to a more effective learning rate schedule, data preparation, or model architecture. The second plot, however, shows recurring validation loss spikes, indicating that the model experienced more difficulty in maintaining stable performance across the validation set. These fluctuations could result from an unstable training process, sensitivity to specific validation batches, or an inadequately tuned learning rate. Despite the differences, both plots demonstrate effective learning, as evidenced by the consistent decrease in training loss and the eventual convergence of validation loss. However, the second plot's instability suggests the need for further investigation into the training dynamics. Strategies such as refining the learning rate schedule, increasing the validation set's size or diversity, or employing techniques like early stopping might help mitigate the observed fluctuations and improve overall stability.

We analyze each experiment into a distinct subparagraph, where we discuss the results and we provide a visual impact related to the mask prediction performed by the proposed models.

4.5. Experiment I Results

In this section, we provide the experimental analysis results related to Experiment I. The testing accuracy in this experiment is equal to 0.8129, the Dice coefficient is equal to 0.788 and the IoU is equal to 0.65.

Figure 8 shows several examples of prediction related to Experiment I.

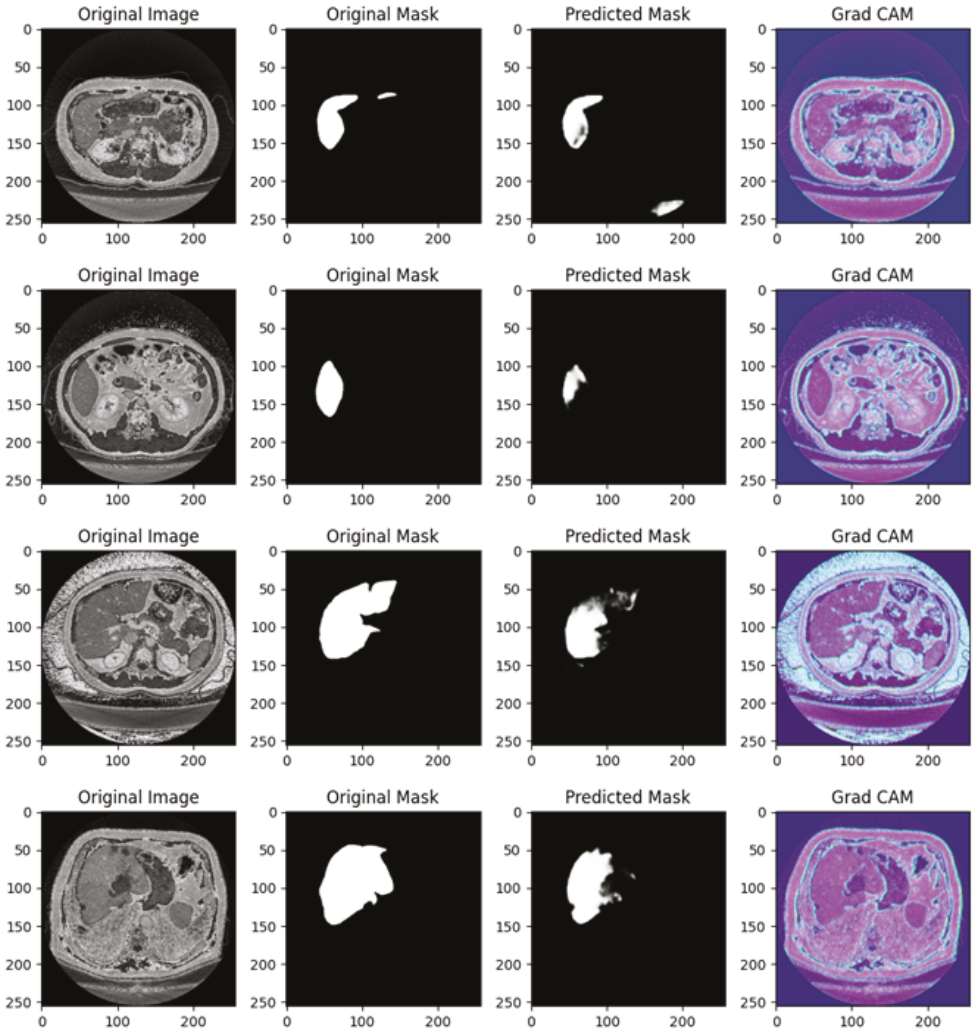


Figure 8. Examples of prediction related to Experiment I, i.e., with images obtained with the model trained on images belonging to the DI dataset with testing images related to the DI dataset.

Figure 8 shows different examples of segmentation performed by the proposed method. In particular, it is related to four different cases, each including the original image, original mask, predicted mask, and a Grad-CAM visualization.

In the following, we provide a detailed analysis for each example of prediction shown in Figure 8:

Example of Prediction 1: The CT scan shows an abdominal cross-section, and the original mask marks a small, distinct organ or lesion in white. The U-Net model's prediction shows partial success; however, the predicted region is smaller and incomplete compared to the original mask, with some background noise visible in the lower right corner. The Grad-CAM heatmap highlights areas the model deems important for making its prediction. The regions in blue and purple have higher activation and focus primarily around the central abdominal organs. However, the important areas do not align perfectly with the true region of interest, suggesting that the model is not fully capturing the relevant structures.

Example of Prediction 2: In a different CT scan showing another cross-sectional view of the abdomen, the original mask isolates a well-defined small structure, likely an organ, in white. The model's prediction in this case is less accurate than in the previous example, with a significant under-prediction. Only a faint, noisy portion of the original mask is captured. The Grad-CAM heatmap indicates that the model is activating over certain areas

but appears to miss the actual region of interest. The model's focus on other parts of the image likely contributes to the poor segmentation in this instance.

Example of Prediction 3: This CT scan shows a broader abdominal region with more complex anatomy. The target organ or lesion is segmented much more prominently and clearly in the original mask. The U-Net model performs better here, capturing a significant portion of the organ or lesion. However, the prediction is still slightly smaller than the original mask, with some blurred edges and missing parts. The Grad-CAM activation map is well aligned with the central portion of the image, focusing on regions relevant to the segmentation task. This explains the improvement in prediction accuracy compared to earlier examples, as the model is now focusing on more pertinent areas of the image.

Example of Prediction 4: In another cross-sectional CT scan of the abdomen, the original mask highlights a medium-sized organ or lesion segmented in white. The prediction is quite accurate, capturing the main structure well, but still slightly under-segmented compared to the original mask. There are minor discrepancies, with some noise around the edges. The Grad-CAM heatmap shows a good focus on the relevant areas, with the model attending to regions near the original mask, which explains the more precise segmentation in this case.

The U-Net model demonstrates a mixed ability to accurately predict the regions of interest. In some cases (like in examples of predictions 3 and 4), the predictions are relatively close to the original mask, but in others (such as examples of predictions 1 and 2), there are clear deficiencies, with significant under-segmentation and background noise. Grad-CAM provides a valuable insight into the decision-making process of the model. In examples of prediction where the segmentation is more accurate (such as examples of predictions 3 and 4), the Grad-CAM heatmaps show a stronger focus on the relevant regions, while in poorer predictions (examples of predictions 1 and 2), the activations are misaligned with the actual target region, which contributes to incorrect segmentation. The quality of predictions improves as the Grad-CAM focus becomes more aligned with the region of interest.

4.6. Experiment II Results

In this section, we provide the experimental analysis results related to Experiment II. The testing accuracy in this experiment is equal to 0.8876, with a Dice coefficient equal to 0.857 and with an IoU equal to 0.75.

Figure 9 shows several examples of prediction related to Experiment II.

Figure 9 in related to a comprehensive analysis of the segmentation predictions of a U-Net model across four different cases, each including the original image, original mask, predicted mask, and a Grad-CAM visualization.

In the following, we provide a detailed analysis for each example of prediction shown in Figure 9; this analysis showcases how well the U-Net network is performing in identifying and segmenting regions of interest in abdominal CT scans, as well as the model's focus during predictions.

Example of Prediction 1: The CT scan shows a cross-sectional view of the abdomen, with the target area, likely a liver or lesion, well-defined and highlighted in white in the original mask. The U-Net model captures a significant portion of the target region, but over-segmentation artifacts are present, with additional highlighted areas not found in the original mask. The prediction is somewhat noisy, and parts of the background are incorrectly included in the segmentation. The Grad-CAM heatmap reveals high activations around the central liver area, indicating that the model is focusing on the correct region. However, significant activation in peripheral areas may explain the incorrect predictions in the background.

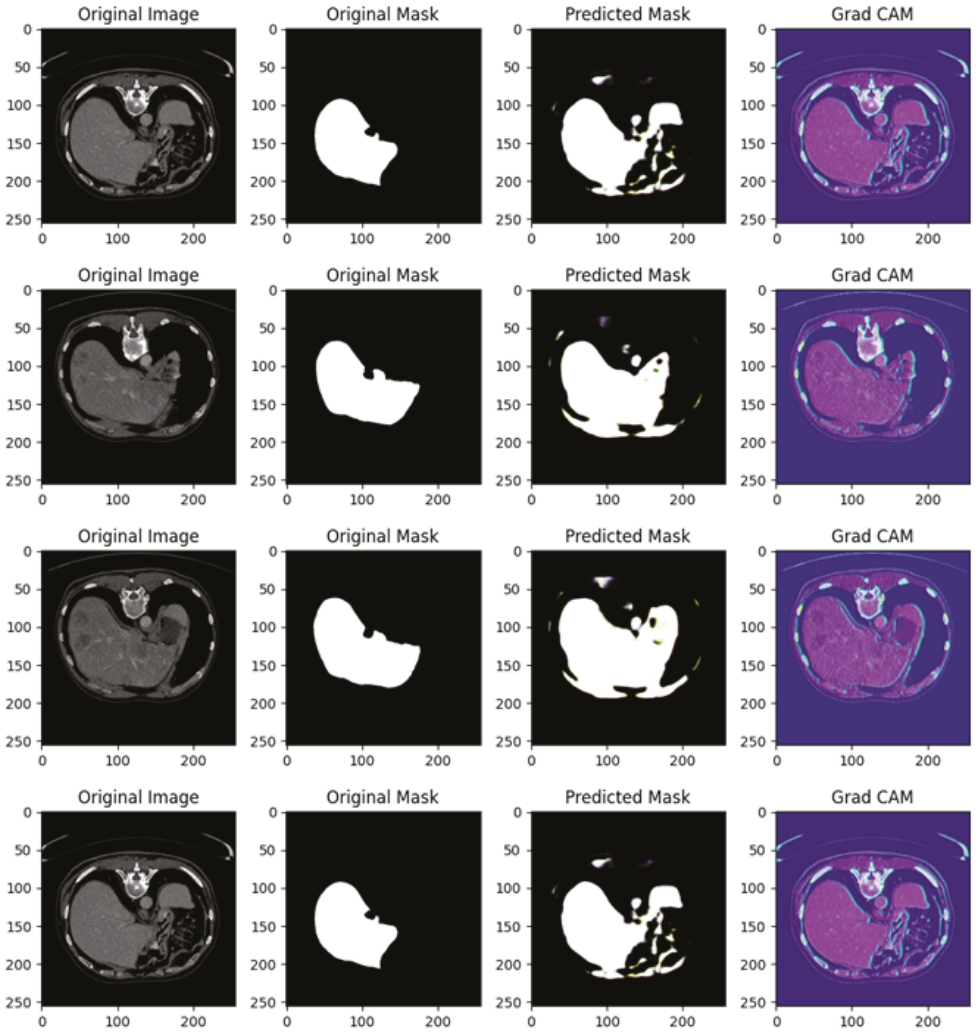


Figure 9. Examples of prediction related to Experiment II, i.e., with images obtained with the model trained on images belonging to the DI dataset with testing images related to the DII dataset.

Example of Prediction 2: In a different abdominal CT scan, showing a similar anatomical structure to the first example, the original mask clearly segments the organ of interest, likely the liver. The U-Net prediction closely resembles the original mask, though small regions around the edges show slight over-segmentation, and there is some minor noise. The model appears to correctly identify the boundaries of the target region, although not perfectly. The Grad-CAM heatmap highlights a strong focus on the liver area, with concentrated activations near the region of interest, reflecting the improved segmentation accuracy compared to the previous example.

Example of Prediction 3: This image shows a cross-sectional view of the abdomen with greater contrast than the previous scans. The original mask outlines a large portion of the target organ accurately. The prediction is highly accurate, with minimal over-segmentation or noise present. The U-Net model performs well, capturing the organ's structure almost perfectly, with slight discrepancies only at the edges. The Grad-CAM heatmap indicates a precise focus on the relevant areas, with the model's attention well concentrated on the organ of interest, which explains the high prediction accuracy. Peripheral activations are minimal, leading to reduced noise.

Example of Prediction 4: Another similar cross-sectional abdominal scan is presented, with the original mask clearly segmenting the target organ. The predicted mask captures the main structure, but, as in the first example, over-segmentation artifacts are present.

The model has overextended the segmentation, including areas outside the original mask, although the general shape of the organ is correctly identified. The Grad-CAM heatmap shows a strong focus on the central region, but some activations are spread across the periphery, particularly in regions not part of the target area. This suggests that the over-segmentation is likely due to the model's attention being scattered across the image.

The U-Net model is generally successful at predicting the organ of interest but tends to over-segment in certain instances (as seen in prediction examples 1 and 4). This could be due to background features or noise influencing the model's predictions. The Grad-CAM visualizations are instrumental in identifying where the model's attention is focused during the segmentation process. In examples showing better predictions (i.e., 2 and 3), the heatmap shows concentrated activations on the target regions. In predictions with more over-segmentation (i.e., 1 and 4), the model's focus is more dispersed, leading to less precise predictions.

The quality of the predictions varies, with examples of predictions 2 and 3 showing near-perfect segmentations, while examples of predictions 1 and 4 display more noise and over-segmentation. The Grad-CAM maps corroborate this by indicating that the model's focus aligns better with the target areas in examples of predictions 2 and 3.

The U-Net model's segmentation performance shows strong potential but is somewhat affected by over-segmentation in some cases. The Grad-CAM visualizations offer valuable insights into the model's decision-making process, indicating that further refinement could be made to improve the focus on the target regions and reduce noise in the predictions. Overall, the model demonstrates good performance, especially in examples of predictions 2 and 3.

4.7. Experiment III Results

In this section, we provide the experimental analysis results related to Experiment III. The testing accuracy in this experiment is equal to 0.8948, with a Dice coefficient equal to 0.888 and an IoU metric with a value of 0.80.

Figure 10 shows several examples of prediction related to Experiment III.

Figure 10 is related to a comprehensive analysis of the segmentation predictions of a U-Net model across four different cases, each including the original image, original mask, predicted mask, and a Grad-CAM visualization.

Example of Prediction 1: The CT scan shows a cross-sectional view of the liver region, with surrounding anatomy clearly visible. In the original mask, the target organ (the liver) is accurately segmented. The U-Net model's prediction closely follows the shape of the original mask, but there is a small under-segmentation issue, particularly around the edges of the liver. While the shape and size are mostly consistent, some fine details are missing. The Grad-CAM heatmap reveals that the model predominantly focuses on the liver region. However, there are some peripheral areas of activation outside the liver, which may explain the minor segmentation inaccuracies.

Example of Prediction 2: Another CT scan provides a slightly different perspective of the abdomen. The original mask segments the target organ over a larger area compared to the first row. The U-Net model's prediction is very close to the original mask, capturing almost the entire organ accurately. Though there is a slight discrepancy in the boundary, the prediction is overall more accurate than the first one. The Grad-CAM heatmap shows that the model's attention is well concentrated on the target organ, with minimal distractions outside the region of interest. This focused attention explains the good quality of the prediction.

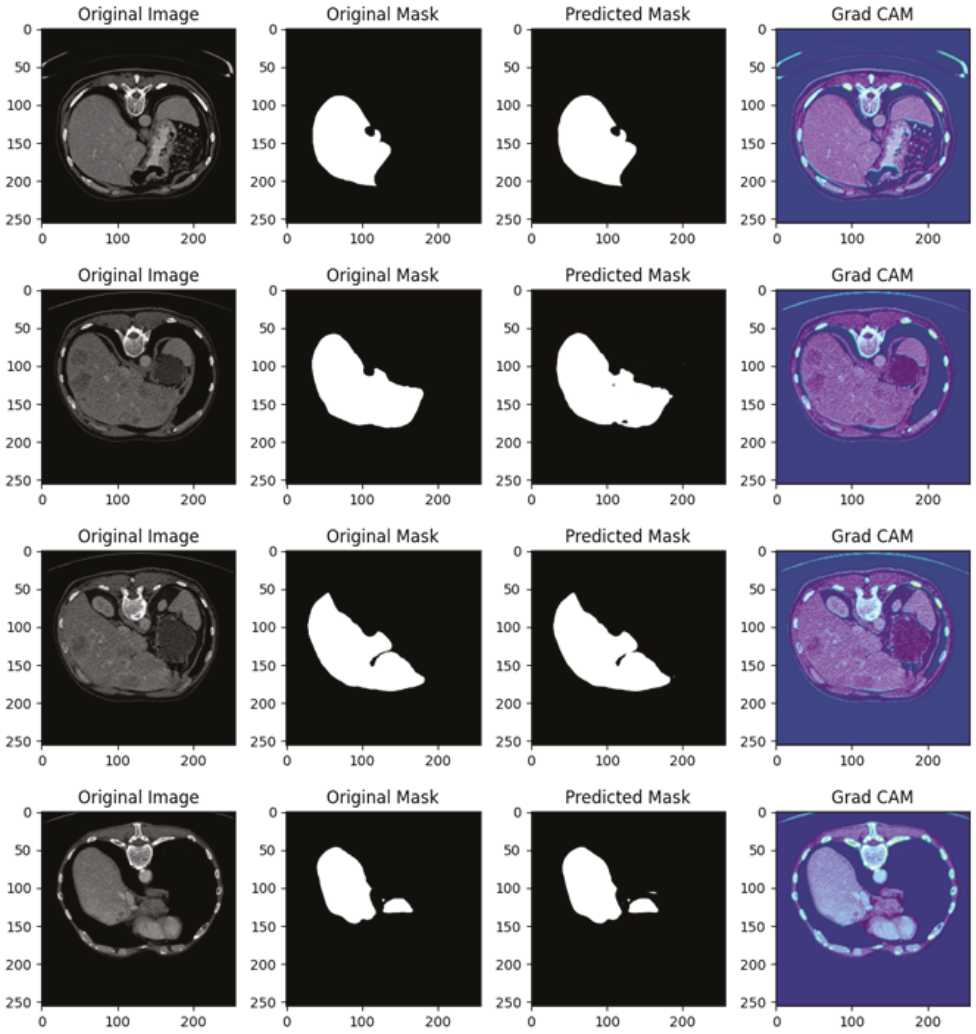


Figure 10. Examples of prediction related to Experiment III, i.e., with images obtained with the model trained on images belonging to the DII dataset with testing images related to the DII dataset.

Example of Prediction 3: This CT scan shows another cross-sectional view of the abdomen, where the original mask highlights a larger organ with precise contours. The U-Net model's prediction performs well, capturing most of the organ. However, there is minor over-segmentation, with the model including some extraneous areas that do not belong to the target organ. Despite this, the overall shape of the organ is well preserved. The Grad-CAM heatmap shows strong activations around the liver area, but there are also peripheral regions activated, which correlates with the over-segmentation observed in the prediction.

Example of Prediction 4: This scan presents a slightly altered view of the abdomen, similar to the previous rows. In the original mask, the liver is clearly outlined. The U-Net prediction shows under-segmentation, where the model fails to capture the entire organ, especially around the edges. The size of the predicted mask is notably smaller than the original mask. The Grad-CAM heatmap indicates that the model's focus is somewhat diffused, with significant attention placed on regions outside the target organ, which likely explains the under-segmentation seen in the predicted mask.

The U-Net model performs well overall, but the predictions show some variability. Prediction examples 2 and 3 display better segmentation results, with only minor deviations in boundary precision, while prediction examples 1 and 4 exhibit under-segmentation issues, where parts of the organ are missed.

In prediction examples 1 and 4, there is under-segmentation, meaning the model fails to capture the full extent of the target organ. In prediction example 3, over-segmentation is observed, where parts outside the target area are mistakenly included.

The Grad-CAM visualizations provide valuable insights into the model's decision-making process. Prediction examples 2 and 3 show that the model focuses well on the relevant regions, leading to more accurate predictions. However, in prediction examples 1 and 4, the Grad-CAM maps reveal that the model is focusing on regions beyond the target, resulting in less precise predictions.

The U-Net model exhibits promising segmentation performance across the images, particularly in prediction examples 2 and 3, where the predictions closely match the original masks. The Grad-CAM visualizations effectively highlight areas where the model's attention is misdirected, offering useful guidance for model tuning and enhancement.

4.8. Experiment IV Results

In this section, we provide the experimental analysis results related to the Experiment IV. The testing accuracy in this experiment is equal to 0.9305, with a Dice coefficient equal to 0.85 and an IoU of 0.85.

Figure 11 shows several examples of prediction related to Experiment IV.

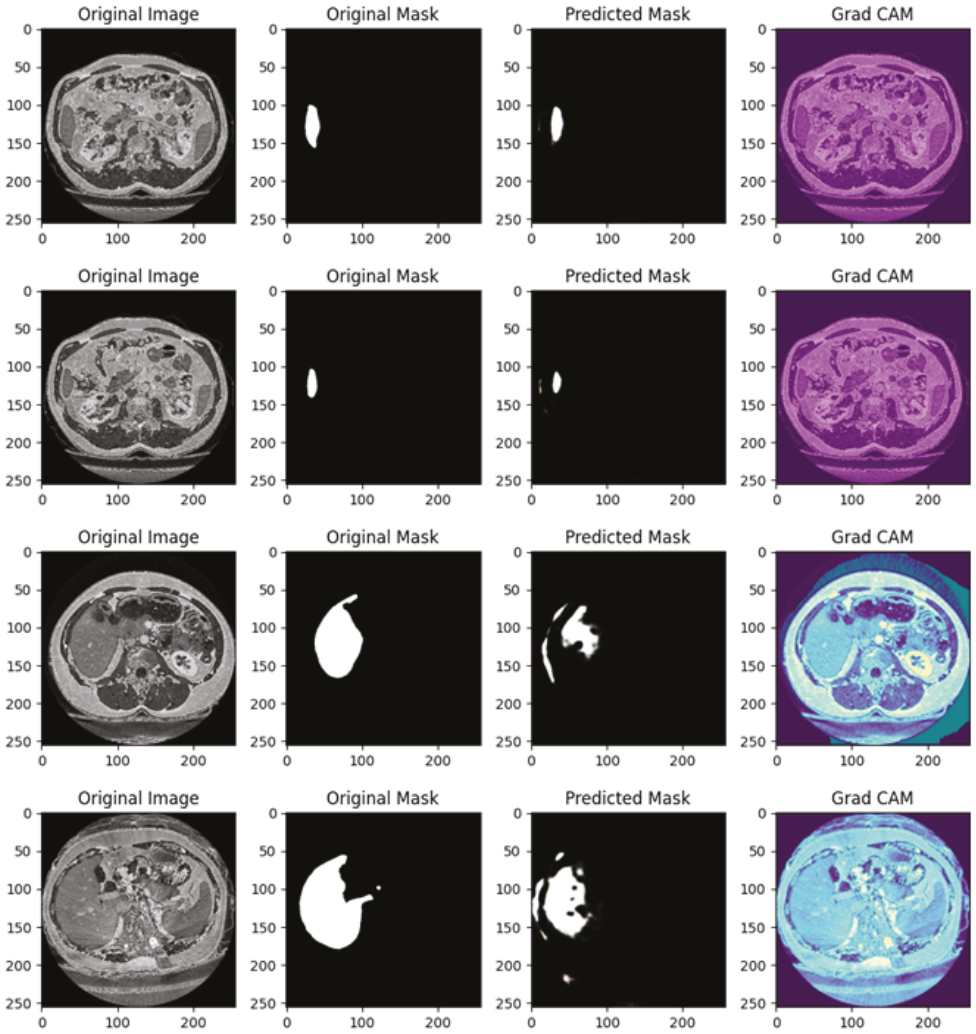


Figure 11. Examples of prediction related to Experiment IV, i.e., with images obtained with the model trained on images belonging to the DII dataset with testing images related to the DI dataset.

Figure 11 is related to a comprehensive analysis of the segmentation predictions of a U-Net model across four different cases, each including the original image, original mask, predicted mask, and a Grad-CAM visualization.

Example of Prediction 1: The CT scan shows an image of the abdomen with the liver as the region of interest. The original mask displays a very small target area that corresponds to a specific structure. However, the U-Net model's prediction only partially captures the structure, showing under-segmentation. It does not fully cover the target area, and the shape of the predicted mask does not perfectly match the original. The Grad-CAM heatmap reveals that the model's attention is somewhat diffuse, focusing on the target region but extending to a broader area. This diffuse focus may explain why the predicted mask is incomplete, as the attention is not entirely centered on the target structure.

Example of Prediction 2: Another CT scan provides a different cross-section of the abdomen. The original mask is less extensive than in the first example, but it clearly outlines the organ structure. The U-Net model's prediction shows some under-segmentation, although it captures more of the structure compared to the first row. Despite this, the boundaries are not well defined, and the predicted region is smaller than the original mask. The Grad-CAM heatmap indicates that the model's attention is focused on the organ of interest, similar to previous heatmaps, with some peripheral regions included. This may explain why the model struggles with precise boundary detection in its predictions.

Example of Prediction 3: This CT scan offers a slightly different view of the abdomen. The original mask segments the organ with accurate contours. The predicted mask is more accurate than in previous rows but still exhibits over-segmentation. The prediction includes areas outside the target structure, suggesting that the model has captured additional regions that were not part of the original mask. The Grad-CAM heatmap shows that the model is strongly focused on the target region, but some activation extends beyond the intended structure. This extra activation corresponds to the over-segmentation observed in the predicted mask.

Example of Prediction 4: Another CT scan presents a different cross-section of the abdomen. The original mask is well defined, covering a larger region compared to earlier examples. However, the U-Net model's prediction exhibits significant over-segmentation. The prediction exceeds the boundaries of the original mask and includes non-target areas. The Grad-CAM heatmap shows that the model is focusing not only on the organ of interest but also on the surrounding areas. This broader focus likely contributes to the over-segmentation observed in the predicted mask.

The U-Net model exhibits mixed performance. In the first two prediction examples, the model struggles with under-segmentation, where it fails to capture the entire organ, leading to incomplete masks. In the third and fourth rows, the model suffers from over-segmentation, where it captures more than just the target area, predicting larger masks than needed. The segmentation inaccuracies vary between under-segmentation (first and second prediction examples) and over-segmentation (third and fourth prediction examples). This suggests the model struggles with precise boundary detection, either missing parts of the target or including too much of the surrounding area.

The Grad-CAM visualizations offer insight into the model's focus areas. In the first two prediction examples, the model's attention is dispersed, contributing to the under-segmentation observed. In the last two prediction examples, the model's attention is broader, extending beyond the intended target, which may explain the over-segmentation. These heatmaps are useful for diagnosing where the model's attention drifts outside the relevant area.

The U-Net model demonstrates varying degrees of segmentation accuracy in this set of images. It struggles with under-segmentation in the first two rows, likely due to diffused

attention seen in the Grad-CAM heatmaps. In the third and fourth rows, the model exhibits over-segmentation, with Grad-CAM visualizations showing excess attention outside the target organ. Refining the model to better focus its attention within the boundaries of the target structure could lead to improved segmentation performance and reduce both under- and over-segmentation issues.

The proposed method demonstrates promising capabilities, but certain areas warrant further analysis and improvement. One interesting challenge in the experimental analysis relates to the potential for overfitting during the training process, which can undermine the robustness and generalizability of the model. The experimental results highlight specific observations about the behavior of the model and its robustness across datasets. The generalization capabilities of the approach are interesting, with accuracies ranging between 0.81 and 0.93 across experiments. This demonstrates its potential for handling variability in imaging conditions also acquired with different medical equipment. However, the validation loss spike observed during training raises concerns about stability. This spike suggests that the model may temporarily overfit to patterns in the training data, and although it recovers, the instability warrants a more detailed analysis. Adjusting the learning rate schedule or incorporating more robust early stopping criteria might address this issue.

Explainability is a critical aspect of the proposed method, as demonstrated through the Grad-CAM visualizations. These heatmaps reveal that the model focuses on anatomically relevant areas, enhancing trust in its predictions. However, in some instances, Grad-CAM outputs indicate that the model's attention extends to irrelevant regions, leading to over-segmentation.

The clinical relevance of the proposed approach is underscored by its ability to integrate into workflows for preoperative planning, disease progression monitoring, and therapy assessment. With high accuracy and explainability, the model has significant potential to assist clinicians in making informed decisions.

To evaluate the performance of the proposed liver segmentation method, several key metrics were considered:

- Dice Coefficient (DSC): The Dice coefficient values range from 0.788 to 0.92, depending on the experimental setup.
- Intersection over Union (IoU)/Jaccard Index: The reported IoU values vary between 0.65 and 0.88, representing the degree of overlap between the predicted and ground truth segmentations.

These metrics provide a comprehensive evaluation of segmentation accuracy, overlap, and boundary precision, ensuring robust performance analysis.

4.9. Explainability Analysis

The adoption of Grad-CAM enhances the prediction explainability of the proposed liver segmentation method by highlighting regions within the input images that significantly influence the predictions of the proposed model.

The integration of Grad-CAM with the U-Net model provides an explainability layer to the proposed liver segmentation method. Grad-CAM highlights regions in the input images that significantly influence the segmentation decisions, offering a dual perspective: an understanding of the model's internal decision-making process and insights into its clinical implications.

From the model point of view, the Grad-CAM outputs demonstrate areas of focus and reveal instances of both alignment and misalignment between the model's attention and the ground truth segmentation. For example, in Figure 8, Grad-CAM outputs for predictions from Experiment I highlight that the model's attention is often concentrated

on the central liver region during successful segmentations. However, the heatmaps also reveal peripheral activations in some cases, such as Example 1, which correspond to over-segmentation in the predicted masks. In Example 3, the Grad-CAM activations align well with the target liver region, resulting in a highly accurate prediction. These instances underscore the importance of targeted attention and reveal opportunities for refining the model to improve its focus.

Similarly, in Figure 9, which pertains to Experiment II, Grad-CAM visualizations demonstrate improved model focus. Examples 2 and 3 show that the heatmaps closely correspond to the actual liver boundaries, resulting in accurate segmentations. However, Example 4 reveals scattered activations in peripheral areas, correlating with over-segmentation. These observations suggest that while the model's generalization capability across datasets is promising, occasional misdirected attention can lead to inaccuracies.

From the clinician's perspective, Grad-CAM enhances trust in the model's predictions by providing visual evidence of the decision-making process. Accurate heatmaps, such as those in Figure 10 (Experiment III, Examples 2 and 3), reassure clinicians that the model bases its segmentation on biologically plausible and anatomically relevant regions. In contrast, the under-segmentation seen in Example 4 is accompanied by Grad-CAM heatmaps showing reduced focus on the liver region and misplaced attention on surrounding structures. This information is crucial for clinicians to identify cases requiring manual review, ensuring that the automated segmentation aligns with clinical standards.

Figure 11 from Experiment IV provides further insights into the model's robustness. Grad-CAM visualizations in Examples 2 and 3 indicate strong attention on the liver, corresponding to precise segmentations. However, Example 1 highlights diffuse activations beyond the liver's boundaries, which explain the under-segmentation observed in the predicted mask. These examples illustrate that while the model performs well overall, certain cases require refinement to enhance its reliability.

Specific examples in the study illustrate these dynamics:

In instances of under-segmentation, such as in Figure 8 (Experiment I, Example 1) and Figure 11 (Experiment IV, Example 1), Grad-CAM maps reveal reduced activation within the liver region and misplaced focus on surrounding structures. This indicates the model's struggle to fully capture the target area, which might stem from insufficient training data or a lack of robust feature representation for specific anatomical variations.

Cases of over-segmentation, such as in Figure 9 (Experiment II, Example 4) and Figure 10 (Experiment III, Example 3), correspond to Grad-CAM outputs showing diffuse activations beyond the liver's anatomical boundaries. This behavior suggests that the model is influenced by non-relevant features, which could be mitigated by incorporating stricter regularization techniques or refining the loss function to penalize erroneous activations.

In summary, the explainability offered by Grad-CAM is not only a valuable diagnostic tool for evaluating model performance but also an essential bridge for fostering trust and collaboration between AI systems and medical professionals. By pinpointing the exact regions influencing segmentation, Grad-CAM ensures that the proposed method is not only accurate but also interpretable and clinically viable.

5. Discussion

Liver segmentation has been a critical area of focus in medical imaging research due to its relevance in diagnosing and treating hepatic diseases. Traditional methods relied heavily on manual segmentation performed by radiologists, which is time-consuming and prone to inter-observer variability. To address these challenges, various computational approaches have been developed over the years.

Initial automated segmentation techniques used rule-based algorithms and statistical shape models, such as active contour models and level sets. While these approaches showed promise, they struggled with complex anatomical variations and noise in medical images, leading to suboptimal segmentation accuracy. The advent of machine learning introduced supervised learning algorithms, including support vector machines (SVMs) and random forests, which improved segmentation performance but required extensive feature engineering.

With the rise of deep learning, convolutional neural networks (CNNs) have become the dominant paradigm for medical image segmentation. The U-Net architecture [30] revolutionized biomedical image segmentation by employing an encoder–decoder structure with skip connections to retain spatial information. U-Net and its variants, such as 3D U-Net and attention U-Net, have been widely adopted for liver segmentation. For instance, Christ et al. [31] developed a cascaded U-Net for liver and tumor segmentation.

Attention mechanisms have further enhanced segmentation performance by enabling models to focus on relevant regions within medical images. Works by Oktay et al. [32] integrated attention gates into the U-Net architecture to refine segmentation outputs, particularly for small and complex structures. Additionally, hybrid approaches combining CNNs with transformer-based architectures have recently emerged, demonstrating superior performance in capturing global context.

Liver segmentation from medical images, such as computed tomography (CT) and magnetic resonance imaging (MRI), plays a crucial role in diagnosis and treatment planning for liver diseases, including liver tumors and cirrhosis. Recent advancements in deep learning have shown significant promise in improving the accuracy and efficiency of liver segmentation, with methods often based on convolutional neural networks (CNNs) and variants such as UNet, VNet, and attention-based mechanisms.

In the context of medical image segmentation, the U-Net model [30] is widely recognized for its effectiveness in medical image segmentation, owing to its encoder–decoder structure that captures both local and global features. Over the years, various enhancements and adaptations of the U-Net model have been proposed to tackle the specific challenges of liver segmentation, such as shape variability, texture heterogeneity, and low contrast in CT images.

The original U-Net [30] laid the foundation for numerous subsequent works in liver segmentation. Following this, several modifications were introduced to improve performance in challenging conditions. For example, Milletari et al. [33] proposed the V-Net, which extended U-Net to 3D segmentation by using volumetric convolutions and a Dice coefficient-based loss function. This approach demonstrated superior performance in 3D liver segmentation tasks.

Authors in Kavur et al. [34] conducted an extensive comparison of liver segmentation algorithms on multiple datasets, highlighting the strengths and limitations of each method. Their study showed that while 2D CNN-based methods, like traditional UNet, are efficient and computationally less expensive, 3D CNNs and hybrid approaches are better suited for capturing volumetric details essential for accurate liver segmentation.

Table 2 directly compares the state-of-the-art research works in the context of liver segmentation.

As shown from the state-of-the-art comparison shown in Table 2, the proposed method is able to obtain interesting performances.

Table 2. Comparison of deep learning-based liver segmentation methods.

Method	Backbone	DSC (%)
U-Net [30]	Encoder-Decoder	91.3
U-Net++ [35]	Dense U-Net	92.1
Attention U-Net [32]	Attention Mechanism	93.2
nnU-Net [36]	Self-Adaptive U-Net	94.6
DeepLabV3+ [37]	ResNet-101	92.5
TransUNet [38]	Transformer	93.8

In medical applications, the explainability of deep learning models is paramount for gaining clinician trust and improving the interpretability of predictions. Gradient-weighted Class Activation Mapping (Grad-CAM) [25] is one of the most widely used techniques for visualizing the regions of an image that contribute most to the model's decisions. While commonly applied in classification tasks, its integration into medical image segmentation models is an emerging research area.

As a matter of fact, there has been limited exploration of explainability in liver segmentation tasks. The proposed method addresses this gap by incorporating Grad-CAM with a U-Net architecture specifically tailored for liver segmentation, providing visual explanations that can aid clinicians in understanding and verifying the segmentation results.

The choice of U-Net with residual connections in this study is well justified, as it offers several advantages for medical image segmentation, particularly in liver CT scans. However, we acknowledge that additional architectural variations, such as attention mechanisms and transformer-based segmentation models, could further enhance performance. Below, we provide a justification for our chosen architecture while discussing potential improvements.

One of the key advantages of U-Net with residual connections is its ability to preserve fine-grained anatomical details. The encoder–decoder structure of U-Net effectively captures both global and local image features, which is particularly beneficial for segmenting the liver, an organ with complex boundaries. Moreover, the incorporation of residual connections improves gradient flow, addressing the issue of vanishing gradients and ensuring that the network learns fine details without losing spatial information.

Another important consideration is the efficiency of training on limited data. Medical datasets are often small, making deep networks susceptible to overfitting. Compared to transformer-based models, U-Net maintains a moderate level of complexity, allowing it to achieve high accuracy even with fewer training samples. The presence of skip connections further enhances the model's performance by retaining low-level spatial features from the encoder and reintroducing them in the decoder, ultimately improving segmentation accuracy despite the limited dataset size.

In addition to its accuracy, U-Net with residual connections offers computational efficiency, a crucial factor for clinical applications. While transformer-based models, such as Swin-UNet, have demonstrated state-of-the-art segmentation results, they require significantly higher computational resources, which may not always be available in real-world medical settings. By contrast, U-Net with residual connections strikes a balance between accuracy and efficiency, making it a practical choice for medical imaging tasks where computational constraints are a concern.

The U-Net model with residual connections and Grad-CAM explainability, employed in this study, is specifically designed for liver segmentation. However, its outputs can be seamlessly integrated with other AI-based diagnostic models, including those for lesion classification and survival prediction. The ability to segment liver structures with high

precision enhances the performance of these models by providing more reliable region-of-interest data and improving interpretability.

Lesion classification models typically depend on either global image features or manually defined regions of interest. The proposed segmentation method automates the extraction of these regions, ensuring that lesion classification models focus solely on liver areas rather than extraneous background structures. This automation reduces the likelihood of false positives and increases classification accuracy, particularly in cases involving complex liver textures or overlapping anatomical structures. Deep learning models such as CNN-based classifiers, including ResNet, VGG, and EfficientNet, can leverage the segmented liver regions to categorize lesions as benign, malignant, or cystic. Transformer-based models, such as Vision Transformers and Swin Transformers, can further enhance classification performance by capturing high-level spatial dependencies within segmented liver regions. Hybrid CNN-RNN models, designed to analyze time-series data, can utilize follow-up scans to track lesion progression or regression. The proposed segmentation approach benefits these classification models by automating lesion localization, improving feature extraction by reducing background noise, and providing explainable segmentation masks that clinicians can verify before proceeding with classification.

Beyond lesion classification, segmentation also plays a crucial role in survival prediction models. These prognostic models typically analyze patient-specific factors such as tumor size, liver volume, lesion count, and radiomic features to estimate survival probabilities. The segmentation outputs contribute to this analysis by extracting quantitative imaging biomarkers, including tumor volume, liver-to-tumor ratio, texture-based radiomic features that capture heterogeneity and intensity distribution, and shape descriptors reflecting tumor boundary irregularities. Machine learning techniques such as Random Forests, Support Vector Machines, and XGBoost can process segmentation-derived features to predict survival time. Meanwhile, deep learning models, including recurrent neural networks and long short-term memory networks, can track tumor growth patterns over multiple time points. Multimodal AI approaches that integrate convolutional neural networks with patient demographics, laboratory results, and genetic data offer a more comprehensive survival analysis. The integration of segmentation data into these models ensures precise tumor and liver delineation, improves interpretability by highlighting regions influencing survival risk, and supports longitudinal tracking of disease progression for personalized prognosis.

Despite these advantages, several challenges must be addressed to optimize integration with lesion classification and survival prediction models. Data standardization remains a significant concern, as variations in imaging protocols across different datasets require harmonization before integration. Additionally, survival prediction models may need segmentation beyond the liver, particularly for detecting vascular invasion or distant metastases. Another challenge is the computational complexity introduced by combining segmentation with classification or survival analysis models. Optimization techniques such as model pruning and the development of lightweight architectures could help mitigate these computational demands.

Future advancements in this area could focus on enhancing model fusion by developing end-to-end pipelines that integrate liver segmentation with lesion classification models. The exploration of multimodal AI approaches that combine imaging data with omics and clinical biomarkers could further refine survival prediction. Expanding the segmentation model to multi-phase CT and MRI data could provide more detailed liver lesion characterization and enable dynamic tumor assessment.

Overall, the proposed explainable U-Net segmentation model establishes a strong foundation for integration with lesion classification and survival prediction systems. By

automating liver region extraction and improving interpretability, it enhances AI-driven diagnostic workflows, leading to more precise, reliable, and clinically valuable predictions.

In the following, we compare the proposed method with state-of-the-art deep learning models and highlight its advantages over recent developments.

U-Net is widely used for biomedical segmentation due to its encoder–decoder architecture with skip connections, making it effective even when trained on limited datasets. However, it struggles with fine-grained details and complex boundaries, leading to suboptimal segmentation in cases involving tumors, lesions, or low-contrast regions. Furthermore, its lack of interpretability makes it difficult for clinicians to trust AI-generated segmentation results.

Enhancement in Our Approach: The integration of residual connections improves feature propagation and prevents vanishing gradients, resulting in better boundary preservation. Additionally, the use of Grad-CAM enhances explainability, allowing medical professionals to visualize which regions influence the segmentation decision.

Attention U-Net introduces attention gates to focus on relevant liver regions, reducing false positives. However, it is computationally more expensive, making real-time clinical deployment challenging, and still operates as a black box without explicit interpretability.

Enhancement in Our Approach: While attention mechanisms improve feature selection, our approach maintains computational efficiency by using residual connections instead of attention gates. Furthermore, Grad-CAM provides post hoc interpretability, which is absent in Attention U-Net.

Transformer-based models, such as TransUNet and Swin-UNet, leverage self-attention mechanisms to capture global context, making them highly effective for segmenting complex structures. They are also more robust to variations in imaging conditions. However, they require large datasets and extensive computational resources, limiting their usability in clinical settings. Furthermore, they remain less interpretable compared to CNN-based models.

Enhancement in Our Approach: While transformers excel in capturing long-range dependencies, our U-Net with residual connections achieves comparable accuracy at a lower computational cost. Moreover, Grad-CAM ensures explainability, which is often lacking in transformer-based segmentation models.

Hybrid models, such as CoTr and UNETR, combine CNNs for local feature extraction with transformers for global context modeling, improving segmentation generalization across datasets. However, these models suffer from high training complexity and memory requirements, which limits their real-world feasibility. Moreover, interpretability remains an issue.

Enhancement in Our Approach: Our method provides a balance between CNN-based efficiency and interpretability, ensuring robust segmentation without high computational costs. The proposed method builds on these advancements by integrating Grad-CAM with a U-Net model specifically tailored for liver segmentation. Unlike previous works, this study evaluates the model's robustness and explainability across multiple datasets acquired under varying conditions, ensuring both high performance and clinical applicability. By addressing gaps in the literature, such as the limited exploration of cross-dataset generalization and detailed Grad-CAM analyses, this research contributes a comprehensive and interpretable solution for liver segmentation.

The primary objective of the model is whole-liver segmentation, without explicitly distinguishing between normal hepatic parenchyma and pathological regions such as tumors, cysts, or transient hepatic attenuation differences (THADs). The missed lesions in Figures 5 and 9 are likely due to several factors. First, the datasets used (DI and DII) focus on liver boundaries rather than tumor annotations, preventing the model from learning lesion-

specific features. Additionally, many liver lesions exhibit low contrast on native CT scans, making them difficult to detect without arterial or venous phase enhancement. Finally, since the model was trained on single-phase CT images, it may struggle to generalize to lesions that appear differently across multiple imaging sequences. Given that lesion segmentation is critical for liver cancer diagnosis, the absence of this capability limits the direct clinical applicability of the model in oncology and hepatology.

To enhance the ability of the model to detect liver lesions, several improvements should be considered. One possible approach is incorporating multiphasic CT data by training the model on datasets such as LiTS, MSD, or 3D-IRCADb. This would allow the model to capture enhancement dynamics across arterial, venous, and delayed phases, thereby improving lesion segmentation accuracy. A multi-input deep learning model could be developed to fuse information from different phases, enabling a more comprehensive representation of liver lesions.

Another enhancement involves adapting the model for lesion segmentation. Instead of restricting the model to whole-liver segmentation, future iterations should incorporate multi-class segmentation that differentiates between normal liver tissue, benign lesions such as cysts, hemangiomas, THAD, and focal nodular hyperplasia (FNH), as well as malignant lesions including hepatocellular carcinoma (HCC), cholangiocarcinoma, and metastases. This extension could be implemented using a modified U-Net or a transformer-based segmentation model trained on datasets with annotated liver lesions.

Explainability is another crucial factor in improving lesion detection. Grad-CAM could be extended to highlight regions in the liver that contribute to an abnormal classification, providing radiologists with a visual representation of AI-based decisions. This would facilitate validation and enhance trust in automated segmentation results, making the model more interpretable and clinically applicable.

The proposed approach can seamlessly integrate into clinical workflows, assisting in preoperative planning, disease progression monitoring, and therapy assessment. Grad-CAM provides interpretability by highlighting areas of the image influencing the segmentation decisions. Improvements and explanations from the explainability point of view include the comparison of Grad-CAM outputs for correctly segmented and missegmented regions. It can highlight discrepancies to analyze failure cases and refine the model and to exploit Grad-CAM heatmaps to ensure that the highlighted regions correspond to liver edges and internal structures, confirming biologically plausible segmentation.

6. Conclusions and Future Work

Clinical imaging techniques, such as CT, currently represent essential tools that support clinicians in diagnosing diseases and planning appropriate interventions. This is especially crucial in malignant conditions like hepatocellular carcinoma, where precise image segmentation—specifically the delineation of the liver and tumor—serves as a critical first step in optimizing diagnosis, staging, and treatment strategies. Such treatments may include liver transplantation, surgical resection, radiotherapy, portal vein embolization, and various other procedures. Consequently, the effectiveness of segmentation methods can significantly impact both diagnostic accuracy and treatment outcomes.

In this paper, we proposed a deep learning-based approach to accurately segment the liver in medical images, addressing a pressing need in the diagnosis and treatment planning of hepatic diseases. Our approach leverages a U-Net architecture enhanced with residual connections, allowing for the capture of fine-grained anatomical details. We also prioritize prediction explainability by utilizing Grad-CAM, which highlights critical areas in the image that correspond to the segmentation results.

Experimental results reveal that the proposed method achieves an accuracy range of 0.81 to 0.93, demonstrating its potential as a reliable and efficient solution for automated liver segmentation. Moreover, we obtained a Dice index equal to 0.92 and a value related to IoU of 0.88.

As future work, we will consider different segmentation models, with the aim to try to improve the proposed performances. Moreover, we plan to adopt the proposed explainable segmentation method to such other organs as, for instance, the lung, with the aim of understanding whether it is also possible to provide explainable segmentation in other contexts. Furthermore, another interesting direction is to explore hybrid architectures that combine the strengths of convolutional neural networks and transformers. For instance, incorporating self-attention mechanisms within the U-Net framework could improve the model's ability to capture long-range dependencies, especially when dealing with complex liver boundaries. Models like TransUNet, which integrate transformers into segmentation tasks, have shown promising results in other domains, and applying similar techniques to liver segmentation could enhance performance.

Although lesion segmentation is a crucial task, whole-liver segmentation remains valuable for defining the region of interest (ROI) in subsequent lesion detection and classification. It plays a significant role in preoperative planning, particularly for liver resection and transplantation. Additionally, volumetric assessment is essential for evaluating cirrhosis progression, steatosis, and hepatocellular carcinoma (HCC) treatment response. The proposed method ensures accurate delineation of liver boundaries, providing a strong foundation for lesion-specific segmentation in clinical applications.

To enhance the clinical applicability of this study, future work can focus on integrating lesion segmentation within the existing framework. One approach involves fine-tuning the model for tumor segmentation by incorporating annotated liver lesion datasets, such as those from the Liver Tumor Segmentation Challenge (LiTS). Adapting the current U-Net model to segment both normal liver tissue and pathological regions would enable a more comprehensive analysis. A multi-class segmentation strategy could also be implemented, allowing the network to distinguish between healthy parenchyma, benign lesions, and malignant tumors.

Beyond dataset expansion, incorporating attention mechanisms could improve lesion enhancement. While the current model achieves precise whole-liver segmentation, it may struggle to highlight small or low-contrast tumors. The addition of attention-based modules, such as Attention U-Net or Squeeze-and-Excitation (SE) blocks, would enable the model to focus on lesion-prone regions, improving sensitivity in detecting abnormalities. Furthermore, extending Grad-CAM explainability to highlight lesion areas that most influence the segmentation decision would enhance clinical trust by ensuring alignment between AI-based predictions and radiologists' expectations.

The proposed method can also serve as a preprocessing step for lesion classification models. By integrating liver segmentation, lesion segmentation, and lesion classification, the workflow can evolve into a complete AI-driven diagnostic pipeline. This would facilitate tumor characterization, support treatment planning, and contribute to prognostic AI models for survival prediction. Combining segmentation with deep learning-based lesion classification and radiomics would provide a clinically meaningful framework for liver disease assessment.

While this study primarily focuses on whole-liver segmentation, future efforts will emphasize lesion differentiation to further improve its clinical relevance. Expanding the model to include liver lesion segmentation by leveraging publicly available annotated datasets will be a key objective. Additionally, attention-based mechanisms will be explored to enhance the detection of small or indistinct lesions. Finally, integrating lesion

segmentation with classification models will contribute to the development of a more comprehensive AI-driven diagnostic workflow, improving the utility of automated liver analysis in clinical practice.

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

Funding: This work has been partially supported by EU DUCA, EU CyberSecPro, SYNAPSE, PTR 22-24 P2.01 (Cybersecurity) and SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the EU-NextGenerationEU projects, by MUR-REASONING: foRmal mEthods for computAtional analySis for diagnOsis and progNosis in imagING-RIN, e-DAI (Digital ecosystem for integrated analysis of heterogeneous health data related to high-impact diseases: innovative model of care and research), Health Operational Plan, FSC 2014–2020, PRIN-MUR-Ministry of Health, the National Plan for NRRP Complementary Investments $D^3$4 Health: Digital Driven Diagnostics, prognostics and therapeutics for sustainable Health care, Progetto MolisCTe, Ministero delle Imprese e del Made in Italy, Italy, CUP: D33B22000060001, FORESEEN: FORmal mEthodS for attack dEtEction in autonomous driviNg systems CUP N.P2022WYAEW and ALOHA: a framework for monitoring the physical and psychological health status of the Worker through Object detection and federated machine learning, Call for Collaborative Research BRiC-2024, INAIL.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in Kaggle website and freely for research purposes (https://www.kaggle.com/datasets/, accessed on 27 February 2025).

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

References

1. Rockey, D.C.; Caldwell, S.H.; Goodman, Z.D.; Nelson, R.C.; Smith, A.D. Liver biopsy. *Hepatology* **2009**, *49*, 1017–1044.
2. Bravo, A.A.; Sheth, S.G.; Chopra, S. Liver biopsy. *N. Engl. J. Med.* **2001**, *344*, 495–500.
3. Gotra, A.; Sivakumaran, L.; Chartrand, G.; Vu, K.N.; Vandenbroucke-Menu, F.; Kauffmann, C.; Kadoury, S.; Gallix, B.; de Guise, J.A.; Tang, A. Liver segmentation: Indications, techniques and future directions. *Insights Imaging* **2017**, *8*, 377–392. [PubMed]
4. Carson, M.D.; Nejak-Bowen, K. Wnt/β-Catenin Signaling in Liver Pathobiology. *Annu. Rev. Pathol. Mech. Dis.* **2024**, *20*, 59–86.
5. Heimann, T.; Van Ginneken, B.; Styner, M.A.; Arzhaeva, Y.; Aurich, V.; Bauer, C.; Beck, A.; Becker, C.; Beichel, R.; Bekes, G.; et al. Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Trans. Med. Imaging* **2009**, *28*, 1251–1265. [PubMed]
6. Mak, L.Y.; Liu, K.; Chirapongsathorn, S.; Yew, K.C.; Tamaki, N.; Rajaram, R.B.; Panlilio, M.T.; Lui, R.; Lee, H.W.; Lai, J.C.T.; et al. Liver diseases and hepatocellular carcinoma in the Asia-Pacific region: Burden, trends, challenges and future directions. *Nat. Rev. Gastroenterol. Hepatol.* **2024**, *21*, 834–851.
7. Campadelli, P.; Casiraghi, E.; Esposito, A. Liver segmentation from computed tomography scans: A survey and a new algorithm. *Artif. Intell. Med.* **2009**, *45*, 185–196.
8. Wang, L.; Fatemi, M.; Alizad, A. Artificial intelligence techniques in liver cancer. *Front. Oncol.* **2024**, *14*, 1415859.
9. Gul, S.; Khan, M.S.; Bibi, A.; Khandakar, A.; Ayari, M.A.; Chowdhury, M.E. Deep learning techniques for liver and liver tumor segmentation: A review. *Comput. Biol. Med.* **2022**, *147*, 105620.
10. Adam, R.; Piedvache, C.; Chiche, L.; Adam, J.P.; Salamé, E.; Bucur, P.; Cherqui, D.; Scatton, O.; Granger, V.; Ducreux, M.; et al. Liver transplantation plus chemotherapy versus chemotherapy alone in patients with permanently unresectable colorectal liver metastases (TransMet): Results from a multicentre, open-label, prospective, randomised controlled trial. *Lancet* **2024**, *404*, 1107–1118.

11. Germain, T.; Favelier, S.; Cercueil, J.P.; Denys, A.; Krausé, D.; Guiu, B. Liver segmentation: Practical tips. *Diagn. Interv. Imaging* **2014**, *95*, 1003–1016.
12. Zhao, X.; Wang, L.; Zhang, Y.; Han, X.; Deveci, M.; Parmar, M. A review of convolutional neural networks in computer vision. *Artif. Intell. Rev.* **2024**, *57*, 99.
13. Bilic, P.; Christ, P.; Li, H.B.; Vorontsov, E.; Ben-Cohen, A.; Kaissis, G.; Szeskin, A.; Jacobs, C.; Mamani, G.E.H.; Chartrand, G.; et al. The liver tumor segmentation benchmark (lits). *Med Image Anal.* **2023**, *84*, 102680. [PubMed]
14. Mharib, A.M.; Ramli, A.R.; Mashohor, S.; Mahmood, R.B. Survey on liver CT image segmentation methods. *Artif. Intell. Rev.* **2012**, *37*, 83–95.
15. Khang, A.; Abdullayev, V.; Litvinova, E.; Chumachenko, S.; Alyar, A.V.; Anh, P. Application of Computer Vision (CV) in the Healthcare Ecosystem. In *Computer Vision and AI-Integrated IoT Technologies in the Medical Ecosystem*; CRC Press: Boca Raton, FL, USA, 2024; pp. 1–16.
16. Marasinghe, R.; Yigitcanlar, T.; Mayere, S.; Washington, T.; Limb, M. Computer vision applications for urban planning: A systematic review of opportunities and constraints. *Sustain. Cities Soc.* **2024**, *100*, 105047.
17. Huang, P.; Zhou, X.; He, P.; Feng, P.; Tian, S.; Sun, Y.; Mercaldo, F.; Santone, A.; Qin, J.; Xiao, H. Interpretable laryngeal tumor grading of histopathological images via depth domain adaptive network with integration gradient CAM and priori experience-guided attention. *Comput. Biol. Med.* **2023**, *154*, 106447.
18. Qu, Y.; Zhou, X.; Huang, P.; Liu, Y.; Mercaldo, F.; Santone, A.; Feng, P. CGAM: An end-to-end causality graph attention Mamba network for esophageal pathology grading. *Biomed. Signal Process. Control* **2025**, *103*, 107452.
19. Santone, A.; Cesarelli, M.; Colasuonno, E.; Bevilacqua, V.; Mercaldo, F. A Method for Ocular Disease Diagnosis through Visual Prediction Explainability. *Electronics* **2024**, *13*, 2706. [CrossRef]
20. He, H.; Yang, H.; Mercaldo, F.; Santone, A.; Huang, P. Isolation Forest-Voting Fusion-Multioutput: A stroke risk classification method based on the multidimensional output of abnormal sample detection. *Comput. Methods Programs Biomed.* **2024**, *253*, 108255.
21. Mercaldo, F.; Zhou, X.; Huang, P.; Martinelli, F.; Santone, A. Machine learning for uterine cervix screening. In Proceedings of the 2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE), Taichung, Taiwan, 7–9 November 2022; pp. 71–74.
22. Ramana, B.V.; Babu, M.S.P.; Venkateswarlu, N. A critical study of selected classification algorithms for liver disease diagnosis. *Int. J. Database Manag. Syst.* **2011**, *3*, 101–114.
23. Chen, E.L.; Chung, P.C.; Chen, C.L.; Tsai, H.M.; Chang, C.I. An automatic diagnostic system for CT liver image classification. *IEEE Trans. Biomed. Eng.* **1998**, *45*, 783–794. [PubMed]
24. Wu, K.; Chen, X.; Ding, M. Deep learning based classification of focal liver lesions with contrast-enhanced ultrasound. *Optik* **2014**, *125*, 4057–4063.
25. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
26. Zhou, X.; Tang, C.; Huang, P.; Mercaldo, F.; Santone, A.; Shao, Y. LPCANet: Classification of laryngeal cancer histopathological images using a CNN with position attention and channel attention mechanisms. *Interdiscip. Sci. Comput. Life Sci.* **2021**, *13*, 666–682.
27. Zhou, X.; Tang, C.; Huang, P.; Tian, S.; Mercaldo, F.; Santone, A. ASI-DBNet: An adaptive sparse interactive resnet-vision transformer dual-branch network for the grading of brain cancer histopathological images. *Interdiscip. Sci. Comput. Life Sci.* **2023**, *15*, 15–31.
28. Huang, P.; Li, C.; He, P.; Xiao, H.; Ping, Y.; Feng, P.; Tian, S.; Chen, H.; Mercaldo, F.; Santone, A.; et al. MamlFormer: Priori-experience guiding transformer network via manifold adversarial multi-modal learning for laryngeal histopathological grading. *Inf. Fusion* **2024**, *108*, 102333.
29. Gulia, A.; Vohra, R.; Rani, P. Liver patient classification using intelligent techniques. *Int. J. Comput. Sci. Inf. Technol.* **2014**, *5*, 5110–5115.
30. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; proceedings, part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
31. Christ, P.F.; Elshaera, M.E.; Ettlinger, F.; Tatavarty, S.; Bickel, M.; Bilic, P.; Rempfler, M.; Armbruster, M.; Hofmann, F.; D'Anastasi, M.; et al. Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields. In Proceedings of the MICCAI, Athens, Greece, 17–21 October 2016.
32. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
33. Milletari, F.; Navab, N.; Ahmadi, S.A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.

34. Kavur, A.E.; Gezer, N.S.; Barış, M.; Aslan, S.; Conze, P.-H.; Groza, V.; Pham, D.D.; Chatterjee, S.; Ernst, P.; Özkan, S.; et al. CHAOS Challenge–Combined (CT-MR) Healthy Abdominal Organ Segmentation. *Med Image Anal.* **2021**, *69*, 101950.
35. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med Imaging* **2019**, *39*, 1856–1867.
36. Isensee, F.; Jaeger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-adapting framework for U-Net-based medical image segmentation. *Nat. Methods* **2021**, *18*, 203–211.
37. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
38. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.

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

Multiphase Computed Tomography Scan Findings for Artificial Intelligence Training in the Differentiation of Hepatocellular Carcinoma and Intrahepatic Cholangiocarcinoma Based on Interobserver Agreement of Expert Abdominal Radiologists

Nakarin Inmutto, Suwalee Pojchamarnwiputh and Wittanee Na Chiangmai *

Department of Radiology, Faculty of Medicine, Chiang Mai University, Chiang Mai 50200, Thailand; nakarin.i@cmu.ac.th (N.I.); suwalee.poj@cmu.ac.th (S.P.)
* Correspondence: wittanee.n@cmu.ac.th

Abstract: Background/Objective: Hepatocellular carcinoma (HCC) and intrahepatic cholangiocarcinoma (ICC) are the most common primary liver cancer. Computed tomography (CT) is the imaging modality used to evaluate liver nodules and differentiate HCC from ICC. Artificial intelligence (AI), machine learning (ML), and deep learning (DL) have been used in multiple studies in the field of radiology. The purpose of this study was to determine potential CT features for the differentiation of hepatocellular carcinoma and intrahepatic cholangiocarcinoma. **Methods**: Patients with radiological and pathologically confirmed diagnosis of HCC and ICC between January 2013 and December 2015 were included in this retrospective study. Two board-certified diagnostic radiologists independently reviewed multiphase CT images on a picture archiving and communication system (PACS). Arterial hyperenhancement, portal vein thrombosis, lymph node enlargement, and cirrhosis appearance were evaluated. We then calculated sensitivity, specificity, the likelihood ratio for diagnosis of HCC and ICC. Inter-observed agreement of categorical data was evaluated using Cohen's kappa statistic (k). **Results**: A total of 74 patients with a pathologically confirmed diagnosis, including 48 HCCs and 26 ICC, were included in this study. Most of HCC patients showed arterial hyperenhancement at 95.8%, and interobserver agreement was moderate (k = 0.47). Arterial enhancement in ICC was less frequent, ranging from 15.4% to 26.9%, and agreement between readers was substantial (k = 0.66). The two readers showed a moderate agreement of cirrhosis appearance in both the HCC and ICC groups, k = 0.43 and k = 0.48, respectively. Cirrhosis appeared in the HCC group more frequently than the ICC group. Lymph node enlargement was more commonly seen in ICC than HCC, and agreement between the readers was almost perfect (k = 0.84). Portal vein invasion in HCC was seen in 14.6% by both readers with a substantial agreement (k = 0.66). Portal vein invasion in ICC was seen in 11.5% to 19.2% of the patients. The diagnostic performance of the two radiologists was satisfactory, with a corrected diagnosis of 87.8% and 94.6%. The two radiologists had high sensitivity in diagnosing HCCs (95.8% to 97.9%) and specificity in diagnosing ICCs (95.8% to 97.9%). **Conclusions:** Cirrhosis and lymph node metastasis could be ancillary and adopted in future AI training algorithms.

Keywords: hepatocellular carcinoma; intrahepatic cholangiocarcinoma; artificial intelligence; computed tomography; interobserver agreement

1. Introduction

Hepatocellular carcinoma (HCC) and intrahepatic cholangiocarcinoma (ICC) are the most common primary liver cancers, accounting for approximately 85% [1]. These two cancers are a cause of a considerable part of oncologic mortality, and they are also a diagnostic and therapeutic challenge for healthcare systems worldwide. Underlying chronic hepatic inflammation with viral hepatitis B or C is a risk factor for HCC, but less for ICC. The correct diagnosis of HCC and ICC is crucial because the treatment of the two entities is different, and percutaneous biopsy is not widely available in Thailand. Rapid and reliable detection and diagnosis of HCC and ICC may allow for earlier treatment onset and better outcomes for these patients.

The role of imaging in the management of HCC and ICC has significantly evolved and expanded beyond the plain radiological confirmation of the tumor based on the typical appearance in a multiphase contrast-enhanced tomography (CT) examination. Both HCC and ICC have characteristic imaging appearances on multiphase CT and magnetic resonance imaging (MRI) [2]. CT is the imaging modality used to evaluate liver nodules and differentiate HCC from ICC. As cross-sectional imaging has become more widely available and of higher quality, the reliance on invasive diagnostic biopsies has declined, positioning imaging-based diagnosis as a more integral part of medical decision-making. This is because these two liver cancers have different enhancing patterns on multiphase CT. Typical HCC shows arterial hyperenhancement, washout appearance in the portovenous phase, and the presence of an enhancing capsule. Vascular invasion, particularly of the portal vein, is an imaging finding of advanced HCC. The American College of Radiology proposed the Liver Imaging-Reporting and Data System (LI-RADS) for standardized imaging diagnosis criteria in patients at risk of HCC [3]. This system is used as a system for patients developing HCC, and this algorithm guides radiologists in differentiating between benign and HCC masses. The LI-RADS is expected to become an international standard in the imaging diagnosis of liver masses. LI-RADS 5 lesions are diagnostic of HCC without the need for histologic confirmation, as shown in Figure 1. In contrast to HCC, typical ICC shows rim-like enhancement at the tumor periphery during the arterial phase and progressive centripetal enhancement in a later phase, as shown in Figure 2. Intrahepatic bile duct dilatation is another radiological finding of ICC. The characteristics and differences between typical HCC and ICC CT findings are described in Table 1.

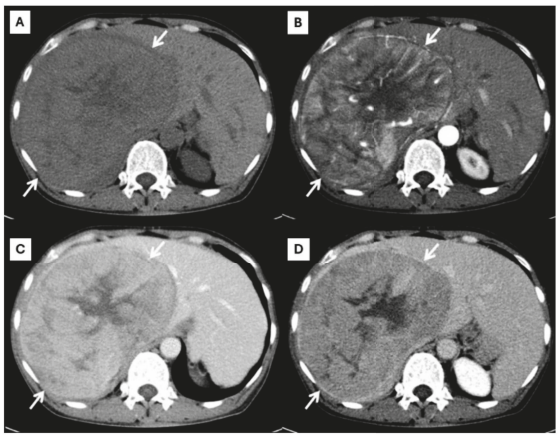


Figure 1. CT findings of typical HCC. Pre-contrast (**A**), arterial (**B**), portal venous (**C**), and delayed (**D**) post-contrast sequences demonstrate a mass with non-rim arterial hyperenhancement ((**B**), arrow), washout appearance in later phase ((**C**), arrow), and the presence of an enhancing capsule ((**D**), arrow).

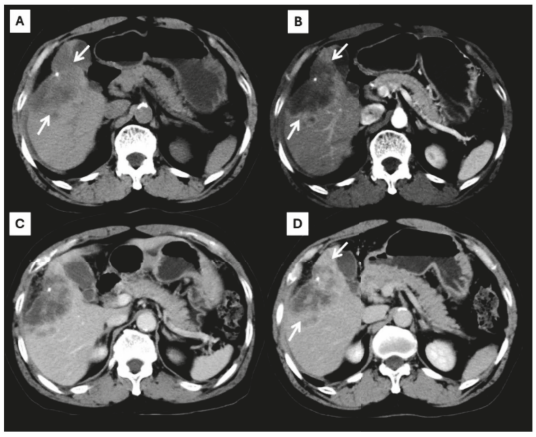


Figure 2. CT findings of typical CCA. Pre-contrast (**A**), arterial (**B**), portal venous (**C**), and delayed (**D**) post-contrast sequences demonstrate a mass with rim-like enhancement at the tumor periphery during the arterial phase ((**B**), arrow) and progressive centripetal enhancement in the later phase ((**C**,**D**), arrow).

Table 1. The characterization and difference of typical hepatocellular carcinoma and intrahepatic cholangiocarcinima detected using multiphase contrast-enhanced tomography.

	Hepatocellular Carcinoma	**Intrahepatic Cholangiocarcinima**
Later arterial phase	Arterial phase hyperenhancement	Rim-like enhancement
Portavenous phase	Non-peripheral washout	Progressive peripheral enhancement
Delayed phase	Capsule enhancement	Progressive peripheral enhancement

However, an atypical enhancement pattern of HCC and ICC has often been seen in daily practice, making diagnosis of HCC or ICC challenging [4–10] (Figure 3). A liver nodule in a high-risk patient might have an enhancing pattern similar to ICC. On the other hand, small ICCs have enhancing patterns mimicking HCC, as shown in Figure 4. One study showed that the accuracy of CT for distinguishing non-HCC malignancies from HCC was 79.9% [11]. Therefore, the diagnosis and differential diagnosis of HCC and ICC is sometimes difficult and requires significant experience from radiologists.

Artificial intelligence (AI) was developed many years ago. Machine learning (ML) is a subset of AI whereby a system can develop the ability to "learn" from data without being explicitly programmed. There are three main types of ML: supervised, unsupervised, and reinforcement learning. Supervised learning uses labeled datasets to make classifications, predictions, and regression. Unsupervised learning uses unlabeled datasets to identify not readily apparent patterns and classify cases. Reinforcement learning is an area of machine learning that handles sequential decision-making problems in a situation of uncertainty. Reinforcement learning optimizes sequential decisions by finding the best strategy, divided into supervised and unsupervised methods [12].

Supervised learning deals with annotated data with input–output pairs and common techniques including linear regression, logistic regression, decision trees, k-nearest neighbor, support vector machine (SVM), random forest (RF), naive Bayes classification, and gradient boosting [13]. ML and deep learning (DL) have been used in multiple studies in the field of medicine. Multi-layered neural network algorithms, including convolutional neural networks (CNNs), have proven of great use in the analysis of radiology images [14,15]. Machine learning algorithms have demonstrated exceptional accuracy in classifying various

diseases through radiological imaging. Two studies have investigated the usefulness of DL-based liver tumor classification from a multiphase contrast-enhanced CT scan [16,17].

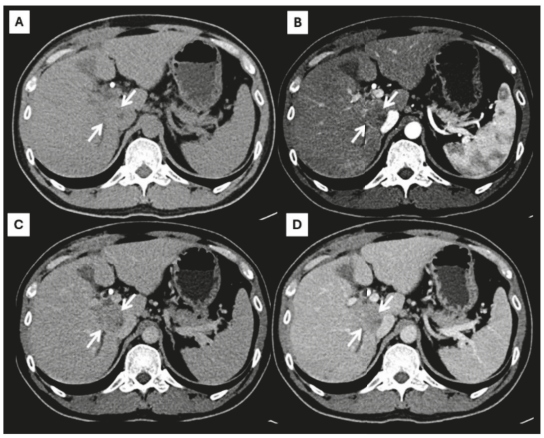


Figure 3. CT findings of atypical HCC. Pre-contrast (**A**), arterial (**B**), portal venous (**C**), and delayed (**D**) post-contrast sequences demonstrate a mass with rim-like enhancement at the tumor periphery during the arterial phase ((**B**), arrow) and progressive centripetal enhancement in the later phase ((**C**,**D**), arrow).

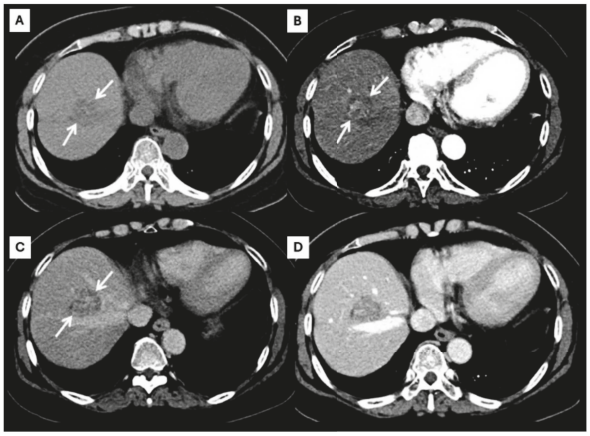


Figure 4. CT findings of atypical ICC. Pre-contrast (**A**), arterial (**B**), portal venous (**C**), and delayed (**D**) post-contrast sequences demonstrate a mass with non-rim arterial hyperenhancement ((**B**), arrow) and progressive centripetal enhancement in the later phase ((**C**,**D**), arrow).

AI models have been trained on CECT datasets to extract key imaging biomarkers that help distinguish HCC from ICC and other liver lesions. Several studies have demonstrated the effectiveness of AI in liver tumor classification. Yasaka et al. developed a CNN model that is capable of differentiating HCC, ICC, hemangiomas, and metastases based on multiphase CT, achieving a 92% area under the receiver operating characteristic curve (AUC) [18]. Similarly, Ponnoprat et al. applied deep learning techniques to multiphase CT scans, reporting an 88% classification accuracy in distinguishing HCC from ICC [18]. Wang et al. further advanced AI applications by proposing an interpretable radiomics-based system that identifies radiological features in liver tumors, demonstrating an 82.9% sensitivity in detecting HCC [19].

The noninvasive differentiation of HCC from ICC remains challenging [2,4–6,20]. Therefore, we attempted to determine the frequency and reliability of CT features that are commonly used by radiologists for the diagnosis of HCC and ICC. Then, we used these CT features to train the AI dataset for differential diagnosis HCC and ICC. Finally, we improved the accuracy of AI applications in the field of hepatology.

2. Materials and Methods

This retrospective study was conducted in accordance with the Declaration of Helsinki and was approved by the Research Ethics Committee, study code: RAD-2562-06135. Informed consent was not required due to the retrospective nature of this study.

2.1. Patient Selection

Patients with radiologically and pathologically confirmed diagnoses of HCC and ICC between January 2013 and December 2015 were included in this retrospective study. The diagnosis of ICC was confirmed based on histopathology (biopsy or resection). The diagnosis of HCC was confirmed based on radiological findings or histopathology. The patient selection flow chart is presented in Figure 5. The inclusion criteria were patients with a diagnosis of HCC and ICC in the hospital database, patients with a pre-diagnosis multiphase CT scan in a picture archiving and communication system (PACS), and pathologically confirmed diagnosis of HCC or ICC in the Pathology Department database. The exclusion criteria were poor images quality or improper phases of CT scan, the pathological diagnosis of metastasis or other malignancies, and CT or pathological report of perihilar or extrahepatic cholangiocarcinoma.

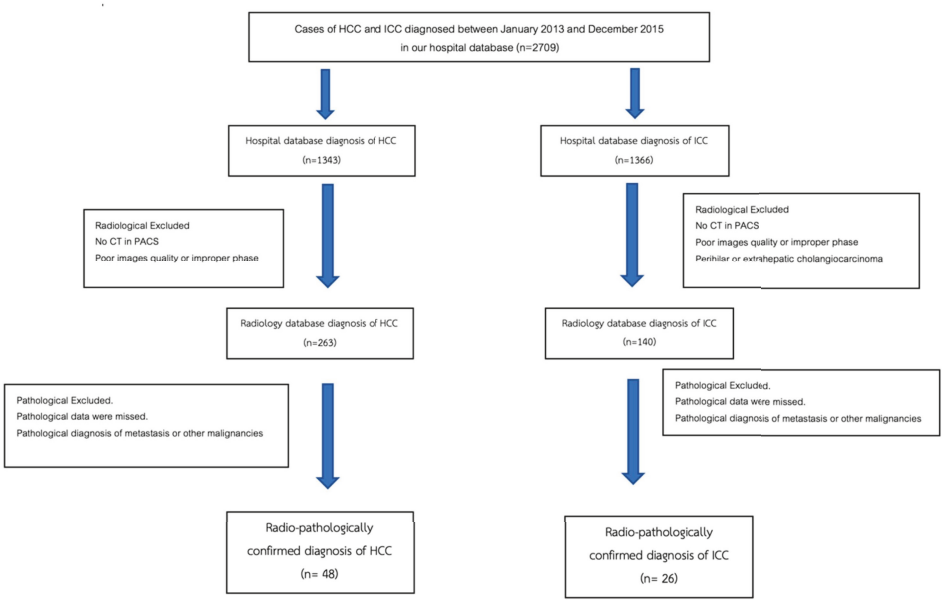


Figure 5. Flowchart of hepatocellular carcinoma and intrahepatic cholangiocarcinoma patient selection process.

2.2. Imaging Acquisition

CT images of the liver were obtained using one of two multidetector CT scanners (SOMATOM FORCE and SOMATOM DEFINITION, Siemens Medical Solutions, Forchheim, Germany). The technique included pre-contrast imaging and multiphase post-contrast imaging following injection of 100 mL of non-ionic iodinated contrast media (350 mg/mL) through the antecubital vein. The hepatic arterial phase (HAP), portovenous phase (PVP), and delayed phase (DP) of scanning began 35, 80, and 180 s after the injection of the contrast media, respectively. The images were reconstructed in 2 mm slices in axial, coronal, and sagittal views.

2.3. Imaging Interpretation

CT images were independently reviewed using Fuji PACS and using Synapse workstation version 5, https://www.fujifilm.com/th/en/healthcare/healthcare-it/it-imaging/

pacs, accessed on 6 March 2025, by two expert board-certified Diagnostic Radiologists, Suwalee Pojchamarnwiputh and Wittanee Na Chiangmai, who had more than 20 years of experience in abdominal radiology. The readers were blinded to both radiological and pathological reports. Each reader independently viewed and evaluated the images, then measured the tumor size on the most visible imaging phase. Each tumor was evaluated in terms of arterial hyperenhancement, portal vein invasion, and final diagnosis. The presence of cirrhosis and lymph node metastasis was evaluated.

The following formal radiological criteria were applied:

Arterial hyperenhancement: Defined as non-rim, unequivocal enhancement of the lesion compared to the surrounding liver parenchyma during the arterial phase, according to the Liver Imaging Reporting and Data System (LI-RADS) guidelines.

Portal vein invasion: Defined as the presence of a soft tissue thrombus or filling defect within the portal vein, with or without vein expansion, showing enhancement characteristics similar to the tumor on post-contrast imaging.

Cirrhosis: Identified based on morphological features such as surface nodularity, lobar or segmental volume redistribution (e.g., caudate lobe hypertrophy or right lobe atrophy) and secondary signs of portal hypertension (e.g., splenomegaly, portosystemic collaterals).

Lymph node metastasis: Defined as lymph nodes with a short-axis diameter of >1 cm or nodes exhibiting central necrosis, irregular margins, or abnormal enhancement, located in periportal, peripancreatic, or para-aortic regions.

2.4. Histopathologic Examination

The histopathologic report served as the reference standard of this study. Biopsy or surgical specimens of the hepatic nodule were fixed with 10% formalin and embedded in paraffin. The tissue slices were stained with hematoxylin–eosin and evaluated by pathologists. No reexamination of pathological slices was done.

2.5. Statistics

Statistical analyses were performed using the software SPSS (IBM SPSS Statistics for Windows, Version 19. Armonk, NY, USA: IBM Corp).

Quantitative data were presented as means and standard deviations or medians and interquartile ranges depending on the data distribution. Categorical variables were analyzed using the chi-square test. The arterial hyperenhancement, portal vein thrombosis, lymph node enlargement, and cirrhosis appearance were calculated and presented with the sensitivity, specificity, and likelihood ratio for diagnosis of HCC and ICC.

Inter-observed agreement of categorical data was evaluated using Cohen's kappa statistic with corresponding 95% confidence intervals. According to Landis and Koch, values of less than 0.00 indicate no agreement, 0.01–0.20 indicate slight agreement, 0.21–0.40 indicate fair agreement, 0.41–0.60 indicate moderate agreement, 0.61–0.80 indicate substantial agreement, and 0.81–1.00 indicate almost perfect agreement [21]. Significance was defined as $p < 0.05$.

3. Results

3.1. Participants

A total of 74 patients with a pathologically confirmed diagnosis, including 48 HCCs and 26 ICC, were included in this study. The mean age of patients with HCC was 52 years (range 21–76), while the mean age of those ICC was 57 years (range 35–76).

3.2. CT Features of HCC and ICC and Interobserver Agreement

Frequent CT features and interobserver agreement between the two readers are demonstrated in Tables 2 and 3. Arterial enhancement in HCC was observed in 95.8% of cases by both readers. The agreement between the readers was moderate (k = 0.47). Arterial enhancement in ICC was identified less often, in 15.4% of cases by Reader 1 and 26.9% by Reader 2. The agreement between the readers was substantial (k = 0.66).

Table 2. Frequency of computed tomography (CT) features of histopathologically proven hepatocellular carcinoma (HCC) and intrahepatic cholangiocarcinoma (ICC) detected by expert radiologists.

CT Features	HCC (*n* = 48) Frequency	ICC (*n* = 26) Frequency
Arterial enhancement		
Reader 1	95.8%	15.4%
Reader 2	95.8%	26.9%
Cirrhosis appearance		
Reader 1	70.8%	42.3%
Reader 2	72.9%	61.5%
Lymph node metastasis		
Reader 1	14.6%	50.0%
Reader 2	25.0%	57.7%
Portal vein invasion		
Reader 1	14.6%	11.5%
Reader 2	14.6%	19.2%

Table 3. Interobserver agreement among expert radiologists regarding computed tomography (CT) features of histopathologically proven hepatocellular carcinoma (HCC) and intrahepatic cholangiocarcinoma (ICC).

	HCC		ICC	
CT Features	Kappa	Agreement	Kappa	Agreement
Arterial enhancement	0.47	Moderate	0.66	Substantial
Cirrhosis appearance	0.43	Moderate	0.48	Moderate
Lymph node metastasis	0.29	Fair	0.84	Almost perfect
Portal vein invasion	0.66	Substantial	0.70	Substantial
Final diagnosis	0.65	Substantial	0.52	Moderate

The two readers showed moderate agreement of cirrhosis appearance in both HCC and ICC. Cirrhosis appeared in the HCC group more frequently than in the ICC group (70.8–72.9% vs. 42.3–61.5%)

Lymph node enlargement in HCC was not commonly seen, in 14.6% of the cases by Reader 1 and 25.0% by Reader 2. The agreement between the readers was fair (k = 0.29). Lymph node enlargement in ICC was more commonly seen, in 50.0% of the cases by reader 1 and 57.7% by Reader 2. The agreement between the readers was almost perfect (k = 0.84).

Portal vein invasion in HCC was seen in 14.6% of the cases by both readers with substantial agreement (k = 0.66). Portal vein invasion in ICC was seen in 11.5% of the cases by Reader 1 and 19.2% by Reader 2. The agreement of both readers was substantial agreement (k = 0.70).

3.3. Correct Diagnosis

The diagnostic performance of the two radiologists was satisfactory, with a corrected diagnosis of 87.8% and 94.6% and a *p* value of 0.108. Both radiologists shared a common correct diagnosis of 65 out 74 patients, including 46 HCCs and 19 ICCs. Two HCCs

and 7 ICCs, which were 4.15% of HCCs and 26.9% of ICCs, were misdiagnosis by both radiologists. The diagnostic performance of the two expert radiologists in the diagnosis of HCC and ICC is summarized in Table 4.

Table 4. Diagnostic performance of two expert radiologists in the diagnosis of hepatocellular carcinoma (HCC) and intrahepatic cholangiocarcinoma (ICC).

	HCC				ICC			
Diagnosis	**Sensitivity**	**Specificity**	**Positive Likelihood Ratio**	**Negative Likelihood Ratio**	**Sensitivity**	**Specificity**	**Positive Likelihood Ratio**	**Negative Likelihood Ratio**
Reader 1	0.979	0.885	8.513043	0.024	0.885	0.979	42.142	0.117
Reader 2	0.958	0.731	3.561338	0.057	0.731	0.958	17.404	0.280

4. Discussion

According to LI-RADS, the percentage of HCC detected based on LR-5 criteria is 95%, while the percentage of malignancy is 98% [3]. Most HCC cases showed arterial hyperenhancement (95.8%), and interobserver agreement was moderate. The presence of arterial enhancement in our study was higher than in prior studies, likely due to the HCCs in our study being larger than in other studies. The presence of arterial enhancement was related to arterial blood flow, which increases in the early stage of hepatocarcinogenesis in well-differentiated and moderately differentiated HCC. Poorly differentiated HCCs have decreased arterial blood flow, resulting in a typical enhancement pattern of some HCCs.

On the other hand, some ICCs show arterial enhancement. Using the presence of arterial hyperenhancement alone may lead to misdiagnosis. Interobserver agreement regarding presence or absence of arterial enhancement was substantial.

Two expert radiologists had high sensitivity in diagnosed HCCs, with a high positive likelihood ratio and substantial agreement. In addition, ICCs were diagnosed with lower sensitivity but higher specificity and a positive likelihood ratio.

Cirrhosis can be associated with an over 30-fold increase in HCC risk and a 10-fold to 20-fold increase in ICC risk. Cirrhosis appearance has been used as a radiological finding for the prediction of HCC. Our study demonstrated cirrhotic appearance in patients with HCCs more than in patient with ICCs; around 70% versus 40 to 60%. Interobserver agreement of cirrhotic appearance was moderate. We favored diagnosing HCC in patients with cirrhotic appearance.

Lymph node enlargement or lymph node metastasis is more commonly seen in ICC than in HCC. Our study revealed lymph node metastasis in about 50% of ICC patients, with almost perfect agreement between the two expert radiologists. We almost diagnosed ICCs when CT showed lymph node enlargement.

Machine learning methods have been widely adopted in classifying liver cancer [13, 15,17–19,22–26]. Supervised learning with a set of data labeled for algorithms such as support vector machines (SVMs) has been used in differentiated liver cancer. Ponnoprat et al. achieved 88% classification accuracy of differentiated HCCs and ICCs based on multiphase CT scans [16]. Deep learning is currently being investigated for liver cancer classification. There are still several limitations due to selection bias of training datasets, small sample sizes, and lack of internal and external validation. A convolutional neural network (CNN) is a kind of deep learning method that does not require the definition of specific radiological features to learn. It interprets a large of training images and may discover additional differential features not yet identified in current radiological practice. Yasaka used a CNN algorithm for differentiation of liver masses into five classes and found

that the median area under the receiver operating characteristic curve for differentiating malignant and benign lesions was 0.92 [18].

Wang et al. proposed a proof-of-concept interpretable deep learning system for clinical radiology by identifying and scoring radiological features [27–29]. They used 14 radiological features such as arterial phase hyperenhancement, central scar, enhancing rim, progressive centripetal filling, progressive hyperenhancement, thin-walled mass, and washout pattern to classify liver lesions into six classes: benign cyst, cavernous hemangioma, focal nodular hyperplasia, HCC, ICC, and metastasis. Their model achieved 76.5% positive predictive value and 82.9% sensitivity in identifying the correct radiological features [28].

Our study has two limitations. First, this was a single-center design that employed only two raters, which may affect generalizability. Future multicenter studies with a larger pool of radiologists are recommended to enhance the external validity of retrospective case–control studies, which may be subject to a selection bias. Second, in our study population, the ratio of HCC to ICC cases differed from the actual clinical distribution.

Our study highlights the importance of multiphase CT imaging in differentiating HCC and ICC based on key radiological features and interobserver agreement among expert radiologists. The results indicate that arterial hyperenhancement, portal vein invasion, lymph node metastasis, and cirrhosis appearance are crucial imaging biomarkers that can aid in distinguishing these two malignancies. These findings provide a foundation for the development of AI models aimed at improving automated classification and reducing diagnostic variability.

Although this study does not involve direct AI model development, the results are intended to contribute to future AI training datasets by identifying reproducible imaging features with validated interobserver agreement, which could assist AI developers in refining liver cancer classification algorithms.

Given the promising role of AI in liver tumor differentiation, future studies should focus on integrating deep learning-based models trained on large, multi-center datasets to enhance generalizability. Additionally, further external validation of AI models is essential to ensure robust performance across diverse patient populations. Prospective studies should also be conducted to evaluate the real-world impact of AI-assisted diagnosis in clinical practice, including its potential role in LI-RADS classification, prognosis prediction, and treatment response assessment. Ultimately, incorporating AI into liver cancer diagnostics could enhance radiologist workflow efficiency, improve early detection, and contribute to personalized treatment strategies.

5. Conclusions

HCC and ICC share overlapping CT hallmark features, making differentiation challenging. This study highlights the role of ancillary imaging features, such as cirrhosis and lymph node metastasis, in improving diagnostic accuracy. By incorporating these features into AI training algorithms, AI-based models have the potential to enhance precision in distinguishing HCC from ICC.

As AI continues to evolve in radiology, future research should prioritize the development of large, multi-center datasets to ensure model generalizability and robustness.

These advancements could lead to more consistent diagnoses, improved early detection, and personalized treatment strategies for patients with liver malignancies.

Author Contributions: Conceptualization, N.I., W.N.C. and S.P.; methodology, N.I.; validation, N.I., W.N.C. and S.P.; formal analysis, N.I., W.N.C. and S.P.; writing—original draft preparation, N.I., W.N.C. and S.P.; writing—review and editing, N.I., W.N.C. and S.P.; funding acquisition, N.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Faculty of Medicine, Chiang Mai University, Thailand. (Grant No. 150/2562).

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Chiang Mai University (protocol code RAD-2562-06135, approval date 13 March 2019).

Informed Consent Statement: Patient consent was waived due the retrospective nature of this study.

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. Massarweh, N.N.; El-Serag, H.B. Epidemiology of Hepatocellular Carcinoma and Intrahepatic Cholangiocarcinoma. *Cancer Control* **2017**, *24*, 1073274817729245. [CrossRef] [PubMed]
2. Lo, E.C.; Rucker, A.N.; Federle, M.P. Hepatocellular Carcinoma and Intrahepatic Cholangiocarcinoma: Imaging for Diagnosis, Tumor Response to Treatment and Liver Response to Radiation. *Semin. Radiat. Oncol.* **2018**, *28*, 267–276. [CrossRef] [PubMed]
3. Marrero, J.A.; Kulik, L.M.; Sirlin, C.B.; Zhu, A.X.; Finn, R.S.; Abecassis, M.M.; Roberts, L.R.; Heimbach, J.K. Diagnosis, Staging, and Management of Hepatocellular Carcinoma: 2018 Practice Guidance by the American Association for the Study of Liver Diseases. *Hepatology* **2018**, *68*, 723–750. [CrossRef]
4. Chen, L.D.; Ruan, S.M.; Liang, J.Y.; Yang, Z.; Shen, S.L.; Huang, Y.; Li, W.; Wang, Z.; Xie, X.Y.; Lu, M.D.; et al. Differentiation of intrahepatic cholangiocarcinoma from hepatocellular carcinoma in high-risk patients: A predictive model using contrast-enhanced ultrasound. *World J. Gastroenterol.* **2018**, *24*, 3786–3798. [CrossRef]
5. Kim, R.; Lee, J.M.; Shin, C.I.; Lee, E.S.; Yoon, J.H.; Joo, I.; Kim, S.H.; Hwang, I.; Han, J.K.; Choi, B.I. Differentiation of intrahepatic mass-forming cholangiocarcinoma from hepatocellular carcinoma on gadoxetic acid-enhanced liver MR imaging. *Eur. Radiol.* **2016**, *26*, 1808–1817. [CrossRef]
6. Ichikawa, S.; Isoda, H.; Shimizu, T.; Tamada, D.; Taura, K.; Togashi, K.; Onishi, H.; Motosugi, U. Distinguishing intrahepatic mass-forming biliary carcinomas from hepatocellular carcinoma by computed tomography and magnetic resonance imaging using the Bayesian method: A bi-center study. *Eur. Radiol.* **2020**, *30*, 5992–6002. [CrossRef]
7. Kim, Y.Y.; Park, M.S.; Aljoqiman, K.S.; Choi, J.Y.; Kim, M.J. Gadoxetic acid-enhanced magnetic resonance imaging: Hepatocellular carcinoma and mimickers. *Clin. Mol. Hepatol.* **2019**, *25*, 223–233. [CrossRef]
8. Vigano, L.; Lleo, A.; Muglia, R.; Gennaro, N.; Sama, L.; Colapietro, F.; Roncalli, M.; Aghemo, A.; Chiti, A.; Di Tommaso, L.; et al. Intrahepatic cholangiocellular carcinoma with radiological enhancement patterns mimicking hepatocellular carcinoma. *Updates Surg.* **2020**, *72*, 413–421. [CrossRef]
9. Sano, S.; Yamamoto, Y.; Sugiura, T.; Okamura, Y.; Ito, T.; Ashida, R.; Ohgi, K.; Aramaki, T.; Nakanuma, Y.; Uesaka, K. The Radiological Differentiation of Hypervascular Intrahepatic Cholangiocarcinoma from Hepatocellular Carcinoma with a Focus on the CT Value on Multi-phase Enhanced CT. *Anticancer Res.* **2018**, *38*, 5505–5512. [CrossRef]
10. Mendiratta-Lala, M.; Park, H.; Kolicaj, N.; Mendiratta, V.; Bassi, D. Small intrahepatic peripheral cholangiocarcinomas as mimics of hepatocellular carcinoma in multiphasic CT. *Abdom. Radiol.* **2017**, *42*, 171–178. [CrossRef]
11. Ludwig, D.R.; Fraum, T.J.; Cannella, R.; Ballard, D.H.; Tsai, R.; Naeem, M.; LeBlanc, M.; Salter, A.; Tsung, A.; Shetty, A.S.; et al. Hepatocellular carcinoma (HCC) versus non-HCC: Accuracy and reliability of Liver Imaging Reporting and Data System v2018. *Abdom. Radiol.* **2019**, *44*, 2116–2132. [CrossRef]
12. Carreras, J.; Hamoudi, R. Anomaly Detection and Artificial Intelligence Identified the Pathogenic Role of Apoptosis and RELB Proto-Oncogene, NF-kB Subunit in Diffuse Large B-Cell Lymphoma. *BioMedInformatics* **2024**, *4*, 1480–1505. [CrossRef]
13. Su, T.H.; Wu, C.H.; Kao, J.H. Artificial intelligence in precision medicine in hepatology. *J. Gastroenterol. Hepatol.* **2021**, *36*, 569–580. [CrossRef]
14. Yang, Y.J.; Bang, C.S. Application of artificial intelligence in gastroenterology. *World J. Gastroenterol.* **2019**, *25*, 1666–1683. [CrossRef]
15. Jimenez Perez, M.; Grande, R.G. Application of artificial intelligence in the diagnosis and treatment of hepatocellular carcinoma: A review. *World J. Gastroenterol.* **2020**, *26*, 5617–5628. [CrossRef]
16. Ponnoprat, D.; Inkeaw, P.; Chaijaruwanich, J.; Traisathit, P.; Sripan, P.; Inmutto, N.; Na Chiangmai, W.; Pongnikorn, D.; Chitapanarux, I. Classification of hepatocellular carcinoma and intrahepatic cholangiocarcinoma based on multi-phase CT scans. *Med. Biol. Eng. Comput.* **2020**, *58*, 2497–2515. [CrossRef]
17. Shi, W.; Kuang, S.; Cao, S.; Hu, B.; Xie, S.; Chen, S.; Chen, Y.; Gao, D.; Chen, Y.; Zhu, Y.; et al. Deep learning assisted differentiation of hepatocellular carcinoma from focal liver lesions: Choice of four-phase and three-phase CT imaging protocol. *Abdom. Radiol.* **2020**, *45*, 2688–2697. [CrossRef]

18. Yasaka, K.; Akai, H.; Abe, O.; Kiryu, S. Deep Learning with Convolutional Neural Network for Differentiation of Liver Masses at Dynamic Contrast-enhanced CT: A Preliminary Study. *Radiology* **2018**, *286*, 887–896. [CrossRef]
19. Wang, M.; Fu, F.; Zheng, B.; Bai, Y.; Wu, Q.; Wu, J.; Sun, L.; Liu, Q.; Liu, M.; Yang, Y.; et al. Development of an AI system for accurately diagnose hepatocellular carcinoma from computed tomography imaging data. *Br. J. Cancer* **2021**, *125*, 1111–1121. [CrossRef]
20. You, M.W.; Yun, S.J. Differentiating between hepatocellular carcinoma and intrahepatic cholangiocarcinoma using contrast-enhanced MRI features: A systematic review and meta-analysis. *Clin. Radiol.* **2019**, *74*, 406.e9–406.e18. [CrossRef]
21. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [PubMed]
22. Maruyama, H.; Yamaguchi, T.; Nagamatsu, H.; Shiina, S. AI-Based Radiological Imaging for HCC: Current Status and Future of Ultrasound. *Diagnostics* **2021**, *11*, 292. [CrossRef]
23. Kawka, M.; Dawidziuk, A.; Jiao, L.R.; Gall, T.M.H. Artificial intelligence in the detection, characterisation and prediction of hepatocellular carcinoma: A narrative review. *Transl. Gastroenterol. Hepatol.* **2022**, *7*, 41. [CrossRef] [PubMed]
24. Sato, M.; Tateishi, R.; Yatomi, Y.; Koike, K. Artificial intelligence in the diagnosis and management of hepatocellular carcinoma. *J. Gastroenterol. Hepatol.* **2021**, *36*, 551–560. [CrossRef] [PubMed]
25. Yang, C.M.; Shu, J. Cholangiocarcinoma Evaluation via Imaging and Artificial Intelligence. *Oncology* **2021**, *99*, 72–83. [CrossRef]
26. Wu, Y.; White, G.M.; Cornelius, T.; Gowdar, I.; Ansari, M.H.; Supanich, M.P.; Deng, J. Deep learning LI-RADS grading system based on contrast enhanced multiphase MRI for differentiation between LR-3 and LR-4/LR-5 liver tumors. *Ann. Transl. Med.* **2020**, *8*, 701. [CrossRef]
27. Hamm, C.A.; Wang, C.J.; Savic, L.J.; Ferrante, M.; Schobert, I.; Schlachter, T.; Lin, M.; Duncan, J.S.; Weinreb, J.C.; Chapiro, J.; et al. Deep learning for liver tumor diagnosis part I: Development of a convolutional neural network classifier for multi-phasic MRI. *Eur. Radiol.* **2019**, *29*, 3338–3347. [CrossRef]
28. Wang, C.J.; Hamm, C.A.; Savic, L.J.; Ferrante, M.; Schobert, I.; Schlachter, T.; Lin, M.; Weinreb, J.C.; Duncan, J.S.; Chapiro, J.; et al. Deep learning for liver tumor diagnosis part II: Convolutional neural network interpretation using radiologic imaging features. *Eur. Radiol.* **2019**, *29*, 3348–3357. [CrossRef]
29. Oestmann, P.M.; Wang, C.J.; Savic, L.J.; Hamm, C.A.; Stark, S.; Schobert, I.; Gebauer, B.; Schlachter, T.; Lin, M.; Weinreb, J.C.; et al. Deep learning-assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver. *Eur. Radiol.* **2021**, *31*, 4981–4990. [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

YOLOSAMIC: A Hybrid Approach to Skin Cancer Segmentation with the Segment Anything Model and YOLOv8

Sevda Gül [1], Gökçen Cetinel [2,*], Bekir Murat Aydin [2], Devrim Akgün [3] and Rabia Öztaş Kara [4]

[1] Department of Electronics and Automation, Adapazarı Vocational School, Sakarya University, 54050 Serdivan, Türkiye; gulsevda@sakarya.edu.tr
[2] Electrical and Electronics Engineering Department, Engineering Faculty, Sakarya University, 54050 Serdivan, Türkiye; murataydin@sakarya.edu.tr
[3] Faculty of Computer and Information Sciences, Software Engineering, Sakarya University, 54050 Serdivan, Türkiye; dakgun@sakarya.edu.tr
[4] Department of Dermatology, Sakarya University Training and Research Hospital, 54050 Serdivan, Türkiye; rabiaoztaskara@sakarya.edu.tr
* Correspondence: gcetinel@sakarya.edu.tr

Abstract: Background/Objective: The rising global incidence of skin cancer emphasizes the urgent need for reliable and accurate diagnostic tools to aid early intervention. This study introduces YOLOSAMIC (YOLO and SAM in Cancer Imaging), a fully automated segmentation framework that integrates YOLOv8 for lesion detection, and the Segment Anything Model (SAM)-Box for precise segmentation. The objective is to develop a reliable segmentation system that handles complex skin lesion characteristics without requiring manual intervention. **Methods**: A hybrid database comprising 3463 public and 765 private dermoscopy images was built to enhance model generalizability. YOLOv8 was employed to localize lesions through bounding box detection, while SAM-Box refined the segmentation process. The model was trained and evaluated under four scenarios to assess its robustness. Additionally, an ablation study examined the impact of grayscale conversion, image blur, and model pruning on segmentation performance. **Results**: YOLOSAMIC demonstrated high segmentation accuracy, achieving Dice and Jaccard scores of 0.9399 and 0.9112 on the public database and 0.8990 and 0.8445 on the hybrid dataset. **Conclusions**: The proposed YOLOSAMIC framework provides a robust, fully automated solution for skin lesion segmentation, eliminating the need for manual annotation. Integrating YOLOv8 and SAM-Box enhances segmentation precision, making it a valuable decision-support tool for dermatologists.

Keywords: skin lesion segmentation; object detection; YOLOv8; SAM-Box; artificial intelligence

1. Introduction

Skin cancer is the fifth most commonly reported cancer in the world as of 2020, according to the World Health Organization, with a total of 1.20 million cases of skin cancer, excluding melanoma [1]. Skin cancers are usually classified into two main types: melanoma, which originates from melanocytes, and nonmelanoma skin cancer, which arises from keratinocytes. Nonmelanoma skin cancer is diagnosed depending on its severity as basal cell carcinoma (BCC) or squamous cell carcinoma (SCC) [2]. Globally, 2 to 3 million nonmelanoma skin cancers and 132 thousand melanoma skin cancers are reported to occur each year [3]. Although 75% of reported skin cancer cases are BCC and 20% are SCC, invasive melanoma is responsible for 80% of skin cancer-related deaths [4,5].

Early diagnosis, timely intervention, and follow-up screening programs are essential to reduce mortality, morbidity, and treatment costs. However, manual diagnosis is time-consuming and requires the opinion of a dermatologist. The lack of a sufficient number of dermatologists and challenging lesion characteristics make early diagnosis difficult. In the absence of dermatologists, patients are examined by general practitioners, and their diagnostic accuracy varies between 24 and 70%. On the other hand, the agreement between dermatologists and practitioners is around 57% [6,7]. As a result, computer-aided diagnostic and decision-support systems have gained importance.

Literature Review

Recent advancements in image processing, machine learning, and artificial intelligence techniques provide promising results in early skin cancer detection. Identifying the skin lesion area is one of the fundamental steps in designing a helpful decision support system for experts. For this purpose, various segmentation methods have been proposed in recent years [8,9]. The recent and valuable studies performed to improve the accuracy and efficiency of automated skin lesion segmentation were investigated to show the contributions of the proposed research.

The study in [10] used the International Skin Imaging Collaboration (ISIC) 2019 database, which is divided into two main sets: training and testing. The training set consists of 1440 benign and 1197 malignant images, while the test set contains 360 benign and 300 malignant images. Each image in the database was standardized to size 224×224. A skin cancer detection method was introduced using the proposed Fractional Student Psychology Based Optimization-based Deep Q Network (FSPBO-based DQN) to detect skin lesions early. The accuracy, sensitivity, and specificity percentages were 92.364%, 93.20%, and 92.63%, respectively [10].

The study in [11] introduced a hyper-parameter optimized Fully Convolutional Encoder–Decoder Network (FCEDN) for dermoscopy image segmentation. The novel Exponential Neighborhood Grey Wolf Optimization (EN-GWO) algorithm was employed to optimize network hyper-parameters. EN-GWO incorporates a neighborhood-based searching strategy, combining individual and global haunting strategies of wolves to strike a balance between exploration and exploitation. The study compared EN-GWO with four variants of Grey Wolf Optimization (GWO), Genetic Algorithm (GA), and Particle Swarm Optimization (PSO) for hyperparameter optimization on the ISIC 2016 and ISIC 2017 databases. The proposed model achieved a Jaccard index of 96.41% and 86.85%, a Dice coefficient of 98.48% and 87.23%, and an accuracy of 98.32% and 95.25% for the ISIC 2016 and ISIC 2017 databases.

Mingzhe and his friends introduced SkinSAM, a finely tuned model derived from the SAM, showcasing remarkable segmentation performance. The evaluation was conducted on the HAM10000 database, encompassing 10,015 dermoscopy images. At the same time, larger models outperformed the smaller ones; the fine-tuned model demonstrated the most significant enhancement, achieving a mean pixel accuracy of 0.945, a mean dice score of 0.8879, and a mean Intersection over Union (mIoU) score of 0.7843. Vascular lesions exhibited the most precise segmentation results among the various lesion types [12].

The study managed by Akyel et al. combined two databases from ISIC and PH2 to explore a novel approach (LinkNet-B7) for denoising and segmenting skin cancer images. LinkNet-B7, an adaptation of the LinkNet architecture incorporating EfficientNetB7 as the encoder, demonstrated a 6% improvement in success rates compared to LinkNet with identical databases and parameters. The training accuracy achieved for noise removal and lesion segmentation was notably high, reaching 95.72% and 97.80%, respectively [13].

Then, the paper in [14] suggested a novel approach named Multiscale Attention U-Net (MSAU-Net), designed for skin lesion segmentation. The enhancement involved integrating an attention mechanism into the typical U-Net architecture at the network's bottleneck, effectively capturing hierarchical representations. The attention module orchestrated a non-linear aggregation of multi-level representations, allowing for selective adjustment of representative features. Additionally, the model utilized a Bidirectional Convolutional Long Short-Term Memory (BDC-LSTM) structure to emphasize standard discriminative features while suppressing less informative ones. The resulting features were integrated into each decoding path block to emphasize crucial regions. The proposed network was thoroughly evaluated on three public skin lesion databases, ISIC 2017, ISIC 2018, and PH2, to prove its efficacy in segmentation tasks. The Jaccard values achieved were 0.9576, 0.956, and 0.9617 for ISIC 2017, ISIC 2018, and PH2 databases, respectively.

Another study proposed a melanoma segmentation approach, including U-net and LinkNet deep learning networks, coupled with transfer learning and fine-tuning techniques. Experiments conducted on three publicly available databases (PH2, ISIC 2018, and DermIS) have shown promising results, with U-net demonstrating notable performance. Specifically, the average Dice coefficient achieved was 0.923 on the PH2 database, 0.893 on ISIC 2018, and 0.879 on the DermIS database. These findings indicate significant success for U-net across the evaluated databases [9].

A pioneering Multi-class Dilated D-Net (MD2N) framework for segmenting and classifying diverse skin cancer types was given in [15]. In the encoder phase of MD2N, a downsampling ratio was employed to mitigate feature information losses and effectively discern small skin lesion patches. Consequently, the model captured richer feature information from skin lesion areas of varying sizes. Evaluation results demonstrated the model's performance, achieving an accuracy of 97.462%, precision of 93.627%, recall of 99.721%, F1 score of 97.182%, and specificity of 93.289%.

Reis et al. proposed InSiNet, a convolutional neural network based on deep learning for detecting benign and malignant skin lesions. Under consistent conditions, the method's performance was assessed using images from the ISIC 2018, ISIC 2019, and ISIC 2020 databases. Comparative computation time and accuracy analysis were conducted between InSiNet and various machine learning techniques, including GoogleNet, DenseNet-201, ResNet152v2, Efficient-NetB0, RBF-support vector machine, logistic regression, and random forest. The results demonstrated that InSiNet achieves accuracy rates of 94.59%, 91.89%, and 90.54% on the ISIC 2018, ISIC 2019, and ISIC 2020 databases, respectively [16].

An innovative Convolutional Neural Network (CNN) framework utilizing atrous convolutions for automatic lesion segmentation is given in [17]. The architecture was constructed based on the concept of atrous/dilated convolutions, which are known for their effectiveness in semantic segmentation. The deep neural network was built, beginning with a comprehensive design approach. It incorporated fundamental elements such as convolutional layers, batch normalization, leaky ReLU layers, and adjusted hyperparameters to optimize performance. The network underwent testing on three benchmark databases: ISIC 2016, ISIC 2017, and ISIC 2018. The proposed network achieved an average Jaccard index of 90.4% on ISIC 2016, 81.8% on ISIC 2017, and 89.1% on ISIC 2018 databases, respectively.

The CNNs and the integration of handcrafted features have shown promising results in the automated detection of skin lesions. The SNC-Net framework, which merges handcrafted image features with deep learning models to improve skin cancer detection, is proposed in [18]. This hybrid approach has significantly improved lesion classification, achieving higher precision and recall rates in distinguishing between benign and malignant skin lesions. The ISIC 2019 database achieved state-of-the-art performance with an accuracy

of 97.81%, a precision of 98.31%, a recall of 97.89%, and an F1 score of 98.10%. The proposed model outperformed EfficientNetB0, MobileNetV2, DenseNet-121, and ResNet-101 baseline models and state-of-the-art classifiers. Furthermore, the method's robustness was validated through an ablation study [18].

Iqbal et al. developed a deep convolutional neural network for multi-class classification of skin lesions, demonstrating a remarkable balance between efficiency and accuracy [19]. In another study, the authors explored the application of transfer learning in medical imaging by detecting synovial fluid in knee MRIs, highlighting the adaptability of deep learning techniques to different medical domains [20]. These studies' efforts in integrating CNNs with handcrafted features emphasize the importance of robust model architecture for accurate and efficient lesion segmentation.

The systematic review in [21] analyzed 164 studies on deep learning applications for skin cancer detection, highlighting AlexNet, ResNet-50, VGG-16, and GoogLeNet as the leading architectures for achieving optimal classification results. Multiclassification techniques are identified as a growing trend, and the availability of robust public databases is emphasized as essential for advancing research in this field. Then, another recent systematic literature review examines federated learning and transfer learning techniques widely applied in skin cancer classification, analyzing their performance based on metrics such as true positive rate, true negative rate, area under the curve, and accuracy. The study reviewed 86 articles published in reputable forums between January 2018 and July 2023, utilizing data from seven prominent databases. A taxonomy summarizing malignant and non-malignant skin cancer classes is also presented. The review identifies key limitations and challenges in existing research, providing insights into future opportunities for automated melanoma and nonmelanoma classification, thereby guiding researchers in advancing this critical healthcare area [22].

Thus, the main contributions of the proposed study can be emphasized as follows:

- Hybrid Integration of Object Detection and Segmentation: Unlike conventional approaches that rely solely on segmentation networks, the proposed study combines YOLOv8's precise lesion localization with SAM-Box's advanced segmentation capabilities, achieving fully automated lesion extraction without manual intervention.
- Extensive Hybrid Database for Clinical Generalizability: The constructed hybrid dataset introduces real-world challenges, including lesions in diverse skin tones, images with hair occlusions, and complex backgrounds to enhance model generalization beyond public databases.
- Evaluation Across Multiple Training Strategies: The system was tested under four distinct training/testing scenarios. The results demonstrate superior generalization when incorporating hybrid databases, providing insights into optimizing model training strategies for clinical applications.
- State-of-the-Art Performance in Automated Skin Lesion Segmentation: The proposed approach achieves exceptional segmentation accuracy, with Dice and Jaccard indices of 0.9399 and 0.9112 on public databases and 0.8990 and 0.8445 on hybrid databases. These results ensure the system's real-time inference capability.
- Comprehensive Ablation Study for Performance Analysis: The presented study conducts a detailed ablation analysis, evaluating the impact of grayscale images, image blur, and model pruning on segmentation performance. The ablation study provides critical insights into the robustness and computational efficiency of YOLOv8 under different conditions.
- Benchmarking YOLOv8 Against YOLOv11 for Computational Efficiency: The performances of YOLOv8 and YOLOv11 were evaluated regarding segmentation accuracy, computational efficiency, and real-time applicability. The results indicate that YOLOv8

achieves higher segmentation accuracy with lower computational overhead. This balance between accuracy and efficiency makes YOLOv8 a more suitable candidate for real-time clinical applications.

2. Materials and Methods

In this section, the constructed database was first introduced. Then, the segmentation process and the algorithms utilized were explained to help the study's methodology be understood better.

2.1. Comprehensive Database

The extended database of the study was built by combining public and private dermoscopy images. A total of 4228 images were utilized in the study, comprising 3463 images acquired from public databases and 765 images sourced from private databases. The comprehensive database encompasses diverse dermoscopy images corresponding to various lesion types. The database was designed to include lesion images representing real-world challenges to ensure a thorough and reliable evaluation of the proposed segmentation process. These lesions are in individuals with dark skin tones, where the similarity between the lesion and the surrounding skin makes detection difficult, as well as lesions partially obscured by hairs. The database also features small lesions, images with complex backgrounds, lesions that span the entire imaging field, and those closely matching the skin tone. By integrating these varied scenarios, the segmentation method is rigorously tested against different realistic conditions to evaluate its capabilities.

The public portion of the database was obtained from ISIC, which focuses on improving the early detection and diagnosis of skin cancer through digital imaging. This collaborative project engages researchers, clinicians, and industry experts to create and share a comprehensive skin image database for research and education. The annual ISIC challenges serve as competitive platforms, inviting researchers and data scientists to devise algorithms facilitating the automated analysis of skin lesions. The ISIC challenges from 2016 to 2018 contributed to establishing the study's public database [23].

In the context of the segmentation task, 7723 images were sourced from the ISIC databases from 2016 to 2018. The presented study removed duplicated images from the combined ISIC databases, resulting in 4075 unique images. Notably, due to the absence of liquid usage during image acquisition in the private database (which was the primary focus of the developed segmentation process), images with bubbles were systematically excluded from the ISIC databases. Consequently, a public database containing 3463 images and ground truth masks was compiled. The private database for the study, which included 765 images (752 h $\times$ 582 w), was constructed by a dermatologist from Sakarya University Training and Research Hospital with the permission of the ethics committee. The dermatologist manually determined the lesion area for the images in the private database, thus creating ground truth masks.

Fifteen benign and malignant skin lesion types included melanoma, melanocytic nevus, lentigo, seborrheic keratosis, angioma, actinic keratosis, Bowen's disease, trichilemmal cyst, keratoacanthoma, hemangioma, malignant melanoma, basal cell carcinoma, pyogenic granuloma, Spitz nevus, and Merkel cell carcinoma. The training, validation, and test sets were assembled by integrating images from public and private databases. These databases were systematically divided into proportions of 80%, 10%, and 10%, respectively. The database preparation process is illustrated in Figure 1.

The hybrid database in this study combines public and private dermoscopic images to improve model generalizability and reflect real-world clinical challenges. However, private clinical images introduce additional complexities that can impact segmentation

performance. These images exhibit more significant variability in resolution, illumination conditions, and background artifacts compared to standardized public databases. Additionally, the private database includes a broader range of lesion types, including rare and atypical cases, which may not be adequately represented in public databases, potentially affecting the model's generalization ability. Another important factor is skin tone diversity, as the private database encompasses a broader spectrum of skin tones, influencing image contrast and segmentation accuracy. These factors contribute to the observed performance differences when training on public versus hybrid databases. To further enhance the robustness of the YOLOv8 and SAM-Box system, ongoing efforts are focused on expanding the dataset with additional clinical images from diverse sources, ensuring broader applicability across different patient demographics and clinical settings.

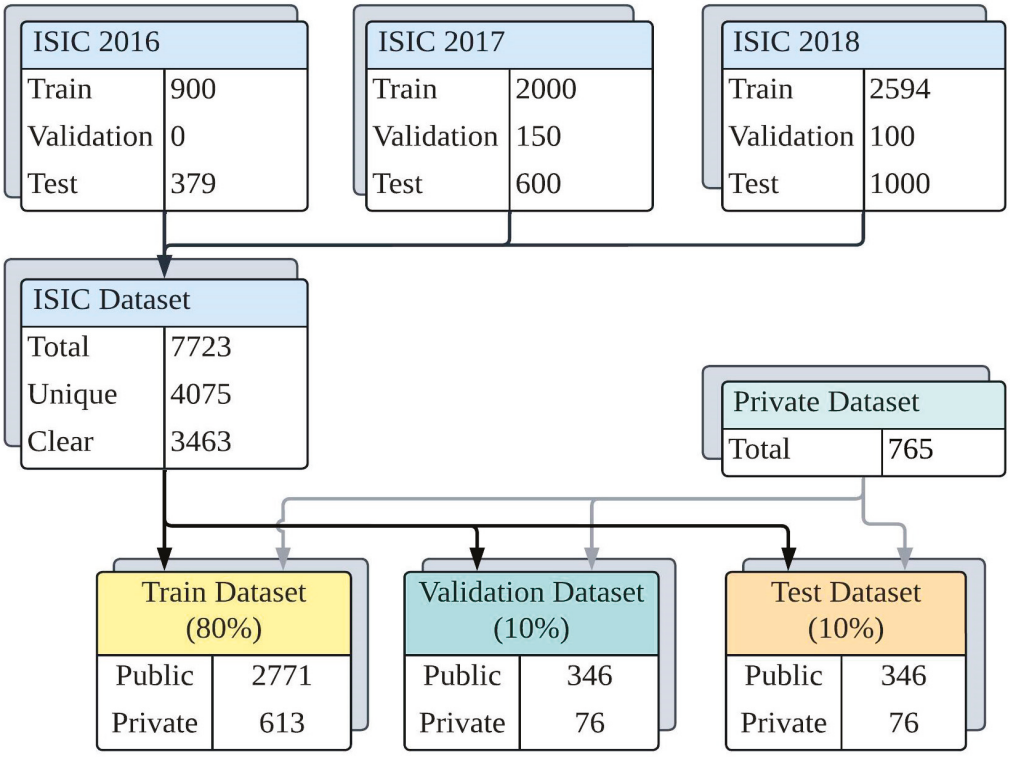


Figure 1. The database of the study.

2.2. YOLOv8 Architecture

Real-time object detection is necessary in many systems, from autonomous vehicles to medical devices. Although various algorithms have been developed for object detection, researchers prefer YOLO due to its advantages, especially in speed and accuracy. In addition, while YOLO gives the user the coordinates of the bounding box where the object is located, it can also undertake the classification task. In response to its interest, YOLO architecture has continued to improve since the day it was first developed. YOLO architecture is used in medical applications such as lesion segmentation, tumor detection, and pill identification to support specialists in deciding the most effective treatment processes [24].

YOLOv8 was introduced in 2023 by the Ultralytics company, which is also the developer of YOLOv5. The YOLOv8 model comes in alternative configurations: nano, small, medium, large, and x-large. Each configuration is designed to meet different computational resources and performance needs to offer options for efficient and accurate object detection tasks. In Figure 2, YOLOv8 architecture was demonstrated. The input layer of the model uses the sigmoid activation for the objectness score. The backbone of YOLOv8 is like that of YOLOv5, with some modifications on the cross-stage partial connections (CSP)

layer referred to as the C2f module, using high-level features and contextual knowledge to strengthen detection ability. The backbone of the YOLOv8 is the CSPDarknet53 feature extractor. A spatial pyramid pooling fast (SPPF) layer is included in the network to speed up the computation by pooling features into a fixed-size map. Batch normalization and SiLU activation are carried out in each convolution block. The head layer is connected to the outputs of the three C2f modules and performs object detection, classification, and regression tasks independently. YOLOv8 is developed using the Pytorch framework, providing 53.9% average precision (AP) on the COCO2017 dataset.

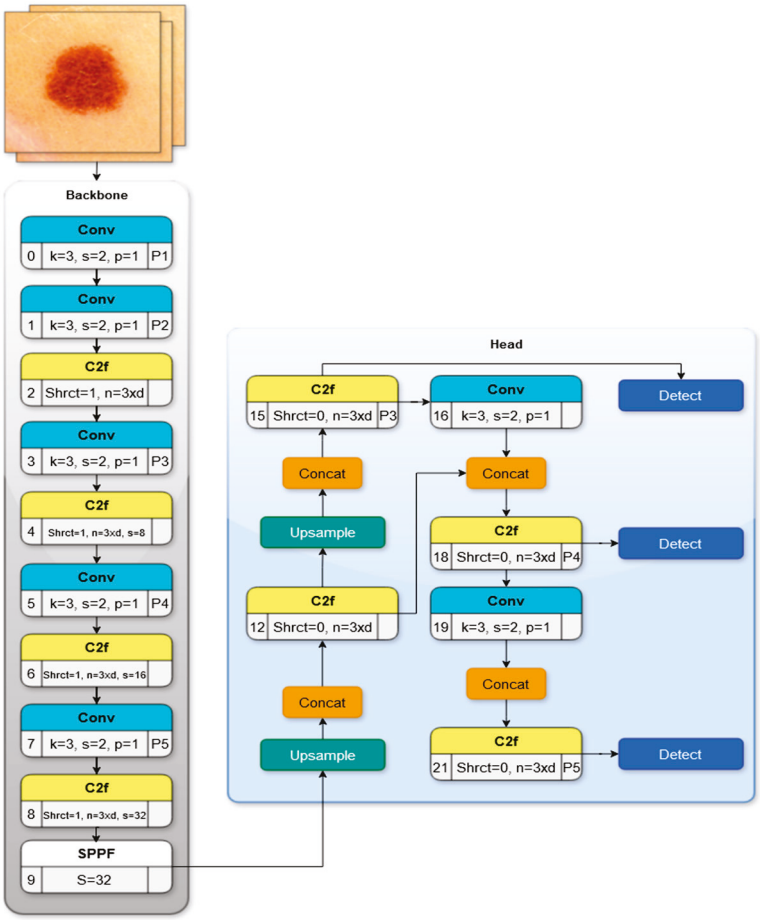


Figure 2. YOLOv8 architecture.

The architectural components and mathematical formulations employed in YOLOv8, including the backbone, neck, and head networks, are summarized below.

Backbone Network: The backbone network in YOLOv8 is responsible for feature extraction from the input images. It utilizes a modified CSPDarknet architecture, which includes the following components:

1. Convolutional layers apply convolution operations to extract fundamental image features such as edges and textures. The convolution operation is defined as

$$y = f(W * x + b) \tag{1}$$

where y represents the output feature map, x is the input feature map, W denotes the convolution kernel, b is the bias term, and f is a non-linear activation function, typically the Rectified Linear Unit (ReLU).

2. Residual blocks facilitate the training of deeper networks by enabling the gradient to flow more effectively through the network. A residual block can be mathematically expressed as

$$y = f(x) + x \tag{2}$$

where $f(x)$ denotes the output of the convolutional layers within the block.

3. CSP layers divide the feature map into two parts and merge them through a cross-stage hierarchy. This design enhances gradient flow and reduces computational complexity.

Neck Network: The neck network in YOLOv8 is designed to enhance the feature maps for object detection at multiple scales. It incorporates the following two networks:

1. The Feature Pyramid Network (FPN) structure allows for multiscale feature extraction by combining low-resolution, semantically rich features with high-resolution, semantically weak features. The merging of features at different scales is mathematically expressed as

$$P_i = F_i + UpSample(P_{i+1}) \tag{3}$$

where P_i represents the feature map at level i, and F_i is the feature map from the backbone network.

2. The Path Aggregation Network (PANet) improves information flow by incorporating a bottom-up path aggregation. The PANet equation can be given as follows:

$$P_i' = F_i + DownSample(P_{i-1}') \tag{4}$$

where $P_i{}'$ denotes the enhanced feature map at level i.

Head Network: The head network in YOLOv8 generates the final predictions, including the class labels, bounding box coordinates, and confidence scores. YOLOv8 employs an anchor-free detection head architecture.

1. Unlike the previous YOLO versions that use predefined anchor boxes in anchor-free prediction, YOLOv8 directly predicts object centers. The bounding box prediction involves the regression of offsets relative to grid cell centers:

$$t_x = \sigma(x) + c_x \tag{5}$$

$$t_y = \sigma(y) + c_y \tag{6}$$

$$t_w = p_w e^w \tag{7}$$

$$t_h = p_h e^h \tag{8}$$

where t_x and t_y are the predicted center coordinates, t_w and t_h are the predicted width and height, c_x and c_y are the grid cell coordinates, and p_w and p_h are the anchor box dimensions.

2. The network predicts a confidence score indicating the presence of an object and class probabilities for each bounding box using a sigmoid activation function:

$$confidence = \sigma(c) \tag{9}$$

$$class\ probabilities = Softmax(p) \tag{10}$$

where c is the raw confidence score, and p represents the raw class score logits.

The loss function in YOLOv8 combines localization loss, confidence loss, and classification loss to optimize the network. It is expressed as

$$L = \lambda_{coord} L_{coord} + L_{conf} + \lambda_{class} L_{class} \quad (11)$$

where L_{coord} is the localization loss (e.g., smooth L_1 loss for bounding box regression), L_{conf} is the confidence loss (e.g., binary cross-entropy), and L_{class} is the classification loss (e.g., categorical cross-entropy). The hyperparameters λ_{coord} and λ_{class} balance the contributions of each loss component. YOLOv8 is trained using a large-scale dataset with annotated images to ensure the detection of various objects with high accuracy. The training process involves optimizing the network parameters to minimize the combined loss function. Evaluation is performed on a separate validation dataset to assess the model's precision, recall, and mean Average Precision (mAP) performance.

In this study, YOLOv8 (Large) is used to determine the coordinates of the bounding box outlining the area with the lesion. Then, the obtained box coordinates are applied to the SAM-Box input.

2.3. Segment Anything Model

The SAM, introduced in 2023, represents a significant leap in image segmentation. Built upon a robust foundation of masked autoencoders (MAE) and vision transformers (ViT), SAM has been trained on the largest segmentation dataset, SA-1B, which includes over 1 billion masks and 11 million images. The core aim behind SAM's design is to provide a universal segmentation solution capable of performing well across a wide array of image types without requiring retraining or fine-tuning for specific tasks. This flexibility makes SAM a valuable tool for various image segmentation applications, including those in complex fields like medical imaging [25]. SAM operates through three key components, each playing a critical role in its segmentation process:

Image Encoder: The image encoder extracts meaningful features from input images. Using a ViT, the model divides an image into fixed-size patches, processes them through transformer layers, and captures intricate spatial relationships. ViT's architecture excels in handling large and high-resolution images, making it particularly effective for tasks like segmentation, where fine details and precise boundaries are crucial.

SAM offers three versions of ViT architecture, each tailored for different levels of complexity. ViT-B is the base version, designed for general use and optimized for computational efficiency while maintaining strong performance. Conversely, ViT-L is a larger version with more parameters, designed to handle more complex tasks requiring deeper feature extraction. ViT-H is the most advanced version, offering the highest level of accuracy, particularly in tasks where precision is critical.

Flexible Prompt Encoder: One of SAM's distinguishing features is its ability to adapt to various input prompts, which guide the segmentation process. These prompts can be sparse, such as a point, box, or text, or dense, like masks that provide pixel-level annotations. The prompt encoder processes these different prompt types using self-attention mechanisms, enabling SAM to adjust its focus based on user input. This flexibility benefits users who wish to guide the model's attention to specific regions of interest, enhancing segmentation accuracy.

Fast Mask Decoder: The final component in SAM is the fast mask decoder, which generates segmentation masks based on the features extracted by the encoder. This decoder combines cross-attention and self-attention to refine the mask predictions. Cross-attention allows the model to align features from the image and the provided prompt, while self-attention enables the decoder to update and improve its predictions iteratively. This attention-driven refinement process ensures high-quality segmentation outputs, even in

complex objects or boundaries requiring precise delineation. Figure 3 gives the SAM overview as used in the proposed study.

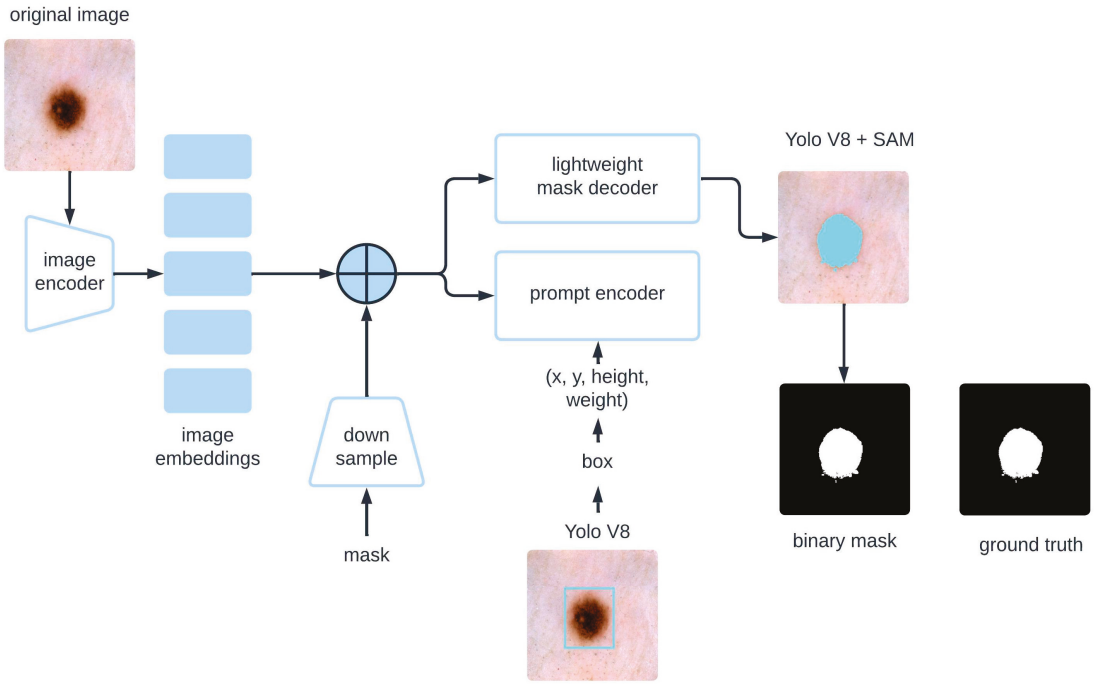


Figure 3. SAM-based scheme of the proposed study.

Since the introduction of SAM, its performance has been investigated in various medical applications. It is still being analyzed whether the SAM can be applied to more challenging medical image segmentation tasks due to the structural complexity, low contrast, and inter-observer variability of medical images. For this purpose, studies summarized under the following subheadings have been carried out: (i) pathology image segmentation, (ii) liver tumor segmentation from contrast-enhanced computed tomography (CECT), (iii) polyps segmentation from colonoscopy images, (iv) brain magnetic resonance imaging (MRI) segmentation, (v) abdominal computed tomography (CT) organ segmentation, (vi) endoscopic surgical instrument segmentation, and (vii) other mixed dataset evaluations. In these segmentation studies, three variations of SAM (SAM-Semantic, SAM-Point, and SAM-Box) were used, and evaluations were made by calculating a standard metric referred to as the Dice coefficient. According to the results of the studies, SAM provides significantly lower Dice values than existing deep learning-based methods used for medical image segmentation, such as U-Net architectures [26–31].

On the other hand, SAM-Box and SAM-Point methods slightly increase the segmentation performance compared to SAM-Semantic because they utilize user-provided prior information for the possible object region. However, requiring user interaction is an essential limitation of SAM-Box and SAM-Point methods in fully automatic segmentation applications. According to the parameter sizes, SAM can be used with three vision transformers: ViT-B, ViT-L, and ViT-H. In this study, the experimental results were achieved by performing all versions of SAM. The ViT-H model was selected as it provided the best performance regarding segmentation accuracy.

After introducing the database and the methods utilized, the flowchart of the proposed system can be summarized to better illustrate the study's contributions. Figure 4 shows the primary stages of the proposed skin lesion segmentation system.

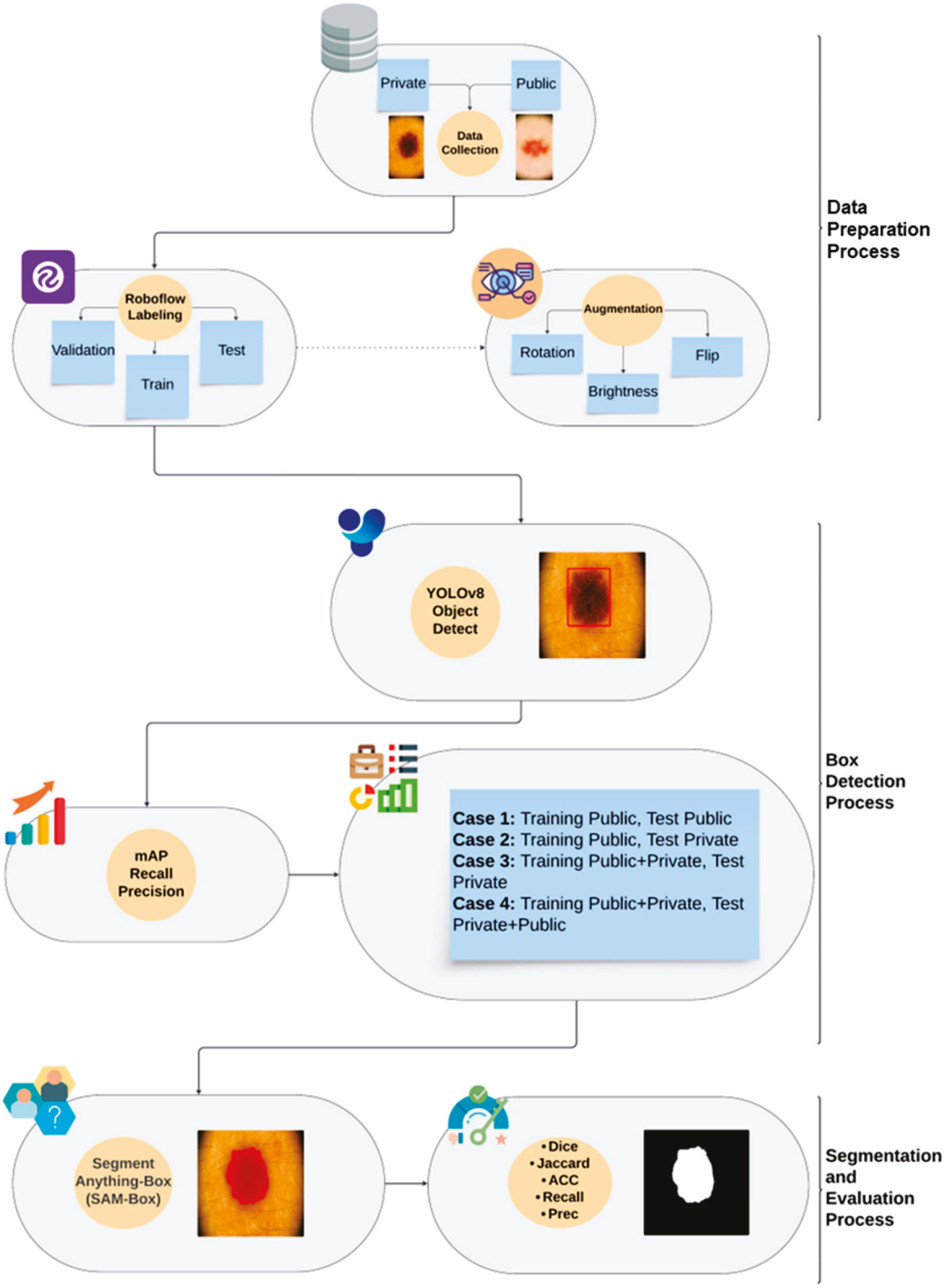


Figure 4. The flowchart of the study.

The main target of the study was to develop a fully automatic system that performs skin lesion segmentation tasks without intervention. As can be seen from the figure, the proposed skin lesion segmentation system consists of three main processes: (i) data preparation, (ii) box detection, and (iii) segmentation and evaluation.

The first step of the data preparation process was to build a comprehensive database consisting of dermoscopy images taken from various skin lesion types. The study database includes public and private lesion images, as illustrated in Figure 1. During the labeling process, only "lesion" regions were marked. This decision was made to focus on segmentation and enable the model to identify lesions accurately. By annotating solely the lesion areas, the aim was to streamline the segmentation task and enhance the model's ability to

detect lesions precisely. Then, the labeled images were used to train, validate, and test the YOLOv8 (Large). Before applying the data augmentation methods to the training images, all database images were resized to 256 × 256. The number of images in the training set has quadrupled through augmentation methods, which include vertical/horizontal flipping, brightness adjustment within the [−25% to +25%] range, and rotation within the [−15° to +15°] range.

During the box detection process, four training/test scenarios were created to evaluate the performance of object detection and segmentation methods reliably. The mAP, precision, and recall metrics were calculated to assess the performance of the YOLOv8 (Large) detector for each scenario. The final step of the system involved the segmentation and evaluation process, which aimed to determine the skin lesion area using SAM-Box automatically. In this step, the bounding box coordinates provided by the YOLOv8 (Large) were inputted into SAM-Box, and skin lesion masks were generated without any manual intervention. Finally, the skin lesion segmentation performance of the entire system was evaluated using five standard metrics by comparing the automatically generated masks with the ground truth masks.

3. Results and Discussions

The study aims to perform skin lesion segmentation without any intervention. The most up-to-date techniques have been used to achieve this goal. The study sequentially conducted data preparation, box detection, segmentation, and evaluation processes.

The performance analysis of the proposed system has been examined under two main headings. The efficacy of the YOLOv8 object detection algorithm was assessed in the first subsection. In contrast, the second subsection focused on the success of the SAM-Box employed for lesion segmentation. All simulations were conducted using Python 3.10.9.

3.1. Object Detection Results

The effectiveness of the YOLOv8 was initially evaluated using commonly employed metrics, followed by an analysis of loss functions to assess its performance. At first, precision, recall, and mAP metrics were considered to evaluate YOLOv8's performance.

Precision is a fundamental metric in evaluating object detection algorithms and measuring the accuracy of positive predictions. It quantifies the ratio of true positive predictions to the sum of true positive and false positive predictions. Precision is crucial in applications where minimizing false alarms is paramount, such as in security and surveillance systems. Recall, synonymous with sensitivity, assesses the ability of an object detection algorithm to identify all instances of a target object. It is calculated as the ratio of true positive predictions to the sum of true positive and false negative predictions. Recall is of utmost importance in applications where missing object instances can have severe consequences, as in the case of autonomous vehicles.

The mAP score, being a single scalar value, facilitates direct model comparisons and benchmarking. The mAP stands out as a comprehensive metric in object detection evaluation. It considers precision at multiple recall levels, providing a nuanced understanding of an algorithm's performance. The mAP score is obtained by calculating the average precision across all object classes, offering a unified measure of detection accuracy. A higher mAP score indicates superior performance, making it a widely adopted metric in benchmarking object detection models.

These metrics offer a comprehensive view of an algorithm's strengths and weaknesses, guiding researchers and practitioners in making informed decisions regarding algorithm selection and deployment. As object detection technology continues to advance, the refinement and standardization of evaluation metrics will play a crucial role in ensuring the

reliability and scalability of these algorithms across diverse applications. Furthermore, object detection models often comprise three main components: objectness prediction (object or not), bounding box regression (box coordinates), and class prediction (classification of the object). Each of these components has its associated loss function. The object loss focuses on determining whether an object is present, the box loss refines the coordinates of the bounding box, and the class loss handles the classification of the object. Combining these three components and their respective loss functions contributes to training a comprehensive object detection model capable of accurately localizing and classifying objects in an image.

As discussed before, the YOLOv8 (Large) model was used to perform the object detection task of the proposed study. According to the literature review, open databases were used in a significant part of the studies conducted for skin lesion segmentation. In the presented study, a private database was created, and a hybrid database was obtained by combining it with public databases. So, the performance of the object recognition algorithm was analyzed in two separate cases during the simulations. In the first case, training was conducted on public databases to compare the presented study with previous studies. In the second case, training was carried out on the hybrid database to assess the system's performance in clinical applications.

The training performance of YOLOv8 was investigated concerning the loss functions, precision, recall, and mAP metrics. Figures 5 and 6 show the performance graphics for "training on public databases" and "training on hybrid database" cases. According to the figures, the loss values vanish to zero as the number of epochs increases. The convergence of class and box loss toward zero collectively enhances the object detection algorithm's ability to accurately classify objects, precisely predicting bounding box coordinates. This convergence is indicative of a more refined and effective object detection model. The dual focal loss (DFL) curve was also given for YOLOv8. DFL considerably improves object detection by addressing class imbalance and handling complex examples more effectively than traditional focal loss. DFL evaluates the model's robustness to real-world challenges and improves its generalization capabilities. A diminishing DFL indicates that the model is becoming increasingly proficient at assigning appropriate emphasis to different examples, resulting in improved performance on challenging instances in the database. As can be seen from the figures, the algorithm training process was completed with high performance.

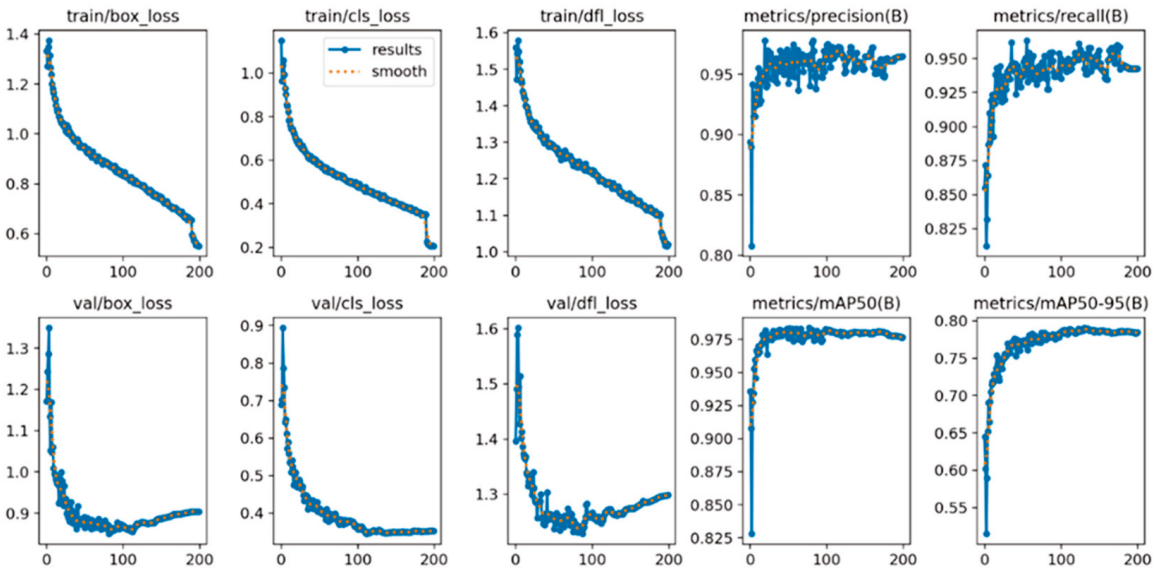


Figure 5. The training performance of YOLOv8 on the public database.

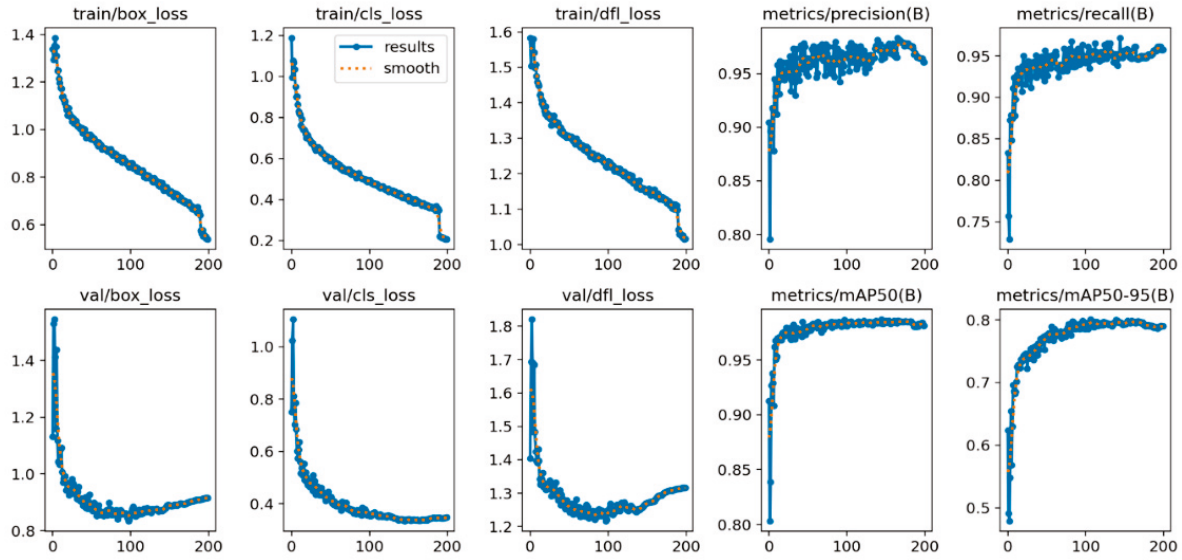


Figure 6. The training performance of YOLOv8 on the hybrid database.

Figure 7 illustrates the lesion detection (test) performance of YOLOv8 for dermoscopy images, including various challenging cases. Specifically, the evaluation considers lesions in individuals with dark skin tones and those partially obscured by hairs. Furthermore, the method's performance is investigated on small lesions, images with complex backgrounds, and lesions that span the entire imaging field. Interpreting segmentation results of challenging cases is crucial to reliably evaluating the performance of the proposed system.

In Figure 7, the columns were bounding boxes obtained by YOLOv8, segmentation masks provided by SAM-Box, binary segmentation masks, ground-truth masks, and difference images of predicted output and ground truth images, respectively. Note that in cases where multiple bounding boxes are detected, the algorithm proceeds with the box having the highest computed similarity score. As can be seen from Figure 7, the segmentation results of the utilized methods meet clinical expectations.

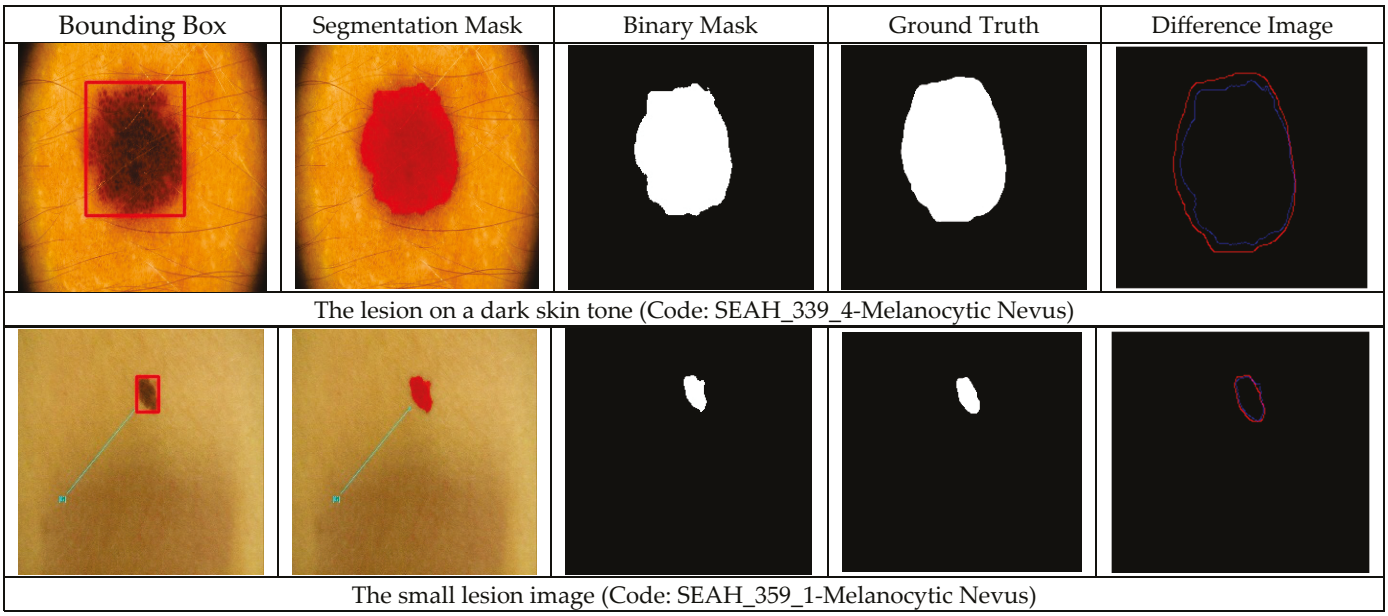


Figure 7. *Cont.*

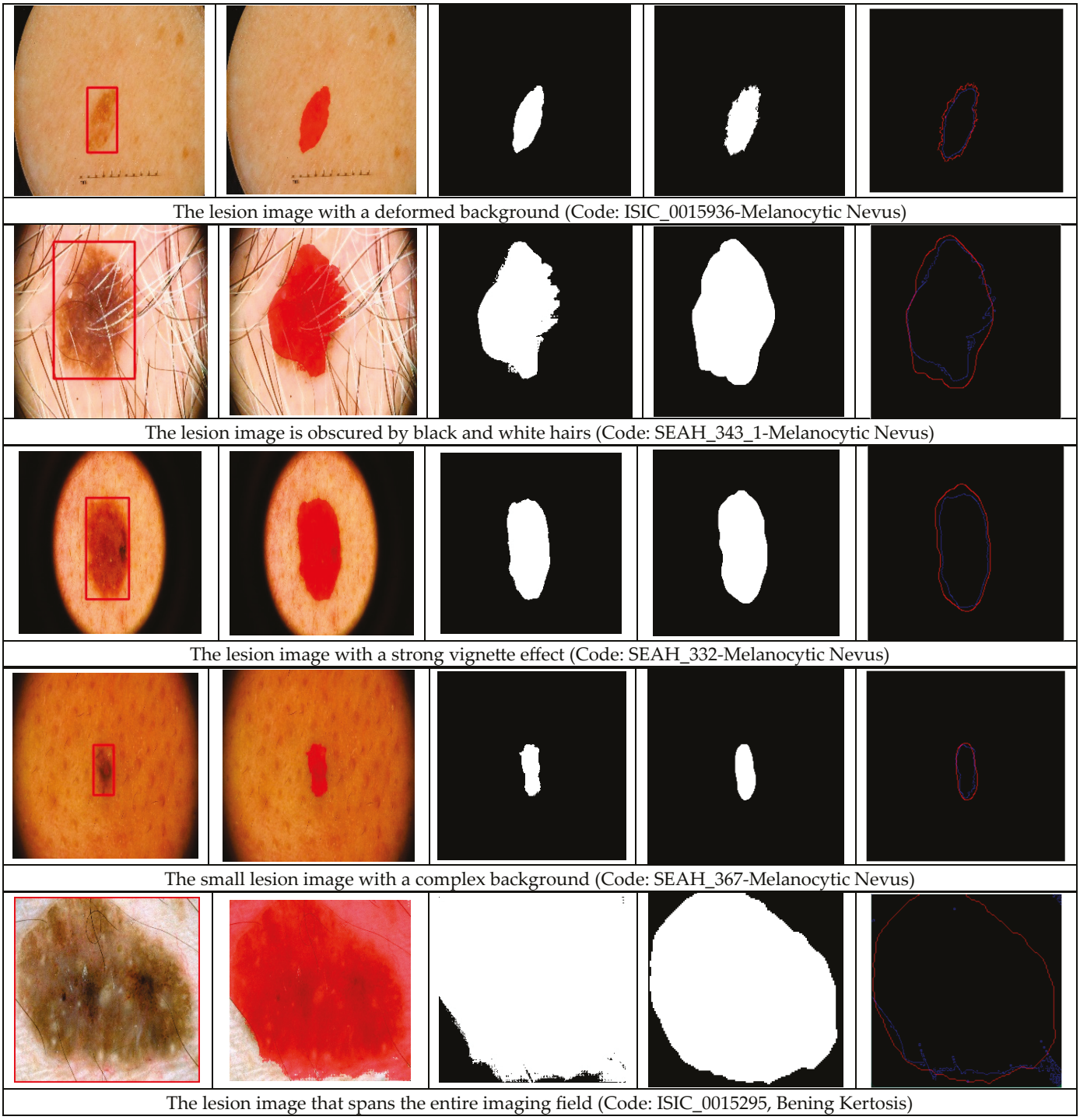


Figure 7. Segmentation results of challenging cases. (Blue and red lines show the predicted and ground-truth lesion boundaries, respectively).

3.2. Segmentation Results

The last process of the presented study is segmentation and evaluation. Accurate segmentation is imperative for biomedical image processing applications, and comparing manual and automatic segmentation results is essential for gauging the reliability of automated algorithms. The proposed study aims to segment skin lesions with the highest accuracy without intervention.

Various metrics were calculated by comparing the predicted masks with the ground truth masks to evaluate the performance of algorithms used for segmenting skin lesions.

The Dice coefficient and Intersection over Union (IoU) are two prevalent similarity measures. Dice is calculated as twice the overlap area between the predicted and ground truth masks, divided by the total area of both masks, yielding a value between 0 and 1, with 1 indicating perfect overlap. IoU or Jaccard index, on the other hand, is the ratio of the intersection area to the union area of the predicted and ground truth masks. These metrics are crucial for assessing the model's performance, especially in datasets with limited examples. The equations for Dice and IoU metrics can be given as follows:

$$Dice = \frac{2|A \cap B|}{|A| + |B|} \tag{12}$$

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{13}$$

$|A|$ and $|B|$ are the number of pixels in the predicted and ground truth segmentation sets. $|A \cap B|$ and $|A \cup B|$ show the intersection and union of these sets, respectively.

On the other side, the pixel-level confusion matrix, which includes true positive (TP), false positive (FP), true negative (TN), and false negative (FN) values, provides a detailed breakdown of the model's predictions. Accuracy, calculated as the ratio of correctly predicted pixels to the total number of pixels, offers an overall measure of the model's correctness. Precision, the proportion of true positive predictions among all positive predictions, indicates the model's ability to avoid false positives. Recall that the proportion of true positive predictions among all actual positives reflects the capacity of the model to detect all relevant cases. These metrics comprehensively evaluate the model's strengths and weaknesses, enabling precise adjustments to enhance performance. The equations of the metrics can be expressed as follows:

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

$$Accuracy = \frac{TP + FP}{TP + TN + FP + FN} \tag{16}$$

In the proposed study, SAM-Box was used for skin lesion segmentation purposes. However, as previously mentioned, the SAM-Box model requires box coordinates to be provided as a prompt. While open-source SAM is widely utilized for segmentation in many applications, prompts are often given manually, and SAM-Semantics, which does not require a prompt, is preferred. When prompts are provided manually, the system loses its automatic functionality, but using SAM-Semantics results in multiple candidate regions from the segmentation. This study evaluated the YOLOv8 (Large) model to obtain region box coordinates automatically.

Figure 8 illustrates the segmentation process applied to the original image through two approaches. The results of applying SAM-Semantics directly to the image are shown in the upper arm. In the lower arm, the outputs are obtained by applying YOLOv8 for object detection, followed by SAM-Box for segmentation, as suggested in the study. This comparison highlights the differences between using SAM-Semantics alone and combining YOLOv8 with SAM-Box for image segmentation. The ground truth mask is provided at the end of the figure to validate the performance of the proposed segmentation system.

The figure shows that SAM-Semantics generates five candidate regions when directly applied to the original image. Identifying which regions correspond to the lesion is crucial, necessitating system intervention. In contrast, the objective of this study was to achieve fully automatic lesion detection without any manual intervention. To this end, YOLOv8

architecture was employed, and bounding box coordinates were utilized as inputs for SAM, facilitating the automatic detection of the lesion region.

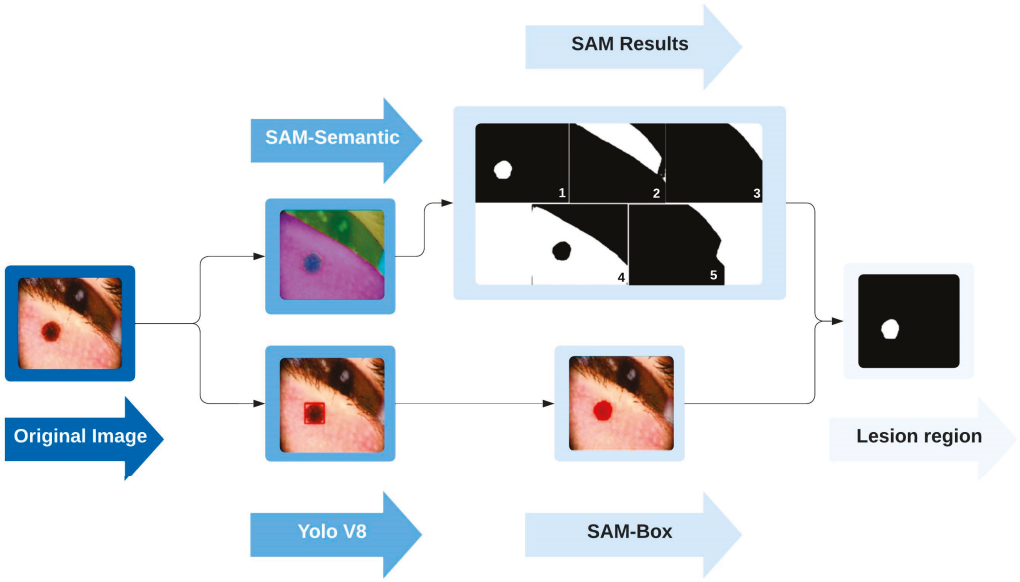


Figure 8. SAM outputs for a challenging skin lesion image.

The image given in Figure 8 shows a lesion located near the eye, sourced from a private database. Components such as the iris and eyelashes were included in acquiring the lesion image. In such cases, segmentation methods operating on the entire image may misinterpret non-lesion regions as lesions, potentially reducing their performance. This study deliberately incorporates images of clinical cases, like those illustrated, into the database. This inclusion aims to ensure that performance evaluation results are reliable for real-world applications, addressing situations commonly encountered in clinical practice and enhancing the robustness of the proposed method for practical implementations.

Table 1 was then compiled to show the entire system's performance for four cases. The first and second cases involve training on the public database, with subsequent testing conducted on both public and private databases. Conversely, the third and fourth cases entail training on a hybrid database consisting of public and private images, followed by testing on private and hybrid image databases. This approach enables a comprehensive evaluation of segmentation performance, capturing a spectrum of conditions encountered across public and private image databases. Conducting a study focused on skin lesion segmentation, a private database is being curated alongside publicly available ones for training. Utilizing both public and private databases offers several advantages. Public data aids in generalizing the model, enabling it to comprehend diverse scenarios effectively.

Conversely, the private database focuses on specific application domains, enhancing the model's performance within those contexts. Employing a mixed training approach that integrates both datasets improves the model's adaptability to general and domain-specific data. This approach promotes diversity in training and bolsters the reliability of the model's predictions.

Table 1 illustrates the performance evaluation of the YOLOv8 in detecting lesions and predicting bounding boxes. The assessment distinguishes between training these algorithms on a public or mixed database. As can be seen from the table, YOLOv8 exhibits robust performance across both training approaches. When trained on the public database, YOLOv8 achieves a test performance of 98.30% mAP, 96.20% precision, and 94.40% recall. Training on the mixed database results in slightly adjusted scores: 98% mAP, 96.90% precision, and 94.30% recall. The results underscore the efficacy of employing a mixed

database strategy, which enhances the method's performance to generalize across diverse lesion detection tasks. The findings highlight improvements in precision and recall metrics, emphasizing the significance of dataset diversity in optimizing algorithmic performance for practical applications in medical imaging.

Table 1. The result of the proposed skin lesion segmentation system for four training/test scenarios.

		Database		YOLOv8 Test Results (%)			SAM-Box Results				
		Public	**Private**	**mAP50**	**Precision**	**Recall**	**Dice**	**Jaccard**	**ACC**	**Recall**	**Precision**
Case 1	**Training**	✓		98.30	96.20	94.40	**0.9399**	**0.9112**	**0.9589**	**0.9308**	**0.9793**
	Test	✓									
Case 2	**Training**	✓					0.8901	0.8129	0.9569	0.9595	0.8424
	Test		✓								
Case 3	**Training**	✓	✓	98.00	96.90	94.30	0.8679	0.7887	0.9492	0.9381	0.8392
	Test		✓								
Case 4	**Training**	✓	✓				0.8990	0.8445	0.9552	0.9282	0.9033
	Test	✓	✓								

Following the evaluation of YOLOv8, SAM-Box has been assessed across four different scenarios, with the highest scores highlighted in bold in the table. A comprehensive analysis using five metrics demonstrates the effectiveness of the proposed method in accurately determining skin lesion regions. This underscores a significant advancement towards developing a fully automated reporting (decision support) system, achieved by integrating a classification step after the segmentation phase to achieve optimal performance. According to the results, the highest scores for YOLOv8 and SAM-Box in Case 1 are as follows: Dice coefficient, Jaccard index, accuracy, recall, and precision, with values of 0.9399, 0.9112, 0.9589, 0.9308, and 0.9793, respectively.

The observation regarding the performance decrease in Case 3 (training on the hybrid database) compared to Case 1 (training on the public database) can be affected by several factors explained below:

1. Variability in Image Quality: Private clinical databases often come from heterogeneous sources, leading to significant variations in image quality, lighting conditions, and resolution compared to standardized public databases. These discrepancies can hinder the model's ability to generalize effectively.
2. Lesion Size and Characteristics: The private database contains lesions with varying sizes, shapes, or appearances that differ from those in the public database, potentially causing lower performance metrics. Additionally, the presence of rare lesion types or atypical characteristics in the private database might not be adequately represented in the public dataset, leading to reduced generalization.
3. Skin Tone Variation: The private database includes a more diverse range of skin tones, which can affect image contrast and model performance. Public databases typically have limited representation of skin tones, which can influence the model's robustness across different populations.

These factors contribute to decreased Dice and Jaccard scores when transitioning from the public to the hybrid database. Regarding the model's generalizability, an external validation on independent clinical databases would be an important step in further assessing the robustness of the YOLOv8 and SAM-Box system. This would help evaluate how well the model performs across different clinical environments and patient demographics, enhancing its applicability and robustness. Therefore, our efforts to collect clinical data from diverse sources are ongoing.

In addition to the performance evaluation of YOLOv8, the latest YOLO: YOLOv11 is considered in the proposed study. A detailed comparison between YOLOv8 and YOLOv11 is performed using the constructed database, and the results support the choice of YOLOv8. The YOLOv11 performance for the skin lesion segmentation task is given in Table 2. Regarding detection performance, YOLOv8 demonstrated superior or comparable results across all metrics. If training was performed on a public database (Cases 1 and 2), YOLOv8 achieved an mAP50 of 98.3%, precision of 96.2%, and recall of 94.4%, while YOLOv11 showed an mAP50 of 98.3%, precision of 95.8%, and recall of 95.4%. These results were consistent across additional cases, highlighting YOLOv8's high accuracy and reliability.

Table 2. The result of the YOLOv11-based lesion segmentation for four training/test scenarios.

		Database		**YOLOv11 Test Results (%)**			**SAM-Box Results**				
		Public	**Private**	**mAP50**	**Precision**	**Recall**	**Dice**	**Jaccard**	**ACC**	**Recall**	**Precision**
Case 1	**Training**	✓		98.3	95.8	95.4	0.8394	0.8433	0.9296	0.8371	0.8827
	Test	✓									
Case 2	**Training**	✓					0.8764	0.7975	0.9443	0.8811	0.8994
	Test		✓								
Case 3	**Training**	✓	✓	98.3	95.8	94	0.8641	0.7874	0.9507	0.8399	0.9117
	Test		✓								
Case 4	**Training**	✓	✓				0.8588	0.7729	0.9442	0.8274	0.9190
	Test	✓	✓								

Regarding segmentation tasks, YOLOv8 outperformed YOLOv11 by a significant margin. For example, in Case 1, YOLOv8 achieved a Dice score of 0.9399 and a Jaccard score of 0.9112, while YOLOv11 provided 0.8394 and 0.7433, respectively. So, the YOLOv8 provides more precise and robust segmentation, which is crucial for the goals of our study.

Furthermore, the performance of YOLOv8 and YOLOv11 is compared to show the method's computational efficiency and memory requirements. The simulations used an MSI laptop with 32 GB RAM and an NVIDIA RTX 4070 8 GB GPU (CUDA). The results are given in Table 3. Regarding computational efficiency, YOLOv8 demonstrated faster inference times, with an average of 6.99 ms per prediction, compared to YOLOv11's 10.42 ms. Lower inference time makes YOLOv8 more suitable for real-time applications if speed is essential. Although YOLOv8 has more parameters (43M) and higher GFLOPs (165.4) than YOLOv11 (25M parameters and 87.3 GFLOPs), the segmentation and total elapsed times were similar: 1.05 s for YOLOv8 versus 1.06 s for YOLOv11, with total times of 164.57 s for YOLOv8 and 163.57 s for YOLOv11. Thus, YOLOv8 efficiently uses computational resources without compromising performance. Based on these findings, YOLOv8 is preferred for our study due to its consistently superior performance in detection and segmentation, faster inference times, and proven reliability in the research community. While YOLOv8 is currently optimal for the given work, we acknowledge the rapid advancements in object detection models and plan to explore newer versions of YOLO and other state-of-the-art models in future research to enhance the performance of our framework further.

Finally, the observation regarding handling challenging cases, such as lesions obscured by hair or those with complex backgrounds, is considered. In such edge cases, the proposed YOLOSAMIC system utilizes the bounding box generated by YOLOv8 (Large) to guide the SAM-Box segmentation process. This pre-localization step significantly improves the model's ability to distinguish lesions from non-lesion regions, especially in cases where lesion boundaries are ambiguous. However, these edge cases still present inherent

challenges, such as the possibility of false positives from non-lesion regions or difficulties in detecting complex lesion boundaries.

Table 3. Computational complexity comparison.

	YOLOv8	YOLOv11
Layers	365	631
Parameters	43.630.611	25.311.251
Gradients	43.630.595	25.311.235
GFLOPs	165.4	87.3
Avarage Segmentation Time (s)	1.05	1.05
Average Prediction Time (ms)	9.88	11.77
Total Time Elapsed (s)	164.57	163.57

So, compared to typical cases, a detailed analysis was performed by computing separate metrics for difficult cases (e.g., lesions obscured by hair, small lesions, and those with complex backgrounds) to better evaluate the system's reliability in such scenarios. These metrics, including the Dice coefficient, Jaccard index, accuracy, recall, precision, and specificity, were calculated explicitly for complex cases, enabling a clear performance comparison. This additional analysis offers more profound insights into the system's robustness and applicability to real-world clinical environments, particularly for cases that may not conform to the idealized conditions seen in typical datasets. Table 4 provides the segmentation results for difficult and typical cases.

Table 4. Segmentation performance for difficult and typical cases.

Typical Cases						
Model	**Dice**	**Jaccard**	**ACC**	**Recall**	**Precision**	**Specificity**
YOLOv8	0.8782	0.7924	0.9604	0.8332	0.9466	0.9816
Difficult Cases						
Model	Dice	Jaccard	ACC	Recall	Precision	Specificity
YOLOv8	0.8198	0.7254	0.8797	0.8292	0.8810	0.8676

Based on the table, the YOLOv8 and SAM-Box-based segmentation approach shows lower performance in previously noted challenging cases than typical ones, which is an expected outcome. A closer look at the table reveals that the decline is more pronounced in Dice and Jaccard scores, indicating a reduced overlap between the ground truth and predicted masks. However, metrics such as accuracy, precision, and specificity derived from the pixel-based confusion matrix do not exhibit a significant decrease. These results suggest that while the method struggles with precise boundary detection in complex cases, its overall classification ability remains relatively stable.

Our study significantly advances skin cancer diagnosis by integrating cutting-edge algorithms. These technologies offer state-of-the-art real-time efficiency and accuracy, outperforming the earlier models. The performance metrics achieved in the presented research are particularly noteworthy. The obtained mAP, precision, and recall values are about 98%, 96%, and 94%. These results indicate high accuracy and reliability in object detection, significantly improving over other models. Compared to existing studies in Table 5, the proposed system excels in balancing accuracy, automation, and clinical applicability. The proposed study integrates a hybrid dataset combining public and private clinical images. This inclusion is pivotal, as it rigorously tests the system against real-world challenges, such as lesions obscured by hair or complex backgrounds. The Dice and Jaccard

indices of 0.9399 and 0.9112 on public datasets highlight the system's superior segmentation precision, surpassing most prior studies. Furthermore, YOLOv8 exhibited higher specificity in all cases. According to the results, YOLOv8 is more effective at minimizing false positives and maximizing true positives, which is vital for our application.

Table 5. Comparison of the proposed system with recent studies.

Reference No.	Data	Model	Dice	Jaccard	Accuracy	Recall	Precision	Specificity
[9]	ISIC 2018 PH2 DermIS	U-net + LinkNet	0.8940	0.8090	0.9690	-	-	
[10]	ISIC	FSPBO-DQN	-	-	0.9236	-	-	
[11]	ISIC 2016 ISIC 2017	FCEDN	0.9848 0.8723	0.9641 0.8685	0.9832 0.9525	-	-	
[12]	HAM10000	SkinSAM	0.8879	0.7843	0.9450	-	-	
[13]	ISIC PH2	LinkNet-B7	0.9675 0.9742	-	-	-	-	
[14]	ISIC 2017 ISIC 2018 PH2	(MSAU-Net)	0.9032 0.8980 0.9377	0.9576 0.9560 0.9617	0.9576 0.9490 0.9617	-	-	
[15]	ISIC	MD^2N	0.9718	-	0.9746	0.9972	0.9363	
[32]	HAM10000	S^2C-DeLeNet	0.9467	0.9007	0.9720	0.9373	0.9593	
Ours	Public	YOLOv8 + SAM-Box	0.9399	0.9112	0.9589	0.9308	0.9793	0.9739
	Hybrid	YOLOv8 + SAM-Box	0.8990	0.8445	0.9552	0.9282	0.9033	0.9735

Table 5 presents a comparative analysis of our method with existing state-of-the-art segmentation models. While not all compared methods were tested on the same dataset, a fair evaluation is ensured by using ISIC 2016–2018 images, which are commonly employed in prior works. Due to ethical constraints, it is impossible to retrain existing SOTA methods on our hybrid dataset, as this would require access to models and training details that are not publicly available. The results highlight the effectiveness of our approach in handling real-world clinical challenges, reinforcing its robustness and generalizability.

Furthermore, our hybrid dataset incorporates public and private clinical images, ensuring a more realistic evaluation that accounts for challenges such as lesion size variability, skin tone diversity, and complex backgrounds. Due to ethical constraints, it is impossible to publicly share our private clinical dataset, as it contains sensitive patient data collected under an institutional-approved protocol. Additionally, re-implementing existing SOTA methods on our hybrid dataset is not feasible, as this would require access to the original models, including source code, hyperparameters, and training pipelines, many of which are unavailable or not reproducible based on published works. Despite this limitation, our results demonstrate that integrating both public and private data improves segmentation performance, underscoring the clinical applicability of the proposed approach.

While some methods achieve high accuracy through manual intervention or task-specific fine-tuning, our approach combines YOLOv8's advanced object detection with SAM-Box's segmentation capabilities to deliver a fully automated workflow. This inno-

vation streamlines the diagnostic process and demonstrates exceptional adaptability to diverse clinical conditions. This study sets a new standard in automation and reliability, paving the way for practical deployment in dermatology practices.

3.3. Ablation Study

This ablation study aims to evaluate the performance of YOLOv8 under various conditions to understand the impact of different model components and settings on their overall effectiveness. Ablation analysis focuses on the ability of YOLOv8 under specific image processing scenarios, including grayscale images, blurred images, and pruned models.

YOLOv8's performance on grayscale images is critical for evaluating how well the model handles reduced color information. The results show that YOLOv8 achieves a Dice score of 0.7938 and a Jaccard index of 0.7012, demonstrating robust segmentation ability even when working with monochromatic images (Table 6). Simulations show that YOLOv8 can effectively differentiate structures in grayscale images, making it suitable for limited color information. In real-world applications, images may suffer from blurring, which can negatively impact segmentation performance. So, YOLOv8 is performed on blurred images to simulate this condition. As shown in Table 6, YOLOv8 achieved a Dice score of 0.8666 and a Jaccard index of 0.7791. These values indicate YOLOv8's strong ability to perform segmentation even under blurred conditions.

Table 6. Ablation results.

	YOLOv8					
	Dice	**Jaccard**	**ACC**	**Recall**	**Precision**	**Specificity**
Gray Image	0.7938	0.7012	0.8547	0.8363	0.8431	0.8370
Blurred Image	0.8666	0.7791	0.9444	0.8325	0.9336	0.9590
%10 Pruned	0.8629	0.7740	0.9440	0.8250	0.9359	0.9592

Additionally, YOLOv8 maintained a high specificity, suggesting that the model minimizes false positives in challenging conditions. Finally, to evaluate YOLOv8's robustness when its model size is reduced, various tests are carried out at a 10% pruning level. The results presented in Table 6 show that YOLOv8 continues to perform well even with pruning, although there is a drop in performance as the pruning level increases. For the 10% pruned model, YOLOv8 achieved a Dice score of 0.8629 and a Jaccard index of 0.77398, demonstrating that the model retains its segmentation ability with minor pruning.

The ablation study demonstrates the effectiveness of YOLOv8 across a variety of conditions. Even with reduced color information, blurred images, and pruned model configurations, YOLOv8 maintains strong segmentation performance. While pruning does impact the model's performance, YOLOv8's robustness to these changes emphasizes its potential for real-world applications, where both computational efficiency and segmentation accuracy are essential.

In the final part of the experiments, a small and hybrid database is constructed, including 500 images to provide a more detailed justification for the choice of YOLOv8 (Large) and SAM's ViT-H variants. First, the bounding box detection performance of YOLOv8 configurations is investigated. Then, the experiments for SAM-Box variants were performed after YOLOv8 (Large) was carried out. The results achieved for YOLOv8 configurations are given in Table 7. A trade-off between detection accuracy and computational feasibility drove the choice of YOLOv8 (Large) over smaller variants. Given the critical importance of precise lesion localization in medical imaging, the large variant provides superior feature extraction capabilities, leading to higher detection performance (94.61% mAP, 93.75% precision, and 92.57% recall) compared to its smaller counterparts. YOLOv8's transformer-based

feature fusion enhances contextual understanding, which is crucial for detecting lesions with subtle differences in contrast.

Table 7. Performance evaluation of YOLOv8 configurations.

Bounding-Box Results (%)			
Model	**mAP50**	**Precision**	**Recall**
YOLOv8n	94.58	93.41	90.24
YOLOv8s	93.88	94.4	87.76
YOLOv8m	93.67	95.03	82.95
YOLOv8l	**94.61**	**93.75**	**92.57**
YOLOv8x	91.51	94.13	83.44

SAM-Box versions were examined after deciding to use the YOLOv8 (Large) model. The results are given in Table 8. According to the experiments, ViT-H was selected for the SAM-Box segmentation stage as it demonstrated the best performance among all available SAM versions. While ViT-B and ViT-L are more computationally efficient, ViT-H offers higher segmentation fidelity, particularly in challenging cases with irregular lesion boundaries. Experimental evaluations confirmed that ViT-H significantly improved segmentation performance, achieving a Dice coefficient of 0.8875 and a Jaccard index of 0.8010 on the constructed database. Furthermore, confusion matrix-based metrics are considerably higher than the remaining versions. The selection of these configurations thus balances computational cost with segmentation accuracy, ensuring optimal performance for real-world clinical applications.

Table 8. Segmentation performance of SAM-Box ViT versions.

SAM-Box Results with YOLOv8 Large Model						
	Dice	**Jaccard**	**ACC**	**Recall**	**Precision**	**Specificity**
SAM-ViT-B	0.8801	0.8012	0.9442	0.8819	0.9091	0.9562
SAM-ViT-L	0.8770	0.7953	0.9486	0.8476	0.9388	0.9736
SAM-ViT-H	**0.8875**	**0.8010**	**0.9531**	**0.8883**	**0.9256**	**0.9793**

In conclusion, the findings of this study contribute to the field of medical image analysis by demonstrating how hybrid deep-learning architecture can enhance segmentation performance, particularly in medical imaging applications. Theoretically, integrating YOLOv8 with SAM-Box shows the potential of combining object detection with segmentation to improve lesion localization and delineation.

From a practical standpoint, the proposed system offers a scalable and efficient approach for automated skin lesion analysis, reducing the reliance on manual annotations and improving diagnostic accessibility. Including public and private datasets ensures generalizability, making this method viable for real-world clinical applications.

4. Conclusions

This study introduces an innovative and fully automated skin lesion segmentation system that synergizes YOLOv8 and SAM-Box for accurate and reliable performance. The proposed method addresses critical challenges in dermatology, including lesion appearance variability and expert dermatologists' scarcity, by combining object detection and segmentation techniques. The system can operate across diverse datasets, including public and private clinical images, emphasizing its robustness and practical use.

The integration of YOLOv8 and SAM-Box achieved satisfactory precision, recall, Dice, and Jaccard indices, demonstrating substantial improvements over existing approaches.

The hybrid dataset, including real-world conditions like hidden lesions and complicated backgrounds, proved the system is reliable and effective. These results emphasize the significant potential of this automated solution to improve diagnostic processes and aid in clinical decision-making.

Furthermore, the critical role of ablation experiments in evaluating the performance of YOLOv8 in medical image segmentation is highlighted. The ablation results emphasize the model's ability to adapt to complex and noisy data. By comparing YOLOv8 with version v11, we have demonstrated that the enhancements in YOLOv8's architecture led to notable improvements in segmentation accuracy and precision.

This study aims to improve segmentation methods and expand the dataset to include a wider variety of lesion types. Future clinical trials will be essential in evaluating how well the model works in real medical settings. Overall, this research represents an important step toward creating accessible, accurate, and scalable tools for early skin cancer detection, supporting global healthcare goals to enhance patient outcomes with the help of technological advancements.

Author Contributions: G.C.: Conceptualization, study design, critical revision of the manuscript, final editing, supervision, and project administration. B.M.A.: Software development and data collection. S.G.: Literature review, software development, results interpretation, and data validation. D.A.: Methodology development, software development, and study design. R.Ö.K.: Data collection, data validation, and statistical analysis. All authors have read and agreed to the published version of the manuscript.

Funding: The study is supported by The Scientific and Technical Research Council of Turkey (TUBITAK) through The Research Support Programs Directorate (ARDEB) with project number 122E629.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Sakarya University Medical Sciences Faculty (protocol code E-71522473-050.01.04-12347-95 and date of 4 April 2022).

Informed Consent Statement: In this study, a hybrid database has been utilized. The hybrid database consists of the open-access ISIC database and a private dataset created specifically for this study, following approval by the relevant ethics committee. The private dataset comprises dermatoscopic images of patients who visited the Dermatology Clinic at Sakarya University Training and Research Hospital. The private dataset includes retrospectively collected images and the corresponding diagnoses made by physicians. No additional data beyond these were collected from the patients. Therefore, there was no need for a voluntary informed consent form, as no extra data were obtained for the purpose of this study.

Data Availability Statement: Data are contained within the article but cannot be shared due to ethical committee decision.

Conflicts of Interest: Authors Dr. Gökçen ÇETİNEL is the coordinator, Dr. Rabia ÖZTAŞ KARA, Dr. Devrim AKGÜN and Dr. Sevda GÜL are the researchers, and Bekir Murat AYDIN is the scholarship student in the project that is supported by The Scientific and Technical Research Council of Turkey (TUBITAK) through The Research Support Programs Directorate (ARDEB) with project number 122E629.

Abbreviations

The following abbreviations are used in this manuscript:

SAM	Segment Anything Model
mAP	Mean Average Precision
ISIC	International Skin Imaging Collaboration
CECT	Contrast-Enhanced Computed Tomography

MRI	Magnetic Resonance Imaging
CT	Computed Tomography
FPN	Feature Pyramid Network
PANet	Path Aggregation Network
ViT	Vision Transformer
CNN	Convolutional Neural Network
SPPF	Spatial Pyramid Pooling Fast
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
IoU	Intersection over Union
BCC	Basal Cell Carcinoma
SCC	Squamous Cell Carcinoma
EN-GWO	Exponential Neighborhood Grey Wolf Optimization
FCEDN	Fully Convolutional Encoder–Decoder Network
FSPBO-DQN	Fractional Student Psychology Optimization-Based Deep Q Network

References

1. World Health Organization, Cancer. Available online: https://www.who.int/news-room/fact-sheets/detail/cancer (accessed on 26 January 2024).
2. Zeng, L.; Gowda, B.H.J.; Ahmed, M.G.; Abourehab, M.A.S.; Chen, Z.; Zhang, C.; Li, J.; Kesharwani, P. Advancements in nanoparticle-based treatment approaches for skin cancer therapy. *Mol. Cancer* **2023**, *22*, 1–50. [CrossRef]
3. Radiation: Ultraviolet (UV) Radiation and Skin Cancer. Available online: https://www.who.int/news-room/questions-and-answers/item/radiation-ultraviolet-(uv)-radiation-and-skin-cancer (accessed on 26 January 2024).
4. Leiter, U.; Keim, U.; Garbe, C. Epidemiology of skin cancer: Update 2019. In *Sunlight, Vitamin D and Skin Cancer*; Reichrath, J., Ed.; Advances in Experimental Medicine and Biology; Springer: Cham, Switzerland, 2020; Volume 1268, pp. 123–139.
5. Urban, K.; Mehrmal, S.; Uppal, P.; Giesey, R.L.; Delost, G.R. The global burden of skin cancer: A longitudinal analysis from the Global Burden of Disease Study, 1990–2017. *JAAD Int.* **2021**, *2*, 98–108. [CrossRef] [PubMed]
6. Liu, Y.; Jain, A.; Eng, C.; Way, D.H.; Lee, K.; Bui, P.; Kanada, K.; de Oliveira Marinho, G.; Gallegos, J.; Gabriele, S.; et al. A deep learning system for differential diagnosis of skin diseases. *Nat. Med.* **2020**, *26*, 900–908. [CrossRef]
7. Lowell, B.A.; Froelich, C.W.; Federman, D.G.; Kirsner, R.S. Dermatology in primary care: Prevalence and patient disposition. *J. Am. Acad. Dermatol.* **2001**, *45*, 250–255. [CrossRef] [PubMed]
8. Qureshi, I.; Yan, J.; Abbas, Q.; Shaheed, K.; Riaz, A.B.; Wahid, A.; Khan, M.W.J.; Szczuko, P. Medical image segmentation using deep semantic-based methods: A review of techniques, applications, and emerging trends. *Inf. Fusion* **2023**, *90*, 316–352. [CrossRef]
9. Araújo, R.L.; de Araújo, F.H.D.; Silva, R.R.V.e. Automatic segmentation of melanoma skin cancer using transfer learning and fine-tuning. *Multimed. Syst.* **2022**, *28*, 1239–1250. [CrossRef]
10. Kumar, K.S.; Suganthi, N.; Muppidi, S.; Kumar, B.S. FSPBO-DQN: SeGAN based segmentation and Fractional Student Psychology Optimization enabled Deep Q Network for skin cancer detection in IoT applications. *Artif. Intell. Med.* **2022**, *129*, 102299. [CrossRef] [PubMed]
11. Mohakud, R.; Dash, R. Skin cancer image segmentation utilizing a novel EN-GWO based hyper-parameter optimized FCEDN. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34*, 9889–9904. [CrossRef]
12. Hu, M.; Li, Y.; Yang, X. SkinSAM: Empowering Skin Cancer Segmentation with Segment Anything Model. *arXiv* **2023**, arXiv:2304.13973. Available online: https://arxiv.org/abs/2304.13973v1 (accessed on 26 January 2024).
13. Akyel, C.; Arıcı, N. LinkNet-B7: Noise Removal and Lesion Segmentation in Images of Skin Cancer. *Mathematics* **2022**, *10*, 736. [CrossRef]
14. Alahmadi, M.D. Multiscale Attention U-Net for Skin Lesion Segmentation. *IEEE Access* **2022**, *10*, 59145–59154. [CrossRef]
15. Kumar, M.D.; Sivanarayana, G.V.; Indira, D.N.V.S.S.S.; Raj, M.P. Skin cancer segmentation with the aid of multi-class dilated D-net (MD2N) framework. *Multimed. Tools Appl.* **2023**, *82*, 35995–36018. [CrossRef]
16. Reis, H.C.; Turk, V.; Khoshelham, K.; Kaya, S. InSiNet: A deep convolutional approach to skin cancer detection and segmentation. *Med. Biol. Eng. Comput.* **2022**, *60*, 643–662. [CrossRef] [PubMed]
17. Kaur, R.; Hosseini, H.G.; Sinha, R.; Lindén, M. Automatic lesion segmentation using atrous convolutional deep neural networks in dermoscopic skin cancer images. *BMC Med. Imaging* **2022**, *22*, 1–13. [CrossRef] [PubMed]

18. Naeem, A.; Anees, T.; Khalil, M.; Zahra, K.; Naqvi, R.A.; Lee, S.-W. SNC_Net: Skin cancer detection by integrating handcrafted and deep learning-based features using dermoscopy images. *Mathematics* **2024**, *12*, 1030. [CrossRef]
19. Iqbal, I.; Younus, M.; Walayat, K.; Kakar, M.U.; Ma, J. Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images. *Comput. Med. Imaging Graph.* **2021**, *88*, 101843. [CrossRef] [PubMed]
20. Iqbal, I.; Shahzad, G.; Rafiq, N.; Mustafa, G.; Ma, J. Deep learning-based automated detection of human knee joint's synovial fluid from magnetic resonance images with transfer learning. *IET Image Process.* **2020**, *14*, 1990–1998. [CrossRef]
21. Mendes, J.; Vardasca, R. Systematic review of deep learning techniques in skin cancer detection. *BioMed Inform.* **2024**, *4*, 2251–2270. [CrossRef]
22. Riaz, S.; Naeem, A.; Malik, H.; Naqvi, R.A.; Loh, W.-K. Federated and transfer learning methods for the classification of melanoma and nonmelanoma skin cancers: A prospective study. *Sensors* **2023**, *23*, 8457. [CrossRef] [PubMed]
23. ISIC | International Skin Imaging Collaboration. Available online: https://www.isic-archive.com/ (accessed on 26 January 2024).
24. Terven, J.; Córdova-Esparza, D.-M.; Romero-González, J.-A. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1680–1716. [CrossRef]
25. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.; et al. Segment Anything. *arXiv* **2023**, arXiv:2304.02643. Available online: https://arxiv.org/abs/2304.02643v1 (accessed on 23 October 2023).
26. He, S.; Bao, R.; Li, J.; Stout, J.; Bjornerud, A.; Grant, P.E.; Ou, Y. Computer-Vision Benchmark Segment-Anything Model (SAM) in Medical Images: Accuracy in 12 Datasets. *arXiv* **2023**, arXiv:2304.09324. Available online: https://arxiv.org/abs/2304.09324v3 (accessed on 26 January 2024).
27. 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] [PubMed]
28. Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; Wang, B. Segment anything in medical images. *Nat. Commun.* **2024**, *15*, 1–9. [CrossRef]
29. Zhang, K.; Liu, D. Customized Segment Anything Model for Medical Image Segmentation. *arXiv* **2023**, arXiv:2304.13785. Available online: https://arxiv.org/abs/2304.13785v2 (accessed on 26 January 2024).
30. Huang, Y.; Yang, X.; Liu, L.; Zhou, H.; Chang, A.; Zhou, X.; Chen, R.; Yu, J.; Chen, J.; Chen, C. Segment Anything Model for Medical Images? *arXiv* **2023**, arXiv:2304.14660. Available online: https://arxiv.org/abs/2304.14660v4 (accessed on 25 October 2023). [CrossRef]
31. Zhang, Y.; Jiao, R. Towards Segment Anything Model (SAM) for Medical Image Segmentation: A Survey. *arXiv* **2023**, arXiv:2305.03678. Available online: https://arxiv.org/abs/2305.03678v3 (accessed on 26 January 2024).
32. Alam, M.J.; Mohammad, M.S.; Hossain, M.A.F.; Showmik, I.A.; Raihan, M.S.; Ahmed, S.; Mahmud, T.I. S2C-DeLeNet: A parameter transfer based segmentation-classification integration for detecting skin cancer lesions from dermoscopic images. *Comput. Biol. Med.* **2022**, *150*, 106148. [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

Promising Results About the Possibility to Identify Prostate Cancer Patients Employing a Random Forest Classifier: A Preliminary Study Preoperative Patients Selection

Eliodoro Faiella [1,2], Matteo Pileri [1,2], Raffaele Ragone [1,2,*], Anna Maria De Nicola [1,2], Bruno Beomonte Zobel [1,2], Rosario Francesco Grasso [1,2] and Domiziana Santucci [1,2]

[1] Unit of Radiology and Interventional Radiology, Fondazione Policlinico Universitario Campus Bio-Medico, Via Alvaro del Portillo, 00128 Rome, Italy; e.faiella@policlinicocampus.it (E.F.); matteo.pileri@unicampus.it (M.P.); annamaria.denicola@unicampus.it (A.M.D.N.); b.zobel@policlinicocampus.it (B.B.Z.); r.grasso@policlinicocampus.it (R.F.G.); d.santucci@policlinicocampus.it (D.S.)

[2] Research Unit of Radiology, Department of Medicine and Surgery, Università Campus Bio-Medico di Roma, Via Alvaro del Portillo, 00128 Rome, Italy

* Correspondence: rragone3@gmail.com

Abstract: Objective: This study evaluates the accuracy of a Machine Learning model of Random Forest (RF) type, using MRI data and radiomic features to predict lymph node involvement in prostate cancer (PCa). **Methods**: Ninety-five patients who underwent mp-MRI, prostatectomy, and lymphadenectomy at the Fondazione Policlinico Campus Bio-medico Radiological Department from 2016 to 2022 were analyzed. Radiomic features were extracted from T2-weighted, DWI, and ADC sequences and processed using a Random Forest (RF) model. Clinical data such as PSA levels and Gleason scores were also considered. **Results**: The RF model demonstrated significant accuracy in predicting lymph node involvement, achieving 84% accuracy for nodules in the peripheral zone (80% for predicting positive lymph node involvement and 85% for negative lymph node involvement) and 87% for those in the transitional zone (86% for predicting positive lymph node involvement and 88% for negative lymph node involvement). In the peripheral zone, key features included ADC shape maximum 2D diameter row and T2 noduloglcm difference variance, while in the transitional zone, DWI glcm difference average and DWI glcm Idm were important. DWI and ADC sequences were particularly crucial for accurate lymph node assessment. First-order features emerged as the most significant in whole-gland analysis, indicating fundamental differences in tumor composition and density critical for identifying malignancies with higher metastatic potential. **Conclusions**: AI-driven radiomic analysis, especially using DWI- and ADC-derived features, effectively predicts lymph node involvement in PCa patients, in particular in negative linfonode status patients, offering a promising tool for preoperative linfonode sparing patient selection. Further validation with larger cohorts is needed. Some limitations of this study are a relatively small sample size and it being a retrospective study.

Keywords: prostate cancer; lymph node involvement; multiparametric MRI; artificial intelligence; radiomics; Random Forest model

1. Introduction

Prostate cancer (PCa) is among the most prevalent malignancies and the second leading cause of cancer-related deaths in males. In 2023, Europe witnessed 1,261,990 cases

of cancer-related deaths [1]. The identification of pelvic lymph node metastasis (PLNM), present in approximately 15% of newly diagnosed prostate cancer patients, is a crucial prognostic factor that correlates with biochemical recurrence and distant metastases after curative treatment. Therefore, precise pre-treatment detection of PLNM in localized PCa cases could dramatically influence clinical decision-making processes, the formulation of treatment strategies, and the prediction of patient outcomes [2].

Advancements in artificial intelligence (AI) have propelled radiomics to the forefront of innovation, concentrating on the extraction of quantitative imaging features that non-invasively predict the behavior of nodules and tumors. This approach has the potential to overcome certain limitations in diagnostic accuracy associated with human interpretation. Radiomics has recently emerged as a critical tool for providing a more quantitative and objective evaluation of medical imaging. Through the analysis of quantitative attributes of tumor intensity and morphology, radiomic features can potentially reveal intratumoral histopathological characteristics, thus offering valuable prognostic insights for cancer management [3]. Among different ML algorithms, Random Forest has demonstrated a high effectiveness for handling high-dimensional datasets due to several advantages such as feature subset selection, dimensionality reduction without explicit feature elimination, resistance to overfitting, and parallel processing.

Magnetic Resonance Imaging (MRI) has been extensively adopted in radiomic analyses, particularly for PCa, owing to its ability to extract advanced features from images with high contrast and various weighting types. In radiomics studies, there are several type of radiomics features that can be taken out from a MRI image, based on the purpose of the specific study: the most common used are intensity, texture, shape of a region of interest, and gray intensity, in different types of weighted image [4]. The prostate is distinguishable into two regions on MRI scans: the peripheral zone (PZ), which exhibits hyperintensity on T2-weighted MRI owing to its glandular-ductal tissue, and the transition zone (TZ), which comprises stromal tissue and appears hypointense on T2-weighted images. Additionally, DWI/ADC sequences are crucial for delineating target areas, highlighted by the hyperintensity/hypointensity of suspicious nodules. Both T2WI and DWI/ADC sequences offer numerous features that can be leveraged to extract a variety of information.

The aim of this study is to evaluate the accuracy of an AI model of RF type in predicting lymph node involvement in prostate cancer using radiomic features from MRI sequences. The study also seeks to identify the most informative MRI sequences and radiomic features that contribute to accurate predictions, enhancing preoperative assessment and potentially reducing unnecessary invasive procedures.

2. Materials and Methods

This retrospective study analyzed 95 patients who underwent multiparametric magnetic resonance imaging (mp-MRI) followed by prostatectomy and lymphadenectomy at the Radiological Department of our hospital between 2016 and 2022. Inclusion criteria mandated that patients had both procedures. Clinical data collected included age, pre-mp-MRI PSA levels, MRI characteristics (signal intensity on T2-weighted images, DWI/ADC maps, and PI-RADS score), and histological tumor details (Gleason Score, TNM staging, capsule invasion, seminal vesicle invasion, neurovascular bundle involvement, and histological type). The cohort was divided into two groups based on lymph node status: 35 patients with at least one positive lymph node (positive group) and 60 patients with no lymph node metastasis (negative group).

The research protocol adhered to the ethical standards laid down by current legislation and the Declaration of Helsinki. The need for ethical review and approval was waived

owing to the retrospective nature of this study. All patients signed an informed consensus before MRI execution.

2.1. Magnetic Resonance Imaging

MRI scans were performed using a 1.5 Tesla Magnetom Aera scanner (Siemens®, Erlangen, Germany). The mp-MRI protocol included multiplanar T2-weighted images, echo-planar DWI with b-values of 0, 800, or 1000 s/mm^2 (ADC maps were automatically calculated), dynamic contrast-enhanced imaging (DCE), and axial T1-weighted images.

Patients were scanned in a supine position using an external coil, covering the prostate gland and seminal vesicles from the pelvis to the aortic bifurcation. Preoperative mp-MRI was performed systematically 4–5 weeks post-ultrasound-guided transrectal biopsy.

2.2. Segmentation, Feature Extraction and Selection

All mp-MRI data were transferred to a dedicated workstation. Manual segmentation of prostate nodules was performed on T2-weighted sequences, DWI, and ADC maps using 3D Slicer (v. 5.0.1). The segmentations were performed using a dedicated tool ("segment editor"), creating a mask of the region of interest (ROI) on the three different dimensional planes (axial, sagittal and coronal), and obtained a volume of interest (VOI). As we can see in Figure 1, the ROI is represented by a green zone on the images. Three volumes of interest (VOIs) were extracted per patient, yielding 339 radiomic features per patient after analysis with PyRadiomics software (version 3.0). Features included first-order statistics, shape-based features, and texture features such as those from the Gray-Level Co-occurrence Matrix (GLCM), Gray Level Dependence Matrix (GLDM), Gray-Level Run-Length Matrix (GLRLM), Gray-Level Size Zone Matrix (GLSZM), and Gray-Level Distance Zone Matrix (GLDZM).

In particular:

- Shape features: these characterize the spatial attributes of the region of interest (ROI), including its surface area, overall volume, diameter, elongation, sphericity, and the ratio of surface area to volume.
- First-order statistics (histogram-based features): these parameters describe the distribution of voxel intensities within the image region of interest, utilizing established statistical measures such as energy, entropy, mean, interquartile range, skewness, kurtosis, and uniformity.
- Second-order statistics (textural features): these methods analyze the statistical relationships between adjacent voxels. The significant features include the following:
    - Gray-level Cooccurrence Matrix (GLCM): examines the three-dimensional arrangement of grayscale intensity values within a volumetric image.
    - Gray-Level Run-length Matrix (GLRLM): evaluates adjacent voxels sharing identical gray-level values, assessing the length of gray-level runs in multiple orientations.
    - Gray-Level Size-Zone Matrix (GLSZM): measures the regions of adjacent voxels with identical gray-level intensity within a three-dimensional image.
    - Neighboring Gray-Tone Difference Matrix (NGTMD): determines the contrast between a voxel's intensity and the mean intensity of surrounding voxels within a defined radius.
    - Gray-level Dependence Matrix (GLDM): evaluates the quantity of interconnected voxels within a specified range that are reliant on the central voxel.

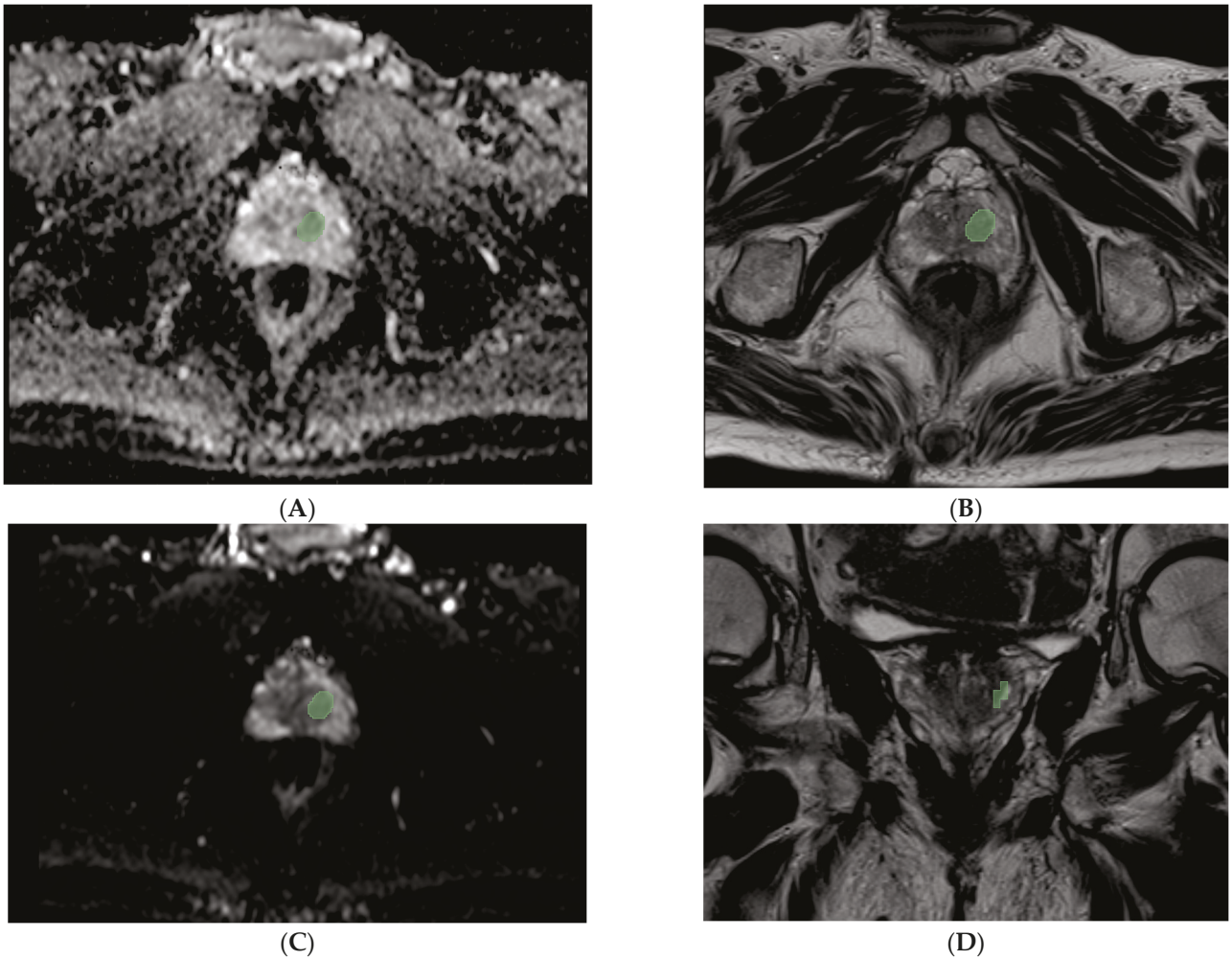


Figure 1. Prostatic nodule (green zone) in ADC map (**A**), axial T2 (**B**), DWI (**C**), and coronal T2 (**D**).

Coefficient value for all selected features was calculated considering the Random Forest model (RF importance) both for positive and negative prediction and calculated for both the peripheral and transitional prostatic zone and considering all features. RF importance value (based on Gini impurity) was reported for T2 sequences, ADC, and DWI, and was considered for the 5 most- and least-significant features. For the analysis, we just used all features.

2.3. Radiomics Analysis, Random Forest Model Development, and Features Identification

The study utilized a Random Forest (RF) algorithm for its efficacy in handling high-dimensional datasets, as we demonstrated in our previous article [5] through the comparison with two other ML classifiers. Separate RF models were constructed for each imaging sequence (DWI, ADC, and T2-weighted images). The model's performance was evaluated using metrics such as accuracy, precision, recall, F1-Score, and Area Under the Receiver Operating Characteristic Curve (AUC). The most impactful features within the AI models were identified and their RF importance values reported, particularly for T2 sequences, ADC, and DWI. ChatGPT version 4.0 was used for statistical analysis [6].

In the Random Forest model, feature importance is a measure of how much each variable contributes to improving the model's ability to make accurate predictions. To calculate feature importance in Random Forest, we used the Gini impurity reduction method, which is one of the most common techniques for assessing how much a feature contributes to improving the model.

Before using the Gini method, the dataset was split to divide the dataset into a training set and a validation set. The training set is used to train the model, while the validation set is used to test its performance during training, allowing for monitoring of potential overfitting.

So, it then continues with Gini method. It is as follows: a feature is used to make a split in the data within a decision tree, and it calculates how much this split reduces the impurity in the node. Gini impurity measures how mixed the classes are within a group: If a group contains mostly data from one class (e.g., patients with or without lymph node invasion), the impurity is low. If the group is highly mixed (containing both patients with and without lymph node invasion), the impurity is high.

The goal of each decision tree is to reduce the impurity at each node. The feature that causes the greatest reduction in impurity, by better separating the classes, is considered the most important. This process is repeated for all the trees in the Random Forest, and the final importance of each feature is calculated by summing the total reduction in impurity caused by that feature across all the trees. The features that contribute most to reducing impurity are those that the model considers most significant for making accurate predictions.

The selected features are the ones that led to the best results in the Random Forest model in terms of accuracy, precision, recall. The selection process, based on improving class separation through Gini impurity reduction, identified these features as the ones that provided the best balance between reducing error and improving model interpretability.

3. Results

The study included 95 patients, of whom 35 had metastatic nodules and 60 did not. The cohort consisted entirely of Caucasian individuals with a mean age of 60.3 y, with a median PSA level of 4.5 ng/mL and a median Gleason score of 7. Tumors were predominantly located in the peripheral zone (63 patients) and less frequently in the transitional zone (32 patients) (Table 1).

Table 1. Patients' characteristics.

Category	Number
Number of patients	95
Patients with metastatic nodule at lymphadenectomy	35
Patients without metastatic nodule at lymphadenectomy	60
Race	Caucasian
Age (Mean)	60.3 (40–80)
PSA [ng/mL] (Median, range)	4.5 (1.0–7.0)
Period of mp-MRI	2016–2023
Gleason grade (Median)	7
Tumor target zone peripheral	63
Tumor target zone transition	32
Disease grade	T3 (36) 37.89% T2 (30) 31.58% T3–5 (29) 30.5%

A single RF model was created considering only patients with prostate cancer. Subsequently, the same model was applied exclusively to patients with prostate cancer in the

transition zone (TZ) and the peripheral zone (PZ). The reported results (in Tables 3–7) correspond to the validation subset. The model exhibited high overall accuracy but showed a notable number of false positives (80) and false negatives (20). This indicates a tendency to favor the majority class, which could explain the lower AUC despite strong overall classification performance (Table 2).

Table 2. The confusion matrix reveals the disparity between accuracy and AUC, illustrating how misclassification of the minority class can affect the model's overall effectiveness.

Accuracy	Precision (Class 0)	Precision (Class 1)
84%	80%	88%
87%	75%	80%

The Random Forest (RF) model demonstrated significant accuracy in predicting lymph node involvement, achieving 84% accuracy for nodules in the peripheral zone (PZ) and 87% for those in the transitional zone (TZ) (Table 3). Specifically, in the PZ, the accuracy for predicting positive lymph node involvement (PL) was 80%, and for negative lymph node involvement (NL) it was 85%. In the TZ, the accuracy for PL was 86%, and for NL it was 88% (Table 4).

Table 3. Summary of the performance of the Random Forest model applied for the two groups based on the location inside the prostate gland. PZ indicates the nodules of the peripheral zone. TZ includes the nodules of the transitional zone.

Group	n	Accuracy	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1-Score (Class 0)	F1-Score (Class 1)
PZ	63	84%	80%	88%	85%	83%	82%	85%
TZ	32	87%	75%	80%	77%	78%	76%	79%

Table 4. The accuracy of RF model for calculating the positive lymph node involvement (PL) and negative lymph node involvement (NL) in peripheral zone (PZ) and transitional zone (TZ).

		Accuracy	AUC
PZ	**PL**	80%	75%
	NL	85%	83%
TZ	**PL**	86%	78%
	NL	88%	84%

The radiomic feature analysis of the study resulted in the extraction of three volumes of interest (VOIs) per patient—derived from the tumor segmentation on T2 sequences, DWI, and ADC maps. A total of 113 radiomic features were selected from each VOI, yielding 339 features per patient after analysis with the PyRadiomics software and subsequent processing by GPT-4. Key radiomic features significantly influenced the predictive performance of the RF models. In the PZ, the most influential features for positive LN prediction included ADC_shape_Maximum_2D_Diameter_Row and T2_noduloglcm_Difference_Variance, while for negative LN prediction, features like DWI_shape_Maximum_2D_Diameter_Row and ADC_glrlm_Long_Run_High_Gray_Level_Emphasis were crucial (Table 5). In the TZ, important features for positive LN prediction were DWI_glcm_Difference_Average and DWI_glcm_Idm, whereas for negative LN prediction, DWI_glcm_Id and DWI_glcm_Idmn were significant (Table 6). For the entire

gland, the top features for positive LN prediction were DWI_firstorder_Total_Energy and DWI_firstorder_Energy, and for negative LN prediction, DWI_glcm_Idmn and DWI_gldm_Large_Dependence_Emphasis were most influential (Table 7).

Table 5. Top 5 most influential radiomic features in terms of RF importance for positive and negative LN in the PZ.

	Radiomic Features	RF Importance
Positive LN	ADCshapeMaimum2DDiameterRow	0.03
	T2_noduloglcmDifferenceVariance	0.02
	T2_noduloshapeElongation	0.02
	ADCglrlmGrayLevelNonUniformity	0.02
	T2_noduloglcmId	0.02
Negative LN	DWIshapeMaimum2DDiameterRow	−0.71
	ADCglrlmLongRunHighGrayLevelEmphasis	−0.68
	ADCglcmCorrelation	−0.66
	DWIshapeMaimum3DDiameter	−0.63
	DWIgldmLargeDependenceEmphasis	−0.62

Table 6. Top 5 most influential radiomic features in terms of RF importance for positive and negative LN in the TZ.

	Radiomic Features	RF Importance
Positive LN	DWIglcmDifferenceAverage	0.05
	DWIglcmIdm	0.04
	ADCfirstorderTotalEnergy	0.03
	DWIglrlmGrayLevelNonUniformityNormalized	0.03
	ADCgldmGrayLevelVariance	0.02
Negative LN	DWIglcmId	−0.56
	DWIglcmIdmn	−0.53
	DWIglcmIdn	−0.50
	DWIglrlmGrayLevelNonUniformityNormalized	−0.49
	DWIglcmIdm	−0.48

Table 7. Top 5 most influential radiomic features in terms of RF importance for positive and negative LN in the whole gland.

	Radiomic Features	RF Importance
Positive LN	DWIfirstorderTotalEnergy	0.04
	DWIfirstorderEnergy	0.03
	T2_noduloglcmDifferenceAverage	0.02
	T2_noduloglszmLargeAreaLowGrayLevelEmphasis	0.02
	DWIglrlmShortRunLowGrayLevelEmphasis	0.02
Negative LN	DWIglcmIdmn	−0.45
	DWIgldmLargeDependenceEmphasis	−0.43
	DWIglcmIdn	−0.41
	T2_noduloshapeMinorAisLength	−0.38
	DWIglcmId	−0.35

4. Discussion

MRI has become an essential non-invasive modality for assessing lymph node (LN) status, although its effectiveness can be influenced by the expertise of radiologists [7,8]. The T2-weighted sequence plays a crucial role by providing high-resolution anatomical details, particularly useful for delineating the transitional zone and identifying tumor boundaries [9,10]. DWI is a functional imaging method that uses the random translational motion of water molecules to allow noninvasive assessment of biological tissues. Tissue characterization is made possible by the quantitative expression of the degree of diffusion limitation through the computation of ADC maps [9,11,12]. The exclusion of post-contrast dynamic contrast-enhanced (DCE) sequences from this study is based on their use within the PI-RADS framework, where they are primarily recommended for indeterminate cases. This decision aligns with current guidelines, which suggest that DCE sequences provide limited additional diagnostic value in routine assessments [13].

Our study aimed to enhance the prediction of LN involvement by extracting and analyzing radiomic features from MRI sequences and incorporating them into a sophisticated Random Forest (RF) analytical model. The radiomics features were combined with various clinical data points, such as laboratory results, PSA levels, and Gleason scores, to determine the most informative MRI sequence for LN status determination.

A recent review conducted by Huynh et al. highlights the growing body of literature advocating for the incorporation of radiological and radiomic features to improve prognostic models in PCa [14] and demonstrated superior clinical utility in predicting LN involvement, with 33 studies detailing the current scope of MRI-derived radiomics in PCa management. These studies employ various MRI sequences, feature extraction methods, and validation techniques to develop radiomic models with impressive discriminative abilities, as evidenced by the AUC values reported. For instance, a study by Damascelli et al. utilized a combined T2-weighted and ADC model that yielded an AUC of 0.90 for LN involvement prediction, supporting the integrative approach [15]. Another study by Liu et al. demonstrated the predictive strength of machine learning models in determining LN status with AUCs up to 0.86, reinforcing the potential of such technologies in clinical practice [16]. In support of this claim, Hou et al. developed a model based on T2W, DWI/ADC sequences with an AUC of 0.92 for assessing lymph node involvement in internal testing and 0.72 for external testing [17]. However, despite these promising developments in radiomic models for LN prediction in PCa, the labor-intensive process of volume of interest (VOI) acquisition remains a significant barrier to widespread clinical adoption, preventing radiologists from performing radiomic analysis [18].

The radiomic features offer a more profound analysis of prostate cancer that traditional modalities [19].

The radiomic features considered in this study, particularly those derived from DWI and ADC sequences, demonstrated significant predictive power for lymph node involvement in prostate cancer. Our analysis showed that radiomic features were notably more impactful and significant in predicting negative lymph node involvement, especially in the transitional zone. The RF importance scores highlighted that features such as "DWI_glcm_Id" and "DWI_glcm_Idmn" were critical for predicting negative lymph node involvement, with an accuracy reaching 85% in the transitional zone (Table 3). This indicates a higher predictive capability for patients with negative lymph nodes, potentially allowing for the avoidance of unnecessary lymphadenectomies, thus supporting the use of MRI as a non-invasive screening tool.

The features selected and previously extracted from MRI sequences (T2W, DWI, and ADC) were chosen because they are directly linked to the biological and physiological characteristics of the tumor and surrounding tissues, which are crucial for diagnosing

and staging prostate cancer. Including these features provides potentially more useful diagnostic information than other features that may be less relevant to lymph node staging.

RF is highly effective for high-dimensional datasets due to its ability to automatically identify relevant features and handle noisy or irrelevant data. Its ensemble approach reduces overfitting, bias, and variance while capturing non-linear relationships without extensive preprocessing. RF is scalable, supports both classification and regression tasks, and handles imbalanced datasets effectively. Its robustness and versatility make it suitable for applications in bioinformatics, finance, and radiomics analysis. The model's poor performance, particularly in terms of precision and recall for class 0 and class 1, is likely due to the dataset's imbalance. With only 95 patients and 35 positive cases of lymph node involvement, the model tends to favor the majority class (class 0), leading to missed detections (low recall) and false positives (low precision) for class 1. Additionally, the small dataset increases the risk of overfitting, as the model may rely excessively on specific features that do not generalize well.

In our previous study we explored three algorithms (RF, Logistic regression, and SVM) in order to predict LNS in PCa patients, and we obtained the significantly best performance using RF model [5].

The importance of these radiomic features was particularly pronounced when data from the peripheral and transition zones were analyzed separately rather than collectively. This differentiation underscores the value of individualized zone-specific analysis in enhancing the accuracy of predictive models. The majority of significant features were from the DWI and ADC sequences, not T2-weighted images, reflecting the lesion intrinsic superior contrast resolution of DWI in the transition zone, which makes it preferable for assessing tumor characteristics in this region.

GLCM (Gray-Level Co-occurrence Matrix) features, which capture textural information, were prominently featured among the top predictive indicators. This aligns with findings by Prata et al. who highlighted the critical role of textural features in radiomic models for PCa [20]. Our results show that GLCM features such as "DWI_glcm_Difference_Average" and "DWI_glcm_Id" had high RF importance scores, emphasizing their utility in capturing the intricate textural nuances of tumor heterogeneity that are crucial for accurate lymph node status prediction (Tables 4 and 5).

Additionally, shape-based radiomic features also played a significant role in the predictive models. Features like "ADC_shape_Maximum_2D_Diameter_Row" and "T2_nodulo_shape_Elongation" were among the most influential for positive lymph node prediction, while "DWI_shape_Maximum_2D_Diameter_Row" was critical for negative lymph node prediction (Table 4). These shape features help characterize the geometric properties of the tumor, providing essential insights into tumor morphology that are complementary to textural features.

When evaluating the predictive value of lymph node involvement for the whole gland, first-order radiomic features emerged as the most significant, particularly in predicting positive lymph node status. This finding suggests that basic statistical measures of voxel intensity within the tumor, such as mean, variance, and skewness, are highly informative. These features likely reflect fundamental differences in tumor composition and density that are critical for identifying malignancies with higher metastatic potential. The prominence of first-order features in whole-gland analysis underscores their utility in capturing overall tumor characteristics that are predictive of positive lymph node involvement.

These findings suggest that radiomic analysis, particularly with DWI and ADC-derived features, can significantly enhance the preoperative assessment of prostate cancer, reducing the need for invasive procedures in patients with negative lymph nodes. This

aligns with previous studies that advocate for the integration of radiomic features to improve diagnostic accuracy and prognostic models in prostate cancer management.

Our findings align also with pioneering studies like those of Faiella et al. and Bourbonne et al., which showcased the efficacy of a synergistic clinical-radiomic methodology [19,21]. By condensing an extensive array of features into a focused subset, their work spotlighted the diagnostic significance of T2-weighted sequence features. However, our analysis goes a step further by demonstrating the discriminative strength of DWI features, which not only complement but also enrich the predictive capacity of our model. The incorporation of these DWI features, along with the proven T2-weighted features, echoes the strategy recommended by Huynh et al., championing a combined radiomic–clinical model. This combination not only elevates the predictive accuracy but also weaves a more detailed tapestry of the tumor's behavioral patterns [14]. By incorporating both DWI and T2-weighted imaging features, we provide a nuanced understanding of the tumor environment, laying down an advanced framework for prognostic modeling. This integrated imaging strategy paves the way for precision medicine, offering enhanced tools for clinical decision-making and personalized patient care strategies, particularly in the predictive modeling of lymph node involvement in prostate cancer.

Zheng et al. [22] developed an integrated radiomics model (IRM) to predict lymph node involvement. This model combined histopathological examination, radiomics features extracted from a prostate index lesion, and clinical characteristics, utilizing only a support vector machine (SVM). The IRM demonstrated strong predictive performance, achieving an area under the curve (AUC) of 0.915 in the testing set. Moreover, the IRM's AUC significantly outperformed existing clinical nomograms, which ranged from 0.698 to 0.724, with a statistically meaningful difference ($p \leq 0.05$).

In another research, Liu et al. [16] introduced two preoperative models for evaluation of PLNM, utilizing multivariate logistic regression and incorporating both radiological and radiomics characteristics. These models demonstrated an AUC of 0.89–0.90. Additionally, their research indicated that a DWI-based radiomics nomogram, combining LN radiomics signature with quantitative radiological features, shows potential for predicting PLNM in PCa patients, especially for normal-sized LNM.

Despite promising results, it is important to acknowledge the limitations of our study, including its retrospective nature, modest sample size, and single-institution dataset. These factors may limit the generalizability of our findings and should be taken into consideration when interpreting our results. Future research should aim to validate our findings in a larger and more diverse population to reinforce the utility of this model in clinical practice, also considering including DCE sequences to potentially enhance predictive models for lymph node involvement in PCa.

5. Conclusions

This study underscores the potential of AI-driven radiomic analysis for improving the prediction of lymph node involvement in prostate cancer, especially when leveraging DWI and ADC sequences. The radiomic model demonstrated high accuracy, particularly for negative lymph node predictions in the peripheral zone, suggesting that MRI-based radiomics could serve as an effective screening tool to reduce unnecessary invasive procedures. Shape-based features, along with textural features, were crucial in the predictive models, highlighting the multifaceted nature of tumor characterization necessary for accurate lymph node assessment. Further validation with larger, multi-institutional cohorts is necessary to confirm these findings and enhance clinical decision-making and personalized treatment strategies in prostate cancer management. Perspective strategies include higher patient sample, multi-cohort study design, possibly with patients differing on

clinic-anamnestic characteristics and employing different MR scans (3T vs. 1.5T) and deep learning techniques application, and longitudinal studies assessment in order to assess the impact of predictive models on clinical outcomes. Even more validation with a larger dataset and open-source sharing would be appreciated in the future.

Author Contributions: Conceptualization, E.F., D.S. and R.R.; methodology, D.S. and R.R.; software, M.P. and R.R.; validation, A.M.D.N., M.P. and B.B.Z.; formal analysis, R.R. and M.P.; investigation, D.S.; resources, B.B.Z. and R.F.G.; data curation, R.R. and A.M.D.N.; writing—original draft preparation M.P. and R.R.; writing—review and editing, D.S.; visualization, R.R.; supervision, E.F., D.S. and R.R.; project administration, E.F. and D.S.; funding acquisition, E.F., B.B.Z. and R.F.G. 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 according to the guidelines of the Declaration of Helsinki. Ethical review and approval were waived for this study, due to the retrospective nature of the study.

Informed Consent Statement: Patient consent was waived due to the retrospective nature of this study.

Data Availability Statement: All datasets generated for this study are included in the manuscript.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. Malvezzi, M.; Santucci, C.; Boffetta, P.; Collatuzzo, G.; Levi, F.; La Vecchia, C.; Negri, E. European cancer mortality predictions for the year 2023 with focus on lung cancer. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **2023**, *34*, 410–419. [CrossRef]
2. Bianchi, L.; Gandaglia, G.; Fossati, N.; Suardi, N.; Moschini, M.; Cucchiara, V.; Bianchi, M.; Damiano, R.; Schiavina, R.; Shariat, S.F.; et al. Pelvic lymph node dissection in prostate cancer: Indications, extent and tailored approaches. *Urologia* **2017**, *84*, 9–19. [CrossRef] [PubMed]
3. Mayerhoefer, M.E.; Materka, A.; Langs, G.; Häggström, I.; Szczypiński, P.; Gibbs, P.; Cook, G. Introduction to Radiomics. *J. Nucl. Med.* **2020**, *61*, 488–495. [CrossRef] [PubMed]
4. Smith, C.P.; Czarniecki, M.; Mehralivand, S.; Stoyanova, R.; Choyke, P.L.; Harmon, S.; Turkbey, B. Radiomics and radiogenomics of prostate cancer. *Abdom. Radiol. N. Y.* **2019**, *44*, 2021–2029. [CrossRef]
5. Santucci, D.; Ragone, R.; Vergantino, E.; Vaccarino, F.; Esperto, F.; Prata, F.; Scarpa, R.M.; Papalia, R.; Beomonte Zobel, B.; Grasso, F.R.; et al. Comparison between Three Radiomics Models and Clinical Nomograms for Prediction of Lymph Node Involvement in PCa Patients Combining Clinical and Radiomic Features. *Cancers* **2024**, *16*, 2731. [CrossRef] [PubMed] [PubMed Central]
6. Meo, A.S.; Shaikh, N.; Meo, S.A. Assessing the accuracy and efficiency of Chat GPT-4 Omni (GPT-4o) in biomedical statistics: Comparative study with traditional tools. *Saudi Med. J.* **2024**, *45*, 1383–1390. [CrossRef] [PubMed] [PubMed Central]
7. Calabrese, A.; Santucci, D.; Landi, R.; Beomonte Zobel, B.; Faiella, E.; de Felice, C. Radiomics MRI for lymph node status prediction in breast cancer patients: The state of art. *J. Cancer Res. Clin. Oncol.* **2021**, *147*, 1587–1597. [CrossRef]
8. Bedrikovetski, S.; Dudi-Venkata, N.N.; Maicas, G.; Kroon, H.M.; Seow, W.; Carneiro, G.; Moore, J.W.; Sammour, T. Artificial intelligence for the diagnosis of lymph node metastases in patients with abdominopelvic malignancy: A systematic review and meta-analysis. *Artif. Intell. Med.* **2021**, *113*, 102022. [CrossRef]
9. Santucci, D.; Faiella, E.; Cordelli, E.; Calabrese, A.; Landi, R.; de Felice, C.; Beomonte Zobel, B.; Grasso, R.F.; Iannello, G.; Soda, P. The Impact of Tumor Edema on T2-Weighted 3T-MRI Invasive Breast Cancer Histological Characterization: A Pilot Radiomics Study. *Cancers* **2021**, *13*, 4635. [CrossRef] [PubMed]
10. Caglic, I.; Barrett, T. Diffusion-weighted imaging (DWI) in lymph node staging for prostate cancer. *Transl. Androl. Urol.* **2018**, *7*, 814–823. [CrossRef] [PubMed]
11. Eiber, M.; Beer, A.J.; Holzapfel, K.; Tauber, R.; Ganter, C.; Weirich, G.; Krause, B.J.; Rummeny, E.J.; Gaa, J. Preliminary results for characterization of pelvic lymph nodes in patients with prostate cancer by diffusion-weighted MR-imaging. *Invest. Radiol.* **2010**, *45*, 15–23. [CrossRef]
12. De Felice, C.; Cipolla, V.; Guerrieri, D.; Santucci, D.; Musella, A.; Porfiri, L.M.; Meggiorini, M.L. Apparent diffusion coefficient on 3.0 Tesla magnetic resonance imaging and prognostic factors in breast cancer. *Eur. J. Gynaecol. Oncol.* **2014**, *35*, 408–414. [PubMed]

13. Turkbey, B.; Rosenkrantz, A.B.; Haider, M.A.; Padhani, A.R.; Villeirs, G.; Macura, K.J.; Tempany, C.M.; Choyke, P.L.; Cornud, F.; Margolis, D.J.; et al. Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2. *Eur. Urol.* **2019**, *76*, 340–351. [CrossRef]
14. Huynh, L.M.; Hwang, Y.; Taylor, O.; Baine, M.J. The Use of MRI-Derived Radiomic Models in Prostate Cancer Risk Stratification: A Critical Review of Contemporary Literature. *Diagnostics* **2023**, *13*, 1128. [CrossRef] [PubMed]
15. Damascelli, A.; Gallivanone, F.; Cristel, G.; Cava, C.; Interlenghi, M.; Esposito, A.; Brembilla, G.; Briganti, A.; Montorsi, F.; Castiglioni, I.; et al. Advanced Imaging Analysis in Prostate MRI: Building a Radiomic Signature to Predict Tumor Aggressiveness. *Diagnostics* **2021**, *11*, 594. [CrossRef]
16. Liu, X.; Tian, J.; Wu, J.; Zhang, Y.; Wang, X.; Zhang, X.; Wang, X. Utility of diffusion weighted imaging-based radiomics nomogram to predict pelvic lymph nodes metastasis in prostate cancer. *BMC Med. Imaging* **2022**, *22*, 190. [CrossRef] [PubMed]
17. Hou, Y.; Bao, J.; Song, Y.; Bao, M.-L.; Jiang, K.-W.; Zhang, J.; Yang, G.; Hu, C.-H.; Shi, H.-B.; Wang, X.-M.; et al. Integration of clinicopathologic identification and deep transferrable image feature representation improves predictions of lymph node metastasis in prostate cancer. *EBioMedicine* **2021**, *68*, 103395. [CrossRef] [PubMed]
18. Han, C.; Ma, S.; Liu, X.; Liu, Y.; Li, C.; Zhang, Y.; Zhang, X.; Wang, X. Radiomics Models Based on Apparent Diffusion Coefficient Maps for the Prediction of High-Grade Prostate Cancer at Radical Prostatectomy: Comparison With Preoperative Biopsy. *J. Magn. Reson. Imaging JMRI* **2021**, *54*, 1892–1901. [CrossRef]
19. Faiella, E.; Vaccarino, F.; Ragone, R.; D'Amone, G.; Cirimele, V.; Piccolo, C.L.; Vertulli, D.; Grasso, R.F.; Zobel, B.B.; Santucci, D. Can Machine Learning Models Detect and Predict Lymph Node Involvement in Prostate Cancer? A Comprehensive Systematic Review. *J. Clin. Med.* **2023**, *12*, 7032. [CrossRef] [PubMed]
20. Prata, F.; Anceschi, U.; Cordelli, E.; Faiella, E.; Civitella, A.; Tuzzolo, P.; Iannuzzi, A.; Ragusa, A.; Esperto, F.; Prata, S.M.; et al. Radiomic Machine-Learning Analysis of Multiparametric Magnetic Resonance Imaging in the Diagnosis of Clinically Significant Prostate Cancer: New Combination of Textural and Clinical Features. *Curr. Oncol.* **2023**, *30*, 2021–2031. [CrossRef] [PubMed]
21. Bourbonne, V.; Jaouen, V.; Nguyen, T.A.; Tissot, V.; Doucet, L.; Hatt, M.; Visvikis, D.; Pradier, O.; Valéri, A.; Fournier, G.; et al. Development of a Radiomic-Based Model Predicting Lymph Node Involvement in Prostate Cancer Patients. *Cancers* **2021**, *13*, 5672. [CrossRef] [PubMed]
22. Zheng, H.; Miao, Q.; Liu, Y.; Mirak, S.A.; Hosseiny, M.; Scalzo, F.; Raman, S.S.; Sung, K. Multiparametric MRI-based radiomics model to predict pelvic lymph node invasion for patients with prostate cancer. *Eur. Radiol.* **2022**, *32*, 5688–5699. [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

Evaluation of the Performance of a YOLOv10-Based Deep Learning Model for Tooth Detection and Numbering on Panoramic Radiographs of Patients in the Mixed Dentition Period

Ramazan Berkay Peker [1,*] and Celal Oguz Kurtoglu [2]

[1] Department of Dentomaxillofacial Radiology, Faculty of Dentistry, Trakya University, Edirne 22030, Turkey
[2] Department of Computer Engineering, Faculty of Engineering, Cukurova University, Adana 01330, Turkey; celaloguz.kurtoglu@gmail.com
* Correspondence: rberkaypeker@trakya.edu.tr

Abstract: This study evaluated the performance of a YOLOv10-based deep learning model in detecting and numbering teeth in the panoramic radiographs of pediatric patients in the mixed dentition period. **Methods:** Panoramic radiographic images from 200 pediatric patients in the mixed dentition period, each with at least 10 primary teeth and underlying permanent tooth germs, were included in the study. A total of 8153 teeth in these panoramic radiographs were manually labeled. The dataset was divided for the development of a YOLOv10-based artificial intelligence model, with 70% used for training, 15% for testing, and 15% for validation. **Results:** The precision, recall, mAP50, mAP50-95, and F1 score of the model for tooth detection and numbering were found to be 0.90, 0.94, 0.968, 0.696, and 0.919, respectively. **Conclusions:** YOLOv10-based deep learning models can be used to accurately detect and number primary and permanent teeth in the panoramic radiographs of pediatric patients in the mixed dentition period, which can support clinicians in their daily practice. Future works may focus on model optimization across varied pediatric cases to enhance clinical applicability.

Keywords: artificial intelligence; YOLO; tooth classification; panoramic radiography; pediatric dentistry

1. Introduction

Panoramic radiographs (PRs), introduced in the 1950s, have become a significant and widely used diagnostic technique in dentistry [1]. One of the primary advantages of PRs is their ability to capture the maxilla, mandible, temporomandibular joints, and associated anatomical structures within a single image [2,3]. However, the superimposition and distortion of anatomical structures due to the two-dimensional nature of PRs can hinder accurate evaluation [4]. Despite these limitations, the advantages of PRs—such as the relatively low radiation dose, the ability to image a broad area within a short clinical timeframe, and procedural simplicity—make them a frequent choice as an initial diagnostic tool in pediatric dentistry [3].

Accurate interpretation of PRs is essential for correct diagnosis and, consequently, appropriate treatment. In the mixed dentition period, particularly in children, identifying and numbering both erupted and unerupted teeth is critical for treatment planning and prioritizing necessary interventions. However, manual numbering of erupted and unerupted

teeth is a time-consuming process. Furthermore, the accuracy of this numbering varies with the experience of the examining dentist [5,6].

Artificial intelligence (AI) is a broad term for computer systems designed to perform tasks that typically require human intelligence, such as visual perception and decision-making [3]. The use of AI-based methods to assist dentists in interpreting images enables faster identification and classification of data. These applications also help eliminate errors associated with human factors, such as fatigue and lack of experience [6,7].

In dental research, machine learning (ML), a subset of AI, focuses on developing algorithms that can learn from data to make predictions or decisions without explicit programming and to improve performance as they encounter more data [3]. Deep learning (DL), a subfield of ML, involves techniques that automatically learn from datasets through the use of convolutional neural networks (CNNs), a type of deep learning model (DLM). Rather than following explicit instructions, CNNs learn by creating models based on data analysis [5,8].

You Only Look Once (YOLO), a DL algorithm, can be directly and independently trained on validation data, enabling it to detect multiple objects in an image in real time without the need for a reference dataset. Due to its high detection speed, accuracy, and generalization capability, YOLO is among the most popular CNNs for object segmentation and detection [9,10].

YOLOv10, developed by researchers at Tsinghua University and released in May 2024, is considered a significant advancement in real-time object detection [11]. This new architecture aims to balance accuracy and computational efficiency to address the challenges of object detection. YOLOv10 achieves this through innovative training strategies, architectural modifications, and a combination of model variants. The core concept involves using "consistent dual assignment" during training, which allows the model to learn from rich supervision signals while eliminating the need for computationally expensive non-maximum suppression (NMS) during inference. Extensive evaluations have demonstrated that YOLOv10 surpasses previous YOLO versions and other state-of-the-art ML models in terms of accuracy–efficiency balance [11].

Several studies have employed DLMs for tooth detection and numbering in PRs [6,12–14]. However, few studies have focused on the mixed dentition period [5]. To our knowledge, no study has investigated the use of YOLOv10 for tooth segmentation and detection during the mixed dentition period. This study evaluates the effectiveness of YOLOv10 in the automatic detection and numbering of teeth in pediatric patients in the mixed dentition period, thereby contributing to both dentomaxillofacial radiology and pediatric dentistry. The study hypothesizes that YOLOv10 will successfully detect and number both primary and permanent teeth.

2. Materials and Methods

The preparation of this manuscript followed the procedures outlined in the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) and the Standards for the Reporting of Diagnostic Accuracy Studies (STARD). This study was approved by the Ethics Committee of Trakya University Faculty of Medicine (Approval number: TÜTF-GOBEAK 2024/450). All the procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and the Declaration of Helsinki. Informed consent was waived because the study involved a retrospective analysis of previously collected and de-identified panoramic radiographs.

2.1. Overview of the Study Workflow

To provide a clear and concise representation of the methodology, a flowchart summarizing the study workflow is presented in Figure 1. This flowchart outlines the key steps, including patient selection, data preprocessing, labeling, model training, validation, and performance evaluation. It is intended to improve the readability and reproducibility of the study by visually guiding the reader through the methodological process.

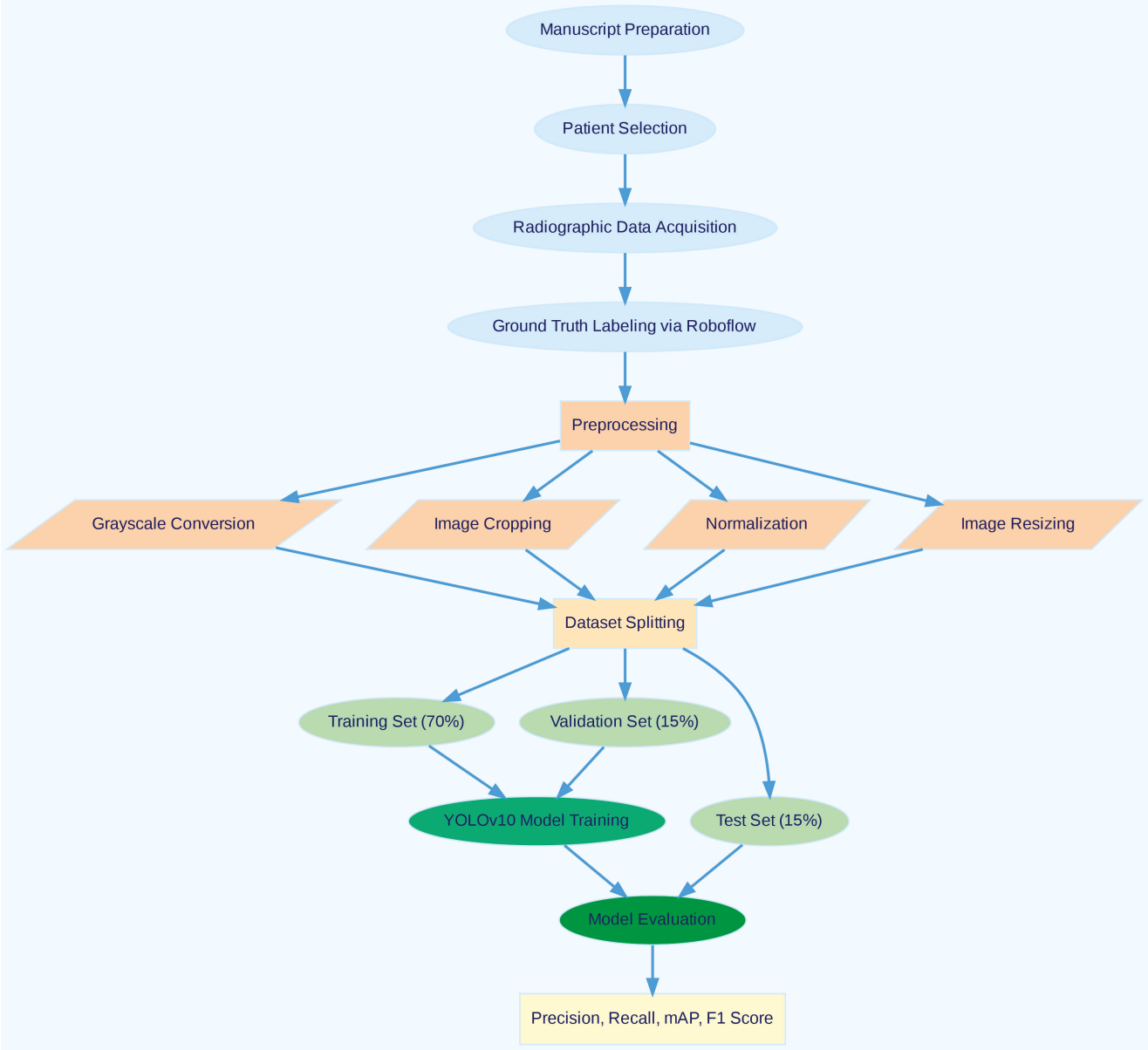


Figure 1. Workflow of the study. The flowchart illustrates the key steps undertaken in this research, including patient selection, data preprocessing, labeling, YOLOv10 model training, validation, and performance evaluation. Each step is detailed to provide a clear and concise overview of the methodology, ensuring reproducibility and aiding in the interpretation of the study results.

2.2. Patient Selection

In this retrospective study, we analyzed the PRs of pediatric patients aged 5 to 13, which were obtained from the archives of the Department of Oral and Maxillofacial Radiology at (Blinded for Review) University Faculty of Dentistry from 2019 to 2024. Patients with PRs containing artifacts (e.g., motion or metal artifacts), those with cystic or tumoral formations in the jaw, developmental abnormalities (e.g., cleft lip or palate), or positioning

errors during PR acquisition were excluded from the study. A total of 200 patient images, comprising PRs classified as Grade 1 or Grade 2 according to the UK National Radiological Protection Board criteria, were included in the study [12]. These PRs contained at least 10 primary teeth with underlying permanent tooth germs. The presence of caries, restorative fillings, root canal treatments, tooth rotation, or supernumerary teeth in the included cases was not considered an exclusion criterion.

2.3. Radiographic Dataset

The PRs were acquired by a single operator using standardized patient positioning on a PaX-Flex radiography device (Vatech, Seoul, Republic of Korea) with settings of 50–90 kV and 4–10 mA and an exposure time of 10.1 s.

2.4. Ground Truth Labeling

The labeling process enables ML or DL models to derive meaningful contexts related to objects in an image. Within these contexts, the classes and regions of the objects are identified. In this study, labeling was carried out by R.B.P., an oral and maxillofacial radiologist with seven years of experience (Figure 2). Two weeks after the initial labeling, all PRs were reevaluated by R.B.P. The labeling process was conducted on the Roboflow website (https://roboflow.com/, accessed on 19 November 2024) using the bounding box method. The software used for dataset preparation and labeling, including Roboflow, offers both free and paid plans depending on the user's needs. Labels were created in TXT format to be compatible with YOLOv10.

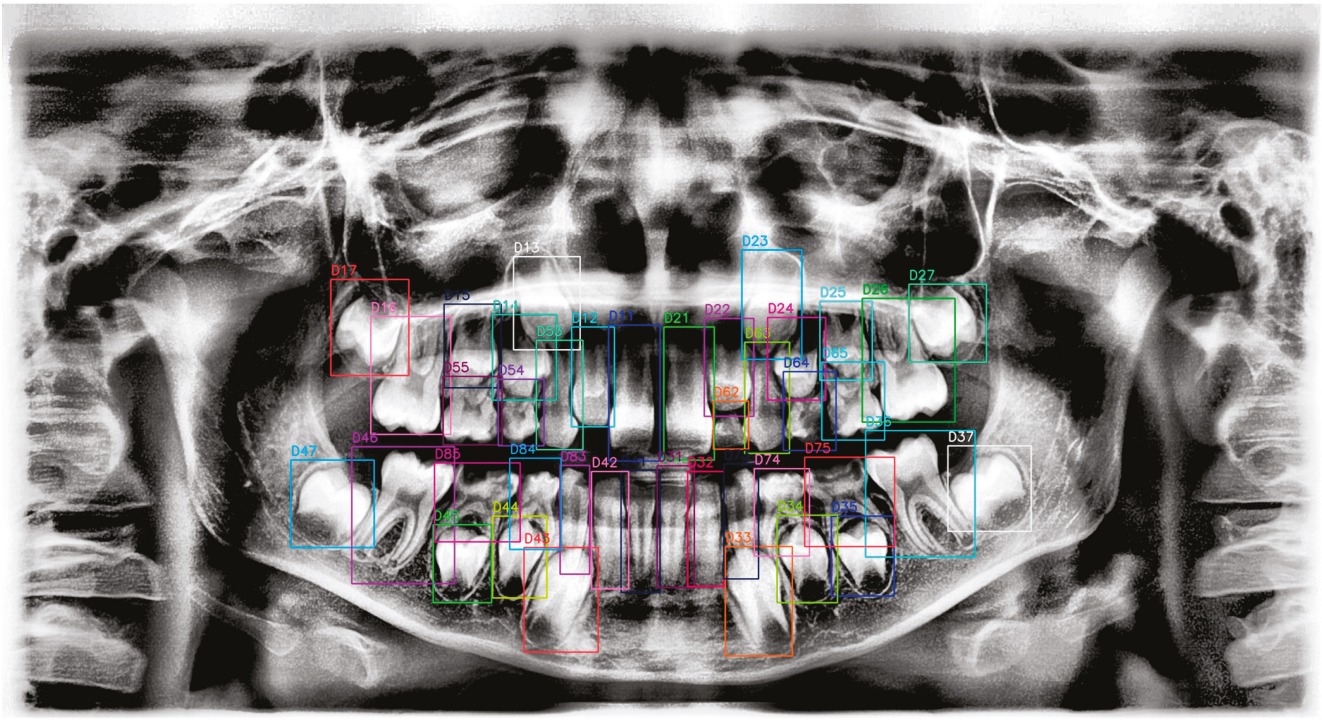


Figure 2. Panoramic radiograph with tooth annotations by R.B.P. for YOLOv10 training. This figure displays a panoramic radiograph where each tooth has been manually annotated with labels and bounding boxes by R.B.P. for the purpose of training the YOLOv10 model. The annotations include distinct colors and identifiers for each tooth, providing a comprehensive dataset for model training and ensuring accurate tooth localization and classification. This manually labeled image serves as ground truth data to enhance the YOLOv10 model's performance in tooth detection tasks.

Before labeling, the names, genders, and birth dates of the patients were removed from the PRs to ensure full anonymity. The numbering of primary and permanent teeth in the labeling process was carried out according to the Fédération Dentaire Internationale (FDI) tooth numbering system.

2.5. YOLOv10

YOLOv10 is an advanced version of the YOLO series of real-time object detection models designed to enhance both accuracy and efficiency. Building on the strengths of its predecessors, it incorporates innovative features that make it particularly effective for edge applications in which computational resources are limited. Previous YOLO models often involved steps like NMS, which could introduce delays and make seamless integration challenging. YOLOv10 addresses these issues by offering a training strategy that focuses on efficiency and accuracy and does not require NMS.

By using consistent dual-assignment methods, YOLOv10 enables training without NMS, achieving both competitive performance and low latency. Additionally, it proposes a comprehensive design strategy that optimizes the various components of YOLO for both efficiency and accuracy. This reduces computational load and enhances performance. YOLOv10 is offered in multiple versions to meet users' different computational requirements and performance needs. These versions vary in their number of parameters, allowing users to select the model suited to their hardware and specific application scenarios.

In this study, the open-source and freely available Extra Large (x) version of YOLOv10 was used to achieve the highest possible accuracy for non-commercial purposes.

2.5.1. Analysis of the YOLOv10 Architecture

The YOLOv10 architecture introduces several significant innovations by building on the strengths of previous versions. The main components of the model are as follows:

Backbone: The backbone used for feature extraction in YOLOv10 is equipped with an enhanced version of Cross Stage Partial Network (CSPNet). This structure improves gradient flow and reduces computational load.

Neck: The neck is designed to aggregate features at different scales and transfer them to the head. It uses Path Aggregation Network (PAN) layers to effectively combine these features.

One-to-Many Head: During training, it generates multiple predictions for each object, providing more supervision signals and increasing learning accuracy.

One-to-One Head: During the inference stage, it provides a single optimal prediction for each object. This approach eliminates the need for NMS, reducing latency and enhancing efficiency.

2.5.2. Key Features of YOLOv10

NMS-free training: YOLOv10 eliminates the need for NMS by using consistent dual-assignment methods. This approach reduces latency during inference and provides faster results.

Holistic model design: The model is comprehensively optimized for both efficiency and accuracy. These optimizations include elements such as lightweight classification heads, spatial-channel separated downsampling, and rank-guided block design.

Enhanced model capabilities: YOLOv10 integrates large-kernel convolutions and partial self-attention modules. These features enhance performance without significantly increasing computational costs.

2.6. Preprocessing

Various image preprocessing techniques were applied to enhance data quality and optimize model performance. The preprocessing tasks described in the manuscript, such as grayscale image conversion, image cropping, normalization, and resizing, were performed using automated scripts within the software environment. These tasks were implemented programmatically to minimize manual intervention and ensure consistency across the dataset. The operator's role in the preprocessing stage was limited to verifying the quality of the input images and ensuring that they met the inclusion criteria. By automating most of the preprocessing steps, the workflow was designed to save time and reduce human effort compared to fully manual preprocessing methods. This balance between automation and human oversight ensured efficiency and consistency throughout the preprocessing workflow. The detailed implementation of these techniques is described below:

- *Grayscale image loading:* Although panoramic radiographs are typically acquired in grayscale, all images were explicitly converted to a single-channel grayscale format during preprocessing. This step was undertaken to ensure uniformity across the dataset by eliminating potential inconsistencies in image format or metadata, which might arise due to differences in file handling or storage protocols.
- *Image cropping:* Specific regions of the images were cropped to focus solely on important areas. This approach eliminated distracting elements, allowing the model to work with only the necessary data during the training process.
- *Contrast enhancement:* The contrast of the images was increased to make the details more prominent. This process enabled the model to learn important features better in low-contrast images.
- *Normalization:* Pixel values were normalized to the (0, 1) range. Normalization helped achieve more consistent results by reducing the impact of brightness and contrast variations across images.
- *Image resizing:* Images were resized to 640 $\times$ 640 pixels. This standardized sizing ensured consistency in the model's inputs and eliminated data mismatches during the training process.

2.7. Dataset Splitting

The dataset, which consisted of 200 images (with a total of 8153 labels) used to train and evaluate the YOLOv10 DLM, was split in the following manner:

- *Training set:* 70% of the total dataset, comprising 140 images, was used to train the YOLOv10 DLM. This portion was allocated to allow the model to learn various object features and improve its overall performance.
- *Test set:* 15% of the dataset, consisting of 30 images, was set aside for independent testing of the model. This portion was used to assess the model's performance on real-world data and evaluate its accuracy.
- *Validation set:* The validation set, comprising 15% of the dataset, was used to fine-tune the model's hyperparameters and monitor performance. During training, validation metrics such as precision, recall, and loss values were evaluated at each epoch to assess the model's ability to generalize to unseen data. Early stopping was implemented to halt training if no improvement was observed in validation loss for 10 consecutive epochs, thereby minimizing the risk of overfitting. The validation set was independent of both the training and test datasets, ensuring unbiased evaluation of the model.

Data splitting ensured a balanced and effective evaluation of YOLOv10 during the training, testing, and validation processes, allowing for accurate measurement and optimization of the model's overall performance.

2.8. YOLOv10 Model Training

YOLOv10 stands out as one of the most advanced and powerful object detection models. In this study, YOLOv10 was trained on an Nvidia RTX 4070 Ti GPU, and the parameters used during the training process were examined.

2.8.1. Training Process and Parameters

- *model:* Set to 'None', meaning that no specific model file was designated. In this case, a default model configuration file or a pretrained model was not used.
- *data:* Specified as 'None', indicating that no configuration file for dataset to be used in training was defined.

2.8.2. Training Duration and Early Stopping

- *epochs:* Training was set to a total of 200 epochs, meaning the model was trained on the 70% training subset 200 times. This iterative process allowed the model to optimize its parameters by repeatedly learning from the training data.
- *time:* Set to 'None', indicating that no maximum training time was restricted. Training was completed based on the specified number of epochs.
- *patience:* Early stopping was applied if no improvement was observed for 10 consecutive epochs. This aimed to prevent the model from overfitting.

2.8.3. Batch and Image Size

- *batch:* The batch size used during training was set to 16. This represents the number of samples processed by the model at each step, which affects processing speed.
- *imgsz:* The image size was set to 640 pixels, meaning that all training images would be resized to this dimension.

2.8.4. Saving and Caching

- *save:* Set to 'True', indicating that the model would be saved during training.
- *save period:* Set to '−1', meaning the model would not be saved at the end of each epoch.
- *cache:* Set to 'False', indicating that the dataset would not be cached in memory.

2.8.5. Optimization and Other Parameters

- *optimizer:* Set to 'Auto', meaning that the most suitable optimization method was automatically selected based on model performance.
- *cos_lr:* Set to 'False', meaning that a cosine curve scheduler was not applied to the learning rate.
- *amp:* Set to 'True', indicating that Automatic Mixed Precision training was enabled to enhance computational efficiency.
- *fraction:* The entire dataset (1.0) was used for training.
- *freeze:* Set to 'None', meaning that no layers of the model were frozen, and all layers were trained.

Parameters such as training duration, batch size, and image size were selected to optimize the model's performance and accuracy.

YOLOv10 was trained on a dataset with 52 classes. The training process lasted for a total of 200 epochs to maximize model performance. An early stopping strategy was applied to prevent overfitting and achieve optimal results. The early stopping criterion was set to terminate training if no improvement was observed within the last 10 epochs.

Throughout the training process, the model was carefully monitored, and the best results were achieved at epoch 71. After this epoch, no further improvements in model

performance were observed, leading to an automatic termination of the training process. The model with the best performance, recorded at epoch 71, was saved as "best.pt" with the corresponding weights. This approach optimized model performance while enhancing the efficiency of the training process and minimizing the risk of overfitting. The final results demonstrate that YOLOv10 performed effectively across a wide range of classes and that the early stopping strategy contributed significantly to the training process.

2.9. Metrics of Model Performance

The performance of the YOLOv10 for tooth detection and numbering was evaluated using the following metrics: precision, recall, mean Average Precision at Intersection over Union (IoU) threshold of 0.5 (mAP50), mean Average Precision across IoU thresholds from 0.5 to 0.95 (mAP50-95), and F1 score. Precision measures the proportion of correctly identified teeth among all detections, while recall indicates the proportion of true positives identified. mAP50 and mAP50-95 provide comprehensive measures of detection accuracy across varying thresholds, and the F1 score represents the harmonic mean of precision and recall, offering a balanced evaluation of model performance.

Various metrics were used to comprehensively evaluate model performance. During the training process, train/box_loss, train/cls_loss, and train/dfl_loss losses were analyzed. These losses were used to measure the accuracy of the model's box predictions, class predictions, and detailed feature predictions, respectively.

The model's prediction accuracy and overall performance were evaluated using metrics such as metrics/precision, metrics/recall, metrics/mAP50, and metrics/mAP50-95. Precision and recall assess the accuracy and coverage of the model's positive predictions, while mAP50 and mAP50-95 measure its overall success rate across various threshold values.

During the validation phase, val/box_loss, val/cls_loss, and val/dfl_loss values were examined. These validation losses were used to assess the model's performance on real-world data outside of the training set.

To optimize the training process, learning rates lr/pg0, lr/pg1, and lr/pg2 were set. By adjusting learning rates for different parameter groups, the aim was to improve the speed and efficiency of the model's learning process.

To analyze the model's performance more comprehensively, the F1 score and confusion matrix were utilized. The F1 score provided the harmonic mean of precision and recall, while the confusion matrix helped visualize the model's correct and incorrect predictions in detail.

3. Results

The weighted average performance metrics of the YOLOv10 DLM for tooth detection and numbering—precision, recall, mAP50, mAP50-95, and F1 score—were 0.90, 0.94, 0.968, 0.696, and 0.919, respectively (Table 1, Figures 3–9). These results were obtained exclusively on the test set, meaning a dataset completely independent of the training data was used to evaluate the model's true performance.

The mAP50 for tooth 51 was found to be 0.712, significantly lower than the mAP50 values of other teeth, all of which exceeded 0.9 (Table 1). Further analysis revealed that the YOLOv10 model incorrectly predicted tooth 51 in certain cases, as illustrated in Figure 10. In the left panoramic radiograph in Figure 10, the ground truth labels are displayed, while in the right panoramic radiograph, the model's predictions are shown. While the model demonstrated strong performance for most teeth, the misclassification of tooth 51 highlights a limitation in its ability to detect this specific tooth accurately.

Table 1. Precision, recall, mAP50, mAP50-95, and F1 scores by tooth number.

Tooth No.	Precision (P)	Recall (R)	mAP50	mAP50-95	F1 Score
11	0.964	1.000	0.995	0.779	0.982
12	0.995	0.897	0.977	0.673	0.943
13	0.908	1.000	0.993	0.783	0.951
14	0.966	0.958	0.992	0.733	0.962
15	0.962	1.000	0.995	0.731	0.980
16	0.839	1.000	0.986	0.840	0.913
17	0.979	1.000	0.995	0.822	0.989
18	0.793	1.000	0.978	0.752	0.885
21	0.967	0.933	0.990	0.736	0.950
22	0.993	0.933	0.990	0.744	0.962
23	0.959	0.967	0.993	0.767	0.963
24	0.967	0.974	0.994	0.720	0.971
25	0.966	0.947	0.987	0.687	0.956
26	0.829	0.967	0.980	0.856	0.893
27	0.907	0.978	0.988	0.790	0.941
28	1.000	0.938	0.995	0.726	0.968
31	0.936	0.970	0.993	0.760	0.953
32	0.977	1.000	0.995	0.781	0.988
33	0.994	0.900	0.984	0.812	0.944
34	1.000	0.938	0.995	0.769	0.968
35	0.965	0.921	0.989	0.800	0.943
36	0.880	0.975	0.991	0.889	0.925
37	0.786	1.000	0.968	0.804	0.880
38	0.967	0.900	0.908	0.652	0.932
41	0.932	1.000	0.983	0.729	0.965
42	0.986	0.933	0.986	0.722	0.959
43	0.871	1.000	0.989	0.785	0.931
44	0.960	1.000	0.993	0.816	0.980
45	0.898	0.967	0.989	0.779	0.931
46	0.916	1.000	0.974	0.854	0.956
47	0.947	1.000	0.972	0.808	0.973
48	0.929	0.900	0.905	0.580	0.914
51	0.656	0.800	0.712	0.337	0.721
52	0.681	0.715	0.835	0.502	0.697
53	0.907	0.926	0.948	0.635	0.916
54	0.914	0.972	0.992	0.692	0.942
55	0.963	0.988	0.959	0.719	0.975
61	0.816	0.898	0.962	0.400	0.855
62	0.911	1.000	0.995	0.588	0.953
63	0.922	0.912	0.969	0.628	0.917
64	0.946	0.921	0.985	0.714	0.933
65	0.805	0.868	0.908	0.668	0.835
71	0.660	1.000	0.995	0.0995	0.795
72	1.000	0.675	0.984	0.492	0.806
73	0.941	0.960	0.972	0.705	0.950
74	0.799	0.865	0.940	0.696	0.830
75	0.975	0.963	0.991	0.770	0.969
81	0.450	1.000	0.995	0.597	0.621
82	0.871	0.751	0.926	0.444	0.807
83	0.892	0.760	0.900	0.621	0.821
84	0.886	1.000	0.965	0.731	0.940
85	0.863	1.000	0.936	0.688	0.926

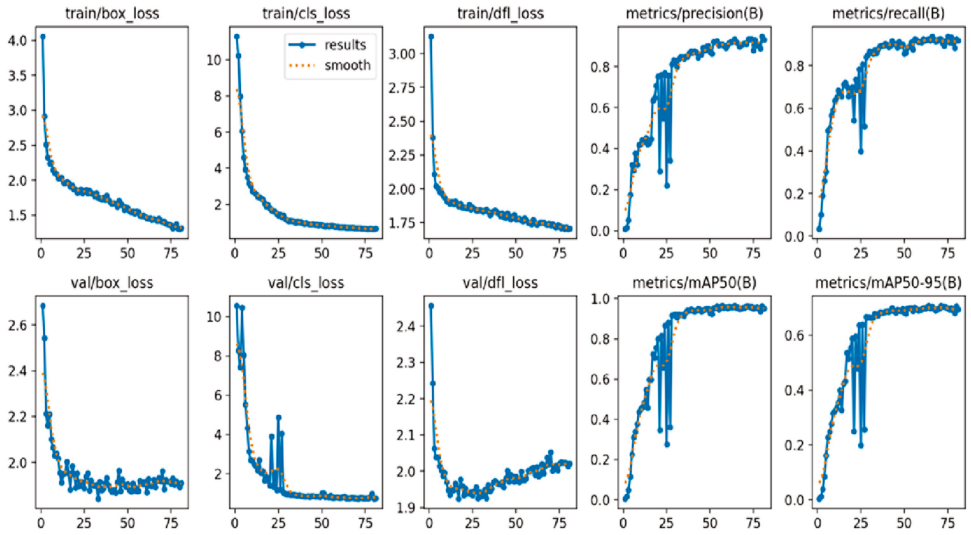


Figure 3. Training and validation metrics for YOLOv10. This figure displays the training and validation loss metrics, including box loss, classification loss, and distribution focal loss, alongside performance metrics of precision, recall, mAP50, and mAP50-95 across 75 epochs. The top row represents training metrics, while the bottom row shows validation metrics. The solid blue line indicates raw results, and the dotted orange line represents the smoothed trend, helping to visualize the model's performance improvement over the training process.

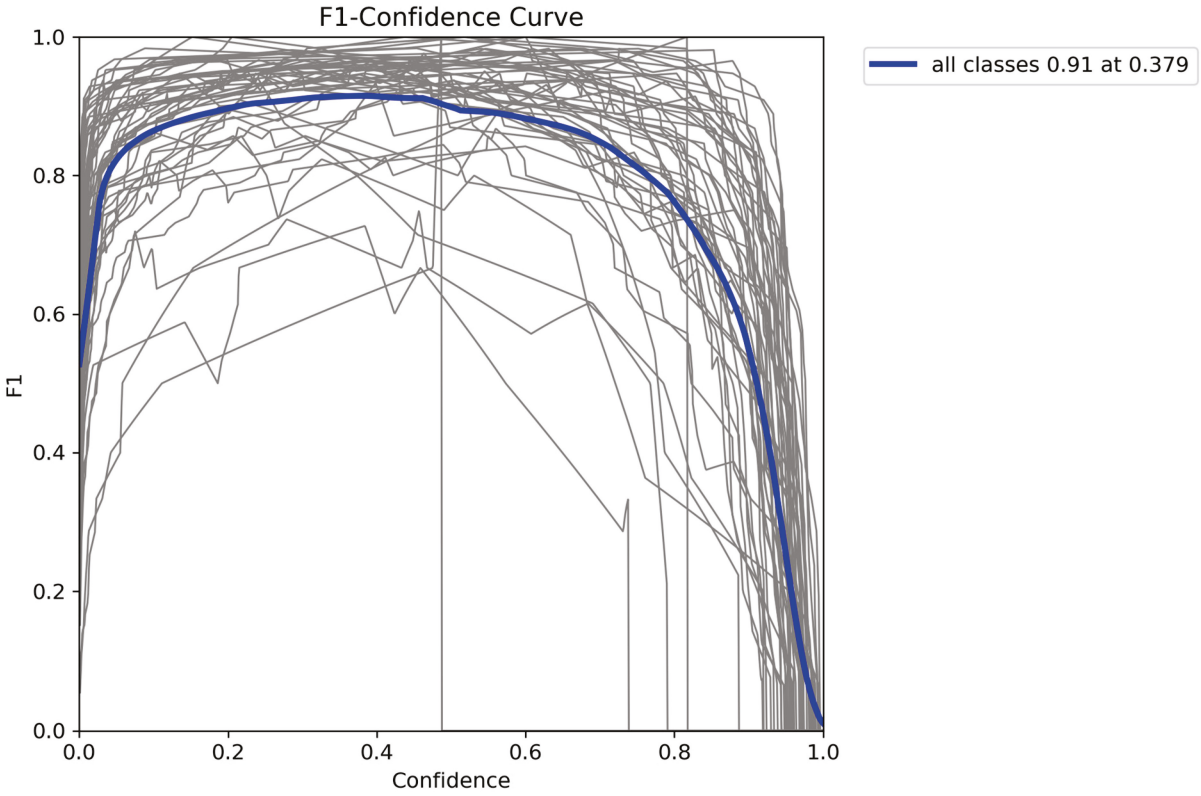


Figure 4. F1-confidence curve for YOLOv10. This figure illustrates the F1-Confidence curve for the YOLOv10 across all classes. The blue line represents the average F1 score across different confidence thresholds, peaking at an F1 score of 0.91 at a confidence level of 0.379. The gray lines depict F1 scores for individual classes, showing the model's performance variance across classes. This curve provides insights into the optimal confidence threshold that maximizes the F1 score, indicating the balance between precision and recall for effective tooth detection and classification.

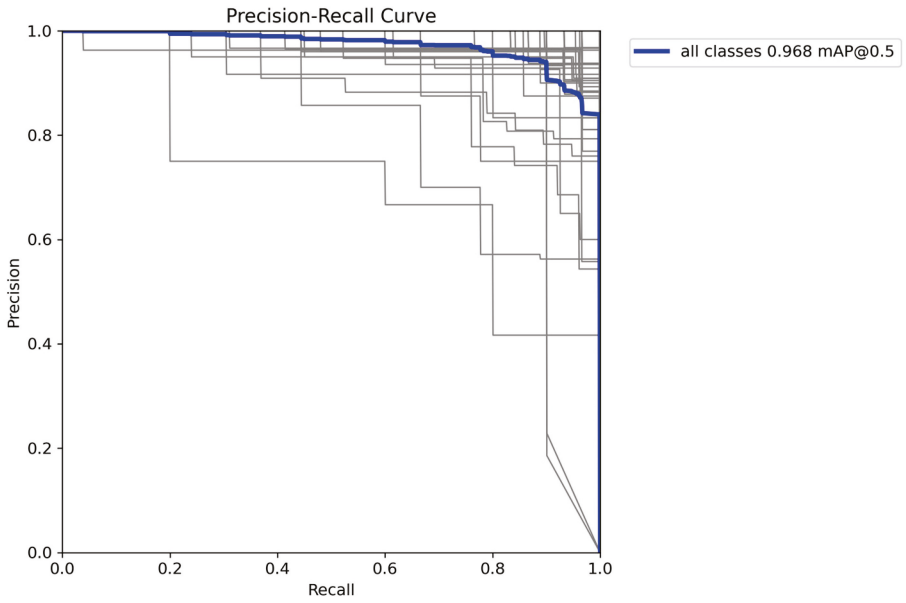


Figure 5. Precision-recall curve for YOLOv10. This figure shows the precision-recall curve for the YOLOv10 across all classes, with an overall mAP (mean Average Precision) of 0.968 at an IoU (Intersection over Union) threshold of 0.5. The blue line represents the average performance across all classes, while the gray lines illustrate the precision-recall relationship for individual classes. This curve highlights the model's ability to maintain high precision as recall increases, indicating strong detection accuracy for the tooth classification task.

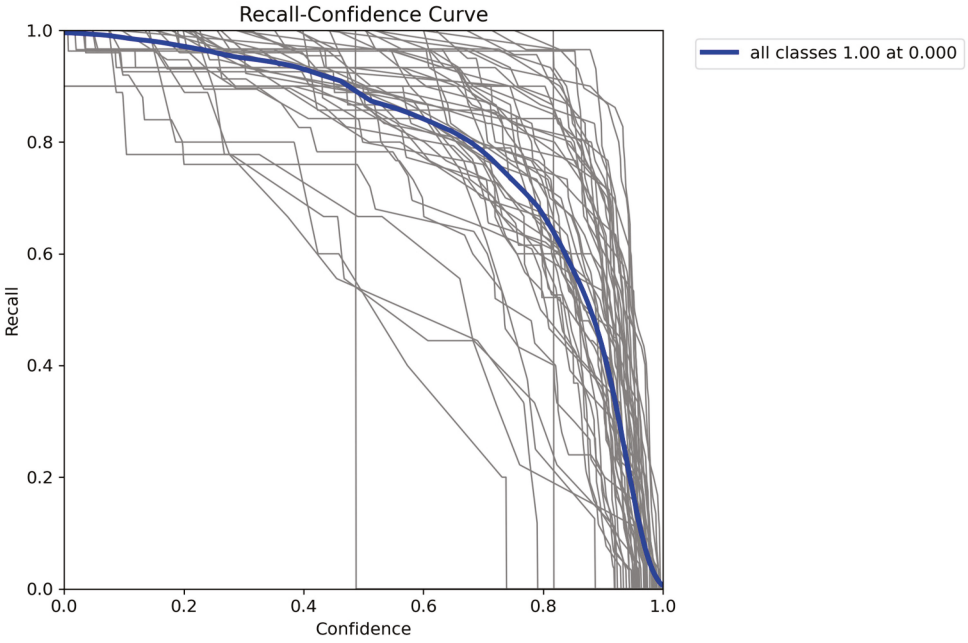


Figure 6. Recall-confidence curve for YOLOv10. This figure shows the recall-confidence curve for the YOLOv10 across all classes. The blue line represents the average recall at different confidence thresholds, achieving a maximum recall of 1.00 at a confidence level of 0.000. The gray lines depict the recall for individual classes, illustrating the model's behavior as confidence levels increase. This curve is useful for analyzing how changes in confidence thresholds impact the model's recall, providing insights into its detection capabilities and threshold selection for optimal performance.

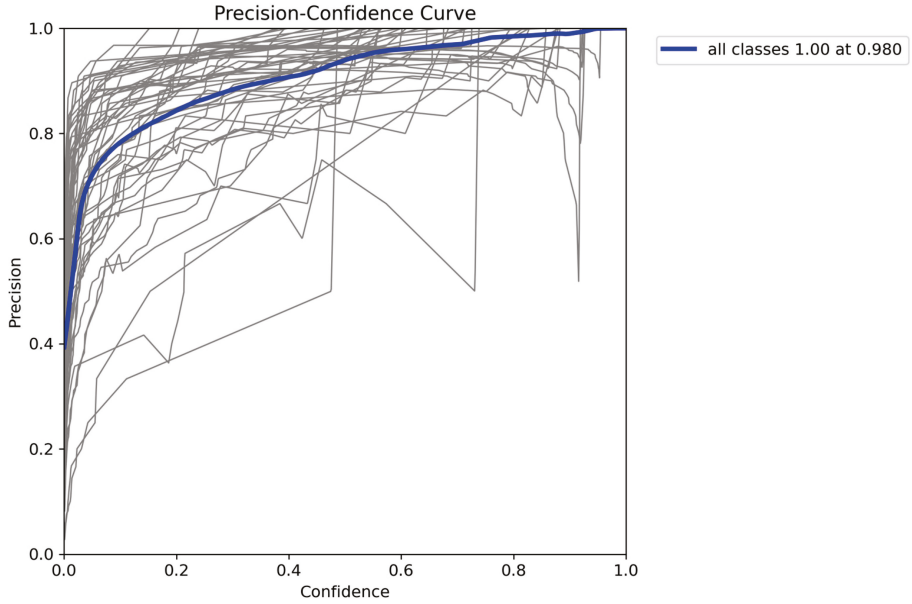


Figure 7. Precision-Confidence Curve for YOLOv10 Model. This figure illustrates the precision-confidence curve for the YOLOv10 across all classes. The blue line represents the average precision at different confidence thresholds, reaching a maximum precision of 1.00 at a confidence level of 0.980. The gray lines show precision values for individual classes, highlighting variations in performance across different confidence levels. This curve helps in understanding how adjusting the confidence threshold impacts precision, providing insights into optimizing the model's performance for tooth detection and classification tasks.

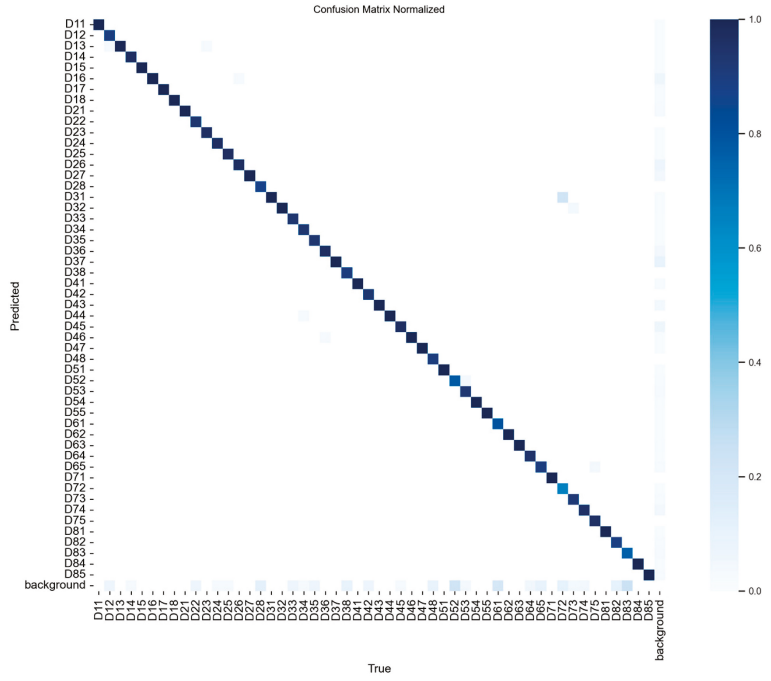


Figure 8. Normalized confusion matrix for tooth detection and classification with YOLOv10. This figure presents the normalized confusion matrix for the YOLOv10's tooth detection and classification performance across various tooth classes. Each cell represents the model's prediction accuracy for a specific class, with diagonal cells indicating correct predictions and off-diagonal cells showing misclassifications. The color intensity corresponds to the proportion of predictions, with darker shades indicating higher accuracy. The background class is also included, highlighting the model's ability to distinguish teeth from non-tooth regions. This matrix provides insight into the model's classification accuracy and error patterns across different tooth types.

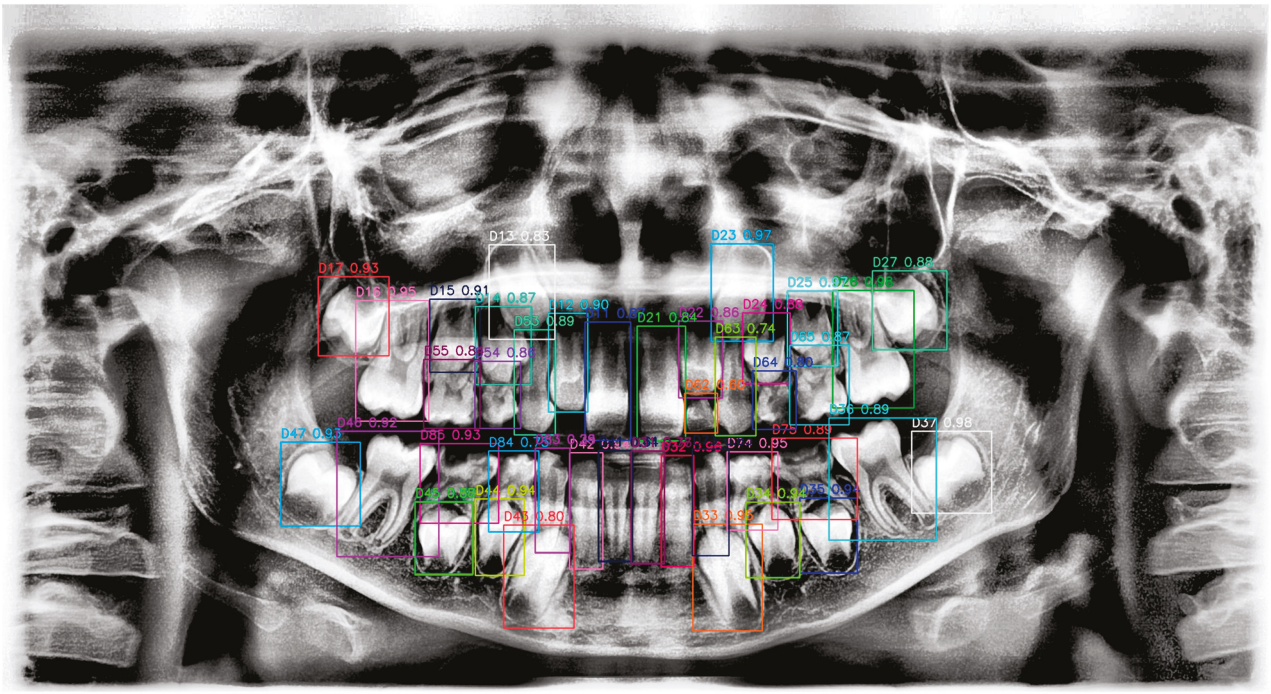


Figure 9. Detected teeth with confidence scores on panoramic radiograph using YOLOv10. This figure shows a panoramic radiograph with bounding boxes indicating the detected tooth regions, labeled by tooth identifiers and confidence scores. Each bounding box represents the YOLOv10's prediction, with different colors used to distinguish individual teeth. The confidence score for each detection is displayed alongside the corresponding tooth label, indicating the model's certainty in identifying each tooth. This visualization demonstrates the model's capability in accurately localizing and classifying teeth within a complex radiographic image.

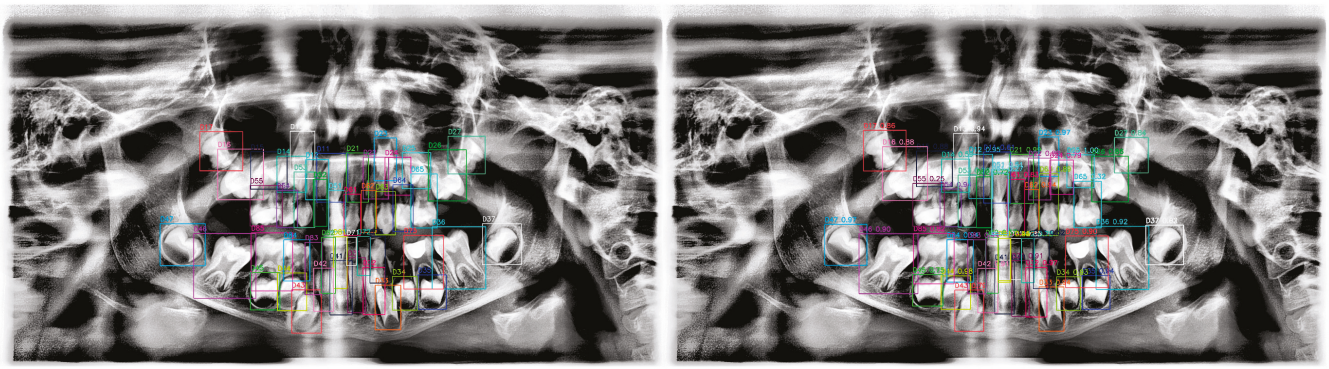


Figure 10. (**Left**) Ground truth labels for the panoramic radiograph. (**Right**) Predictions by the YOLOv10 model. The misclassification of tooth 51 is indicated. Each bounding box represents YOLOv10's predictions, with different colors distinguishing individual teeth. Confidence scores indicate the model's certainty for each detection.

4. Discussion

With the advancement of AI and DL technologies, there is growing interest in integrating these technologies into dentistry. AI and DLMs play a diagnostic and supportive role by enhancing diagnostic accuracy through the analysis of radiographic images. This positively impacts the success of treatments and long-term prognosis [15]. Tooth detection and numbering, which are among the initial steps of accurate diagnosis and treatment, can be a time-consuming and error-prone process when performed manually. Conducting these procedures with the assistance of DLMs not only reduces the clinical workload but also minimizes potential errors.

Numerous studies have utilized various imaging techniques for tooth detection and numbering [5,14,16,17]. Although studies have demonstrated the effectiveness of DLMs [18], they have primarily focused on permanent teeth, with significantly fewer studies targeting mixed dentition.

Ahn et al. compared different DLMs for detecting mesiodens in the PRs of children in the primary or mixed dentition stages and reported that they accurately detected the actual locations of mesiodens in many cases [19]. Similarly, a study examining the detection of supernumerary teeth using CNNs found that they yielded positive results in the early mixed dentition period [20]. Recent studies using earlier versions of YOLOv10, such as YOLOv3 and YOLOv4, have also shown that these DLMs are highly successful in detecting both mesiodens and permanent teeth in PR images obtained from pediatric patients [21,22].

Despite these advancements, research focusing specifically on tooth detection and numbering in pediatric PRs remains limited. Kaya et al. examined YOLOv4's performance in detecting and numbering primary and permanent teeth in pediatric PRs and reported an mAP of 92.22%, mAR of 94.44%, and weighted F1 score of 0.91, noting that this model could be used as a high-accuracy, fast method for automated tooth detection and numbering [23]. Similarly, Beşer et al. reported precision, recall, F1 score, and mAP-0.5 values all exceeding 0.98 using YOLOv5 for the same purpose [5]. These findings underline the potential of YOLO models in automating tooth detection with high accuracy. In comparison, our study using YOLOv10 achieved a weighted F1 score of 91%, comparable to earlier YOLO models while demonstrating enhanced efficiency and fewer parameters in its architecture. This improvement in efficiency without compromising accuracy makes YOLOv10 particularly well-suited for clinical applications.

In addition to tooth detection and numbering, YOLOv10 can be adapted for other clinical applications such as identifying supernumerary teeth, impacted teeth, or pathologies, making it a versatile tool in pediatric and general dentistry. Furthermore, it can assist in orthodontic treatment planning by monitoring tooth development and positioning. In trauma cases, the model could be used to detect missing or fractured teeth rapidly, aiding emergency dentistry workflows. Such versatility highlights its potential to address a variety of diagnostic and treatment challenges in dental practice.

Other studies have also highlighted the strong performance of different DLM architectures in tooth detection tasks. For instance, Putra et al. used YOLOv4 for automated permanent tooth numbering and reported an F1 score of 93.44%, demonstrating its high reliability and faster detection times compared to manual methods [24]. Similarly, Bilgir et al. employed Faster R-CNN and achieved a sensitivity of 95.59% and an F1 score of 96.06% for tooth numbering, confirming the robustness of DLMs for automated dental diagnostics [25]. Our findings align with these studies, reaffirming that YOLOv10 can perform comparably or better in mixed dentition cases, where the anatomical complexity poses additional challenges.

In our study, we aimed to build on this foundation by evaluating the YOLOv10 model's performance. The weighted F1 score of 91% achieved in our study confirms the model's effectiveness in detecting and numbering both primary and permanent teeth in pediatric patients. YOLOv10's novel architecture, utilizing fewer parameters, enables faster detection and numbering compared to earlier versions, making it a practical choice for clinical applications [11].

However, certain limitations should be acknowledged. The study was conducted at a single center, using PRs obtained exclusively from patients in the Thrace region. While the uniform imaging parameters ensured consistency, the limited population size constrained the sample size, which included only PRs with at least 10 primary teeth and underlying permanent tooth germs to ensure accuracy. This selective approach, while enhancing data

reliability, reduced the dataset's diversity. Furthermore, the emphasis on the F1 score over accuracy was necessary to address the imbalance in the dataset, ensuring a more robust evaluation.

Using data from multiple imaging machines could address these limitations and significantly enhance the model's generalizability. Different imaging devices often vary in parameters such as resolution, contrast, and exposure settings, which could initially introduce variability in the model's performance. However, training the model on datasets obtained from multiple machines would expose it to a broader range of imaging conditions and artifacts, improving its robustness and adaptability. Such an approach would make the model more suitable for deployment in diverse clinical environments, where imaging equipment and protocols differ significantly. While this could add complexity to preprocessing workflows, it would ultimately enhance the model's applicability and usability in real-world scenarios.

Additionally, ensuring global accessibility and ease of use is crucial for the adoption of YOLOv10 in clinical settings. Open-source availability facilitates integration into workflows, while user-friendly interfaces with graphical tools can minimize the need for technical expertise. Cloud-based solutions could further enhance accessibility by eliminating hardware requirements, particularly in resource-limited settings. Providing comprehensive documentation, tutorials, and support resources would also empower practitioners to implement and utilize the model effectively. Future efforts should focus on these aspects to ensure seamless integration into diverse clinical practices.

Future research should focus on expanding the sample size and incorporating data from multiple centers with diverse populations. Such efforts would enhance the generalizability of the findings and further validate YOLOv10's potential as a reliable tool for automated tooth detection and numbering in pediatric dentistry.

5. Conclusions

The results of our study demonstrate that YOLOv10 can be effectively used for tooth detection and numbering. Although it uses fewer parameters than previous versions, YOLOv10 shows comparable or even higher accuracy in performance. The application of YOLOv10 in tooth detection and numbering is expected to prevent unnecessary time loss for clinicians. We believe that this study will serve as a foundation for future research with larger sample sizes and segmentation analyses involving YOLOv10.

Author Contributions: Conceptualization, R.B.P. and C.O.K.; methodology, R.B.P.; software, C.O.K.; validation, R.B.P. and C.O.K.; formal analysis, C.O.K.; investigation, R.B.P.; resources, R.B.P.; data curation, R.B.P.; writing—original draft preparation, R.B.P. and C.O.K.; writing—review and editing, R.B.P. and C.O.K. 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 of Trakya University Faculty of Medicine (protocol code TÜTF-GOBEAK 2024/450, approval date 04.11.2024).

Informed Consent Statement: Informed consent was waived because the study involved a retrospective analysis of previously collected and de-identified panoramic radiographs. These radiographs were anonymized prior to analysis, and no identifiable patient information was included. As the data were entirely de-identified, individual consent to participate or for publication was not required, in accordance with the ethical standards and guidelines.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to patient confidentiality and ethical restrictions.

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

Abbreviations

DLM	Deep learning model
PR	Panoramic radiograph
AI	Artificial intelligence
ML	Machine learning
YOLO	You Only Look Once
DL	Deep learning
CNN	Convolutional neural network
CLAIM	Checklist for Artificial Intelligence in Medical Imaging
STARD	Standards for the Reporting of Diagnostic Accuracy Studies
FDI	Fédération Dentaire Internationale
NMS	Non-maximum suppression
CSPNet	Cross Stage Partial Network
PAN	Path Aggregation Network
IoU	Intersection over Union

References

1. Shah, N.; Bansal, N.; Logani, A. Recent advances in imaging technologies in dentistry. *World J. Radiol.* **2014**, *6*, 794–807. [CrossRef]
2. Angelopoulos, C.; Bedard, A.; Katz, J.O.; Karamanis, S.; Parissis, N. Digital panoramic radiography: An overview. *Semin. Orthod.* **2004**, *10*, 194–203. [CrossRef]
3. Bumann, E.E.; Al-Qarni, S.; Chandrashekar, G.; Sabzian, R.; Bohaty, B.; Lee, Y. A novel collaborative learning model for mixed dentition and fillings segmentation in panoramic radiographs. *J. Dent.* **2024**, *140*, 104779. [CrossRef]
4. Jacobs, R.; Quirynen, M. Dental cone beam computed tomography: Justification for use in planning oral implant placement. *Periodontology 2000* **2014**, *66*, 203–213. [CrossRef] [PubMed]
5. Beser, B.; Reis, T.; Berber, M.N.; Topaloglu, E.; Gungor, E.; Kılıc, M.C.; Duman, S.; Çelik, Ö.; Kuran, A.; Bayrakdar, I.S. YOLO-V5 based deep learning approach for tooth detection and segmentation on pediatric panoramic radiographs in mixed dentition. *BMC Med. Imaging* **2024**, *24*, 172. [CrossRef]
6. Leite, A.F.; Gerven, A.V.; Willems, H.; Beznik, T.; Lahoud, P.; Gaêta-Araujo, H.; Vranckx, M.; Jacobs, R. Artificial intelligence-driven novel tool for tooth detection and segmentation on panoramic radiographs. *Clin. Oral Investig.* **2021**, *25*, 2257–2267. [CrossRef] [PubMed]
7. Bağ, İ.; Bilgir, E.; Bayrakdar, I.; Baydar, O.; Atak, F.M.; Çelik, Ö.; Orhan, K. An artificial intelligence study: Automatic description of anatomic landmarks on panoramic radiographs in the pediatric population. *BMC Oral Health* **2023**, *23*, 764. [CrossRef] [PubMed]
8. Heo, M.-S.; Kim, J.-E.; Hwang, J.-J.; Han, S.-S.; Kim, J.-S.; Yi, W.-J.; Park, I.-W. Artificial intelligence in oral and maxillofacial radiology: What is currently possible? *Dentomaxillofacial Radiol.* **2021**, *50*, 20200375. [CrossRef] [PubMed]
9. Kong, H.-J.; Yoo, J.-Y.; Lee, J.-H.; Eom, S.-H.; Kim, J.-H. Performance evaluation of deep learning models for the classification and identification of dental implants. *J. Prosthet. Dent.* **2023**. [CrossRef] [PubMed]
10. Bonfanti-Gris, M.; Herrera, A.; Paraíso-Medina, S.; Alonso-Calvo, R.; Martínez-Rus, F.; Pradíes, G. Performance evaluation of three versions of a convolutional neural network for object detection and segmentation using a multiclass and reduced panoramic radiograph dataset. *J. Dent.* **2024**, *144*, 104891. [CrossRef]
11. Hussain, M.; Khanam, R. In-Depth Review of YOLOv1 to YOLOv10 Variants for Enhanced Photovoltaic Defect Detection. *Solar* **2024**, *4*, 351–386. [CrossRef]
12. Public Health England (PHE); Faculty of General Dental Practice (UK) (FGDP[UK]). *Guidance Notes for Dental Practitioners on the Safe Use of X-Ray Equipment*, 2nd ed.; PHE and FGDP(UK): London, UK, 2020.
13. Lee, J.-H.; Han, S.-S.; Kim, Y.H.; Lee, C.; Kim, I. Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **2020**, *129*, 635–642. [CrossRef]
14. Tuzoff, D.V.; Tuzova, L.N.; Bornstein, M.M.; Krasnov, A.S.; Kharchenko, M.A.; Nikolenko, S.I.; Sveshnikov, M.M.; Bednenko, G.B. Tooth detection and numbering in panoramic radiographs using convolutional neural networks. *Dentomaxillofac. Radiol.* **2019**, *48*, 20180051. [CrossRef]
15. Ong, S.-H.; Kim, H.; Song, J.-S.; Shin, T.J.; Hyun, H.-K.; Jang, K.-T.; Kim, Y.-J. Fully automated deep learning approach to dental development assessment in panoramic radiographs. *BMC Oral Health* **2024**, *24*, 426. [CrossRef]

16. Görürgöz, C.; Orhan, K.; Bayrakdar, I.S.; Çelik, Ö.; Bilgir, E.; Odabaş, A.; Aslan, A.F.; Jagtap, R. Performance of a convolutional neural network algorithm for tooth detection and numbering on periapical radiographs. *Dentomaxillofacial Radiol.* **2022**, *51*, 20210246. [CrossRef]
17. Jang, T.J.; Kim, K.C.; Cho, H.C.; Seo, J.K. A Fully Automated Method for 3D Individual Tooth Identification and Segmentation in Dental CBCT. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 6562–6568. [CrossRef]
18. Sadr, S.; Rokhshad, R.; Daghighi, Y.; Golkar, M.; Kheybari, F.T.; Gorjinejad, F.; Kojori, A.M.; Rahimirad, P.; Shobeiri, P.; Mahdian, M.; et al. Deep learning for tooth identification and numbering on dental radiography: A systematic review and meta-analysis. *Dentomaxillofacial Radiol.* **2023**, *53*, 5–21. [CrossRef]
19. Ahn, Y.; Hwang, J.J.; Jung, Y.-H.; Jeong, T.; Shin, J. Automated Mesiodens Classification System Using Deep Learning on Panoramic Radiographs of Children. *Diagnostics* **2021**, *11*, 1477. [CrossRef]
20. Mine, Y.; Iwamoto, Y.; Okazaki, S.; Nakamura, K.; Takeda, S.; Peng, T.; Mitsuhata, C.; Kakimoto, N.; Kozai, K.; Murayama, T. Detecting the presence of supernumerary teeth during the early mixed dentition stage using deep learning algorithms: A pilot study. *Int. J. Paediatr. Dent.* **2022**, *32*, 678–685. [CrossRef]
21. Ha, E.-G.; Jeon, K.J.; Kim, Y.H.; Kim, J.-Y.; Han, S.-S. Automatic detection of mesiodens on panoramic radiographs using artificial intelligence. *Sci. Rep.* **2021**, *11*, 23061. [CrossRef]
22. Kaya, E.; Gunec, H.G.; Aydin, K.C.; Urkmez, E.S.; Duranay, R.; Ates, H.F. A deep learning approach to permanent tooth germ detection on pediatric panoramic radiographs. *Imaging Sci. Dent.* **2022**, *52*, 275–281. [CrossRef]
23. Kaya, E.; Gunec, H.G.; Gokyay, S.S.; Kutal, S.; Gulum, S.; Ates, H.F. Proposing a CNN Method for Primary and Permanent Tooth Detection and Enumeration on Pediatric Dental Radiographs. *J. Clin. Pediatr. Dent.* **2022**, *46*, 293–298. [CrossRef]
24. Putra, R.H.; Astuti, E.R.; Putri, D.K.; Widiasri, M.; Laksanti, P.A.M.; Majidah, H.; Yoda, N. Automated permanent tooth detection and numbering on panoramic radiograph using a deep learning approach. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **2023**, *137*, 537–544. [CrossRef]
25. Bilgir, E.; Bayrakdar, I.Ş.; Çelik, Ö.; Orhan, K.; Akkoca, F.; Sağlam, H.; Odabaş, A.; Aslan, A.F.; Ozcetin, C.; Kıllı, M.; et al. An artificial ıntelligence approach to automatic tooth detection and numbering in panoramic radiographs. *BMC Med. Imaging* **2021**, *21*, 124. [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

Automated Detection and Differentiation of Stanford Type A and Type B Aortic Dissections in CTA Scans Using Deep Learning

Hung-Hsien Liu [1], Chun-Bi Chang [2], Yi-Sa Chen [1], Chang-Fu Kuo [3], Chun-Yu Lin [4,5], Cheng-Yu Ma [6,7,8,*] and Li-Jen Wang [2,*]

[1] Department of Medical Imaging and Intervention, New Taipei City Municipal Tucheng Hospital, New Taipei City 236043, Taiwan; jerryliu0502@gmail.com (H.-H.L.)
[2] Department of Medical Imaging and Intervention, Chang Gung Memorial Hospital, Taoyuan City 333423, Taiwan; cooler@cgmh.org.tw
[3] Center for Artificial Intelligence in Medicine, Chang Gung Memorial Hospital, Taoyuan City 333423, Taiwan; zandis@gmail.com
[4] Department of Medicine, College of Medicine, Chang Gung University, Taoyuan City 333323, Taiwan; b9002078@cgmh.org.tw
[5] Department of Cardiothoracic and Vascular Surgery, New Taipei City Municipal Tucheng Hospital, New Taipei City 236043, Taiwan
[6] Department of Artificial Intelligence, Chang Gung University, Taoyuan City 333323, Taiwan
[7] Artificial Intelligence Research Center, Chang Gung University, Taoyuan City 333323, Taiwan
[8] Division of Rheumatology, Allergy and Immunology, Chang Gung Memorial Hospital, Taoyuan City 333423, Taiwan
* Correspondence: cyma@cgu.edu.tw (C.-Y.M.); lijenwang0918@gmail.com (L.-J.W.)

Abstract: Background/Objectives: To develop and validate a model system using deep learning algorithms for the automatic detection of type A aortic dissection (AD), and differentiate it from normal and type B AD patients. Methods: In this retrospective study, a deep learning model is developed, based on aortic computed tomography angiography (CTA) scans of 498 patients using training, validation and test sets of 398, 50 and 50 patients, respectively. An independent test set of 316 patients is used to validate and evaluate its performance. Results: Our model comprises two components. The first one is an objection detection model, which can identify the aorta from CTA. The second one is a dissection classification model, which can automatically detect the presence of aortic dissection and determine its type based on Stanford classification. Overall, the sensitivity and specificity for Type A AD were 0.969 and 0.982, for Type B AD were 0.946 and 0.996 and for normal cases were 0.988 and 1.000, respectively. The average processing time per CTA scan was 7.9 $\pm$ 2.8 s. (mean $\pm$ standard deviation). Conclusions: This deep learning automatic model can accurately and quickly detect type A AD patients, and could serve as an imaging triage in an emergency setting and facilitate early intervention and surgery to decrease the mortality rates of type A AD patients.

Keywords: deep learning; aortic dissection; computed tomography

1. Introduction

Aortic dissection (AD) is a life-threatening medical condition, which is primarily due to a tear in the aortic intima, which creates a true and false lumen. The annual incidence of acute AD in the general population is estimated to range from 2.6 to 3.5 per 100,000 people [1,2]. AD can be divided into two categories according to the widely used Stanford classification. The involvement of the ascending aorta is classified as Stanford type A AD, while Stanford type B AD does not involve the ascending aorta. It is clinically

important to rapidly distinguish acute Standford type A AD, because it is a cardiovascular surgical emergency. Without surgical intervention, mortality rates are about 1 to 2 percent per hour after symptom onset [3]. The computed tomography angiography (CTA) of the aorta (aortic CTA) can provide detailed assessment of extension of the dissection, and differentiate type A and B AD and thereby guide therapy, which has been reported with a pooled sensitivity of 100% and a specificity of 98% in a meta-analysis [4]. Indeed, aortic CTA has become a crucial imaging tool in emergency patients presented with atypical chest pain and can be served as the first line in diagnosing dissection in today's medical practice. [5,6]. Nevertheless, despite its critical role, the report turnaround time had a wide range from 0.4 to 628.5 h [7]. The turnaround time could be long due to the complexities of image interpretation and the workload of radiologists, especially in a busy emergency department with a long waiting list. In emergency cases, such as type A AD, if left unnoticed, it could potentially have life-threatening consequences. This highlights the need for developing an automatic algorithm to incorporate clinical workflow, assisting in the early identification of AD.

Deep learning with convolutional neural networks has been shown to succeed in medical imaging segmentation tasks [8], and could be used to develop algorithms for object detections and classifications in images. Deep learning models have been developed to detect AD with a comparable diagnostic performance to radiologists, and showed the potential to support clinical practice [9–11]. Raj et al. created an algorithm that could detect AD in the abdomen [12]. In contrast to their study, we intend to develop a deep learning model focusing on AD in the thorax, which could potentially be a very useful tool for automatic detection of the presence of AD for aortic CTA, followed by notifications to clinical physicians and radiologists for confirmation, which facilitates the early initiation of treatment of type A AD patients. Thus, we aimed to develop an automatic artificial intelligence (AI) system using deep learning algorithms as a triage tool to detect and classify AD in this study, in order to reduce type A AD mortality by reducing the time gaps between the end time of aortic CTA examination and the start time of intervention or surgery.

Our proposed system consists of two primary deep learning models aimed at achieving our objectives. The first model is an object detection system designed to locate and crop the aorta area from each slice within a set of CTA images. For this task, we selected RetinaNet [13] due to its effectiveness in handling high class imbalance between the foreground and background instances typical in dense object detection tasks. Introduced by Lin et al., RetinaNet is particularly recognized for its use of focal loss, a specialized loss function that reduces the loss in well-classified examples, allowing the model to focus more on challenging and misclassified instances. This approach enhances the detection capability, especially when dealing with small or infrequent objects. RetinaNet, as a quintessential single-stage detector, integrates the strengths of a feature pyramid network (FPN) within a coherent network structure, contributing to its efficiency in object detection tasks. With its ability to focus on hard-to-detect objects, RetinaNet is highly adaptable and efficient for various datasets and applications, particularly in real-time scenarios that demand both processing speed and prediction accuracy.

The second part of our system is a classification model that identifies whether each aorta patch contains a dissection. For this purpose, we employed EfficientNet-B0, a CNN architecture that maximizes performance relative to resource usage in image classification tasks. Proposed by Tan and Le in 2019 [14], EfficientNet introduces a compound scaling method, which systematically balances and scales depth, width and resolution, thereby improving accuracy while reducing computational costs. Unlike traditional methods that resize dimensions independently, EfficientNet scales all three aspects simultaneously, leading to more effective parameter utilization and an overall improved performance. The

EfficientNet-B0 model, a baseline of this architecture, utilizes depth-wise separable convolutions and squeeze-and-excitation blocks. EfficientNet has demonstrated state-of-the-art accuracy in the ImageNet dataset, with a significantly smaller memory footprint and computational cost, marking a crucial advancement in deep learning. This makes EfficientNet particularly advantageous for deployment in resource-constrained environments and diverse applications.

2. Materials and Methods

2.1. Study Design

This retrospective study was conducted at a single medical center after approval obtained by the Institutional Review Board Committee (IRB) of Chang Gung Memorial Hospital (IRB number: 202100479B0) issued on 29 March 2021. The requirement for informed consent from the patients were waived by the IRB. The dataset collection and experiments were conducted in agreement with the approved ethical guidelines and regulations.

2.2. Dataset

From January 2018 to December 2020, there were 2243 patients receiving aortic CTA scans selected from the picture archiving and communication system. Only one exam per patient was admitted to the study to prevent duplicates. A total of 1718 patients met the inclusion criteria, which were as follows: (1) clinical presentation of atypical chest pain, (2) having had an aortic CTA performed and (3) age over 20 years. Four radiologists with at least 5 years of experience reviewed the clinical information and images and excluded the cases based on the following criteria: (1) missing data or medical record (2) suboptimal CTA study showing poor opacification or visualization of aorta, (3) presence of intra-aortic device or surgical change at the time of the study, (4) thrombus in the aorta and (5) other co-existing radiological findings in the thorax, such as pleural effusions, lung mass and consolidation. Figure 1 shows examples of AD and normal cases in our study.

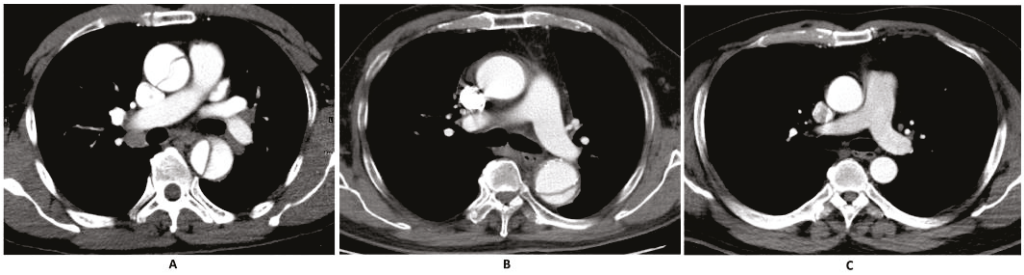


Figure 1. Example of aortic CTA in our study. (**A**) showed aortic dissection, and (AD) involved the ascending and descending aorta, which was the typical imaging presentation of Stanford type A AD. (**B**) displayed an intimal flap in the descending aorta, suggesting a Stanford type B AD. (**C**) was a normal case.

2.3. CT Image Acquisition

A non-gated CTA of aorta was performed using 100 milliliters (mL) of iodinated contrast medium (Omnipaque, 350 mg Iodine/mL), followed by 50 mL of saline flush. The arterial phase images were acquired by bolus tracking, with a threshold of 160 Hounsfield units at the ascending aorta. The scan range was set from the lower neck to the groins. The size of all images are 512 $\times$ 512 pixels.

2.4. Aorta Labeling

Aortic CTA axial images of each patient were analyzed by 4 radiologists for the absence and presence of type A and B AD according to the Stanford classification, which served as ground truth. Each axial image of aortic CTA of chest region of each patient were also

labeled by the 4 radiologists as the following: (1) normal ascending aorta, (2) normal aortic arch, (3) normal descending thoracic aorta, (4) dissected ascending aorta, (5) dissected aortic arch and (6) dissected descending thoracic aorta, using a rectangular bounding box for each. The window level was set to a center of 40 and width of 400 Hounsfield units. The ascending aorta began at the level of aortic valve while the descending thoracic aorta ended at the last axial image of the lung parenchyma which remained present. Joint meetings reviewing the labeling were held prior to further analysis.

2.5. Main Framework

The code that was used to run the model is available at: https://github.com/brainma/aortic_dissection_det_pytorch/tree/main (accessed on 18 September 2024).

In our system (Figure 2), the first step used an object detection model to identify three parts of the aorta—namely, the ascending aorta, aortic arch and descending aorta—from the axial view of a CTA scan. Subsequently, we extracted and cropped each detected aorta into an individual image. These cropped patches were then resized into 100 × 100 pixels and sent to a classification model to determine the presence or absence of a dissection in each slice. After classification, the aortic patches were arranged in a linear sequence reflecting their anatomical structure, beginning with the ascending aorta, followed by the aortic arch and ending with the descending aorta. A sliding window analysis, encompassing a span of seven consecutive images, was employed to scrutinize the sequence. If dissection was detected in at least five of the seven images with involvement of the ascending aorta, the system would classify the condition as type A AD and with involvement of the ascending aorta, type B AD would be categorized. If dissection was not detected in at least five of the seven images, the patient's case would be considered normal. For the objection detection model, we utilized a RetinaNet [13] architecture with a ResNet50 [15] encoder as its backbone. Optimization was performed using the Adam algorithm with an initial learning rate of 0.00001, and training was conducted in batches of 72. For the classification model, an EfficientNet B0 [14] architecture was employed. The Adam optimization algorithm is again utilized, this time with a learning rate of 0.001, and the model processes data in batches of 512.

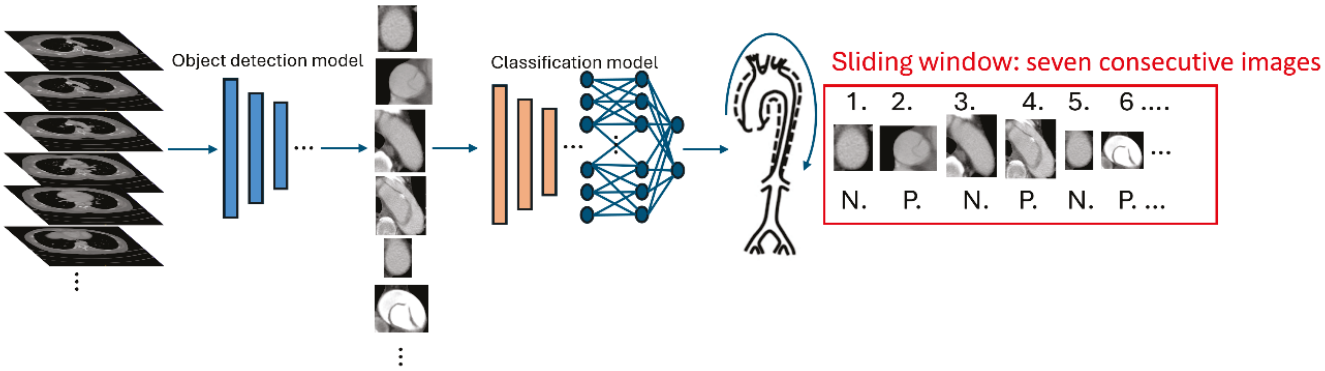


Figure 2. Diagram shows deep learning algorithms. An objection detection model would identify the ascending aorta, aortic arch, and descending aorta from CTA scan. These cropped patches were then sent to the classification model to determine the presence of aortic dissection. After classification, the cropped patches were arranged in a sequence from ascending to descending aorta to determine their Stanford type. The "P." and "N." labels under each cropped patch indicate whether the patch is positive or negative for the presence of dissection.

2.6. Statistical Analysis

For the objection detection model, the precision and recall of each of the 3 aorta locations (ascending aorta, aortic arch and descending aorta) were analyzed using the

labeling of 4 radiologists as the ground truths. For the dissection classification model, precision, recall and accuracy, as well as false positive and negative rates, were analyzed in addition to the F1 score and area under a receiver operating characteristics curve (ROC). For the diagnostic performance of the whole algorithm, including detection and dissection models, the sensitivity, specificity, positive predict value and negative predict value of Stanford type A AD, Stanford type B AD and the normal were calculated and demonstrated by the confusion matrix, using the radiologists' diagnosis for each patient as ground truths. A two-sided p value less than 0.05 was considered statistically significant, and a 95% confidence interval was calculated. All statistical analyses were performed using SPSS v26 for Windows.

3. Result

3.1. Patient Selection and Demographic

From January 2018 to December 2020, 2243 patients receiving aortic CTA scans A total of 1718 patients met the inclusion criteria, and 903 patients were excluded. In total, 814 different patients were enrolled in this study. Among them, 86 patients were diagnosed with Stanford type A AD, 59 patients were diagnosed with Stanford type B AD and 669 patients were normal. A total of 498 patients were selected for the development of the model, who were further divided into the following three sets randomly: training set, validation set, and testing set for the training object detection model (Table 1). From the same 498 patients, we cropped all the labeled aorta areas into 31,622 patches which were further randomly divided into the training set, validation set, and testing set for the training classification model (Table 2). A figure of 316 patients constituted an independent external test set. Figure 3 shows the process of participants selection. The patient characteristics of the developing model dataset and the independent external test set are detailed in Table 3. There are no significant differences in age and gender between the two datasets.

Table 1. Datasets for development of the models.

Set	Ascending Aorta	Aortic Arch	Descending Aorta	Images
Training	7191	2709	19,334	21,992 (398 patients)
Validation	927	343	2575	2916 (50 patients)
Test	885	353	2433	2787 (50 patients)

Table 2. Datasets for development of the classification model.

Set	Positive	Negative	Images
Training	1516	25,765	27,281
Validation	303	35,797	3882
Test	262	197	459

Table 3. Characteristics of datasets.

	Model Developing Dataset	Independent Testing Dataset	*p* Value
Age (year)	59.30 ± 13.51	60.32 ± 13.57	0.296
Sex			0.058
Male	277 (55.6%)	197 (62.1%)	
Female	221 (44.4%)	119 (37.7%)	
Aortic dissection			<0.001
Normal	422 (84.7%)	247 (78.2%)	
Type A	54 (10.8%)	32 (10.1%)	
Type B	22 (4.4%)	37 (11.7%)	

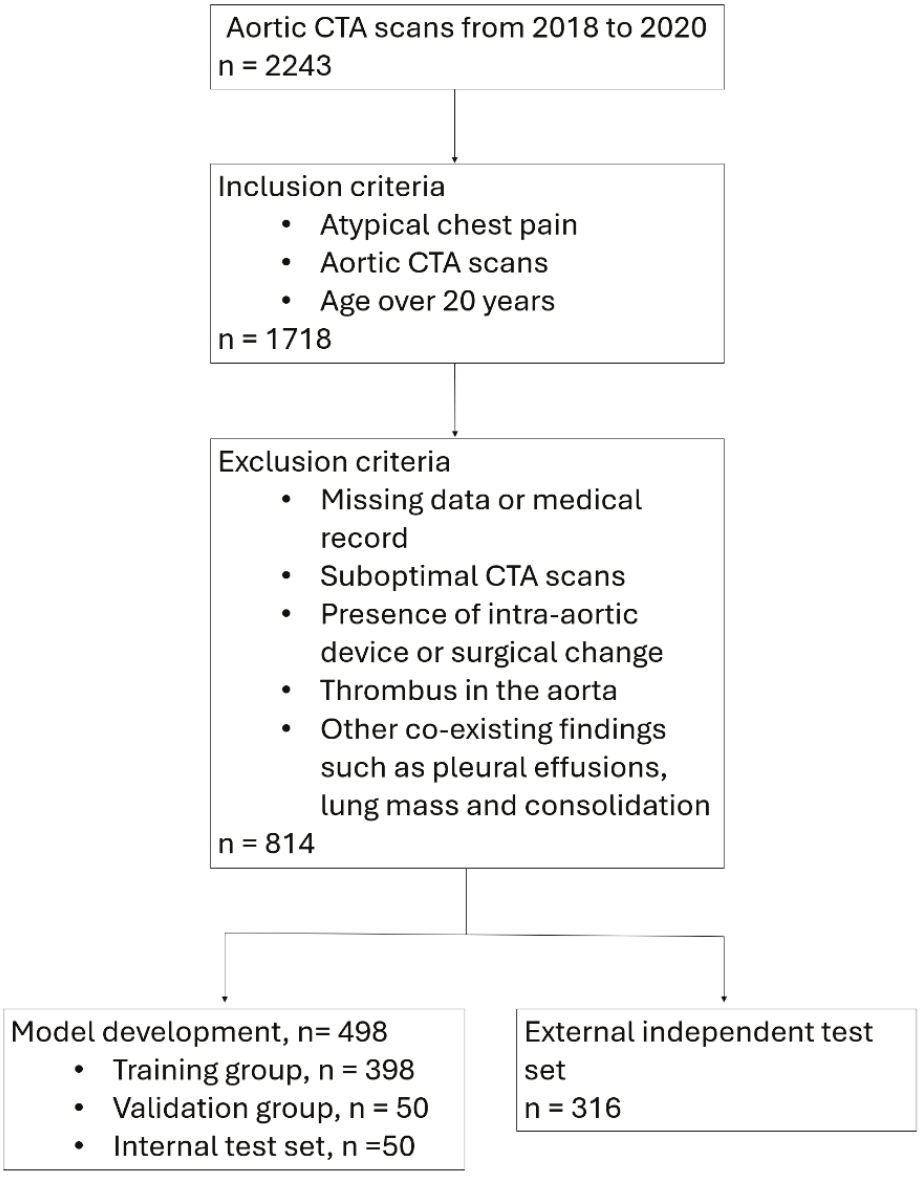


Figure 3. Flowchart shows participants selection. Among the 2243 patients with aortic CTA scans from 2018 to 2020, 1718 patients were included if the following criteria was met: (1) clinical presentation of atypical chest pain, (2) an aortic CTA for evaluating aortic dissection was performed and (3) age over 20 years. A total of 903 patients were excluded because of (1) missing data or medical record, (2) suboptimal CTA scans, (3) presence of intra-aortic device or surgical change, (4) thrombus in the aorta and (5) other co-existing findings. A total of 814 patients were enrolled. A figure of 498 patients were randomly assigned for model development, and 316 patients constitute the independent test set.

3.2. Objection Detection Model

The efficacy of our aorta detection model was as follows: the mAP is 0.938, and the precision values for the ascending aorta, aortic arch and descending aorta were 0.906, 0.975 and 0.977, respectively, while the recall rates for these regions were 0.925, 0.928 and 0.981, respectively.

3.3. Dissection Classification Model

As for the dissection classification model, its performance metrics on the test set showed an accuracy of 0.996. The model demonstrated a false negative rate of 0.004 and a false positive rate of 0.005. Its precision and recall (sensitivity) both matched 0.996. The F1 score aligned with these metrics at 0.996, while the specificity was slightly lower at 0.995, and the area under ROC reached the maximum value of 0.990.

3.4. Whole System Evaluation

Our aortic dissection screening system was further evaluated on an independent test set comprising 316 patients, of which 247 were normal cases, 32 had Type A AD, and 37 had Type B AD. For normal cases, the system achieved a sensitivity of 0.988 and a specificity of 1.000. In detecting Type A AD, the sensitivity was 0.969 with a specificity of 0.982, while for Type B AD, the sensitivity and specificity were 0.946 and 0.996, respectively. The results were consolidated in a confusion matrix presented in Figure 4. The average processing time per aortic CTA study was 7.9 $\pm$ 2.8 s (mean $\pm$ standard deviation).

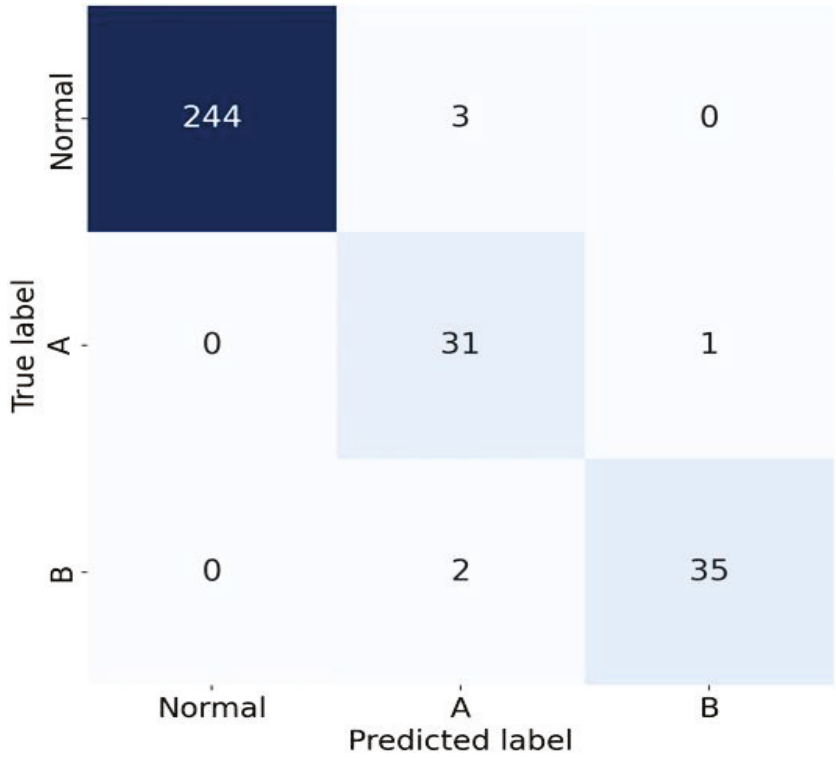


Figure 4. The 3 $\times$ 3 confusion matrix. The case numbers of the whole system's prediction and imaging diagnoses, obtained from the independent test set, were demonstrated in the 3 $\times$ 3 confusion matrix.

4. Discussion

We developed a deep learning-based model that could successfully detect AD and determine its Stanford type. The primary clinical purpose of this model was to serve as a triage tool, with the intention of assisting and flagging potential cases for radiology review. In scenarios where radiology expertise was unavailable or delayed, the model

could provide value by highlighting critical imaging findings for clinicians. As a result, the algorithm needed to be accurate in identifying and classifying AD with a fast-processing time. The accuracy of our model was 0.981, and therefore, it possibly could reach human diagnostic ability. The average processing time was 7.9 s per case, which would be timesaving for clinicians. Wang [16] developed a commercialized automatic product that could both detect AD and intramural hematoma (IMH), and its process time was 37.9 s per study. Nonetheless, the detection of AD and IMH together using AI algorithms would increase many notifications due to IMH detection, which is not an emergency situation. Howard et al. [17] reported that 48.6% of type A AD patients die prior to hospital assessment, and the mortality rate of type A AD has been reported as 0.5% to 1.0% per hour in previous studies [3,18]. An early and prompt diagnosis is crucial, since type A AD could be treated by surgery, thus lowering the mortality rate [18]. Aortic CTA served as the most important diagnostic tool for AD. However, the average final report turnaround time differs. Even in the presence of radiologists dedicated to reports from emergency departments, the median final report turnaround time still took 2.75 h [7]. Our model had a very fast processing time, and could be a triage tool to alert clinicians and radiologists to identify this emergency as soon as possible.

Segmentation is an important step to develop an AD diagnostic model [19]. Cao et al. [20] created an automated type B AD segmentation by manually segmenting the true and false lumen of aorta, which is a time-consuming task. Our model used a rectangular labeling box for aorta detection instead of segmentation, which eased the workload of physicians during model development. Thus, we were able to enroll more than 800 patients in this study. Besides that, various models [21–23] were developed for segmentation and aortic diameter measurement. However, these studies mainly focused on type B AD. Whether the ascending thoracic aorta was involved in AD or not was clinically crucial, because type A AD is a surgical emergency. A two-dimensional (2D) framework was applied in our model so we could simplify the imaging process and lower the demand for computing power. Hata and Yi et al. [9,10] developed aortic dissection detection models with non-enhanced CT studies. Previous studies [10,24,25] often employed a three-dimensional framework, and we demonstrated comparable sensitivity and specificity as shown in Table 4 using a 2D model.

Table 4. Comparison of different deep learning algorithms for aortic dissection detection. AD: aortic dissection.

Study [Reference], Year, Country	Sample Size	Framework	Imaging Modality	Dissection Type Interpretation	Intramural Hematoma Identification	Sensitivity (%)	Specificity (%)
Our model 2024, Taiwan	814 (145 AD)	2D	CTA	Yes	No	96.9 (Type A) 94.6 (Type B) 98.8 (Normal)	98.2 (Type A) 99.6 (Type B) 100.0 (Normal)
Hata et al. [9] 2020, Japan	170 (85 AD)	2D	Non-enhanced CT	No	No	90.0	88.2
Yi et al. [10] 2022, China	452 (185 AD)	3D	Non-enhanced CT	Yes	No	86.2	92.3
Huang et al. [24] 2022, Taiwan	130 (100 AD)	3D	CTA	Yes	No	95.45 (Type A) 79.31 (Type B) 93.53 (Normal)	98.55 (Type A) 94.05 (Type B) 94.12 (Normal)
Harris et al. [25] 2019, USA	34,689 (112 AD)	3D	CTA	No	No	87.8	96
Wang [16] 2024, USA	380 (156 AD)	No detail	CTA	No	Yes	92.9	91.5

To reduce false negatives in detecting type A AD, we employed a sliding window analysis method. Hata et al. [9] proposed a consecutive slices algorithm for AD detection, and they concluded that five consecutive slices may be appropriate in the clinical setting. Huang et al. [24] also used five slices in their AD detection model. Reviewing the original

CT images, we discovered that all false negatives, including three normal cases being misinterpreted as type A AD, were due to a streaking artifact, which was caused by a dense contrast medium in the superior vena cava [26]. The highly attenuated contrast medium may produce a metal-like appearance in the CT scans due to the combination of a beam hardening-artifact, photon starvations and a scatter artifact [27]. These were commonly encountered artifacts because an optimal CTA of the aorta requires a concentrated bolus of contrast medium, injected from venous access in the upper extremities. The artifacts could be minimized by using filtration, calibration correction and beam-hardening correction software [26]. Experienced physicians were generally able to distinguish between artifacts and true intima flaps. One Stanford type A AD case was considered as type B AD by the model. The entry site of intimal flap in this case was in the distal ascending aorta, in proximity to the aortic arch. Further investigations and larger data sets for training may be needed in the future.

This study has several limitations. First, the abdominal aorta was not included, although detecting the full extent of AD using the aortic diameter measurement may guide the treatment strategy for the patient. The exclusion of abdominal aorta in this study was based on the aim of our algorithms for detecting AD involvements in the ascending aorta, Second, thrombus in the aorta was excluded during model training. Thrombosis was usually observed in the chronic stage of AD. Our primary goal was to develop a model for acute AD detection instead of follow-up care. Besides that, thrombus within the aorta could obscure or compromise the visualization of the intimal flap. As a result, we excluded these patients during data selection. Future work may be required involving additional training to address cases with thrombus to enhance our model's clinical applicability. Third, the imaging sets were enrolled from a single institute using CT scanners from several vendors. Also, the imaging protocols for aortic CTA may vary in different hospitals. Further studies of the performance of this AI model system in many institutes will be conducted for further assessment and to increase its generalizability.

The development of a deep learning model capable of automatically detecting aortic dissection and classifying it based on the Stanford classification may have an implication in clinical practice, since early and accurate diagnosis is critical in managing AD. We expected to incorporate our model into the workflow of daily practice to serve as a triage tool to assist both clinicians and radiologists. The system could aid in prioritizing surgical intervention for type A AD, for which urgent treatment is required. Our model was able to determine AD. Future research focusing on enhancing our model may be performed by identifying other acute aortic syndromes, including intramural hematoma and penetrating atherosclerotic ulcers. It is also important to expand our model to incorporate with other imaging modalities, since multimodality imaging, such as aortic CTA, echocardiography and magnetic resonance imaging, has been shown to play an important role in the diagnosis, complications and clinical management of AD [28].

5. Conclusions

This study demonstrates that AD on CTA could be automatically detected by deep learning algorithms with a high accuracy (0.981), with classifications of AD types provided based on Stanford classification. This automatic triage system takes less than 10 s for inferences of AD and its type for each aortic CTA examination. The implementation of these algorithms could potentially accelerate clinical workflow, by enabling rapid and precise diagnosis, which is crucial for conditions that require urgent surgical intervention, such as type A AD. The triage system could provide real-time alerts for clinicians and radiologists once aortic CTA is performed. Future research for the further validation of the model across different healthcare systems may be helpful for managing AD.

Author Contributions: Conceptualization, C.-Y.M. and L.-J.W.; data curation, C.-B.C.; formal analysis, H.-H.L., C.-B.C., C.-Y.L., C.-Y.M. and L.-J.W.; investigation, Y.-S.C.; methodology, H.-H.L., C.-B.C., Y.-S.C., C.-Y.M. and L.-J.W.; project administration, L.-J.W.; resources, C.-Y.L.; software, H.-H.L., Y.-S.C., C.-F.K. and C.-Y.M.; supervision, C.-F.K.; validation, H.-H.L., Y.-S.C., C.-F.K., C.-Y.L., C.-Y.M. and L.-J.W.; visualization, C.-F.K. and L.-J.W.; writing—original draft, H.-H.L., C.-B.C., Y.-S.C. and C.-Y.L.; writing—review and editing, L.-J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by CHANG GUNG MEDICAL FOUNDATION, TAIWAN grant number CORPVVM0061.

Institutional Review Board Statement: This retrospective study was conducted at a single medical center after approval obtained by the Institutional Review Board Committee (IRB) of Chang Gung Memorial Hospital (IRB number: 202100479B0) issued on 29 March 2021.

Informed Consent Statement: The requirement for informed consents from the patients were waived by the IRB due to the retrospective nature of the study and the analysis used anonymous clinical data.

Data Availability Statement: The code that was used to run the model is available at: https://github.com/brainma/aortic_dissection_det_pytorch/tree/main (accessed on 18 September 2024).

Conflicts of Interest: All authors also declare no affiliation with or involvement in any organization or entity with any financial interest (such as honoraria; educational grant; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or material discussed in this manuscript.

References

1. Melvinsdottir, I.H.; Lund, S.H.; Agnarsson, B.A.; Sigvaldason, K.; Gudbjartsson, T.; Geirsson, A. The incidence and mortality of acute thoracic aortic dissection: Results from a whole nation study. *Eur. J. Cardiothorac. Surg.* **2016**, *50*, 1111–1117. [CrossRef] [PubMed]
2. Clouse, W.D.; Hallett, J.W., Jr.; Schaff, H.V.; Spittell, P.C.; Rowland, C.M.; Ilstrup, D.M.; Melton, L.J., 3rd. Acute aortic dissection: Population-based incidence compared with degenerative aortic aneurysm rupture. *Mayo Clin. Proc.* **2004**, *79*, 176–180. [CrossRef] [PubMed]
3. Nienaber, C.A.; Eagle, K.A. Aortic dissection: New frontiers in diagnosis and management: Part I: From etiology to diagnostic strategies. *Circulation* **2003**, *108*, 628–635. [CrossRef] [PubMed]
4. Moore, A.G.; Eagle, K.A.; Bruckman, D.; Moon, B.S.; Malouf, J.F.; Fattori, R.; Evangelista, A.; Isselbacher, E.M.; Suzuki, T.; Nienaber, C.A.; et al. Choice of computed tomography, transesophageal echocardiography, magnetic resonance imaging, and aortography in acute aortic dissection: International Registry of Acute Aortic Dissection (IRAD). *Am. J. Cardiol.* **2002**, *89*, 1235–1238. [CrossRef]
5. Shiga, T.; Wajima, Z.; Apfel, C.C.; Inoue, T.; Ohe, Y. Diagnostic accuracy of transesophageal echocardiography, helical computed tomography, and magnetic resonance imaging for suspected thoracic aortic dissection: Systematic review and meta-analysis. *Arch. Intern. Med.* **2006**, *166*, 1350–1356. [CrossRef]
6. Expert Panel on Vascular Imaging; Gunn, A.J.; Kalva, S.P.; Majdalany, B.S.; Craft, J.; Eldrup-Jorgensen, J.; Ferencik, M.; Ganguli, S.; Kendi, A.T.; Khaja, M.S.; et al. ACR Appropriateness Criteria(R) Nontraumatic Aortic Disease. *J. Am. Coll. Radiol.* **2021**, *18*, S106–S118. [CrossRef]
7. Lamb, L.; Kashani, P.; Ryan, J.; Hebert, G.; Sheikh, A.; Thornhill, R.; Fasih, N. Impact of an in-house emergency radiologist on report turnaround time. *CJEM* **2015**, *17*, 21–26. [CrossRef]
8. 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.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
9. Hata, A.; Yanagawa, M.; Yamagata, K.; Suzuki, Y.; Kido, S.; Kawata, A.; Doi, S.; Yoshida, Y.; Miyata, T.; Tsubamoto, M.; et al. Deep learning algorithm for detection of aortic dissection on non-contrast-enhanced CT. *Eur. Radiol.* **2021**, *31*, 1151–1159. [CrossRef]

10. Yi, Y.; Mao, L.; Wang, C.; Guo, Y.; Luo, X.; Jia, D.; Lei, Y.; Pan, J.; Li, J.; Li, S.; et al. Advanced Warning of Aortic Dissection on Non-Contrast CT: The Combination of Deep Learning and Morphological Characteristics. *Front. Cardiovasc. Med.* **2021**, *8*, 762958. [CrossRef]
11. Cheng, Z.; Zhao, L.; Yan, J.; Zhang, H.; Lin, S.; Yin, L.; Peng, C.; Ma, X.; Xie, G.; Sun, L. A deep learning algorithm for the detection of aortic dissection on non-contrast-enhanced computed tomography via the identification and segmentation of the true and false lumens of the aorta. *Quant. Imaging Med. Surg.* **2024**, *14*, 7365–7378. [CrossRef]
12. Raj, A.; Allababidi, A.; Kayed, H.; Gerken, A.L.H.; Muller, J.; Schoenberg, S.O.; Zollner, F.G.; Rink, J.S. Streamlining Acute Abdominal Aortic Dissection Management-An AI-based CT Imaging Workflow. *J. Imaging Inform. Med.* **2024**, *37*, 2729–2739. [CrossRef] [PubMed]
13. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988. [CrossRef]
14. Tan, M.; Le, Q.V. 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. [CrossRef]
15. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
16. Wang, T.-H. Premarket Notification [510(k)] Summary: K240291. U.S. Food and Drug Administration. Available online: https://www.accessdata.fda.gov/cdrh_docs/pdf24/K240291.pdf (accessed on 18 December 2024).
17. Howard, D.P.; Banerjee, A.; Fairhead, J.F.; Perkins, J.; Silver, L.E.; Rothwell, P.M. Oxford Vascular Study. Population-based study of incidence and outcome of acute aortic dissection and premorbid risk factor control: 10-year results from the Oxford Vascular Study. *Circulation* **2013**, *127*, 2031–2037. [CrossRef] [PubMed]
18. Harris, K.M.; Nienaber, C.A.; Peterson, M.D.; Woznicki, E.M.; Braverman, A.C.; Trimarchi, S.; Myrmel, T.; Pyeritz, R.; Hutchison, S.; Strauss, C.; et al. Early Mortality in Type A Acute Aortic Dissection: Insights From the International Registry of Acute Aortic Dissection. *JAMA Cardiol.* **2022**, *7*, 1009–1015. [CrossRef] [PubMed]
19. Pepe, A.; Li, J.; Rolf-Pissarczyk, M.; Gsaxner, C.; Chen, X.; Holzapfel, G.A.; Egger, J. Detection, segmentation, simulation and visualization of aortic dissections: A review. *Med. Image Anal.* **2020**, *65*, 101773. [CrossRef]
20. Cao, L.; Shi, R.; Ge, Y.; Xing, L.; Zuo, P.; Jia, Y.; Liu, J.; He, Y.; Wang, X.; Luan, S.; et al. Fully automatic segmentation of type B aortic dissection from CTA images enabled by deep learning. *Eur. J. Radiol.* **2019**, *121*, 108713. [CrossRef]
21. Yu, Y.; Gao, Y.; Wei, J.; Liao, F.; Xiao, Q.; Zhang, J.; Yin, W.; Lu, B. A Three-Dimensional Deep Convolutional Neural Network for Automatic Segmentation and Diameter Measurement of Type B Aortic Dissection. *Korean J. Radiol.* **2021**, *22*, 168–178. [CrossRef]
22. Hahn, L.D.; Mistelbauer, G.; Higashigaito, K.; Koci, M.; Willemink, M.J.; Sailer, A.M.; Fischbein, M.; Fleischmann, D. CT-based True- and False-Lumen Segmentation in Type B Aortic Dissection Using Machine Learning. *Radiol. Cardiothorac. Imaging* **2020**, *2*, e190179. [CrossRef]
23. Wobben, L.D.; Codari, M.; Mistelbauer, G.; Pepe, A.; Higashigaito, K.; Hahn, L.D.; Mastrodicasa, D.; Turner, V.L.; Hinostroza, V.; Baumler, K.; et al. Deep Learning-Based 3D Segmentation of True Lumen, False Lumen, and False Lumen Thrombosis in Type-B Aortic Dissection. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine & Biology Society, Guadalajara, Mexico, 1–5 November 2021; Volume 2021, pp. 3912–3915. [CrossRef]
24. Huang, L.T.; Tsai, Y.S.; Liou, C.F.; Lee, T.H.; Kuo, P.P.; Huang, H.S.; Wang, C.K. Automated Stanford classification of aortic dissection using a 2-step hierarchical neural network at computed tomography angiography. *Eur. Radiol.* **2022**, *32*, 2277–2285. [CrossRef]
25. Harris, R.J.; Kim, S.; Lohr, J.; Towey, S.; Velichkovich, Z.; Kabachenko, T.; Driscoll, I.; Baker, B. Classification of Aortic Dissection and Rupture on Post-contrast CT Images Using a Convolutional Neural Network. *J. Digit. Imaging* **2019**, *32*, 939–946. [CrossRef]
26. Barrett, J.F.; Keat, N. Artifacts in CT: Recognition and avoidance. *Radiographics* **2004**, *24*, 1679–1691. [CrossRef]
27. Gupta, A.; Obmann, V.C.; Jordan, M.; Lennartz, S.; Obmann, M.M.; Grosse Hokamp, N.; Zopfs, D.; Pennig, L.; Furtjes, G.; Ramaiya, N.; et al. CT artifacts after contrast media injection in chest imaging: Evaluation of post-processing algorithms, virtual monoenergetic images and their combination for artifact reduction. *Quant. Imaging Med. Surg.* **2021**, *11*, 226–239. [CrossRef]
28. Perone, F.; Guglielmo, M.; Coceani, M.; La Mura, L.; Dentamaro, I.; Sabatino, J.; Gimelli, A. The Role of Multimodality Imaging Approach in Acute Aortic Syndromes: Diagnosis, Complications, and Clinical Management. *Diagnostics* **2023**, *13*, 650. [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

Automatic Segmentation of the Left Ventricle in Apical Four-Chamber View on Transesophageal Echocardiography Based on UNeXt Deep Neural Network

Lingeer Wu [1,†], Yijun Ling [2,†], Ling Lan [1], Kai He [1], Chunhua Yu [1], Zhuhuang Zhou [2] and Le Shen [1,*]

[1] Department of Anesthesiology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China; 13426365656@126.com (L.W.); lanling_1988@163.com (L.L.); harveyhekai@sina.com (K.H.); yuchh@pumch.cn (C.Y.)

[2] Department of Biomedical Engineering, College of Chemistry and Life Science, Beijing University of Technology, Beijing 100124, China; lingyijun@emails.bjut.edu.cn (Y.L.); zhouzh@bjut.edu.cn (Z.Z.)

* Correspondence: shenle@pumch.cn

† These authors contributed equally to this work and share first authorship.

Abstract: Background/Objectives: The automatic left ventricle segmentation in transesophageal echocardiography (TEE) is of significant importance. In this paper, we constructed a large-scale TEE apical four-chamber view (A4CV) image dataset and proposed an automatic left ventricular segmentation method for the TEE A4CV based on the UNeXt deep neural network. **Methods:** UNeXt, a variant of U-Net integrating a multilayer perceptron, was employed for left ventricle segmentation in the TEE A4CV because it could yield promising segmentation performance while reducing both the number of network parameters and computational complexity. We also compared the proposed method with U-Net, TransUNet, and Attention U-Net models. Standard TEE A4CV videos were collected from 60 patients undergoing cardiac surgery, from the onset of anesthesia to the conclusion of the procedure. After preprocessing, a dataset comprising 3000 TEE images and their corresponding labels was generated. The dataset was randomly divided into training, validation, and test sets in an 8:1:1 ratio on the patient level. The training and validation sets were used to train the UNeXt, U-Net, TransUNet, and Attention U-Net models for left ventricular segmentation. The dice similarity coefficient (DSC) and Intersection over Union (IoU) were used to evaluate the segmentation performance of each model, and the Kruskal–Wallis test was employed to analyze the significance of DSC differences. **Results:** On the test set, the UNeXt model achieved an average DSC of 88.60%, outperforming U-Net (87.76%), TransUNet (85.75%, $p < 0.05$), and Attention U-Net (79.98%; $p < 0.05$). Additionally, the UNeXt model had a smaller number of parameters (1.47 million) and floating point operations (2.28 giga) as well as a shorter average inference time per image (141.73 ms), compared to U-Net (185.12 ms), TransUNet (209.08 ms), and Attention U-Net (201.13 ms). The average IoU of UNeXt (77.60%) was also higher than that of U-Net (76.61%), TransUNet (77.35%), and Attention U-Net (68.86%). **Conclusions:** This study pioneered the construction of a large-scale TEE A4CV dataset and the application of UNeXt to left ventricle segmentation in the TEE A4CV. The proposed method may be used for automatic segmentation of the left ventricle in the TEE A4CV.

Keywords: transesophageal echocardiography; left ventricle; segmentation; deep learning

1. Introduction

Echocardiography is a non-invasive medical imaging technique and serves as a primary method for the clinical dynamic assessment of cardiac function. It can be divided into two categories: transthoracic echocardiography (TTE) and transesophageal echocardiography (TEE). By analyzing echocardiographic images, multiple important cardiac function parameters can be quantitatively obtained, with accurate segmentation of the left ventricle serving as the foundation for subsequent metric calculations. Currently, clinicians typically

rely on manual segmentation of the left ventricular contours from echocardiography, a process that is time-consuming and labor-intensive, and heavily dependent on the clinician's personal experience. The measurements of the same patient may exhibit inter-operator differences as well as intra-operator differences. Deep learning techniques have demonstrated significant potential in medical image analysis, offering effective solutions for automatic segmentation of the left ventricle in echocardiography.

In the past decade, advancements in deep learning-based semantic segmentation methods, such as fully connected networks (FCNs) [1], DeepLab [2–4], and U-Net [5–8], have facilitated their increasing applications in medical image segmentation. DeepLab has three versions: V1 [2], V2 [3], and V3 [4]. DeepLab V1 incorporates atrous convolution and conditional random fields (CRFs), using VGG16 as the backbone [2]. DeepLab V2 further incorporates atrous spatial pyramid pooling (ASPP) with ResNet-101 as the backbone [3]. DeepLab V3 builds on DeepLab V2 and uses improved ASPP and deeper atrous convolution while removing CRFs [4]. The U-Net series, including U-Net [5], UNet++ [6], nnU-Net [7], and UNet3+ [8], have shown promising performance in medical image segmentation. Recently, researchers have combined attention mechanisms with convolutional neural networks (CNNs) for medical image segmentation to enhance the model's global perception capabilities. Notable models include DANet [9], CBAM [10], and Attention U-Net [11]. Additionally, the advantages of Transformers in capturing long-range dependencies have led to their integration with CNNs, resulting in models like TransUNet [12], Swin Transformer [13], and CvT [14]. Researchers have improved network architectures based on the aforementioned methods for TTE left ventricular segmentation tasks. Liu et al. [15] designed a U-Net-based model, DPS-Net, achieving a dice similarity coefficient (DSC) of 93.2% (end diastole) and 92.8% (end systole) on the CAMUS (Cardiac Acquisitions for Multi-structure Ultrasound Segmentation) dataset [16]. Moradi et al. [17] proposed MFP-U-Net, which enhanced U-Net's skip connection structures by merging features from each layer of the decoder with those of the encoder, achieving a DSC of $(94.5 \pm 0.12)\%$ on the CAMUS dataset. Zeng et al. [18] proposed the MAEF-Net, which employed multiple attention mechanisms to capture heartbeat features while suppressing noise, enabling fully automated left ventricle segmentation. Their model achieved a DSC of $(93.1 \pm 2.22)\%$ on the EchoNet-Dynamic dataset [19]. Zhu et al. [20] presented SAM-Att, which integrated an attention module into the segment anything model (SAM), resulting in a DSC of 93.49% on the CAMUS dataset. Lal et al. [21] enhanced U-Net by incorporating residual modules and modifying the skip connection structure, achieving a DSC of 95.59% on the CAMUS dataset for left ventricular segmentation. Wan et al. [22] employed a semi-supervised learning approach for left ventricular segmentation, attaining DSCs of 93.05% and 93.93% on the CAMUS and EchoNet-Dynamic datasets, respectively.

Although deep learning techniques have been extensively studied for automatic left ventricle segmentation in the apical four-chamber view (A4CV) on TTE, demonstrating the desired performance [15,17,18,20–22], research on deep learning methods for segmenting the left ventricle in the TEE A4CV remains limited. Few studies involve applications of deep learning in automatic left ventricular segmentation in TEE. Kang et al. [23] introduced an attention mechanism and a residual feature aggregation module in U-Net for segmenting the left ventricle in the long-axis view of the mid-esophagus, achieving a DSC of $(89.9 \pm 1.7)\%$. However, the dataset was too small, containing only 120 TEE images from nine patients. Ahn et al. [24] employed Res-U-Net++ [25], combining features extracted from 158 TEE images (number of patients not described) for left ventricular segmentation, resulting in a DSC of 82.08%. The small number of images in existing deep learning methods on left ventricle segmentation in TEE may limit their further development. Moreover, the current deep learning models for left ventricular segmentation in TEE had large parameter sizes, which may lead to longer inference times. Therefore, it may be needed to construct a large-scale TEE image dataset. In addition, UNeXt [26] is a recently introduced deep neural network based on U-Net, which shows promising performance in image segmentation

with a reduced number of network parameters. However, the UNeXt [26] model has not been applied to TEE image segmentation.

In this paper, we constructed a large-scale TEE A4CV image dataset, as there is no publicly available dataset for TEE A4CV, and proposed an automatic segmentation method for the left ventricle in the TEE A4CV based on the UNeXt [26] deep neural network. UNeXt uses U-Net as the backbone network and incorporates a multilayer perceptron (MLP), which could produce promising segmentation performance while reducing both the number of network parameters and computational complexity. Furthermore, UNeXt has not been applied to left ventricular segmentation in the TEE A4CV. Standard TEE A4CV videos were collected from 60 patients undergoing cardiac surgery, from the onset of anesthesia to the conclusion of the procedure. After data preprocessing, a dataset containing 3000 TEE images and corresponding labels was generated. The dataset was randomly divided into training, validation, and test sets, ensuring that images from the same patient appeared in only one subset. U-Net [5], TransUNet [12], and Attention U-Net [11] were utilized for comparison. The training and validation sets were input into the UNeXt, U-Net, TransUNet, and Attention U-Net networks to train left ventricular segmentation models. The test set was used to evaluate the segmentation performance of each model. The experimental results indicated that the proposed method improved segmentation accuracy while simultaneously reducing the number of network parameters and inference time.

2. Materials and Methods

Figure 1 illustrates the process of automatic segmentation of the left ventricle in the TEE A4CV. Initially, a dataset containing 3000 TEE images and their corresponding labels was obtained from the TEE videos of 60 patients after data preprocessing. The dataset was randomly divided into training, validation, and test sets in an 8:1:1 ratio. The training and validation sets were used to train and optimize the model. Subsequently, the test set was input into the trained deep learning network model to obtain 300 segmented images of the left ventricle.

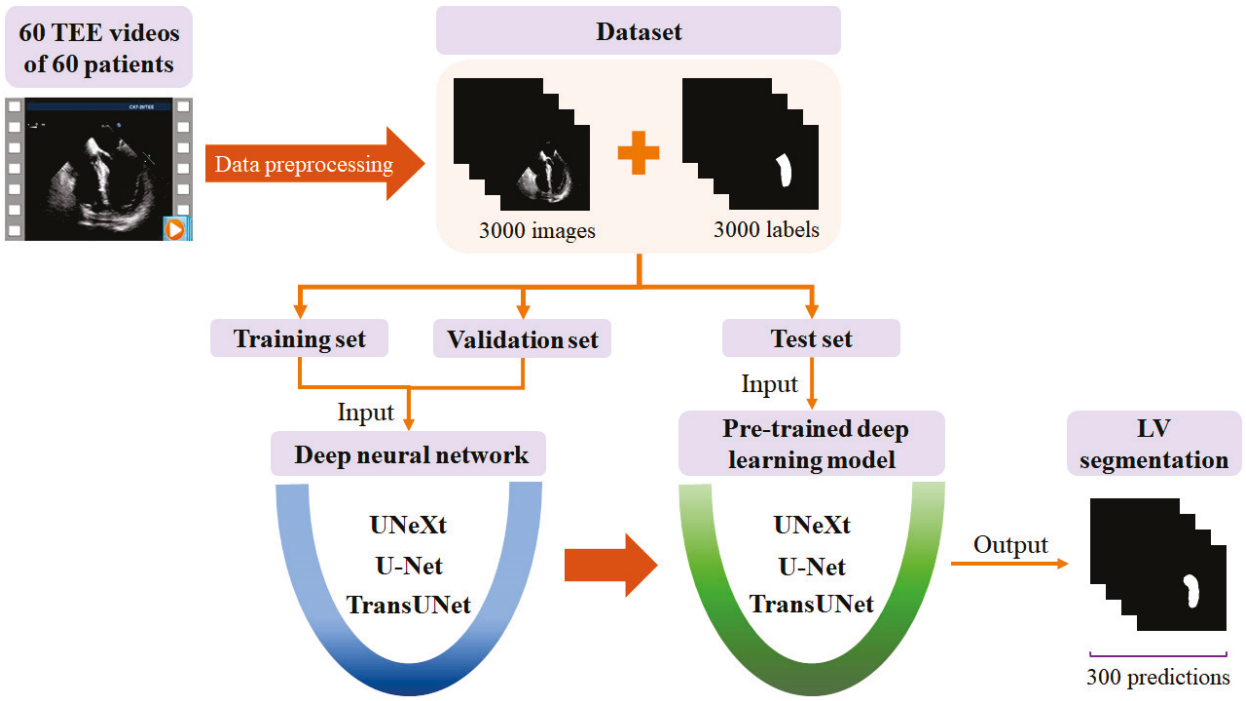


Figure 1. Flow chart of the proposed deep learning-based method for automatic segmentation of the left ventricle (LV) in the transesophageal echocardiography (TEE) apical 4-chamber view.

2.1. TEE Dataset

2.1.1. Data Collection

The TEE data were collected from the Department of Anesthesiology at Peking Union Medical College Hospital, comprising standard TEE A4CV videos from 60 patients undergoing cardiac surgery. The videos were acquired from the onset of anesthesia to the completion of the procedure. This retrospective, single-center study was approved by the Ethics Review Committee of the Peking Union Medical College Hospital, Chinese Academy of Medical Sciences. The video resolution was 800 × 600 pixels, with each video containing at least 1–1.5 cardiac cycles (Figure 2). The ultrasound videos were acquired by two professional anesthesiologists using an ultrasound scanner (IE33; Philips Medical Systems, Andover, MA, USA), with an X7-2t 1.0−5.0 MHz transducer (Philips Medical Systems). Note that TEE acquisition was much more difficult than TTE scanning because it involved the intervention of the ultrasound transducer.

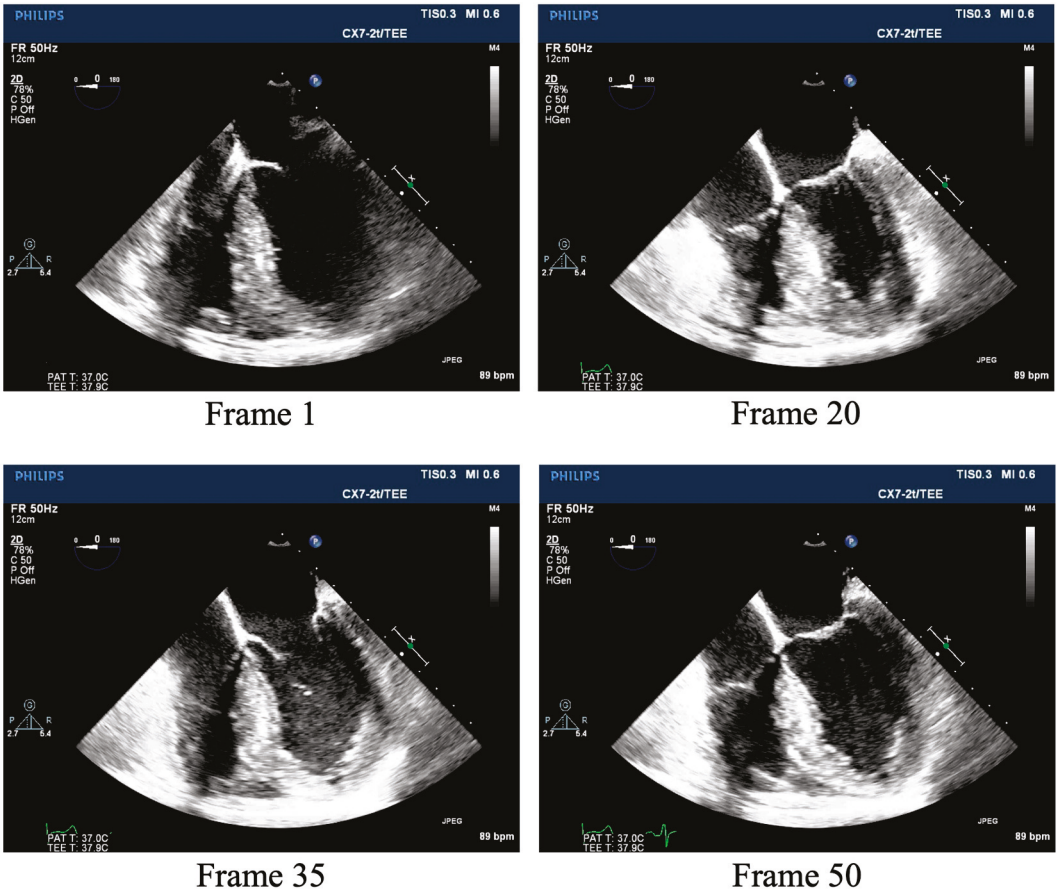


Figure 2. The 1st, 20th, 35th, and 50th TEE A4CV frames of a TEE video of the same patient. TEE: transesophageal echocardiography; A4CV: apical four-chamber view.

2.1.2. Data Preprocessing

For each TEE video, 50 consecutive frames covering at least one complete cardiac cycle were selected. As shown in Figure 3, a binary mask with a fan-shaped region (center: (425.36, 106), radius: 494 pixels, angle: 92.67°) was multiplied with each frame of the images to remove patient identifiers, scanning details, and other irrelevant information, which might interfere with left ventricle segmentation. This resulted in 50 ultrasound images of the region of interest (ROI) for each video (Figure 3). All ROI ultrasound images were manually annotated using the open access LabelMe software (version 4.5.13) to generate corresponding binary masks for the left ventricle (Figure 4). Each binary mask was stored as a ".json" file corresponding to the TEE A4CV image. The manual annotation was

performed by an anesthesiologist and checked by another anesthesiologist, both of whom were experienced in the TEE A4CV. To meet the input requirements of the convolutional neural network (CNN), all grayscale ROI ultrasound images and their corresponding label masks were resized to 512 × 512 pixels using bilinear interpolation. These resized images were then used to form the TEE dataset (Figure 4). The TEE dataset was randomly divided into three subsets in an 8:1:1 ratio, a training set of 48 patients (2400 images), a validation set of 6 patients (300 images), and a test set of 6 patients (300 images), ensuring that images from the same patient appeared only in one subset. The training set was utilized to train the deep learning model, the validation set was used to monitor the model and adjust its optimal parameters, and the test set was used to evaluate the model's segmentation performance.

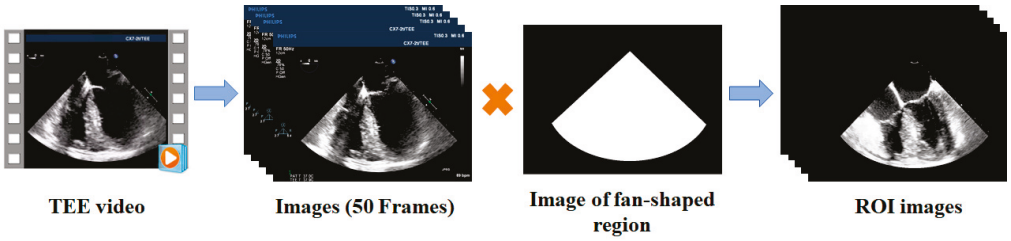


Figure 3. Generation of ROI images. TEE: transesophageal echocardiography; ROI: region of interest.

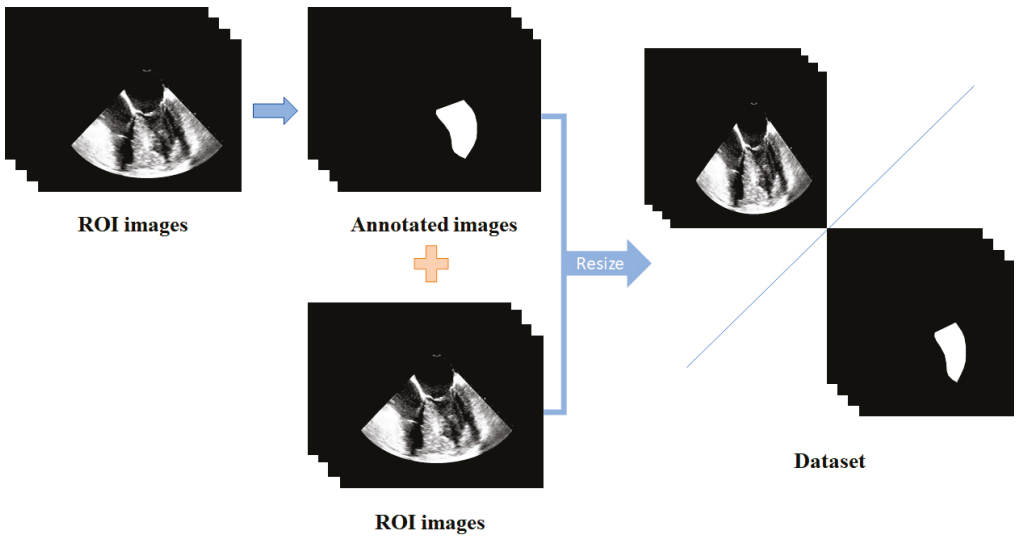


Figure 4. Process of constructing the dataset. ROI: region of interest.

2.2. Deep Learning Network Models

In this study, UNeXt [26] was selected as the model for left ventricular segmentation in TEE (Figure 5). UNeXt uses U-Net as the backbone network and incorporates an MLP. The overall structure of UNeXt follows a U-shaped encoder–decoder architecture. The first three layers of the encoder consist of convolutional layers, which perform three down-sampling operations on the input image. The fourth and fifth layers are MLP modules, where a Shifted-MLP module is introduced before a tokenized MLP (Tok-MLP) module. The Shifted-MLP alters channels before passing them into the MLP to focus the network on learning local features. The Tok-MLP module reduces the number of network parameters and computational complexity by using tokenized MLPs, while incorporating shift operations to extract local information across different axes for better feature representation. A skip connection structure is employed between the encoder and decoder, allowing the model to utilize features at different scales simultaneously. UNeXt can yield satisfying segmentation performance while reducing the number of network parameters, thereby lowering computational costs.

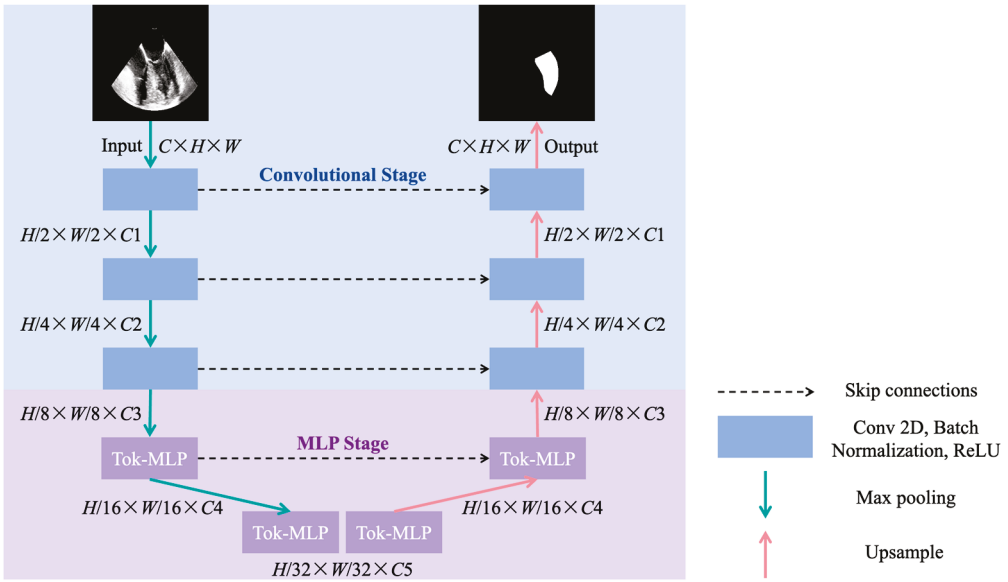


Figure 5. Structure of UNeXt [26]. (*C*, *H*, *W*) represent the channels, height, and width of the TEE image, respectively. Conv: convolutional; ReLU: rectified linear unit; MLP: multilayer perceptron; Tok-MLP: tokenized multilayer perceptron; TEE: transesophageal echocardiography.

As comparison methods, this study also selected U-Net [5], TransUNet [12], and Attention U-Net [11] for TEE left ventricular segmentation. U-Net is a CNN architecture widely used for medical image segmentation tasks. It features a U-shaped network structure composed of an encoder, a decoder, and skip connections (Figure 6). In the encoder, the input image undergoes four convolutional blocks and four max-pooling operations, extracting feature information and reducing spatial resolution. After the fifth convolutional block, the decoder performs four upsampling operations to gradually reconstruct the feature maps extracted by the encoder, restoring them to the original image resolution. The skip connections transfer feature maps from the encoder to the decoder to retain spatial information that may have been lost during downsampling.

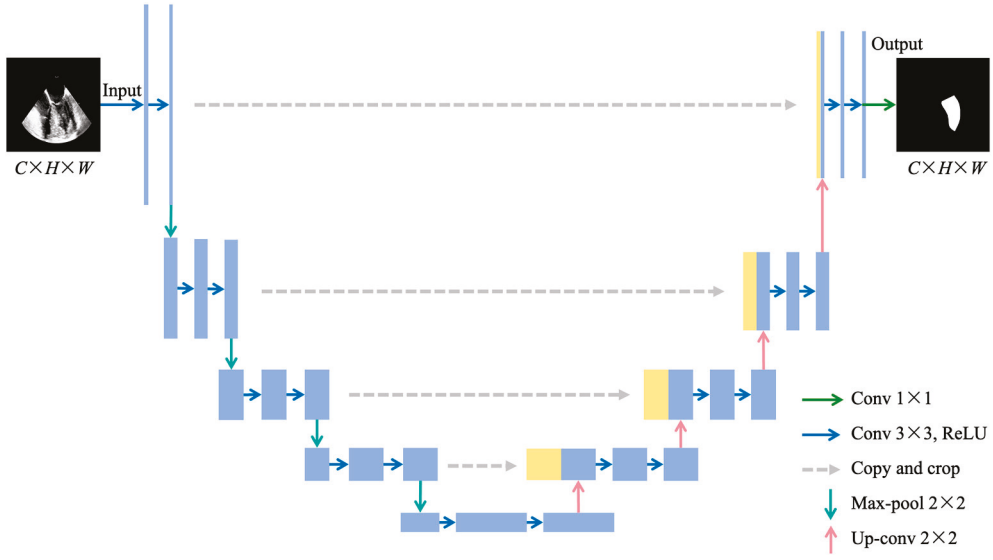


Figure 6. Structure of U-Net. (*C*, *H*, *W*) represent the channels, height, and width of the TEE image, respectively. Conv: convolutional; ReLU: rectified linear unit; TEE: transesophageal echocardiography.

TransUNet [12] builds on the U-Net framework and introduces an attention mechanism, specifically the self-attention module, to better capture global information (Figure 7).

The overall structure of TransUNet is also U-shaped, with an encoder–decoder architecture. The encoder first splits the input TEE image into fixed-size patches, flattens these patches into one-dimensional vectors, and uses a linear transformation to map each vector into a high-dimensional space, forming feature vectors. Positional encoding is then applied, adding the positional encoding vector to each feature vector to retain spatial location information. The self-attention mechanism in the Transformer layers captures global features and long-range dependencies of the feature vectors. The upsampling operations in the decoder restore the sequence data output by the encoder to the original image size, while skip connections leverage features at different scales.

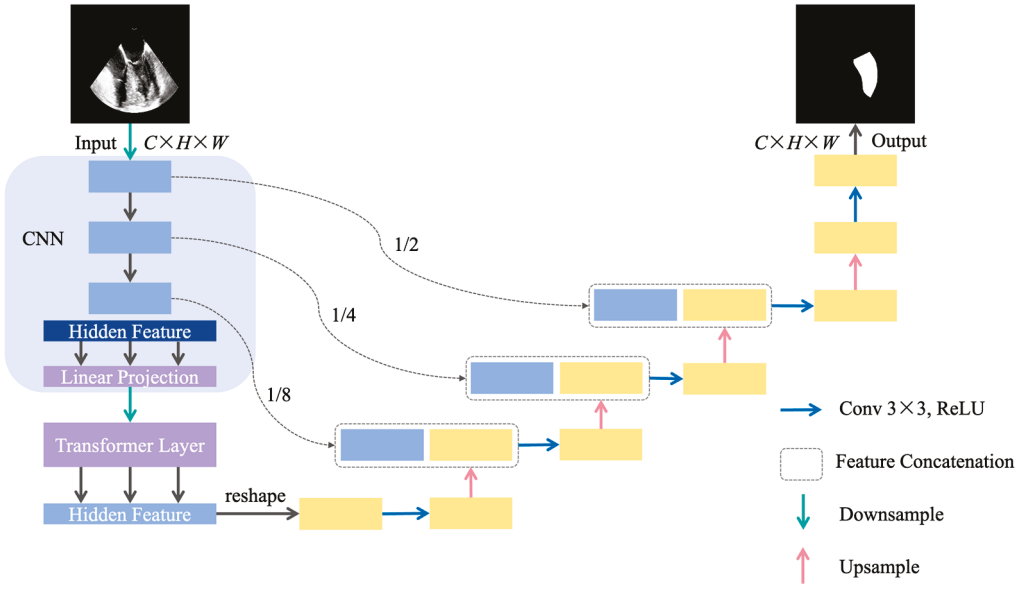


Figure 7. Structure of TransUNet [12]. (*C*, *H*, *W*) represent the channels, height, and width of the TEE image, respectively. Conv: convolutional; ReLU: rectified linear unit; TEE: transesophageal echocardiography.

Attention U-Net [11] is based on the U-Net architecture, with the addition of an attention gate to enhance the representation of features in regions of interest. The network consists of an encoder, a decoder, and skip connections (Figure 8). The encoder, similar to that of U-Net, comprises convolutional layers and max-pooling layers. An attention gate is employed within the skip connections, enabling the fusion of low-level encoder features with high-level decoder features. The generated attention weights are applied to reweight the features, suppressing irrelevant background information and emphasizing target features.

2.3. Experimental Setup

The experiment was conducted on a workstation configured with two Intel (R) Xeon (R) Gold 6132 CPUs @ 2.60 GHz, an NVIDIA TITAN RTX 24 G GPU, and 128 GB of RAM. The deep learning framework used Python (version 3.6) and PyTorch (version 1.10). In the experiments, the batch size was set to 16, and the training process was run for 100 epochs. The Adam optimizer (adaptive moment estimation) was employed, with an initial learning rate (LR) of 10^{-3} and a minimum LR of 10^{-5}. The learning rate adjustment followed a cosine annealing strategy (CosineAnnealing LR), and the momentum was set to 0.9. The input dimensions for the model were set to ($C \times H \times W$) = ($3 \times 512 \times 512$), where C, H, and W represent the number of channels, height, and width of each frame, respectively. The output dimensions were ($C \times H \times W$) = ($1 \times 512 \times 512$).

The binary cross-entropy loss (L_{BCE}) function was selected as the loss function for the experiments. Binary cross-entropy was used to measure the difference between the predicted and actual values:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)], \quad (1)$$

where N represents the total number of pixels in the input image, y_i represents the actual value of the i-th pixel, and p_i represents the predicted probability that the i-th pixel belongs to the left ventricular region.

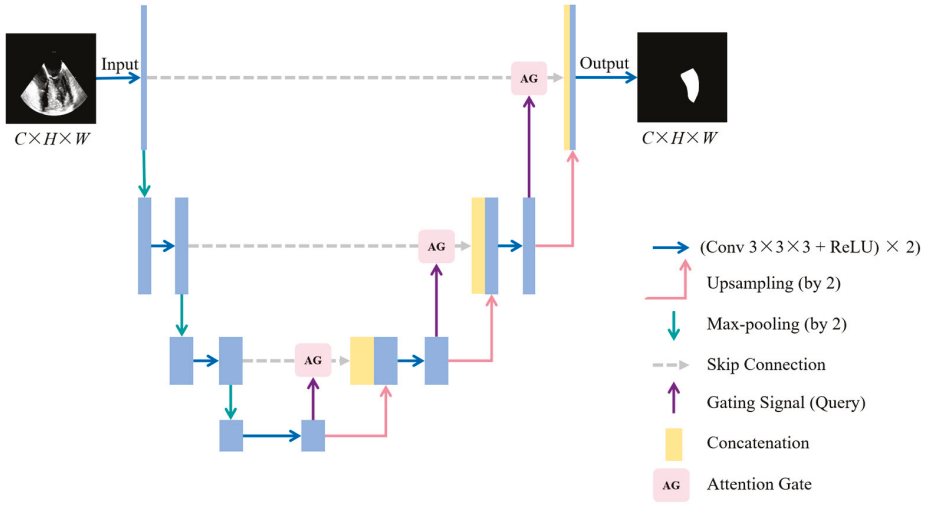


Figure 8. Structure of Attention U-Net. (*C*, *H*, *W*) indicates the channels, height, and width of the TEE image. Conv: convolutional; ReLU: rectified linear unit; TEE: transesophageal echocardiography.

2.4. Evaluation Metrics

The model's performance in predicting the left ventricular region was evaluated using the DSC, which measures the overlap between the model-predicted region and the region manually annotated by human experts. DSC is a similarity metric commonly used to compute the similarity between two regions, with values ranging from 0 to 1. The higher the DSC is, the better the segmentation performance. The DSC was calculated by

$$\mathrm{DSC}(A, B) = \frac{2|A \cap B|}{|A| + |B|}, \quad (2)$$

where A represents the left ventricular region in a TEE A4CV image predicted by the model, and B represents the left ventricular region in the same TEE A4CV image manually annotated by human experts.

The Intersection over Union (IoU) was also employed to evaluate the similarity between the left ventricular region predicted by the model and the manually annotated region provided by human experts. IoU values range from 0 to 1, with higher values indicating a closer match between the predicted and annotated regions. The IoU was calculated by

$$\mathrm{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

2.5. Statistical Analysis

The Kruskal–Wallis test [27] was used to determine whether the DSC values of UNeXt, U-Net, TransUNet, and Attention U-Net on the test set exhibited statistically significant differences. A significance level of $p < 0.05$ was considered a significant difference in DSC. Statistical analyses were conducted using IBM SPSS Statistics 27 (IBM Corp., Armonk, NY, USA).

3. Results

Figure 9 shows the loss curves for UNeXt, U-Net, TransUNet, and Attention U-Net on both the training and validation sets. The loss values of all four deep neural networks steadily decreased on the training set and eventually stabilized. Similarly, the loss curves on the validation set closely followed those of the training set, gradually decreasing and ultimately converging. Before the 60th epoch, the validation loss curves exhibited varying degrees of fluctuation across the four networks. UNeXt showed the smallest fluctuation, while Attention U-Net experienced the largest, likely due to the limited dataset size. Although the dataset contained 3000 TEE images, the number of patients included was relatively small. After the 80th epoch and through the end of training, the loss values for all three models remained stable on both the training and validation sets, indicating convergence without signs of overfitting.

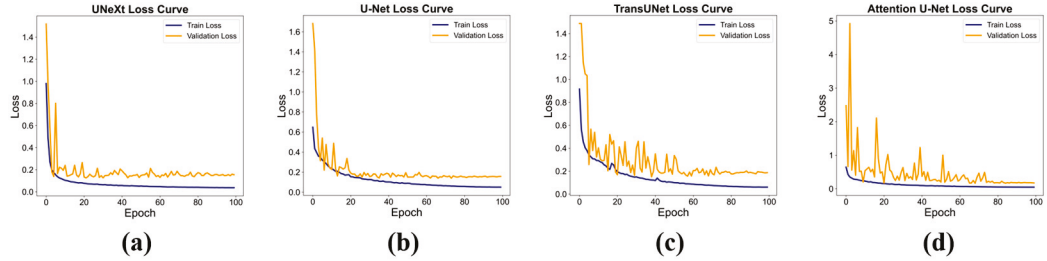


Figure 9. The loss on the training set (Train Loss) and validation set (validation loss) as a function of the training epochs for various deep learning models: UNeXt (**a**), U-Net (**b**), TransUNet (**c**), Attention U-Net (**d**).

Figure 10 illustrates the DSC curves for UNeXt, U-Net, TransUNet, and Attention U-Net on the training and validation sets. The DSC values of all four deep neural networks gradually increased on the training set and stabilized. The DSC curves on the validation set closely followed the training set, gradually increasing and showing no significant changes after reaching their peak. Before the 60th epoch, the validation DSC curves of the four networks exhibited varying degrees of fluctuation. UNeXt had the smallest fluctuation, while Attention U-Net had the largest, likely due to the limited dataset size.

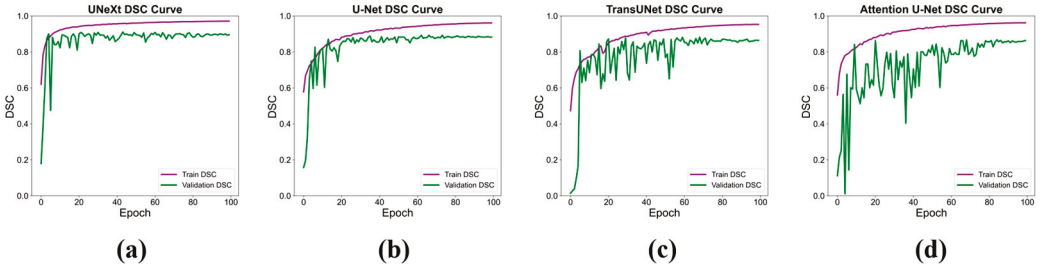


Figure 10. The DSC on the training set (Train DSC) and validation set (validation DSC) as a function of the training epochs for various deep learning models: UNeXt (**a**), U-Net (**b**), TransUNet (**c**), Attention U-Net (**d**). DSC: dice similarity coefficient.

Table 1 presents the left ventricular segmentation performance of UNeXt on the test data, compared to U-Net [5], TransUNet [12], and Attention U-Net [11]. The results showed that UNeXt achieved the highest segmentation accuracy among the four models. The Kruskal–Wallis test indicated significant differences in DSC performance on the test set among UNeXt, U-Net, TransUNet, and Attention U-Net. Pairwise comparisons indicated significant differences in DSC between Attention U-Net and UNeXt, between Attention U-Net and U-Net, between Attention U-Net and TransUNet, and between UNeXt and TransUNet. In contrast, no significant differences were identified between UNeXt and U-Net or between U-Net and TransUNet on the test set.

Table 1. Left ventricular segmentation performance by different deep learning models. DSC: dice similarity coefficient; IoU: Intersection over Union.

Deep Learning Models	DSC (%)	IoU (%)
U-Net [5]	87.76 ± 3.61	76.61 ± 7.45
TransUNet [12]	85.75 ± 5.93	77.35 ± 8.50
Attention U-Net [11]	79.98 ± 9.26	68.86 ± 11.95
UNeXt [26]	88.60 ± 2.92	77.60 ± 6.33

Figure 11 shows the left ventricular segmentation results of UNeXt, U-Net, TransUNet, and Attention U-Net on the test data. Each column represents representative TEE images from different patients. The first row shows the input TEE A4CV images, and rows two to five show the segmentation results of UNeXt, U-Net, TransUNet, and Attention U-Net, respectively. It can be observed that UNeXt's segmentation results closely match the manual annotations by human experts. All three models, U-Net, TransUNet, and Attention U-Net, tended to misidentify regions outside the left ventricle. For example, U-Net and Attention U-Net mistakenly identified areas of the right ventricle (Figure 11, third row, first, third, and fourth columns; fifth row, second and fourth columns). TransUNet showed less sensitivity in detecting the boundaries of the left ventricle (Figure 11, fourth row, third column). In contrast, UNeXt provided more accurate left ventricular identification, with better boundary delineation that more closely matched the anatomical structure of the left ventricle.

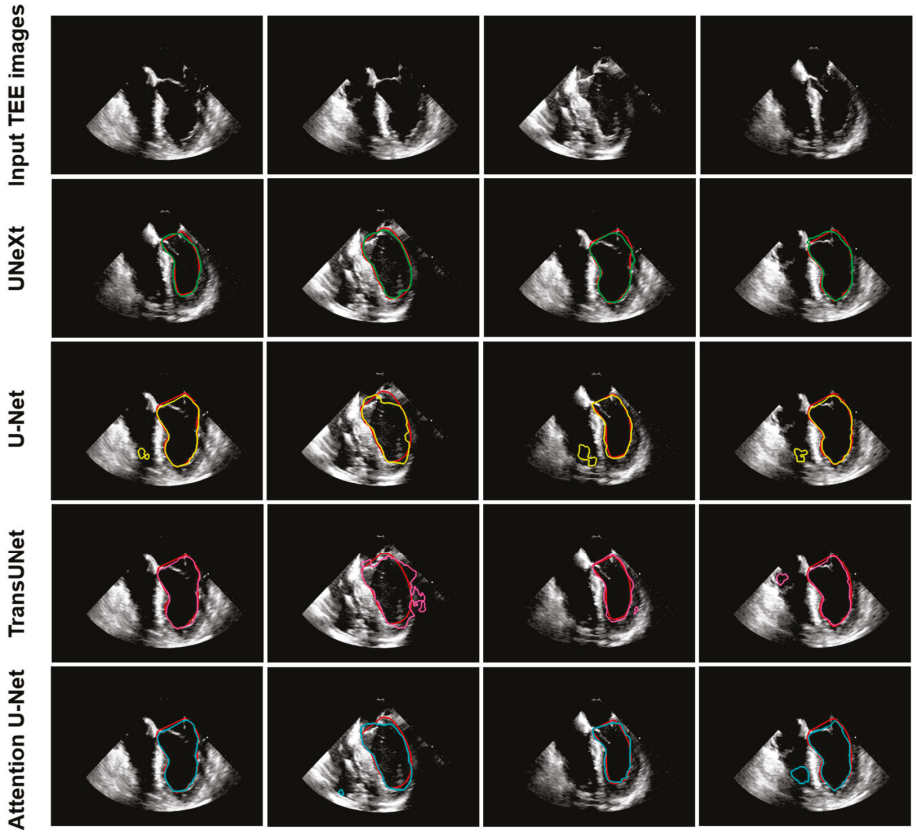


Figure 11. Results of TEE A4CV LV prediction by various deep learning models. Red contours represent manual LV segmentation as the reference standard. Green boundaries represent LV segmentation by UNeXt. Yellow contours represent LV segmentation by U-Net. Pink contours represent LV segmentation by TransUNet. Cyan contours represent LV segmentation by Attention U-Net. TEE: transesophageal echocardiography; LV: left ventricle.

Table 2 presents the parameter counts of the different models for left ventricular segmentation. The comparison shows that the UNeXt model used in this study had signifi-

cantly fewer parameters and floating point operations (FLOPs) than U-Net, TransUNet, and Attention U-Net, making it a more lightweight model.

Table 2. Parameters of different deep learning models. FLOPs: floating point operations.

Deep Learning Models	# of Model Parameters (million)	Inference Time for a Single Image (ms)	FLOPs (giga)
U-Net [5]	7.85	185.12	218.57
TransUNet [12]	79.71	209.08	272.24
Attention U-Net [11]	34.88	201.13	265.92
UNeXt [26]	1.47	141.73	2.28

4. Discussion

In this paper, a large-scale TEE dataset was constructed, and an automatic left ventricle segmentation method for the TEE A4CV based on the UNeXt deep neural network was proposed. Its performance was compared with U-Net, TransUNet, and Attention U-Net. The UNeXt deep neural network, introduced by Valanarasu et al. [26] in 2022, has the advantage of reducing network parameters while maintaining segmentation accuracy, but it had not yet been applied to TEE A4CV left ventricle segmentation. The results demonstrated that UNeXt achieved the best overall performance, with an average DSC for left ventricle segmentation of 88.60%, surpassing U-Net (87.76%), TransUNet (85.75%; $p < 0.05$), and Attention U-Net (79.98%; $p < 0.05$). Additionally, UNeXt had the smallest number of network parameters (1.47 million) and FLOPs (2.28 giga), as well as the shortest average inference time per image (141.73 ms).

TTE is the preferred cardiac ultrasound method in emergency rooms [28]. Despite its widespread use, TTE has certain limitations. Since data are collected externally on the body, it is highly dependent on the operator's skill, leading to inconsistent image quality. In contrast, TEE, where the probe is positioned in the patient's esophagus, allows for better guidance to the target area and more consistently produces high-quality images. TTE image acquisition is also affected by patient-related factors [29], such as obesity, mechanical ventilation, subcutaneous emphysema, and ongoing cardiopulmonary resuscitation (CPR), which may result in poor or unobtainable images. TEE, however, is not affected by factors like subcutaneous emphysema, fat tissue thickness, or surgical dressings, and can still capture high-quality images without interrupting CPR during cardiac arrest. Consequently, TEE is widely used in settings like cardiac surgery rooms and intensive care units, where it continuously monitors cardiac output and other heart function indicators in real time, assisting clinicians in adjusting surgical plans and making decisions during operations. Although deep learning techniques have been extensively studied and have demonstrated strong performance in automatic left ventricle segmentation on the TTE A4CV [15,17,18,20–22], there are limited studies on TEE A4CV left ventricle segmentation using deep learning [23,24]. The reason may be that constructing a large-scale TEE A4CV dataset is somewhat difficult.

As shown in Table 3, the proposed method for automatic left ventricle segmentation in the TEE A4CV based on the UNeXt deep neural network achieved an average DSC of 88.60%, which represents a 6.52% improvement compared to the Res-U-Net++ [25] used by Ahn et al. [24]. The dataset used in this study was collected from 60 patients undergoing cardiac surgery, with standard TEE A4CV videos captured from the induction of anesthesia until the conclusion of the surgery. After data preprocessing, a dataset of 3000 TEE images and corresponding labels was generated. Compared to Kang et al. [23] (120 images from nine patients) and Ahn et al. [24] (158 images), this study significantly increased both the number of patients and the volume of TEE images. However, it should be noted that the image datasets used in the studies shown in Table 3 were different, so the DSC values of different models may not be directly compared. The size and diversity of the dataset are crucial factors that influence the generalizability of the trained models. The small size of a dataset may lead to overfitting, whereas a larger and more varied dataset provides a wider range of training cases, which helps the model learn more robust features that are

applicable to different situations. In this sense, the small dataset used in Kang et al. [23] and Ahn et al. [24] may limit the reliability of their models.

Table 3. The comparison of LV segmentation performance among different methods using TEE data. LV: left ventricle; TEE: transesophageal echocardiography; DSC: dice similarity coefficient.

Authors	Year	Deep Learning Models	Data	DSC (%)
Kang, et al. [23]	2023	Variation of U-Net [5]	128 TEE images of 9 patients	89.90
Ahn, et al. [24]	2023	Res-U-Net++ [25]	158 TEE images (number of patients not described)	82.08
Ours	2024	UNeXt [26]	3000 TEE images of 60 patients	88.60

Although the image quality of TEE is generally clearer than that of TTE, various factors, such as patient characteristics, equipment, technical skill, and environmental conditions, can still lead to challenges. These include strong echo interference from muscle tissue, unclear TEE images, and the presence of artifacts, which posed challenges in this study.

Figure 12 presents several challenging examples of left ventricle segmentation in TEE images. Each column represents TEE images containing different segmentation challenges. The first row shows the original TEE images input into the network models, while rows two to five display the left ventricle segmentation results from the trained UNeXt, U-Net, TransUNet, and Attention U-Net models, respectively. The boundary of the left ventricle is composed of muscular tissue, which appears as a strong echo in the ultrasound image. However, when foreign objects (lesions) are present above the left ventricle, such as in the left atrium, these also produce strong echoes in the ultrasound image. In this scenario, all four models incorrectly identified the boundary of the foreign object as the left ventricle boundary, resulting in inaccurate segmentation (Figure 12, first column). The papillary muscles within the left ventricle also exhibit strong echoes in the ultrasound image, which similarly interferes with the segmentation of the left ventricle. As shown in Figure 12 (second column), UNeXt was least affected by this interference, whereas U-Net, TransUNet, and Attention U-Net mistakenly segmented the papillary muscle as part of the left ventricle boundary, resulting in a significant deviation from the manually annotated contour. When facing incomplete left ventricle edges, U-Net and TransUNet produced segmentation results that, while not entirely accurate, closely matched the manually annotated boundaries, whereas the segmentation results of UNeXt and Attention U-Net differed more significantly (Figure 12, third column). Lastly, during ultrasound scanning, technical factors such as poor contact between the probe and the patient or insufficient coupling gel can result in artifacts, seen as low-echo bands extending from the near field to the far field in the ultrasound image. In the presence of artifacts, U-Net, UNeXt, and Attention U-Net were only slightly affected, whereas TransUNet was more sensitive to their presence (Figure 12, fourth column).

It should be noted that the left ventricle segmentation in TEE images is not the final goal in clinical settings. However, analyzing the cardiac function indices, such as the ejection fraction and cardiac output, relies on accurate segmentation of the left ventricle. Although end-to-end deep neural networks may be used to directly predict these indices, the results may lack clinical interpretability. Segmenting the left ventricle and analyzing the cardiac function indices based on the segmentation may be more in line with the clinical workflow. When each TEE A4CV image in a cardiac cycle was segmented using the proposed method, the dynamic variation in the area of the segmented left ventricle images could be used to localize the end-diastolic and end-systolic frames, which could be used to calculate the end-diastolic volume (EDV) and end-systolic volume (ESV) using methods such as the Simpson bi-plane method. The calculated EDV and ESV could be used to compute the left ventricular ejection fraction or cardiac output (together with heart rate).

In sum, the proposed method may be used for fast, automatic left ventricular segmentation in TEE A4CV images, facilitating dynamic assessment of left ventricle function.

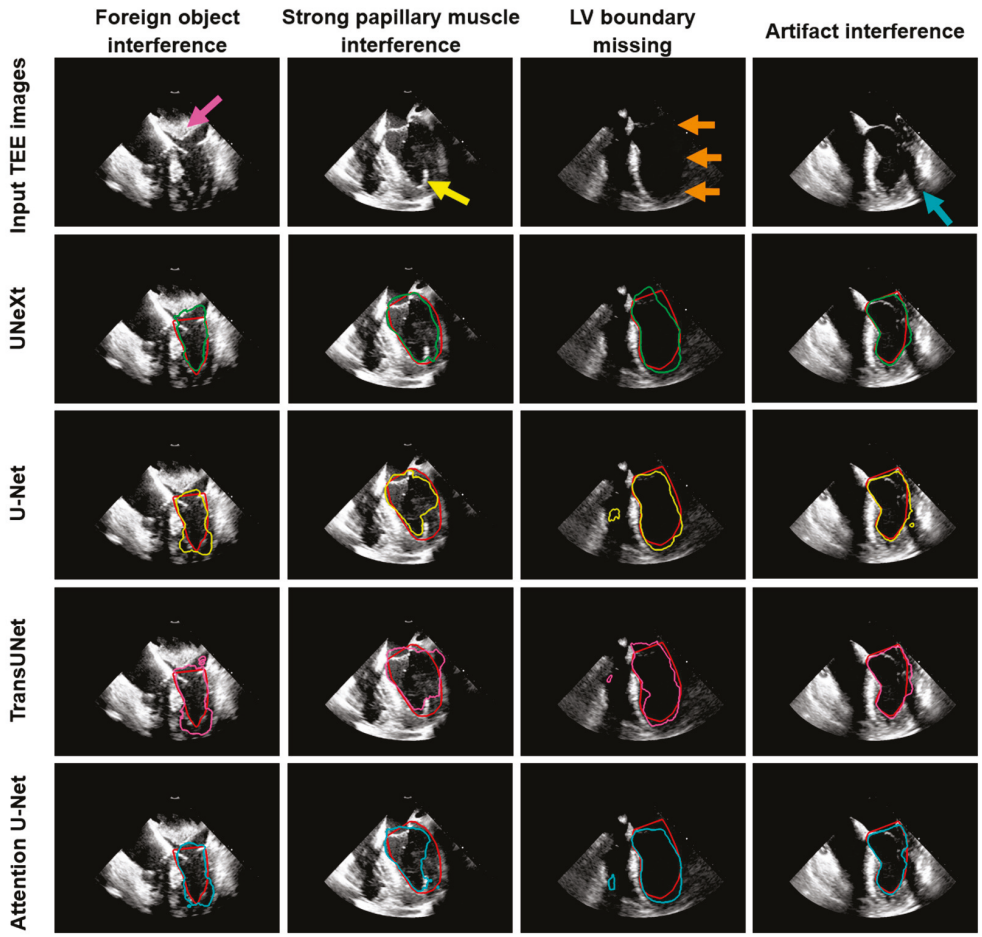


Figure 12. Representative LV segmentation using different deep learning models in TEE images with challenging cases: foreign object interference (column 1), strong papillary muscle interference (column 2), LV boundary missing (column 3), artifact interference (column 4). Red contours represent the manually annotated LV, serving as the ground truth. Green boundaries stand for LV segmentation by UNeXt. Yellow contours represent LV segmentation by U-Net. Pink boundaries indicate LV segmentation by TransUNet. Cyan contours indicate LV segmentation using Attention U-Net. Pink arrows indicate the foreign object. Yellow arrows indicate the papillary muscle. Orange arrows correspond to a missing LV boundary. Cyan arrows correspond to the artifact. LV: left ventricle; TEE: transesophageal echocardiography.

This study has limitations. First, although the constructed dataset contains 3000 TEE images, the number of patients remains relatively small. Second, while the UNeXt network was applied to the task of left ventricle segmentation on the TEE A4CV, the DSC may be increased by improving the network structure. In future work, more patients can be included and improvement on the UNeXt network (e.g., EnsUNet [30]) can be considered to further improve the model performance.

5. Conclusions

In this study, we constructed a large-scale TEE A4CV dataset and proposed an automatic segmentation method for the left ventricle on the TEE A4CV based on the UNeXt deep neural network, which was compared with U-Net, TransUNet, and Attention U-Net networks. A dataset consisting of 3000 TEE images from 60 patients and their corresponding labels was constructed, and the three deep neural networks were trained and validated.

The results indicated that the UNeXt network yielded an average DSC of 88.60% on the test set, outperforming U-Net (87.76%), TransUNet (85.75%; $p < 0.05$), and Attention U-Net (79.98%; $p < 0.05$). Additionally, the UNeXt model had a smaller number of parameters and FLOPs (1.47 million, 2.28 giga) and shorter average inference times per image (141.73 ms) than U-Net (185.12 ms), TransUNet (209.08 ms), and Attention U-Net (201.13 ms). The average IoU of UNeXt (77.60%) was also higher than that of U-Net (76.61%), TransUNet (77.35%), and Attention U-Net (68.86%). The proposed method may serve as a new deep learning approach for automatic segmentation of the left ventricle in TEE A4CV images, with promising segmentation accuracy and reduced inference time.

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

Funding: This work was supported in part by National High Level Hospital Clinical Research Funding, Beijing, China (No. 2022-PUMCH-A-149).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Review Committee of the Peking Union Medical College Hospital (protocol code K2616 and date of approval 9 December 2022).

Informed Consent Statement: Informed consent was waived for this retrospective study.

Data Availability Statement: Data are unavailable due to ethical restrictions.

Acknowledgments: The authors would like to thank the anonymous reviewers for their constructive comments and suggestions.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Chen, C.; Qin, C.; Qiu, H.; Tarroni, G.; Duan, J.; Bai, W.; Rueckert, D. Deep Learning for Cardiac Image Segmentation: A Review. *Front. Cardiovasc. Med.* **2020**, *7*, 25. [CrossRef] [PubMed]
2. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
3. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
4. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
5. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
6. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018; Proceedings 4. pp. 3–11.
7. Isensee, F.; Petersen, J.; Klein, A.; Zimmerer, D.; Jaeger, P.F.; Kohl, S.; Wasserthal, J.; Koehler, G.; Norajitra, T.; Wirkert, S. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv* **2018**, arXiv:1809.10486.
8. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.-W.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1055–1059.
9. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.

10. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 22018; pp. 3–19.
11. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
12. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
13. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
14. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 22–31.
15. Liu, X.; Fan, Y.; Li, S.; Chen, M.; Li, M.; Hau, W.K.; Zhang, H.; Xu, L.; Lee, A.P.-W. Deep learning-based automated left ventricular ejection fraction assessment using 2-D echocardiography. *Am. J. Physiol.-Heart Circ. Physiol.* **2021**, *321*, H390–H399. [CrossRef] [PubMed]
16. Leclerc, S.; Smistad, E.; Pedrosa, J.; Østvik, A.; Cervenansky, F.; Espinosa, F.; Espeland, T.; Berg, E.A.R.; Jodoin, P.-M.; Grenier, T. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. *IEEE Trans. Med. Imaging* **2019**, *38*, 2198–2210. [CrossRef] [PubMed]
17. Moradi, S.; Oghli, M.G.; Alizadehasl, A.; Shiri, I.; Oveisi, N.; Oveisi, M.; Maleki, M.; Dhooge, J. MFP-Unet: A novel deep learning based approach for left ventricle segmentation in echocardiography. *Phys. Medica* **2019**, *67*, 58–69. [CrossRef] [PubMed]
18. Zeng, Y.; Tsui, P.-H.; Pang, K.; Bin, G.; Li, J.; Lv, K.; Wu, X.; Wu, S.; Zhou, Z. MAEF-Net: Multi-attention efficient feature fusion network for left ventricular segmentation and quantitative analysis in two-dimensional echocardiography. *Ultrasonics* **2023**, *127*, 106855. [CrossRef] [PubMed]
19. Ouyang, D.; He, B.; Ghorbani, A.; Yuan, N.; Ebinger, J.; Langlotz, C.P.; Heidenreich, P.A.; Harrington, R.A.; Liang, D.H.; Ashley, E.A. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **2020**, *580*, 252–256. [CrossRef] [PubMed]
20. Zhu, Y.; Xiong, C.; Zhao, H.; Yao, Y. SAM-Att: A Prompt-free SAM-related Model with an Attention Module for Automatic Segmentation of the Left Ventricle in Echocardiography. *IEEE Access* **2024**, *12*, 50335–50346. [CrossRef]
21. Lal, S. TC-SegNet: Robust deep learning network for fully automatic two-chamber segmentation of two-dimensional echocardiography. *Multimed. Tools Appl.* **2024**, *83*, 6093–6111. [CrossRef] [PubMed]
22. Wan, Y.; Li, D.; Li, Z.; Bu, J.; Tong, M.; Luo, R.; Yue, B.; Yu, S. A Semi-supervised Four-Chamber Echocardiographic Video Segmentation Algorithm Based on Multilevel Edge Perception and Calibration Fusion. *Ultrasound Med. Biol.* **2024**, *50*, 1308–1317. [CrossRef] [PubMed]
23. Kang, S.; Kim, S.J.; Ahn, H.G.; Cha, K.-C.; Yang, S. Left ventricle segmentation in transesophageal echocardiography images using a deep neural network. *PLoS ONE* **2023**, *18*, e0280485. [CrossRef] [PubMed]
24. Ahn, H.; Kim, S.J.; Kang, S.; Han, J.; Hwang, S.O.; Cha, K.-C.; Yang, S. Ventricle tracking in transesophageal echocardiography (TEE) images during cardiopulmonary resuscitation (CPR) using deep learning and monogenic filtering. *Biomed. Eng. Lett.* **2023**, *13*, 715–728. [CrossRef] [PubMed]
25. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Johansen, D.; De Lange, T.; Halvorsen, P.; Johansen, H.D. Resunet++: An advanced architecture for medical image segmentation. In Proceedings of the 2019 IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, 9–11 December 2019; pp. 225–2255.
26. Valanarasu, J.M.J.; Patel, V.M. Unext: Mlp-based rapid medical image segmentation network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer Nature: Cham, Switzerland, 2022; pp. 23–33.
27. Kruskal, W.H.; Wallis, W.A. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621. [CrossRef]
28. Arntfield, R.; Pace, J.; Hewak, M.; Thompson, D. Focused transesophageal echocardiography by emergency physicians is feasible and clinically influential: Observational results from a novel ultrasound program. *J. Emerg. Med.* **2016**, *50*, 286–294. [CrossRef] [PubMed]
29. Jaidka, A.; Hobbs, H.; Koenig, S.; Millington, S.J.; Arntfield, R.T. Better with ultrasound: Transesophageal echocardiography. *Chest* **2019**, *155*, 194–201. [CrossRef] [PubMed]
30. Laouamer, I.; Aiadi, O.; Kherfi, M.L.; Cheddad, A.; Amirat, H.; Laouamer, L.; Drid, K. EnsUNet: Enhancing brain tumor segmentation through fusion of pre-trained models. In *International Congress on Information and Communication Technology*; Springer Nature: Singapore, 2024; pp. 163–174.

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

From Binary to Multi-Class Classification: A Two-Step Hybrid CNN-ViT Model for Chest Disease Classification Based on X-Ray Images

Yousra Hadhoud [1], Tahar Mekhaznia [1], Akram Bennour [1,*], Mohamed Amroune [1], Neesrin Ali Kurdi [2], Abdulaziz Hadi Aborujilah [3] and Mohammed Al-Sarem [4]

[1] LAMIS Laboratory, Larbi Tebessi University, Tebessa 12002, Algeria; yousra.hadhoud@univ-tebessa.dz (Y.H.); tahar.mekhaznia@univ-tebessa.dz (T.M.); mohamed.amroune@univ-tebessa.dz (M.A.)
[2] College of Computer Science and Engineering, Taibah University, Medina 41477, Saudi Arabia; nkordi@taibahu.edu.sa
[3] Department of Management Information Systems, College of Commerce & Business Administration, Dhofar University, Salalaha 211, Oman; aborujilah@du.edu.om
[4] Department of Information Technology, Aylol University College, Yarim 547, Yemen; mohsarem@gmail.com
* Correspondence: akram.bennour@univ-tebessa.dz

Abstract: Background/Objectives: Chest disease identification for Tuberculosis and Pneumonia diseases presents diagnostic challenges due to overlapping radiographic features and the limited availability of expert radiologists, especially in developing countries. The present study aims to address these challenges by developing a Computer-Aided Diagnosis (CAD) system to provide consistent and objective analyses of chest X-ray images, thereby reducing potential human error. By leveraging the complementary strengths of convolutional neural networks (CNNs) and vision transformers (ViTs), we propose a hybrid model for the accurate detection of Tuberculosis and for distinguishing between Tuberculosis and Pneumonia. Methods: We designed a two-step hybrid model that integrates the ResNet-50 CNN with the ViT-b16 architecture. It uses the transfer learning on datasets from *Guangzhou Women's and Children's Medical Center* for Pneumonia cases and datasets from *Qatar and Dhaka (Bangladesh) universities* for Tuberculosis cases. CNNs capture hierarchical structures in images, while ViTs, with their self-attention mechanisms, excel at identifying relationships between features. Combining these approaches enhances the model's performance on binary and multi-class classification tasks. Results: Our hybrid CNN-ViT model achieved a binary classification accuracy of 98.97% for Tuberculosis detection. For multi-class classification, distinguishing between Tuberculosis, viral Pneumonia, and bacterial Pneumonia, the model achieved an accuracy of 96.18%. These results underscore the model's potential in improving diagnostic accuracy and reliability for chest disease classification based on X-ray images. Conclusions: The proposed hybrid CNN-ViT model demonstrates substantial potential in advancing the accuracy and robustness of CAD systems for chest disease diagnosis. By integrating CNN and ViT architectures, our approach enhances the diagnostic precision, which may help to alleviate the burden on healthcare systems in resource-limited settings and improve patient outcomes in chest disease diagnosis.

Keywords: chest diseases; X-rays; convolutional neural networks; vision transformers; classification

1. Introduction

Chest diseases, also known as thoracic diseases, encompass a wide range of medical illnesses that affect organs and structures inside the chest cavity. Millions of people around the world suffer from chest disorders which cause discomfort and hinder daily activities [1]. There are more than 200 types of chest diseases, including Pneumonia, Tuberculosis, pleural effusion, cardiomegaly, lung cancer, etc. According to the *World Health Organization* (WHO), the *Chronic Obstructive Pulmonary Disease* (COPD) remains globally the third leading cause

of death [2]. It resulted in 3.5 million deaths in 2019 ("Chronic obstructive pulmonary disease (COPD)", 2023). In 2023, 1.25 million people worldwide died from Tuberculosis (TB) [3]. In 2022, the United States reported 8300 TB cases and over 13 million people with latent TB infection [4]. Similarly, other countries faced significant TB burdens. India has approximately 2.95 million new TB cases, while Indonesia and China reported around 824,000 and 746,000 cases, respectively. South Africa and Nigeria had about 328,000 and 452,000 new cases, respectively [5]. Pneumonia claims a child every 43 s, making it the most significant infectious cause of death in this age group [6].

Diagnosing chest diseases typically entails physical examination combined with laboratory tests or biopsies; however, imaging studies, particularly chest X-rays (CXRs), are the most commonly used as diagnostic tools due to its fastness. Examining CXR images remains a challenge in regard to its subjective nature of interpretation, poor image quality, and low lighting conditions. Moreover, the similar patterns of many thoracic disorders lead to frequent misinterpretations. These challenges emphasize the need for developing *Computer-Aided Diagnosis* (CAD) systems for the early detection and accurate diagnosis of chest diseases.

CAD systems have proven to be highly effective in detecting and classifying chest diseases. With their capacity in extracting complex features, they excel at performing precise analyses of medical images, leading to new avenues for improving patient care and diagnostic accuracy. Consequently, CAD systems have become crucial tools in contemporary healthcare [7]. *Deep learning*, particularly (CNNs), is predominantly used for disease detection and classification. They achieve state-of-the-art results due to their robustness in handling variations in CXR images and their ability to capture hierarchical patterns and structures. However, CNNs primarily analyze the correlation between spatial adjacent pixels, making it difficult to determine relationships between distant pixels. Additionally, the convolution operation causes information loss and leads to overfitting, resulting in poor generalization. Accordingly, recent studies have focused on transformers [8], especially vision transformers (ViTs) [9]. Their architecture consists of a self-attention mechanism and a feed-forward neural network. The related key idea is to treat a given image as a series of patches, similar to how a sentence is treated in NLP. The self-attention mechanism used by ViTs allows for the comprehension of relationships between different regions of an image. *Transformers* are able to process sequences of data, which is advantageous in cases where the context of the whole image is important. In their first appearance, transformers were used in the field of *Natural Language Processing* (NLP), such as text generation, translation, etc. [10]. The groundbreaking results achieved in NLP prompted researchers to take advantage of this success in other fields, such as computer vision.

In this study, we focused on two critical chest diseases, namely Pneumonia and Tuberculosis. Both diseases appear in various forms, impacting different parts of lungs. They manifest with overlapping features in CXR images as pulmonary infiltrates, cavitation air bronchograms, and pleural effusion. This inhibits the distinction between them, hence leading to ambiguous diagnosis, especially in settings with a shortage of trained radiologists. On the other hand, CXR images are more cost-effective compared to the other modalities of early detection and treatment. Also, X-rays are non-invasive compared to laboratory tests such as *Polymerase Chain Reaction* (PCR) and biopsies, which seem painful. They have a risk of complication and call for specific knowledge and equipment. For this reason, we propose a hybrid model that benefits from the power of both CNNs and ViTs. It operates in a parallel manner, ensuring the early and accurate detection of diseases. CNNs are well known for their capacity to evaluate visual data by identifying hierarchical patterns and structures within images. ViTs, on the other hand, allow for capturing relationships between different regions of an image, leading the user to easily understand the global context of the image.

Hence, the main contribution of this study is illustrated by the following:

- The application of feature-level fusion to interpret complex features of X-ray images, especially overlapping characteristics.

- Supporting the parallelism through the application of a refined attention mechanism to both ResNet-50 and ViT-b16 streams after the feature extraction step. This attention mechanism enables the model to prioritize features consistently across diverse samples.
- Providing an effective solution to the significant challenge of data imbalance by combining oversampling and strategic augmentation techniques.

The rest of this paper provides a set of related works in Section 2, followed by a presentation of the proposed approach in Section 3. The experimental process is shown in Section 4, along with an analysis and discussion in Sections 5 and 6, respectively. Finally, we conclude our work in Section 7, specifying future perspectives.

2. Literature Review

In this section, we explore various advancements and methodologies employed in recent studies focusing on CNNs, ViTs, and hybrid models that leverage the strength of both architectures.

2.1. Convolutional Neural Networks

CNNs have gained significant attention for medical image classification tasks. Several studies have since built upon this approach. The authors of [11] developed a lightweight CNN named ChestX-ray 6 for the detection of multiple thoracic diseases, including COVID-19, Pneumonia, pleural effusion, lung opacity, and cardiomegaly. It uses a combination of various datasets. The issue of imbalanced data was addressed with data augmentation. The model was first used in the classification of six kinds of diseases and then as a pretrained model for a binary classification of normal and Pneumonia CXR images. It achieved an accuracy of 80% for the first task and an accuracy of 97.94% with a recall of 98% for the binary classification. Similarly, Ref. [12] proved that their proposed lightweight CNN for epidemic lung disease prediction outperformed six pretrained CNNs, achieving an accuracy of 98.56%. In a previous study [13], we proposed a hybrid model based on a shallow CNN and XGboost model to detect and differentiate between viral and bacterial Pneumonia, achieving an accuracy of 96% for binary classification and 87% for multi-class classification.

A fine-tuned VGG-19 designed by [14] was used for multi-class chest disease classification, assessing lung cancer, lung opacity, Pneumonia, Tuberculosis, and COVID-19. It achieved an accuracy of 93.75%. The authors of [15] fine-tuned ResNet-50, VGG-16, and Inception-V3 to detect COVID-19 from CXR images using multiple combined datasets. The best performance was attained by VGG-16 with an accuracy of 98.30%. However, In Ref. [16], Authors classified COVID-19, non-COVID-19, viral Pneumonia, bacterial Pneumonia, and normal CXR images using a fine-tuned AlexNet model, which yield an accuracy of 93.42%. In [17], the authors trained a CNN model by applying transfer learning to differentiate between viral Pneumonia, bacterial Pneumonia and CXR images. This model achieved an accuracy of 92.8%. H. Malik et al. in [18] classified ten different chest diseases by proposing a novel fusion model with deep convolutional neural networks (DCNNs), following three essential steps: the segmentation of CXR images to isolate the lung region using the Info MGAN network [19], feature extraction using a hand-crafted technique, and the fusion of these feature, which were then fed to the DCNNs. Various machine learning models were used as the last layer of the DCNNs, achieving an accuracy of 98.20%.

In [20], authors identified chest abnormalities using the Vin Dr-CXR dataset [21] by proposing a model based on YOLO V5 for anomaly localization and ResNet50 for the classification, achieving a performance of 80% in terms of Mean Average Precision (MAP) and an F1-score of 76%. In [22], the authors proposed two models for COVID-19 detection using both CXR and Computer Tomography (CT) images. The first model stacked six pretrained models, and the second was a self-designed model based on ECA-Net and EfficientNet V2 using 10-fold cross validation. The results demonstrated that the difference between the performance of the two models was not significant, with an accuracy tending towards 99%.

2.2. Vision Transformers

Vision transformers are a type of neural networks, originally designed for NLP tasks, subsequently developed to benefit from its performance in computer-vision tasks such as classification [23], object localization [24] and segmentation [25]. ViTs leverage the self-attention mechanism, allowing the model to assess a relative performance of different parts within the input data and focus on critical features across the entire image. This task is achieved by splitting the image into smaller patches; each represents a small region of the original image. In so doing, the self-attention calculates the score for each pair of patches. The model, by focusing on informative patches, highlights the connection between a suspicious mass and surrounding healthy tissue in a medical X-ray.

In the field of medical images, ViTs have been used for many clinical purposes, including disease diagnosis and anomaly localization. The work in [26] used a new proposed pretrained Vision Outlooker model [27] (VOLO) for COVID-19 diagnosis, using a large dataset obtained by combining two available datasets [28,29]. The approach achieved an accuracy of 99.67%. For the same task, another fine-tuned ViT model was developed by M. Chetoui et al. [30]; they used the SIIM-FISABIO-RSNA COVID-19 dataset [31], achieving a performance of 96% in terms of sensitivity and specificity.

For multi-class classification, Krishan et al. [32] applied ViTs to classify COVID-19, Pneumonia, and normal CXR images, achieving an accuracy of 97.61%. The authors in [33] applied a ViT model for the classification of CXR images to normal lungs, diseased lungs, and COVID-19-affected lungs. To assess the impact of patches' size on model performance, images were divided into various sizes, and a top accuracy of 95.36% was reached using a patch size of 32×32 pixels.

In [34], a deep model based on a pretrained swin-transformer was proposed for the multi-label classification of thoracic diseases using the chest X-ray-14 dataset [35]. It achieved a score of 81% in terms of Area Under the Curve (AUC). In [36], the authors applied a pretrained ViT model to the ChexPert dataset [37] and demonstrated performance gains above pretrained CNN-based VGG16 and ResNet architectures.

Moreover, Nasser, A.A. et al. [38] classified chest diseases using the merged datasets ChexPert [37] and VinDr-CXR [21] with a two-step approach; the first step involved classifying CXR images into three different classes, namely lung diseases, heart diseases, and normal. The second step involved classifying seven lung and heart pathologies as normal or abnormal by implementing two deep learning models (an ensemble deep CNN and a modified ViT). The modified ViT recorded a better performance with an AUC of 95.13% for the first step, and an average AUC of 99.26% and 99.57% for heart diseases and lung diseases, respectively, in the second step. In [39], the authors employed a ViT model to detect COVID-19 after segmenting the lung region with a U-Net model. The model achieved an AUC more than 95%. In [40], the authors adapted a ViT model to analyze CXR images by introducing an attention region selection module in the encoder layer of ViT. This technique enables the model to focus on most relevant regions. The model achieved an accuracy of 83.4% and a sensitivity of 86.3%. Another model named PneuNet was developed in [41] for Pneumonia and COVID-19 diagnosis. The model applies multi-head attention on channel patches rather than feature patches, achieving an accuracy of 94.96%.

2.3. Hybrid Models

Some works aim to benefit from the advantages of ViTs and CNNs by using a combination of them. The authors in [42] proposed an IEViT (Input-Enhanced Vision Transformer) model by introducing a CNN block in parallel with the ViT block to classify CXR images using four datasets containing different pathologies (COVID-19, Tuberculosis, Pneumonia). The CNN block embeds the input image and feeds it repeatedly to each transformer encoder layer by concatenating the image embedded to the input of each encoder layer. The experiments proved that the IEViT outperformed the original ViT and its variants in all the datasets, with a recall ranging between 93.5% and 100%.

Another hybrid model was proposed by [43] to identify Pneumonia based on CXR images. For the feature extraction step, an ensemble of pretrained CNNs was used, which was then fed to the transformer for the identification phase, achieving a best accuracy of 98.01%. In contrast, Ref. [27] suggested a hybrid model based on EfficientNet and ViT for the detection of Tuberculosis using a large dataset curated by combining several public datasets, recording an accuracy of 97.72%. To ensure the model's performance, each of the EfficientNet and ViT models were tested separately, where the hybrid model achieved a better result.

In [44], the authors developed a model called VGG16-CoordAttention for automatic Tuberculosis detection, which entails adding a coordinated attention mechanism to the VGG16 architecture, achieving an accuracy of 92.73%. In a slightly different manner, Authors in [45] worked on TB detection using lateral CXR images obtained from [37,46] and thus, instead of frontal images. They built a model based on CNN and ViT which reached a performance of 91% in terms of the Matthews Correlation Coefficient (MCC). The authors in [47] developed a ResfEANet model for Tuberculosis detection based on a shallow ResNet. The model was trained from scratch without applying transfer learning. It achieved an accuracy of 97.59%. Additionally, a novel hybrid convolutional–transformer model named HydraViT was proposed by [48] for multi-classification. The model achieved an AUC of 83.8% across all the 14 pathologies.

The previous studies and numerous other existing works highlighted the impressive advancements in the use of CNNs and ViTs for thoracic disease prediction and classification. CNNs have been demonstrated to be effective in detecting abnormalities, achieving high accuracy rates. ViTs have provided a new dimension in medical imaging by capturing global context and long-range dependencies in images. Most of these studies focus on the use of CNNs, or ViTs separately, or using them in a sequential manner. The standalone CNNs [11,12] may have trouble in detecting long-range patterns and overlapped features. Pure ViTs [30] address this issue. However, the combination of both of them resulted in better performance, especially in the multi-class classification. The sequential hybrid model proposed by [43] processes features in a hierarchical manner, which may lead to the loss of valuable information in the transition between architectures. These limitations highlight our use of a parallel habitation technique and feature-level fusion to accurately distinguish between two mimic chest diseases, Tuberculosis and Pneumonia, through CXR images. Tables 1–3 show a comparative study of the performance of CNN and ViT-based approaches according to the existing literature.

Table 1. Performance comparison of CNN-based approaches for chest X-ray classification.

Authors	Approach	Pathology	Classification Method	Performance
[11]	LightWeight CNN	COVID-19, Pneumonia Pleural effusion, Lung opacity, Cardiomegaly	Multi-class classification	Accuracy = 80%
		Pneumonia, Normal	Binary classification	Accuracy = 97.94%
[12]	LightWeight CNN	Lung opacity, Fibrosis, Pneumonia, COVID-19, Tuberculosis	Multi-class classification	Accuracy = 98.56%
[18]	INFO-MGAN for segmentation + Novel handcrafted features with DCNNs	COVID-19, Lung cancer, Atlectasis, Consolidation lung, Tuberculosis, Pneumothorax, Edema, Pneumonia, Pleural thickening	Multi-class classification	Accuracy = 98.20%

Table 1. *Cont.*

Authors	Approach	Pathology	Classification Method	Performance
[20]	Yolo V5 for anomalies localization ResNet-50 for classification	Chest abnormalities	Multi-class classification	F1-score = 76%
[22]	Six stacked pretrained CNNs	COVID-19 with X-ray images and Computer Tomography (CT scan) images	Binary classification	Accuracy ≈ 99%
	ECA-Net + EfficientNet V2			
[17]	Fine-tuned CNN	Viral Pneumonia, Bacterial Pneumonia, Normal	Binary classification	Accuracy = 92.8%
[14]	Fine-tuned VGG-19	Lung cancer, Lung opacity, COVID-19, Pneumonia, Tuberculosis	Multi-class classification	Accuracy = 93.75%

Table 2. Performance comparison of ViT-based approaches for chest X-ray classification.

Authors	Approach	Pathology	Classification Method	Performance
[27]	Volo	COVID-19	Binary classification	Accuracy = 99.67%
[30]	Fine-tuned ViT	COVID-19	Binary classification	Accuracy = 96%
[32]	ViT-b32	COVID-19, Pneumonia, Normal	Multi-class classification	Accuracy = 97.61%
[33]	ViT-b32	Normal Lung, Diseased Lung, COVID-19	2-Step Binary classification	Accuracy = 95.36%
[34]	Pretrained Swin Transformer	14 chest diseases from the chest-14 dataset	Multi-class classification	AUC = 81%
[36]	Pretrained ViT	Pneumonia, Normal	Binary classification	F1-score = 86%
		14 chest diseases from the Chexpert dataset (Irvin et al., 2019 [37])	Multi-class classification	F1-score = 59%
[38]	Modified ViT	Heart diseases and lung diseases	Multi-class classification	AUC = 95.13%
			Binary classification	AUC ≈ 99%
[39]	U-Net for lung segmentation ViT for classification	COVID-19	Binary classification	AUC > 95%
[40]	Adapted ViT	14 chest diseases from chest-14 dataset [49] (Wang, 2017)	Multi-class classification	Accuracy = 83.4%
[25]	PneuNet	Pneumonia, COVID-19	Multi-class classification	Accuracy = 94.96%

Table 3. Comparative analysis of hybrid CNN-ViT models for chest disease detection.

Authors	Approach	Pathology	Classification Method	Performance
[42]	Enhanced Vit: Parallel CNN and ViT	COVID-19, Tuberculosis, and Pneumonia, using different datasets	Multi-class classification	93.5% $\leq$ Recall $\leq$ 100%
[43]	Ensemble CNNs + ViT	Pneumonia	Binary classification	Accuracy = 98.01%
[28]	EfficientNet + ViT	Tuberculosis	Binary classification	Accuracy = 97.72%
[44]	VGG-16 + CoorAttention	Tuberculosis, using merged datasets	Binary classification	Accuracy = 92.73%
[45]	CNN-ViT	Tuberculosis, using lateral and frontal CXR images	Binary classification	Accuracy = 91%
[47]	ResfEANet	Tuberculosis	Binary classification	Accuracy = 97.95%
[48]	Hydra-ViT	14 chest diseases from the chest-14 dataset	Multi-class classification	AUC = 83.8%

3. Methodology

This section presents the working approach in detail, starting with the datasets used as well as the preprocessing and data augmentation phase, and also including the architecture and the different configurations used for the classification process.

3.1. The Proposed Approach

In this paper, we incorporated CNNs and ViTs to benefit from the strength of both by designing a hybrid model based on ResNet-50 and ViT-b16, pretrained on the ImageNet dataset in order to combine the spatial feature extraction capability of ResNet-50 with the global contextual understanding of ViT-b16.

The ResNet-50 introduces the concept of residual learning by applying the skip connections that allow the model to focus on incremental improvements rather than complete transformations. The skip connections can be formulated as:

$$H(x) = F(x, \{W_i\}) + x \tag{1}$$

where $H(x)$ denotes represents the desired underlying mapping, and $F(x)$ and x denote the residual function that captures changes in features through learned weights W_i and the input to the bloc, respectively.

The complete residual block consists of convolutional operations, incorporating batch normalization (BN) and the *Relu* activation function σ as detailed in Equation (2):

$$output = \sigma\left(BN\left(conv\left(\sigma(\,BN(conv(x)))\right)\right)\right) \tag{2}$$

The ViT-b16 splits the images into 16×16 patches p_i and linearly embeds them into a vector z_i of dimension d using a linear projection, which is formulated as:

$$z_i = W_p \cdot p_i + b_i \tag{3}$$

where W_i represents the learnable projection matrix and b denotes the bias. Next, a learnable position embedding is added to retrain spatial information.

$$z_i' = z_i + Position\ embedding \tag{4}$$

These embedded patches are fed to the transformer block encoder, which consists of 12 encoders, each containing 12 attention heads and a Multi-Layer Perceptron (MLP) block with a Gelu activation function. The self-attention of each head is presented as follows:

$$Attention = (Q,\ K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{5}$$

Q, K, and V represent the query, key, and values matrices, respectively, derived from the input through learnable projections.

Finally, a vector with a size of 768 ($16 \times 16 \times 3$) is passed through the classification head, which is a single dense layer with a single neuron that indicates Tuberculosis or normal in our case.

$$Output = \sigma(W_{out} \cdot z^f + b_{out}\Big) \tag{6}$$

The output is a probability, which ranges between 0 and 1. σ represents the sigmoid activation function, W_{out} is the weight matrix of the final dense layer, z^f is the final 768-dimensional feature vector from the ViT-b16 model, and b_{out} represents the bias of term of the final letter.

The input to the entire ensemble model (Figure 1) is an image of $224 \times 224 \times 3$, which is trained end-to-end as a single model by applying feature-level fusion to combine and concatenate the features extracted by each model. This enables the model to optimally combine the features from both streams during training. At this level, a self-attention mechanism was applied in both streams using the attention layer. This allows the model to focus on the most relevant features from each one. The attention layers were followed by a global average pooling to reduce the spatial dimensions, ensuring that the features maps from ResNet-50 match the output shape of ViT-b16. The two models were fine-tuned to maintain the benefits of transfer learning, where the last 10 layers for the ResNet-50 and the last 4 layers for the ViT-b16 are unfrozen, while keeping the earlier layers frozen, preserving the general feature extraction capabilities, while allowing for the adaption to our specific datasets.

The extracted and processed features were passed through two fully connected layers (dense layers) with 512 and 256 neurons, respectively, each one followed by batch normalization and dropout layers, leading to the best generalization and stability, effectively adapting the pretrained features for our task. The final layer represents the output layer for the classification process, which consists of a single neuron with a sigmoid activation function for the binary classification task.

3.2. Material and Methods

3.2.1. Dataset

In our study, we used the following publicly available datasets in Kaggle [50,51]. The Tuberculosis chest X-ray dataset is used, which includes two classes (TB-positive and TB-negative). The images are collected by a team of researchers from Qatar and Dhaka (Bangladesh) universities with their Malaysian collaborators. Only 700 TB images are publicly available, against 3500 normal images.

The Pneumonia dataset was obtained from Guangzhou Women's and Children's Medical Center, which contains 1344 images of normal cases and 3875 images of Pneumonia cases, devised to 2530 samples of bacterial Pneumonia and 1345 samples of viral Pneumonia. The Tuberculosis dataset was first used for a binary classification to detect Tuberculosis and differentiate between normal cases and Tuberculosis cases. Then, we merged the

two datasets for a multi-class classification task to distinguish between viral Pneumonia, bacterial Pneumonia, and Tuberculosis.

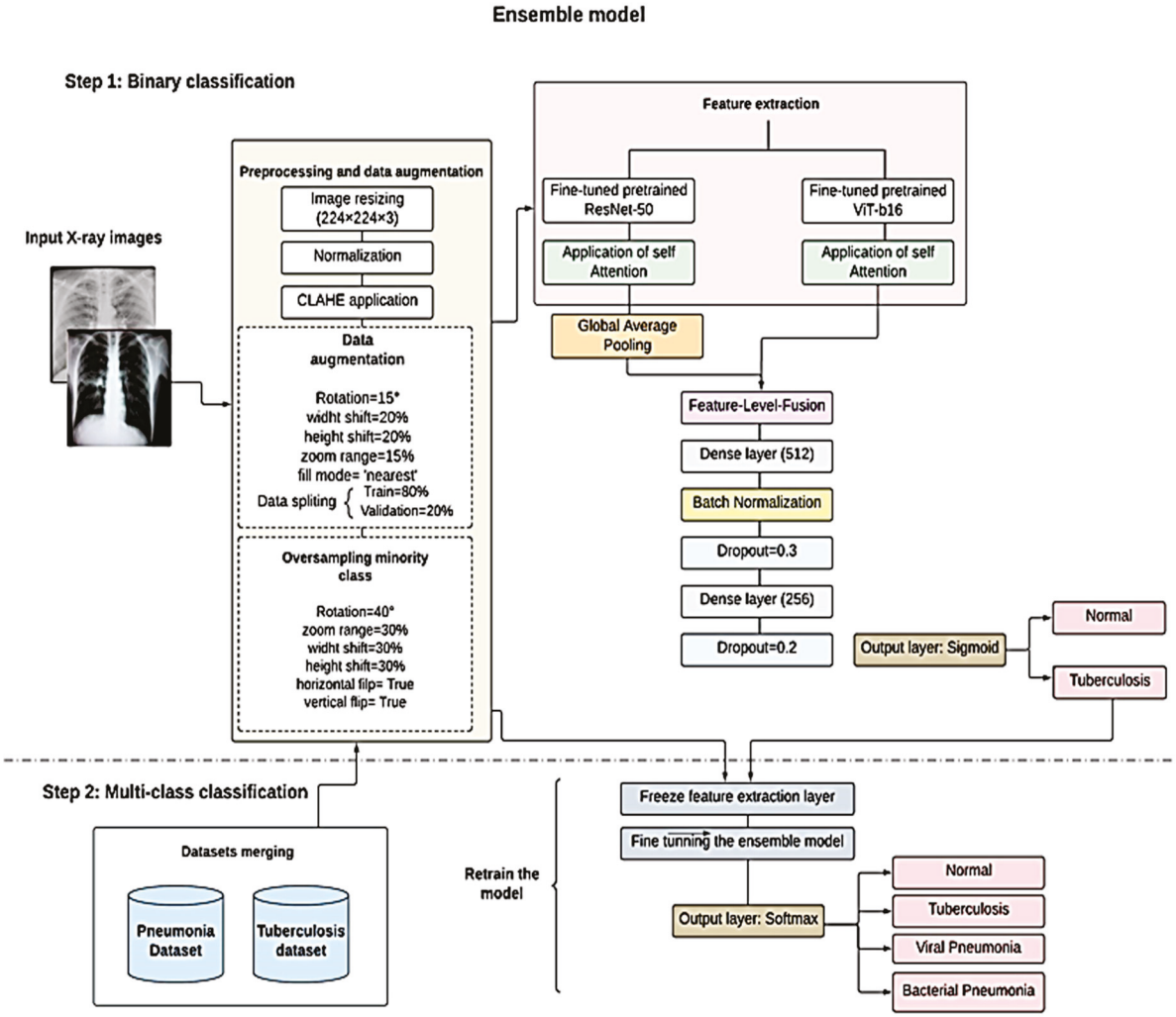
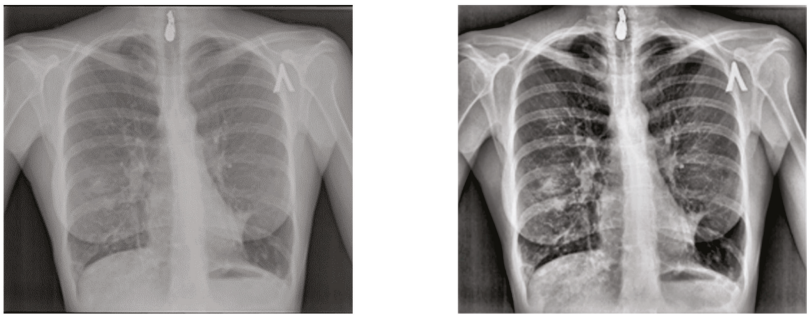


Figure 1. The overall ensemble model architecture.

3.2.2. Preprocessing and Data Augmentation

As a start, the images were resized to 224 × 224 × 3. Next, normalization was applied in order to scale the pixels values to a range [0, 1] that is suitable for model training to perform and generalize better to new data by reducing the influence of outliers and ensuring that all features contribute effectively in the learning process. Then, Contrast Limited Adaptive Histogram Equalization (CLAHE) was used with a cliplimit of 3.5 (Figure 2) to improve the local contrast of images, enhancing visibility in lighter and darker regions. This makes the detection of features like keypoints and edges more effective. The improved contrast makes subtle differences in pixel intensities more pronounced, allowing feature detection algorithms to more easily identify and localize important structures.

Afterward, a general data augmentation process was applied to increase the size of the entire dataset and introduce some variability in the training process, which helps to reduce the possibility of overfitting by encouraging the model to learn more general and transferable features. Four geometric transformations were used, namely a rotation of 15°, a horizontal and vertical translation of 20%, and a zoom range of 15%. These transformations offer new samples that encourage the learning of invariant features. Then, we employed a 'nearest fill' mode to handle the empty spaces created at the image edges by filling any newly created empty pixels with the value of the nearest valid pixel.



(a) (b)

Figure 2. Example of Tuberculosis cases (**a**): the original image; (**b**): the image after the application of CLAHE.

In order to treat the problem of data imbalance, which occurs significantly in both of the two datasets used, we employed an oversampling technique targeted at the minority classes. This strategy involved generating additional samples through the application of more aggressive augmentation, which includes a rotation up to 40°, a zoom range of 30%, horizontal and vertical translations up to 30%, and horizontal and vertical flips. The application of oversampling gives the model more opportunities to learn from minority class samples, which mitigate the risk of developing a bias toward the majority classes.

Finally, the training dataset was randomly split into 'Train' and 'validation' using the following ratio: 80% and 20%, respectively.

4. Experimental Study

4.1. Evaluation Metrics

The performance of the ensemble model was evaluated using the following metrics: accuracy (Equation (7)), which gives an overview about the model's performance across all classes; and recall (Equation (8)), which is better suited for imbalanced datasets by capturing all the instances of the minority classes. This metric is important when the cost of missing positive instance is high. The precision (Equation (9)) metric ensures that the positive predictions made by the model are generally correct. We also used the F1-score metric (Equation (10)), providing a balance between the precision and recall which is useful when both false positives and false negatives are critical. Additionally, the Jaccard score (Equation (11)) and the Dice coefficient (Equation (12)) were also calculated; the first gauges the similarity and diversity of sample sets and illustrates how well the model handles the trade-off between precision and recall, and the second shows the effectiveness of the distinction between the model's classes.

$$Accuracy\ (acc) = (TP + TN)/(TP + TN + FP + FN) \tag{7}$$

$$Recall = TP/(TP + FN) \tag{8}$$

$$Precision = TP/(TP + FP) \tag{9}$$

$$F1\text{-}score = (2 \times Precision \times Recall)/(Precision + Recall) \tag{10}$$

$$Jaccard\ score = (Precision \times Recall)/Precision + Recall - (Precision \times Recall) \tag{11}$$

$$Dice\ coefficient = 2 \times TP/(2 \times TP + FP + FN) \tag{12}$$

where true positives (TPs) indicate the model's ability to correctly detect patients with the disease, which is crucial for timely and accurate treatment. True negatives (TNs) represent the model's ability to correctly identify patients without anomalies. False positives (FPs) occur when the model incorrectly identifies a disease that is not present. Conversely, false negatives (FNs) occur when the model fails to detect the presence of a disease, which is an unacceptable mistake as it delays treatment and potentially worsens patient outcomes.

4.2. Model Training and Experimental Setup

The model was first trained for a binary classification task using the Tuberculosis dataset, with TensorFlow version 2.10.1 on a GeForce 1050 Ti graphic card with dedicated memory. We also used Google Collab Pro for more performant GPUs, namely the T4 and A100 GPUs.

The choice of the *ResNet-50* was confirmed after testing different fine-tuned pretrained CNN models such as VGG-16, DenseNet-121, etc., where the ResNet-50 produced a better result. For the ViT, we only tested the base model instead of the large ViT and huge ViT models, due to the computation costs and time-related efforts with different patch sizes such as 32×32, as the 16×16 patch size enhances the model's performance.

We employed the *RectifiedAdam optimizer* with an initial learning rate of 1×10^{-4} to obtain a more accurate and generalizable deep neural network by rectifying the variance of the adaptive learning rate; we also used binary cross entropy as a loss function. The model was trained for 30 epochs with a batch size of 16. During this training we implemented early stopping based on the validation F1-score to prevent overfitting. The F1-score considers both false positives and false negatives, which is critical in the medical context. We introduced callbacks to automatically adjust training parameters and to save the best model, ensuring the retraining of the optimal model.

The ensemble model was extended to perform multi-class classification for the distinction between Tuberculosis, viral Pneumonia, bacterial Pneumonia, and normal cases, based on the knowledge gained during the binary classification.

Firstly, we merged the dataset of Tuberculosis and the dataset of Pneumonia, as is mentioned above. This merging caused a data imbalance. To address this issue, we kept the data augmentation techniques, including the oversampling of minority classes; then, we implemented the class weights method to give more importance to underrepresented classes during training. We changed the classification head by replacing the output layer with another that contained four neurons, each corresponding to one of the four classes, combined with softmax activation function.

We maintained our original training and optimization methods. However, we changed the loss function with focal categorical entropy, which is particularly useful for multi-class classification and imbalanced data; we also minimized the learning rate value to 1×10^{-5} to facilitate fine-tuning without altering the pre-learned features. In this step, the model trained only for 15 epochs due to computation limits.

5. Results

The proposed hybrid workflow was designed by combining the capabilities of both ResNet-50 and ViT-b16. The model was trained within two steps. In the first step, a binary classification for Tuberculosis detection was applied. In the second step, we performed a multi-class classification to effectively distinguish between Tuberculosis, bacterial Pneumonia, and viral Pneumonia. The model's performance was evaluated over three scenarios: testing the performance of individual CNNs, individual ViTs, and hybrid model-based CNN-ViT. Tables 4–7 present the obtained results according to these scenarios.

For the binary classification step, the hybrid model achieved a performance of 98.97% in terms of accuracy, 99.87% in terms of precision, 99.91% in terms of recall (Figures 3 and 4) and 99.08% in terms of F1-score. The ensemble model outperforms both individual CNNs and ViTs.

Table 4. Confusion matrix for binary classification.

	Normal	**Tuberculosis**
Normal	2444	6
Tuberculosis	2	487

Table 5. Confusion matrix for multi-class classification.

	Normal	Tuberculosis	Viral Pneumonia	Bacterial Pneumonia
Normal	4653	68	43	80
Tuberculosis	8	672	7	13
Viral Pneumonia	14	18	1295	18
Bacterial Pneumonia	38	20	32	2440

Table 6. The classification evaluation performance of the tested deep learning models for binary classification.

Pretrained Models	Accuracy	Precision	Recall	F1-Score	Jaccard Score	Dice Coefficient
VGG-16	0.7756	0.7698	0.3309	0.4628	0.3713	0.4428
DenseNet-121	0.8299	0.8098	0.8355	0.8224	0.7198	0.7100
Xception	0.8754	0.8960	0.8934	0.8947	0.8063	0.7988
ResNet-50	0.9115	0.9210	0.9077	0.9114	0.8464	0.8137
ViT-b16	0.9418	0.9367	0.9549	0.9457	0.8964	0.8511
ViT-b32	0.8909	0.9023	0.8809	0.8915	0.7890	0.8095
Ensemble model	0.9897	0.9987	0.9991	0.9908	0.9838	0.8914

Table 7. The classification evaluation performance of the tested deep learning models for multi-class classification.

Pretrained Models	Accuracy	Precision	Recall	F1-Score	Jaccard Score	Dice Coefficient
VGG-16	0.5518	0.5657	0.5187	0.5412	0.9756	0.6821
DenseNet-121	0.9043	0.8945	0.8898	0.8921	0.8064	0.7508
Xception	0.8514	0.8438	0.8499	0.8468	0.7543	0.8224
ResNet-50	0.9365	0.9411	0.9387	0.9384	0.8839	0.8854
ViT-b16	0.9725	0.9710	0.9766	0.9738	0.9245	0.8975
ViT-b32	0.9156	0.8965	0.9076	0.9020	0.8623	0.8813
Ensemble model	0.9618	0.9680	0.9600	0.9640	0.9032	0.9124

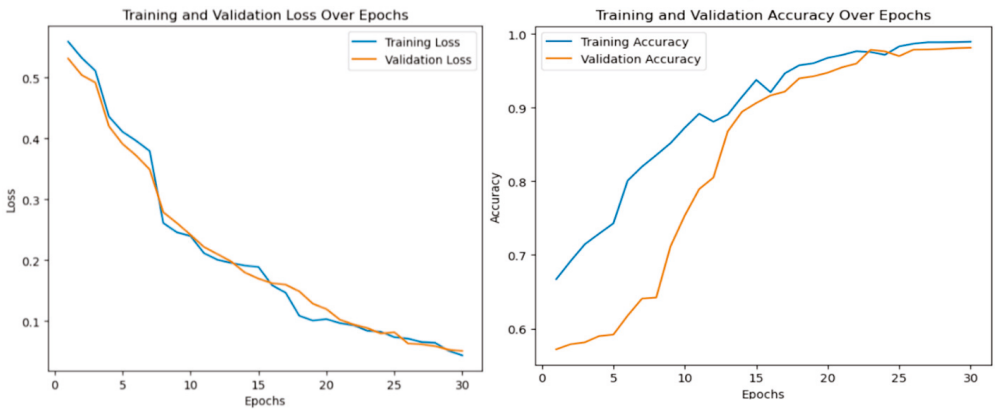


Figure 3. Ensemble model performance in terms of loss and accuracy for the binary classification.

For the multi-class classification step, the hybrid model achieved acceptable performance with an accuracy of 96.18%, a precision of 96.8%, a recall of 96% (Figures 5 and 6), and an F1-score of 96.4%.

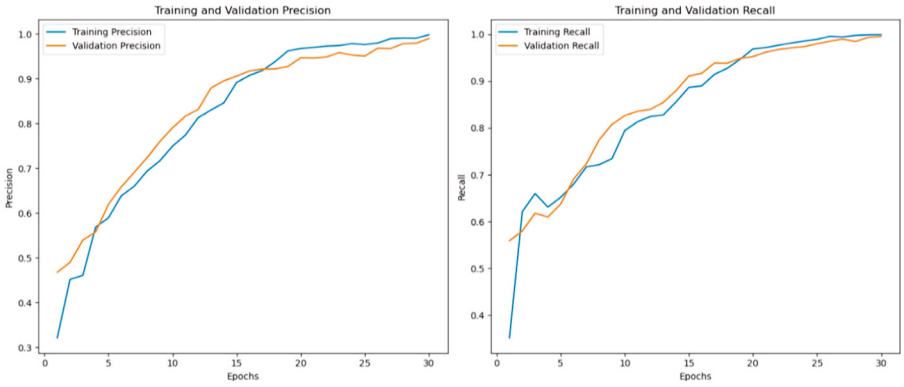


Figure 4. Ensemble model performance in terms of precision recall for the binary classification.

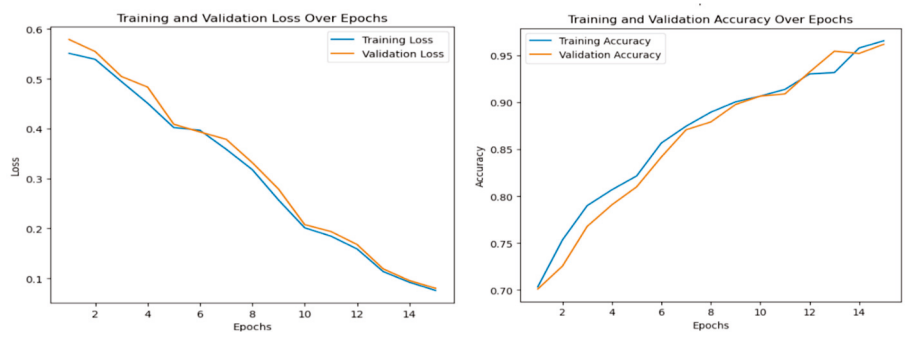


Figure 5. Ensemble model performance in terms of loss and accuracy for the multi-class classification.

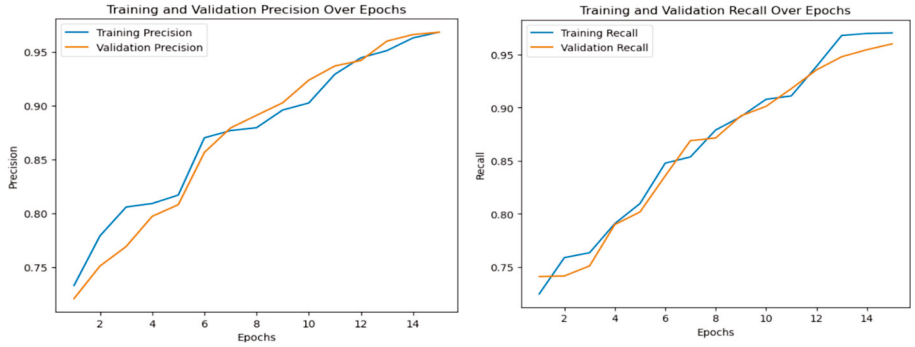


Figure 6. Ensemble model performance in terms of precision and recall for the multi-class classification.

6. Discussion

In this study, we presented a novel hybrid model based on CNN and ViT for Tuberculosis detection as a binary classification task, and the distinction between Tuberculosis, bacterial Pneumonia, viral Pneumonia, and Normal cases.

In binary classification, the obtained results proved that our hybrid model outperforms the standalone CNNs and ViTs. The balanced precision and recall across all classes (Tables 1 and 2) indicates that the model achieves a consistent performance for the underrepresented classes, which is crucial in the medical context. The model's high F1-score suggest its stability and potential application in real-world clinical environments, especially in regions with limited access to radiology experts.

Our model exhibits competitive performance compared to several state-of-the-art studies. For instance, the obtained accuracy of 98.97% for Tuberculosis detection surpasses the hybrid EfficientNet-ViT model proposed by [52] that achieved an accuracy of 97.92%, which proves the robustness of our fine-tuning technique for the ensemble model.

In the case of multi-class classification, ViT-b16 slightly outperformed the hybrid model, which was attributed to the increased complexity of the task. Furthermore, we were not able to train the model for more than 15 epochs due to computation capacity limitations.

Table 8 represents a detailed comparison between our approach and state-of-the-art models that combine CNNs with ViT models. The results indicate that our hybrid model outperforms these models in both binary classification and multi-class classification.

Table 8. Comparative analysis of hybrid CNN-ViT models for chest disease detection.

Authors	Approach	Achieved Result	Authors	Approach	Achieved Results
[42]	Enhanced Vit: Parallel CNN and ViT	93.5% $\leq$ recall $\leq$ 100%	[44]	VGG-16 + CoorAttention	Accuracy = 92.73%
[43]	Ensemble CNNs + ViT	Accuracy = 98.01%	[47]	ResfEANet	Accuracy = 97.95%
[28]	EfficientNet + ViT	Accuracy = 97.72%	[45]	CNN-ViT	Accuracy = 91%
[48]	HydraViT	AUC = 83.8%	Our hybrid model	ResNet50-ViTb16	Accuracy = 98.97% for binary classification

7. Conclusions

In the last decade, chest disease detection-based on chest X-ray images using deep learning models has been an active area of research. Chest X-ray image interpretation can be challenging due to the overlap of anatomical structures, the similarities of symptoms between diseases, and the shortage of radiologists. In this study, we proposed a hybrid model based on ResNet-50 and ViT-b16 for Tuberculosis detection and distinction between Tuberculosis, viral Pneumonia, and bacterial Pneumonia. This approach achieved promising results with an accuracy of 98.97% and an F1-score of 99.08% for binary classification, and an accuracy of 96.4% and an F1-score of 96.18% for multi-class classification. The hybrid model outperformed a lot of state-of-the-art models for both binary classification and multi-class classification.

Despite the promising results achieved by our hybrid model, the limitation of computational resources prevented us from training the model for more than 15 epochs in the case of multi-class classification. With additional computational capacities, it is likely that training for more epochs could lead to further performance improvements. Our future goal is to develop a more computationally efficient version of the model for potential deployment on resource-constrained devices, as well as a model that generalizes efficiently and stably using different datasets.

Author Contributions: Y.H. conceived and designed the work, and was reponsile for the writing—original draft. T.M. and A.B. have contributed to the supervision, data curation, writing—original draft, and validation. M.A. has contributed to the supervision, software, visualization, and validation. N.A.K., A.H.A. and M.A.-S. carried out the critical revision of the article, the acquisition of funds, and the data curation. All authors reviewed the finalized manuscript and were responsible for the writing—review and editing. 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; the study did not involve humans or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data will be available on request.

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

References

1. Levine, S.M.; Marciniuk, D.D. Global Impact of Respiratory Disease. *Chest* **2022**, *161*, 1153–1154. [CrossRef] [PubMed]
2. World Health Organization. Available online: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd) (accessed on 28 May 2024).
3. World Health Organization. Tuberculosis. Available online: https://www.who.int/news-room/fact-sheets/detail/tuberculosis (accessed on 28 May 2024).

4. Schildknecht, K.R.; Pratt, R.H.; Feng, P.-J.I.; Price, S.F.; Self, J.L. Tuberculosis—United States, 2022. *MMWR. Morb. Mortal. Wkly. Rep.* **2023**, *72*, 297–303. [CrossRef] [PubMed]
5. World Health Organization. Global Tuberculosis Report 2024. Available online: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2024 (accessed on 11 October 2024).
6. UNICEF. Pneumonia in Children: What You Need to Know. Available online: https://www.unicef.org/stories/childhood-pneumonia-explained#:~:text=Every%20year,%20pneumonia%20claims%20the,of%20these%20deaths%20are%20preventable (accessed on 28 May 2024).
7. Qin, C.; Yao, D.; Shi, Y.; Song, Z. Computer-aided detection in chest radiography based on artificial intelligence: A survey. *Biomed. Eng. Online* **2018**, *17*, 113. [CrossRef] [PubMed]
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017.
9. 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 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the ICLR 2021—9th International Conference on Learning Representations, Virtual Event, Austria, 3–7 May 2021.
10. Chernyavskiy, A.; Ilvovsky, D.; Nakov, P. Transformers: "The End of History" for Natural Language Processing? In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Cham, Switzerland, 2021. [CrossRef]
11. Nahiduzzaman, M.; Islam, M.R.; Hassan, R. ChestX-Ray6: Prediction of multiple diseases including COVID-19 from chest X-ray images using convolutional neural network [Formula presented]. *Expert Syst. Appl.* **2023**, *211*, 118576. [CrossRef] [PubMed]
12. Sanida, T.; Dasygenis, M. A novel lightweight CNN for chest X-ray-based lung disease identification on heterogeneous embedded system. *Appl. Intell.* **2024**, *54*, 4756–4780. [CrossRef]
13. Hedhoud, Y.; Mekhaznia, T.; Amroune, M. An improvement of the CNN-XGboost model for pneumonia disease classification. *Pol. J. Radiol.* **2023**, *88*, 483–493. [CrossRef]
14. Alshmrani, G.M.M.; Ni, Q.; Jiang, R.; Pervaiz, H.; Elshennawy, N.M. A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images. *Alex. Eng. J.* **2023**, *64*, 923–935. [CrossRef]
15. Guefrechi, S.; Jabra, M.B.; Ammar, A.; Koubaa, A.; Hamam, H. Deep learning based detection of COVID-19 from chest X-ray images. *Multimed. Tools Appl.* **2021**, *80*, 31803–31820. [CrossRef]
16. Ibrahim, A.U.; Ozsoz, M.; Serte, S.; Al-Turjman, F.; Yakoi, P.S. Pneumonia Classification Using Deep Learning from Chest X-ray Images During COVID-19. *Cognit. Comput.* **2024**, *16*, 1589–1601. [CrossRef]
17. 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]
18. Malik, H.; Anees, T.; Chaudhry, M.U.; Gono, R.; Jasinski, M.; Leonowicz, Z.; Bernat, P. A Novel Fusion Model of Hand-Crafted Features With Deep Convolutional Neural Networks for Classification of Several Chest Diseases Using X-ray Images. *IEEE Access* **2023**, *11*, 39243–39268. [CrossRef]
19. Hu, X.; Chen, J.; Chen, Y. RegMamba: An Improved Mamba for Medical Image Registration. *Electronics* **2024**, *13*, 3305. [CrossRef]
20. Pham, V.T.N.; Nguyen, Q.C.; Nguyen, Q.V. Chest X-Rays Abnormalities Localization and Classification Using an Ensemble Framework of Deep Convolutional Neural Networks. *Vietnam J. Comput. Sci.* **2023**, *10*, 55–73. [CrossRef]
21. 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]
22. Huang, M.L.; Liao, Y.C. Stacking Ensemble and ECA-EfficientNetV2 Convolutional Neural Networks on Classification of Multiple Chest Diseases Including COVID-19. *Acad. Radiol.* **2023**, *30*, 1915–1935. [CrossRef]
23. Tanzi, L.; Audisio, A.; Cirrincione, G.; Aprato, A.; Vezzetti, E. Vision Transformer for femur fracture classification. *Injury* **2022**, *53*, 2625–2634. [CrossRef]
24. Shou, Y.; Meng, T.; Ai, W.; Xie, C.; Liu, H.; Wang, Y. Object Detection in Medical Images Based on Hierarchical Transformer and Mask Mechanism. *Comput. Intell. Neurosci.* **2022**, *2022*, 1–12. [CrossRef]
25. Wang, C.; Jin, Y.; Liang, J.; Yang, Y.H.; Nie, S.; Hai, Y. Modif-SegUnet: Innovatively Advancing Liver Cancer Diagnosis and Treatment through Efficient and Meaningful Segmentation of 3D Medical Images. In Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2023, Istanbul, Türkiye, 5–8 December 2023. [CrossRef]
26. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021. [CrossRef]
27. Liu, C.; Yin, Q. Automatic diagnosis of COVID-19 using a tailored transformer-like network. *J. Phys. Conf. Ser.* **2021**, *2010*, 012175. [CrossRef]
28. Cohen, J.P.; Morrison, P.; Dao, L.; Roth, K.; Duong, T.; Ghassem, M. COVID-19 Image Data Collection: Prospective Predictions are the Future. *Mach. Learn. Biomed. Imaging* **2020**, *1*, 1–38. [CrossRef]
29. Chowdhury, M.E.H.; Rahman, T.; Khandakar, A.; Mazhar, R.; Kadir, M.A.; Mahbub, Z.B.; Islam, K.R.; Khan, M.S.; Iqbal, A.; Emadi, N.A.; et al. Can AI Help in Screening Viral and COVID-19 Pneumonia? *IEEE Access* **2020**, *8*, 132665–132676. [CrossRef]
30. Chetoui, M.; Akhloufi, M.A. Explainable Vision Transformers and Radiomics for COVID-19 Detection in Chest X-rays. *J. Clin. Med.* **2022**, *11*, 3013. [CrossRef] [PubMed]

31. SIIM-FISABIO-RSNA. SIIM-FISABIO-RSNA COVID-19 Detection Challenge. Available online: https://siim.org/research-journal/siim-machine-learning-challenges/covid-19-kaggle-challenge/ (accessed on 20 June 2024).
32. Krishnan, K.S.; Krishnan, K.S. Vision Transformer based COVID-19 Detection using Chest X-rays. In Proceedings of the IEEE International Conference on Signal Processing, Computing and Control, Solan, India, 7–9 October 2021. [CrossRef]
33. Than, J.C.M.; Thon, P.L.; Rijal, O.M.; Kassim, R.M.; Yunus, A.; Noor, N.M.; Then, P. Preliminary Study on Patch Sizes in Vision Transformers (ViT) for COVID-19 and Diseased Lungs Classification. In Proceedings of the 2021 IEEE National Biomedical Engineering Conference (NBEC), Kuala Lumpur, Malaysia, 9–10 November 2021; pp. 146–150. [CrossRef]
34. Taslimi, S.; Taslimi, S.; Fathi, N.; Salehi, M.; Rohban, M.H. SwinCheX: Multi-label classification on chest X-ray images with transformers. *arXiv* **2022**, arXiv:2206.04246.
35. National Institutes of Health Chest X-Ray—Chris Crawford. NIH Chest X-Rays Dataset. Kaggle. Available online: https://www.kaggle.com/datasets/nih-chest-xrays/data (accessed on 10 June 2024).
36. Usman, M.; Zia, T.; Tariq, A. Analyzing Transfer Learning of Vision Transformers for Interpreting Chest Radiography. *J. Digit. Imaging* **2022**, *35*, 1445–1462. [CrossRef] [PubMed]
37. 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. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 590–597. [CrossRef]
38. Nasser, A.A.; Akhloufi, M.A. Deep Learning Methods for Chest Disease Detection Using Radiography Images. *SN Comput. Sci.* **2023**, *4*, 388. [CrossRef]
39. Park, S.; Kim, G.; Oh, Y.; Seo, J.B.; Lee, S.M.; Kim, J.H.; Moon, S.; Lim, J.-K.; Ye, J.C. Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification. *Med. Image Anal.* **2022**, *75*, 102299. [CrossRef]
40. Huang, L.; Ma, J.; Yang, H.; Wang, Y. Research and implementation of multi-disease diagnosis on chest X-ray based on vision transformer. *Quant. Imaging Med. Surg.* **2024**, *14*, 2539–2555. [CrossRef]
41. Wang, T.; Nie, Z.; Wang, R.; Xu, Q.; Huang, H.; Xu, H.; Xie, F.; Liu, X.-J. PneuNet: Deep learning for COVID-19 pneumonia diagnosis on chest X-ray image analysis using Vision Transformer. *Med. Biol. Eng. Comput.* **2023**, *61*, 1395–1408. [CrossRef]
42. Okolo, G.I.; Katsigiannis, S.; Ramzan, N. IEViT: An enhanced vision transformer architecture for chest X-ray image classification. *Comput. Methods Programs Biomed.* **2022**, *226*, 107141. [CrossRef]
43. Ukwuoma, C.C.; Qin, Z.; Belal Bin Heyat, M.; Akhtar, F.; Bamisile, O.; Muaad, A.Y.; Addo, D.; Al-antari, M.A. A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images. *J. Adv. Res.* **2023**, *48*, 191–211. [CrossRef]
44. Xu, T.; Yuan, Z. Convolution Neural Network With Coordinate Attention for the Automatic Detection of Pulmonary Tuberculosis Images on Chest X-Rays. *IEEE Access* **2022**, *10*, 86710–86717. [CrossRef]
45. Rajaraman, S.; Zamzmi, G.; Folio, L.R.; Antani, S. Detecting Tuberculosis-Consistent Findings in Lateral Chest X-Rays Using an Ensemble of CNNs and Vision Transformers. *Front. Genet.* **2022**, *13*, 864724. [CrossRef] [PubMed]
46. 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] [PubMed]
47. Ejiyi, C.J.; Qin, Z.; Nnani, A.O.; Deng, F.; Ejiyi, T.U.; Ejiyi, M.B.; Agbesi, V.K.; Bamisile, O. ResfEANet: ResNet-fused External Attention Network for Tuberculosis Diagnosis using Chest X-ray Images. *Comput. Methods Programs Biomed. Updat.* **2024**, *5*, 100133. [CrossRef]
48. Öztürk, Ş.; Turalı, M.Y.; Çukur, T. HydraViT: Adaptive Multi-Branch Transformer for Multi-Label Disease Classification from Chest X-ray Images. *arXiv* **2023**, arXiv:2310.06143. [CrossRef]
49. 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 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 22–25 July 2017. [CrossRef]
50. 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]
51. Chest X-Ray Images (Pneumonia). Kaggle. Available online: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia/data (accessed on 12 November 2024).
52. Duong, L.T.; Le, N.H.; Tran, T.B.; Ngo, V.M.; Nguyen, P.T. Detection of tuberculosis from chest X-ray images: Boosting the performance with vision transformer and transfer learning. *Expert Syst. Appl.* **2021**, *184*, 115519. [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

Early Detection of Parkinson's Disease Using AI Techniques and Image Analysis

Marilena Ianculescu [1], Corina Petean [1], Virginia Sandulescu [1,*], Adriana Alexandru [1,2,*] and Ana-Mihaela Vasilevschi [1]

[1] National Institute for Research and Development in Informatics, 011455 Bucharest, Romania; marilena.ianculescu@ici.ro (M.I.); corina.petean@ici.ro (C.P.); ana.vasilevschi@ici.ro (A.-M.V.)
[2] Faculty of Electrical Engineering, Electronics and Information Technology, Valahia University of Targoviste, 130004 Targoviste, Romania
* Correspondence: virginia.sandulescu@ici.ro (V.S.); adriana.alexandru@ici.ro (A.A.)

Abstract: Background: Parkinson's disease (PD) diagnosis benefits significantly from advancements in artificial intelligence (AI) and image processing techniques. This paper explores various approaches for processing hand-drawn Archimedean spirals in order to detect signs of PD. Methods: The best approach is selected to be integrated in a neurodegenerative disease management platform called NeuroPredict. The most innovative aspects of the presented approaches are related to the employed feature extraction techniques that convert hand-drawn spirals into a frequency spectra, so that frequency features may be extracted and utilized as inputs for various classification algorithms. A second category of extracted features contains information related to the thickness and pressure of drawings. Results: The selected approach achieves an overall accuracy of 95.24% and allows acquiring new test data using only a pencil and paper, without requiring a specialized device like a graphic tablet or a digital pen. Conclusions: This study underscores the clinical relevance of AI in enhancing diagnostic precision for neurodegenerative diseases.

Keywords: Parkinson's disease; image analysis; Parkinson's disease detection; Archimedean spiral

1. Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disorder (after Alzheimer's Disease) [1] and it is the fastest-growing neurological disorder in terms of disability [2]. PD presents a wide range of symptoms, including motor symptoms (such as bradykinesia, tremor, and rigidity) and non-motor symptoms (such as cognitive impairments, sleep disturbances, and pain) [1,2], which progressively worsen over time. Although there is no cure for PD, managing its symptoms is of the essence for maintaining a good quality of life and independent living. For this reason, early diagnosis is crucial.

Diagnosing PD is a challenging task, as there is no single definitive standardized test that clearly detects the presence or absence of the disease. Diagnosis is primarily based on the clinical evaluation performed by a neurologist. The assessment includes an analysis of medical history, observation of motor symptoms, and response to specific medication. Imaging and other tests may be used to rule out other conditions. Even though scales for rating the severity of PD exist, the most widely used are the Unified Parkinson's Disease Rating Scale (UDPRS) and its variants (like MDS UDPRS) [3,4], these must also be applied and assessed under the supervision of specialized personnel. For example, the study presented in [3] shows that there are statistically significant differences in the results of these standardized tests when performed by nurses or by neurologists, depending on the level of expertise of the performer.

PD, even in early stages, is associated with movement disorders like tremor, rigidity, bradykinesia. In order for a person to be diagnosed with PD, bradykinesia and either

rigidity or tremor must be present. These symptoms reflect in the way a person is able to write and draw.

A simple test for PD detection is based on the analysis of hand-drawn Archimedean spirals. The authors of [5] demonstrated that spiral analysis correlates with the motor part of the UDPRS and may be useful in the assessment of Parkinsonian signs. The test is usually performed in a clinical setting, where the drawing is assessed by specialized personnel with the expertise to put the results in perspective and correlate them with other relevant data from medical history and other different tests. There are a multitude of studies that try to automatically detect signs of PD in hand-drawn spirals [5–8].

Early studies focused on visually analyzing the drawn spirals and the behavior of the subjects while drawing, subsequent studies focused on using digital tools to more precisely analyze spiral drawings, and more recent studies have introduced machine learning (ML), including deep learning (DL) techniques for automatically detecting signs of the disease.

Ongoing studies continue to refine the techniques for analyzing spiral drawings and explore their correlation with other clinical measures. The goal is to enhance early detection and monitoring of PD, ultimately improving quality of life and health outcomes for PD patients.

The main aim of this study is to develop a binary classification algorithm that reliably detects signs of PD based on hand-drawn spirals. This algorithm is to be integrated in a complex application for managing neurodegenerative diseases, including PD. This paper presents the steps employed in building the algorithm. It is designed to work only on the spiral images, without requiring additional information like drawing speed, in-air and on-surface timings, number of pauses, and other information that may only be available if the drawing is produced using specialized devices that are not widely available, like digital tablets or pens. This way, tests on new images may be performed using a regular pencil and paper, without needing a graphic tablet or digital pen. This requirement comes from the need to integrate the algorithm in a PD screening tool that must be available to use through the NeuroPredict [9] platform in different testing settings, not just a clinical setting.

NeuroPredict was designed and is currently being developed as one of the results of the project "Advanced Artificial Intelligence Techniques in Science and Applied Domains", a project in its second year of implementation out of a total length of four years. The NeuroPredict platform integrates multiple predictive modeling algorithms customized for three specific disease, PD, Alzheimer's Disease, and Multiple Sclerosis, that rely on data provided by Internet of Medical Things devices integrated in the subject's living environment or worn by the monitored subjects. The platform integrates AI algorithms based on the acquired physiological information and self-reported data and is designed to deliver accurate and timely insights that have the potential to impact neurodegenerative disease management and comprehension. The PD detection algorithm presented in the current work is one of the many tools designed to be integrated in the NeuroPredict platform that may be used for screening and monitoring PD subjects or patients at risk.

The most innovative aspect of the presented work lies in the feature extraction phase, where the spiral images are converted into signals that may be further processed in both the time and frequency domains. The algorithm also uses features related to the weight of the line and its variations. The extracted features are grouped in two main categories called frequency features and pencil features. The features are then used to build multiple ML-based classifiers and the model with best performances in terms of accuracy, precision, and recall is selected to be integrated in NeuroPredict, the platform for managing neurodegenerative diseases.

As previously mentioned, the current work explores multiple approaches for developing the PD classification algorithm, with various feature sets and employing multiple ML techniques, described in Section 2, Materials and Methods. For comparison with other works selected from a literature review, it is important to mention that in the current work,

the best results were achieved using Random Forests (RFs). The average accuracy for the best model is 94.59%, as will be further detailed in the Results Section.

Table 1 presents a summarization of other works in the area of PD detection from hand-drawn spirals, selected by novelty criteria, dataset used, and data acquisition modality.

Table 1. Comparative presentation of other works on PD detection from hand-drawn Archimedean spirals.

Previous Work/Year of Study	Dataset Used	Features	Classification Algorithm Basis	Best Average Accuracy [%]	Test Data May Be Acquired with Pencil and Paper
Proposed method/ 2024	Zham et al. [7]	Frequency-based features combined with pencil content features	RFs	95.24	Yes
[10] 2024	Zham et al.	DL directly on images from the augmented dataset	DenseNet201	94	Yes
[8] 2024	Zham et al.	DL directly on images from the augmented dataset	DenseNet—50 epochs	96	Yes
			DenseNet—100 epochs	89	
[6] 2022	University of California, Irvine Parkinson's disease spiral drawings [11]	Kinematic features	k-Nearest Neighbors	92.38	No
			RFs	92.32	
[12] 2023	Built by the authors for the study: 58 images	Kinematics features Pressure (force) features	CatBoost	94.83	No
			LightGBM	89.65	
			RFs	79.31	
			Logistic regression	84.48	
[13] 2020	Zham et al.	DL directly on images from the augmented dataset	Convolutional Neural Networks	93.33	Yes
[14] 2021	University of California, Irvine Parkinson's disease spiral drawings	Kinematic features	RFs	94.35	No
			Support Vector Machine (SVM)	93.58	
			AdaBoost	96.02	
			Extreme Gradient Boosting	89.80	
[15] 2019	PaHaW	Tremor estimation distance feature	SVM	75.92	No
		Features based on Fourier Transform	SVM	79.63	
		Both above categories	SVM	81.66	
[16] 2018	PaHaW	Kinematic, pressure and spatio-temporal features	Gaussian Naïve Bayes	54.67	No

As seen in Table 1, there is a dominant trend in recent studies for using DL methods, particularly variations of DenseNet, showcasing high accuracy levels. In [8], the authors use DenseNet and ResNet variations to discriminate between healthy and PD patients using the dataset presented in [7] augmented several times, achieving the highest accuracy found in the literature review using DL, 96%. This accuracy is reported for the algorithm trained for 50 epochs and it is stated that the same algorithm, when trained for 100 epochs, achieves a lower average accuracy of only 89%. This might suggest that the high accuracy achieved at 50 epochs was measured before the learning curve stabilized and that the classifier built for 50 epochs might not perform well on new data, or it may suggest that if trained for 100 epochs, overfitting occurs. Learning and loss curves are not provided to

clarify this aspect. Although DenseNet achieves good results, 96% in [8] and 94% in [10], there are disadvantages presented by DL techniques, compared with the classification technique selected for current work (RFs). These drawbacks stem from the fact that they are highly complex models that require significant computing resources for training the models and also for inference on new data.

The main purpose of the work presented in this paper is to develop an algorithm that reliably detects PD from hand-drawn spirals that may be integrated into a complex application for remote monitoring and management of neurodegenerative disease patients and people at risk, including PD patients. For the algorithm to be appropriate for the required usage scenario, it must meet two criteria: (a) demonstrate good discrimination performance and (b) rely on data that may be acquired without specialized equipment. The motivation for (b) is that the proposed PD testing is intended to be used in different environments, not only in clinical settings. Even in clinical settings, it is preferable to avoid the need for dedicated equipment when testing new subjects. To achieve this, the presented algorithm is designed to work directly on images of hand-drawn spirals, without requiring additional data like timing information for each point of the drawing or in-air timings. For the same reason, each work from the literature review has been analyzed to check if it meets this requirement, as indicated in the last column of Table 1. Based on the input data, two major categories are identified: (a) classifiers that work on data collected using only traditional methods (pen and paper) and (b) classifiers that require data recorded using devices that acquire in-air or/and on-surface kinematic variables, which may only be obtained using digital devices like graphic tablets, digital pens, and other devices with tactic screens. The best reported accuracy in the literature review is presented in [14], where multiple classification techniques are used on data recorded with a digital tablet. The extracted features comprise drawing speed, acceleration, and in-air and on-surface time, among others, which are available only when a graphic tablet or a special pencil is used for recording the input data.

The structure of the paper is as follows: Section 2 provides a clear presentation of the steps employed in building the algorithm. It provides a description of the dataset used, the augmentation techniques employed to enrich the dataset, the steps for extracting the frequency features, and the pencil-related features. Afterwards, it presents descriptions of the models built in order to find the most appropriate technique for the dataset and the extracted features by presenting the classification models and the used evaluation metrics, using different approaches: multiple ML techniques and different subsets of the extracted features. Section 3 presents the results obtained with each ML technique on each of the feature subsets. Section 4, Discussions, presents the limitations of the approach and interpretations of the results and puts them in perspective comparing them to other works in the field. The final section, Conclusions, presents the implications of the presented work and future work directions.

2. Materials and Methods

This study presents a novel approach for detecting PD using spiral drawings, which differs from existing methods that rely on digital devices to measure additional real-time characteristics such as speed and pressure. The proposed method exclusively uses traditional tools—paper and pencil—to collect data, focusing on extracting meaningful information from the drawing converted into a signal in the frequency domain and the inherent variations in pencil content, such as pencil thickness and pressure, as part of the drawing's physical attributes. The frequency domain conversion allows analyzing both the signal and noise content, reflecting the specific characteristics of the tremor associated with PD.

The smoothness and consistency of the spiral are reflected in the frequency analysis without needing direct measurement. The proposed image-to-frequency conversion method simplifies the process by analyzing digitized versions of paper-and-pencil drawings, allowing for efficient detection without the need for additional real-time measurements.

The steps for building the PD detection algorithm based on hand drawings of Archimedean spirals are presented in Figure 1a. Figure 1b presents the steps for utilizing the decision algorithm in order to make predictions/decisions on new images. New images are processed in order to compute the relevant features that are fed to the decision algorithm in order to make a decision: the subject that drew the image is healthy or suffers from PD.

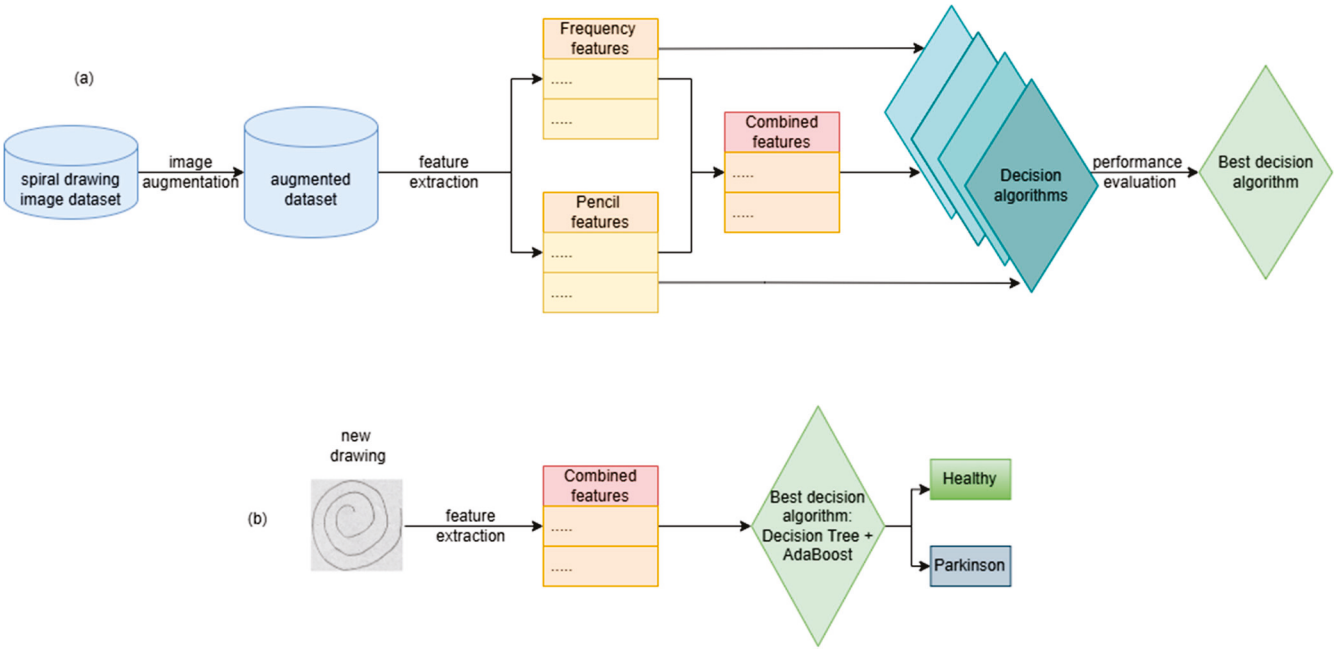


Figure 1. Pipeline for (**a**) building the decision algorithm and (**b**) using the decision algorithm.

The steps depicted in Figure 1a are presented in detail in the following sections.

2.1. Description of the Dataset

The literature review presents only a limited number of datasets containing images of hand-drawn spirals produced by both healthy subjects and patients with PD. Access to the PaHaW [17] dataset, possibly the most used dataset found in the literature review, is conditioned by a license agreement that strictly limits its usage for research purposes. The purpose of this work goes beyond validating the viability of detecting PD from hand-drawn spirals to creating a model that may be used in real-world scenarios to aid in PD screening and diagnosis, so the usage of PaHaW was dismissed. Future work may look at using the dataset for validation of the classification method purposes only. Another dataset frequently used in this area is the University of California, Irvine Parkinson's disease spiral drawings [11], which may be also used in future work for both improvement and validation of the built classification model.

The dataset used in the current work is presented in [7]. It contains drawings of Archimedean spirals produced by 55 subjects: 28 from a control group, healthy subjects, and 27 PD patients. The dataset contains 102 spirals: 51 spirals drawn by subjects from the control group and 51 spirals drawn by PWP subjects. The control group was chosen to be demographically similar with the PWP group, considering gender and age, in order to minimize differences between the groups that may be related to other aspects than PD. For the PD patients who were on levodopa treatment, the experiments were conducted while they were in the "on" stage. As examined using the UDPRS, on the Hoehn and Yahr Modified Scale [18], none of the PD patients were assessed to be in a late stage of the disease.

The drawings were captured using a digital drawing tablet with a dedicated pen in order to obtain the image itself and the drawing speed and pressure as time series with a recording frequency of 133 Hz. Participants were asked to draw guided Archimedean

spirals and waves. In the current study, only the images containing the spirals are used. The guiding spiral consisted of 4.5 revolutions, had a maximum radius of 75 mm, was sketched in the clockwise direction, and was shown to the participants as dots spaced 2 mms apart. The final images contain only the drawings made by the subjects and no dots from the guiding spiral.

2.2. Dataset Augmentation

Different augmentation techniques were considered in order to enrich the dataset. The usual techniques for augmenting image datasets comprise the following: rotation, shifting, blurring, brightening, and scaling with subunitary and supraunitary factors, among others. The only augmentation technique that was deemed appropriate was rotation. Scaling was dismissed, as the dimension of the drawn spiral might contain information related to the targeted condition, as several studies concluded that PD patients tend to draw smaller and more compact spirals (micrographia) [1,8]. Shifting was dismissed because both feature extraction pipelines (for pencil-related features and for frequency features) rely on detecting the center of the spiral, and its position on the drawing relative to the margins of the drawing carries no significance, blurring was dismissed considering that it may hide meaningful information from images, and brightening was dismissed as both pipelines for feature extraction work on the masked image, were original luminosity is removed.

To illustrate the augmentation process, Figure 2 presents a set of images generated through rotation at various angles. As described earlier, each original image was augmented by applying rotations between $-45°$ and $165°$ at 15-degree intervals. Figure 2 showcases examples of augmented images at selected angles, including $-45°$, $0°$, and $45°$. It is important to note that the image at $0°$ represents the original, unaltered image, while the other images demonstrate the effects of rotation at various angles.

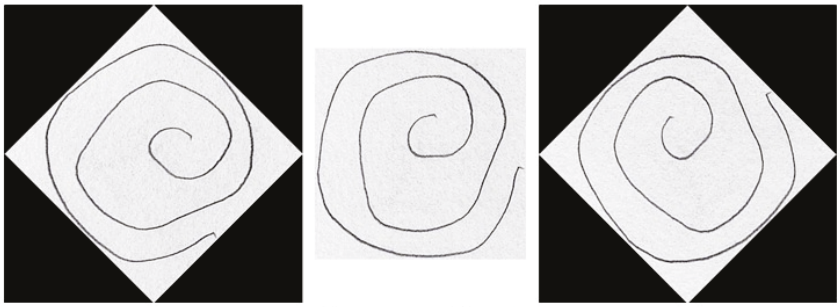


Figure 2. Examples of augmented spiral images rotated at various angles.

Following augmentation, the process continued with the extraction of frequency domain features, pen-related features, and combined features, which were computed and stored in separate comma-separated values (csv) files. Each csv file contained the label for each image and the corresponding feature values.

For each spiral, two main types of features were extracted:

1. Frequency-based features: These features are derived by converting the spiral drawings into frequency spectra using Fast Fourier Transform (FFT). This allows the quantification of the frequency and amplitude of tremors, which are characteristic of PD. Several metrics from the frequency domain are computed, including the peak frequency, peak magnitude, and signal-to-noise ratio (SNR).
2. Pencil-based features: These features are extracted from the grayscale intensity of the spirals. Mean pixel intensity and variance in line thickness are computed.

2.3. Feature Extraction: Frequency Features

2.3.1. Image-to-Frequency Conversion Method

The process for obtaining the frequency information is methodically structured, as illustrated in Figure 3, and each step is clearly outlined in Table 2 to ensure replicability and accuracy in identifying Parkinsonian tremors.

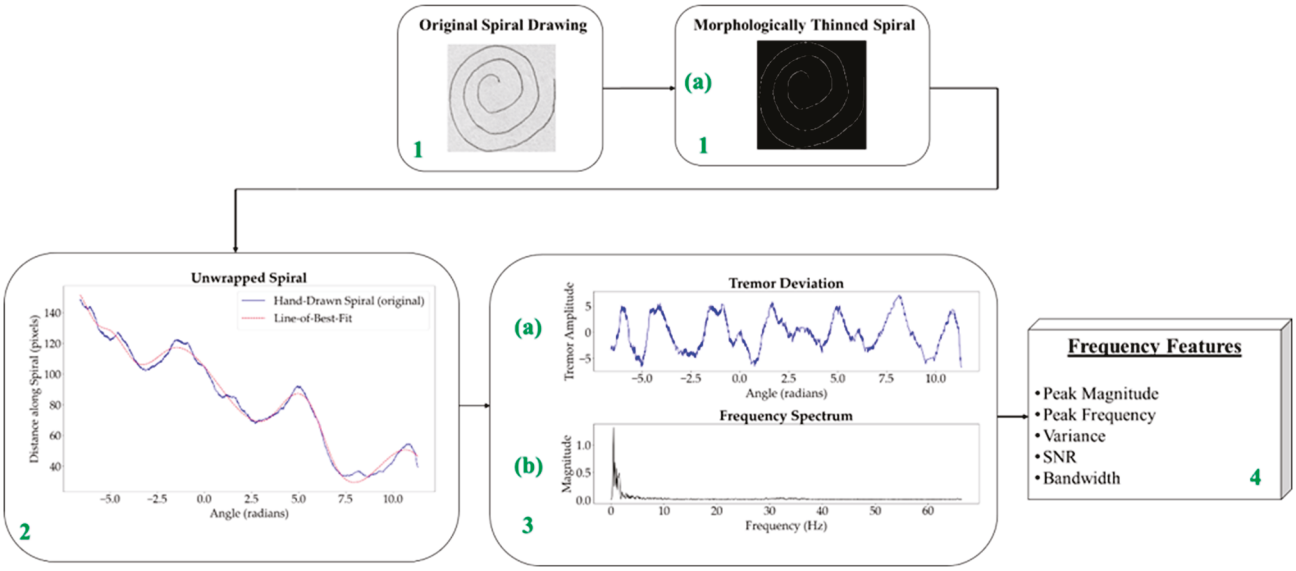


Figure 3. Image-to-frequency conversion process for spiral drawings.

Table 2. Steps for obtaining the frequency domain features.

Step	Description	Reference Block in Figure 3
1	Load the original image of the drawn spiral	Block 1
2	Perform morphological thinning to obtain one-pixel-width spirals	Block 1a
3	Unwrap the spiral and obtain the "Line-of-Best-Fit (LOBF)"	Block 2
4	Compute the tremor deviation by taking the difference between the LOBF and the spiral drawing	Block 3a
5	Convert the tremor deviation signal to the Fourier domain	Block 3b
6	Compute the relevant frequency features: peak magnitude, peak frequency, variance, SNR, and bandwidth	Block 4

To effectively analyze the spiral drawings, the mathematical model of the Archimedean Spiral is used, which is defined in polar coordinates (r, θ) as follows:

$$r = a\theta, \tag{1}$$

where r is the length of the radius from the origin, a is a constant, and θ controls the number of rotations in radians. The radius length is a fundamental component of polar coordinates, indicating the distance from a specific point on the spiral to the origin. As the angle θ increases, the radius r also increases, causing the point to move away from the center in a spiral pattern. The constant a determines the rate at which the spiral expands away from the origin as it rotates around the center. This uniform and proportional increase in the radius with the angle makes the Archimedean spiral particularly suitable for scenarios

where such growth needs to be measured. In the context of the present study, this model provides a solid foundation for quantifying tremor deviations, which are indicative of PD. The mathematical model of the Archimedean spiral [19] serves as the basis for the unwrapping process described in the next section.

Starting with a raw image, as shown in Figure 3 (block 1), the first step involves converting the spiral drawings into single-pixel-width sketches. This morphological thinning simplifies the spirals, making it easier to perform a pixel-by-pixel analysis along the spiral path in order to unravel the drawing into a linear representation. A one-pixel-width sketch facilitates precise measurement of tremor deviations along the spiral, which is essential for accurately identifying Parkinsonian tremors.

2.3.2. Morphological Thinning

To achieve an accurate one-pixel-width contour of the spiral drawings, various image processing techniques can be employed, including skeletonization, medial axis skeletonization, and morphological thinning. In this study, we utilized morphological thinning because it produces a one-pixel-width image with minimal unnecessary branches, which is essential for accurate tremor analysis in PD detection. This method was chosen over other techniques, such as curvature detection or edge detection algorithms, which often produce artifacts like closed loops or double contours—two parallel lines, one on the outer and one on the inner boundary of the shape. These artifacts occur because these algorithms interpret the edges by creating lines on both sides of the actual boundary, leading to complex patterns that complicate the analysis. In contrast, morphological thinning removes redundant pixels that do not contribute to the essential trajectory of the spiral, reducing noise and potential distortions. This step is crucial for preparing the image for further processing, such as unwrapping and detailed frequency analysis, by retaining only the core path of the spiral. Morphological thinning was implemented using the OpenCV library [20], applying the Zhang–Suen algorithm [21] to transform thick lines into thin, one-pixel-width lines. As shown in Figure 4, morphological thinning (Figure 4b) provides a clear and accurate representation of the spiral's core path.

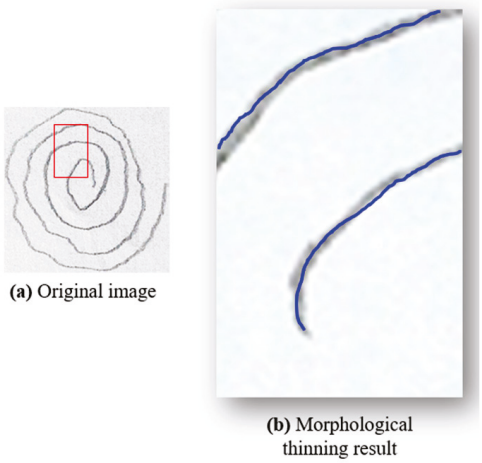


Figure 4. The process of morphological thinning and edge detection results on a spiral drawing showing (**a**) the original image from the dataset with the red rectangle marking the region of interest that is presented in (**b**) the processed image with the selected morphological thinning algorithm.

2.3.3. Spiral Unwrapping

In the process of drawing a spiral, the starting point is defined when the pencil first touches the paper, and the endpoint is reached when the pencil lifts, completing the spiral's path. For studies utilizing graphic tablets to collect spiral drawings, a time component is measured as the drawing progresses, capturing the motion from start to finish, and data are captured as time series. In contrast, for paper-and-pencil sketches, the drawings must

be interpreted considering the positions of pixels without any time-related information. To achieve this, the thinned image is used to define the start point (x_c, y_c) as the central pixel of the spiral. The unwrapping process involves counting each pixel (x_i, y_i) from this start point to the endpoint, calculating the Euclidean distance to the center at each step. The Euclidean distance, defined as

$$d = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}, \quad (2)$$

is computed for each pixel along the spiral path, where $0 \leq i \leq$ total length of the spiral. This iterative counting and distance calculation effectively and accurately "unwraps" the spiral into a linear form, as illustrated in Figure 5. The central pixel, marked as the START point in purple, and the endpoint, marked as the STOP point in red (Figure 5a), represent the boundaries of this unwrapping process.

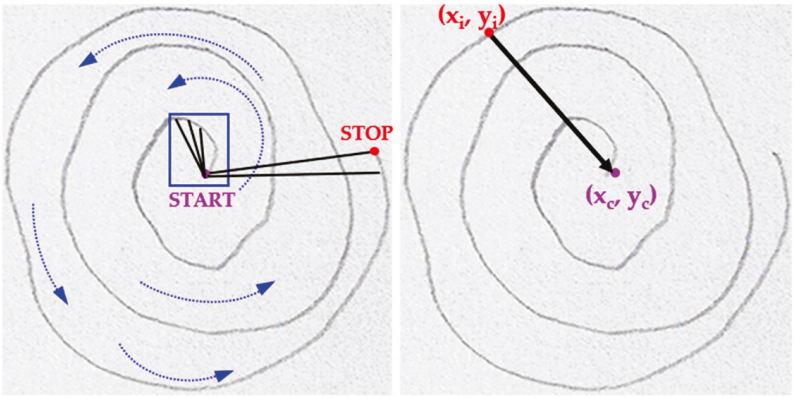


Figure 5. Unwrapping process of the spiral drawing by calculating the distance from the center at each pixel location: (**a**) depicts the START point in purple and the STOP point in red (**b**) shows the distance from an arbitrary point on the spiral (red) to the center of the spiral (purple).

To implement this unwrapping step, the unprocessed image is first converted to grayscale to simplify the analysis, as color information is not relevant in this context. A Gaussian filter is then applied to reduce noise and smooth out fine details that could introduce errors in subsequent processing steps.

Following filtering, adaptive thresholding is used to differentiate the spiral from the background, converting the image into a high-contrast black-and-white representation where the spiral becomes the focal point of analysis. The largest contour representing the spiral's path is extracted. The central pixel is calculated based on this contour as the starting point of the drawing, to serve as a reference point for further analysis. The coordinates of the contour are transformed from Cartesian to polar coordinates, allowing for a clear representation of the spiral's unwound path in terms of radius and angle. This transformation is essential for maintaining the correct directional angles around the center and ensuring a smooth transition in the graphical representation. Finally, the sorted distances and angles form the basis for the angle–distance plot, which visualizes and facilitates further analysis of the unwrapped spiral.

2.3.4. Tremor Deviation

In the analysis of hand tremor, particularly for assessing PD, tremor deviation refers to the measurement of irregular movements that occur when an individual attempts to draw a spiral. To quantify this deviation, an "ideal" reference model of the spiral is required. This reference is obtained by unwrapping the hand-drawn spiral and comparing it to a fitted curve, often referred to as the LOBF. Spline fitting was employed to create a smooth curve that represents the ideal spiral path. This method was chosen for its flexibility

in modeling rounded contours while minimizing the introduction of artifacts, such as excessive oscillations, that could obscure true tremor deviations.

The deviation is then calculated as the difference between the actual unwrapped spiral path and the idealized curve. This approach allows for precise quantifications of deviations, which directly reflect the irregularities in hand movements associated with Parkinsonian tremors.

The tremor deviation analysis utilizes the previously calculated radial distances and polar angles from each contour point of the spiral to the centroid. These measurements serve as the basis for constructing the coordinate pairs needed to analyze tremor characteristics. To smooth the data, cubic spline fitting is applied, with a smoothing parameter s that is carefully adjusted. The cubic spline (degree $k = 3$) provides the necessary flexibility to model the smooth, rounded shape of the spiral while minimizing overfitting or unnecessary oscillations.

The smoothing factor s is dynamically adjusted based on the number of data points in the unwrapped spiral, considering the reduced dataset size after the thinning process. The formula used is as follows:

$$s = (total\ number\ of\ points\ in\ unwrapped\ spiral) \times N \tag{3}$$

where N is a scaling factor that controls the degree of smoothing. This adjustment is necessary to maintain a balance between capturing the essential data features and preventing the loss of significant details due to excessive smoothing. Different values of N were tested to identify the optimal balance that highlights key data characteristics without obscuring the underlying tremor patterns. The selected value for N ultimately influences the smoothness of the LOBF and its ability to accurately represent the tremor deviations.

Table 3 summarizes the impact of various N values on the smoothing process, illustrating how fine-tuning this parameter helps to distinguish between normal and Parkinsonian tremor characteristics effectively.

Table 3. Values tested for the scaling factor N to determine the impact on the smoothing process.

Parameter	Tested Values
Scaling factor (N)	1, 5, 10, 15, 20

Figure 6 illustrates the impact of varying the smoothing parameter s, which is controlled by the scaling factor N, on the unwrapped spiral curve. With lower values of N (and therefore s), the fitted spline closely follows the original spiral, capturing all fine details and minor fluctuations. As N increases, the smoothing effect becomes more pronounced, allowing the spline to highlight the significant features of the spiral while reducing sensitivity to less relevant variations, such as noise or minor fluctuations.

The plots generated for each tested value of N reveal that lower values (e.g., $N = 1$ and $N = 5$) capture every minor fluctuation, including noise, which may not be meaningful for analyzing tremor characteristics. However, as N increases to 10, an optimal balance is achieved: the spline effectively smooths out noise and minor variations while preserving the key features associated with tremor. Further increases in N (e.g., $N = 15$ and $N = 20$) lead to excessive smoothing, where the spline risks omitting clinically significant details, as it no longer follows the path of the original spiral closely enough.

Thus, $N = 10$ is identified as the optimal value, providing a balanced approach that allows the spline to accurately reflect the original spiral's trajectory while applying enough smoothing to distinguish between significant tremor deviations and noise. The resulting curve captures the essential tremor characteristics without over-smoothing, ensuring both fidelity to the original data and clarity in the diagnostic analysis. The tremor deviation is quantified as the difference between the original spiral distances and the distances adjusted by the cubic spline interpolation, as shown in Figure 3 (block 3a). This deviation, plotted as amplitude against angle, reflects variations in hand movement indicative of tremor, with

the *x*-axis representing the angle of each point in radians and the *y*-axis indicating the amplitude of deviation.

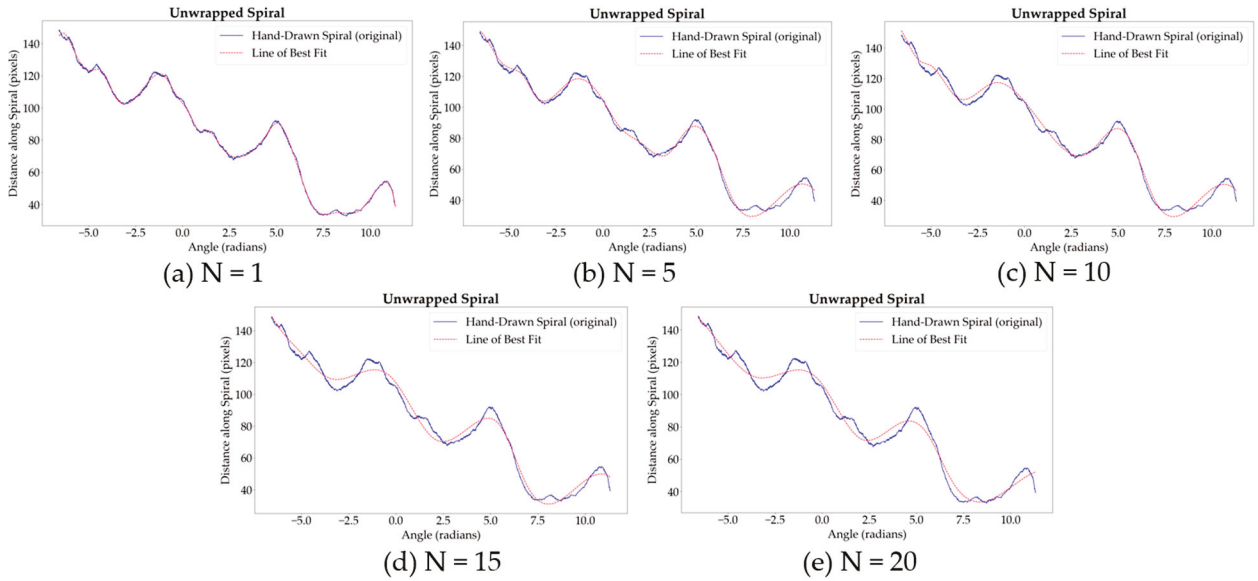


Figure 6. Impact of smoothing degree determined by scaling factor *N*.

2.3.5. Frequency Spectrum Analysis

In this study, the FFT is applied to convert the tremor deviation signal from the time domain into the frequency domain. This transformation decomposes the tremor deviation signal into sinusoidal components of different frequencies, allowing for the identification and characterization of dominant frequencies associated with Parkinsonian tremor.

To perform this analysis, the extracted signal is analyzed in the frequency domain using a varying frequency computed for each image as the number of points that form the spiral, divided by the average time it takes to draw a spiral for a healthy person, approximated at 10 s. The FFT is applied on the tremor deviation signal. The magnitude of the frequency spectrum is determined by computing the modulus of the complex values obtained from the FFT. To ensure the accuracy of the analysis across signals of varying lengths, the magnitude of the spectrum is normalized by dividing it by the length of the signal. This normalization step allows for a consistent interpretation of amplitude regardless of the signal length. Without normalization, differences in signal length could lead to discrepancies in the magnitudes, negatively impacting the ability to identify frequency characteristics associated with Parkinsonian tremor.

The FFT produces a set of frequency components, each corresponding to a specific frequency. The frequency vector contains the actual frequency values for each component in the FFT output, allowing each component to be associated with its respective frequency. As the FFT spectrum is symmetric, with the negative half mirroring the positive half for real-valued signals, only the positive half of the spectrum is retained in order to avoid redundancy. This is achieved by keeping the first half of both the frequency and magnitude vectors, ensuring that the resulting spectrum contains only the relevant frequencies for further signal analysis.

2.3.6. Frequency Feature Calculation

Using the data from the frequency domain, several key features were calculated to analyze the tremor signal associated with PD.

- Peak Magnitude: Peak magnitude represents the highest value of the magnitude spectrum, indicating the maximum amplitude of the frequency components in the signal. This peak is determined as the maximum value in the positive magnitude spectrum.

- Peak Frequency: The peak frequency is the frequency at which the magnitude peak occurs, providing insight into the dominant frequency characteristic of the analyzed signal. It is identified by the index of this peak within the frequency vector.
- SNR: The SNR assesses the quality of the signal relative to the noise level. It is defined as the ratio of the average power of the signal to the average power of the noise. The average power of the signal is computed as the sum of the squares of the values above a noise threshold, divided by the number of signal samples, and similarly for the noise. The noise threshold is set as twice the mean of all values in the positive frequency spectrum, rounded to three decimal places. Components of the magnitude vector that exceed this threshold are considered part of the useful signal, while those below are classified as noise, allowing for a clear separation between signal and background noise.

$$SNR = \sum \frac{(signal\ values)^2}{len(signal)} / \sum \frac{(noise\ values)^2}{len(noise)} - 1 \quad (4)$$

- Variance: The variance of the signal content was calculated for the segment of the frequency spectrum identified above the noise threshold. This calculation provides an indication of the uniformity and stability of the signal by evaluating how much the magnitude of the frequency components deviates from their mean.
- Bandwidth: The bandwidth represents the total number of sample points contained within the signal, specifically the points above the noise threshold. This measure indicates the range of frequencies that significantly contribute to the signal.

These characteristics are illustrated in Figure 7. The peak magnitude represents the highest value of the magnitude spectrum, while the peak frequency is the frequency at which this peak occurs. The *variance*, *SNR*, and *bandwidth* are calculated based on the content of the signal and noise. The *noise threshold* is depicted in orange, separating the significant frequency components (signal) from the background noise. Signal values are above the noise threshold as indicated by the green arrow, indicating the portions of the spectrum considered as signal, while noise values are below the noise threshold, as indicated by the red arrow.

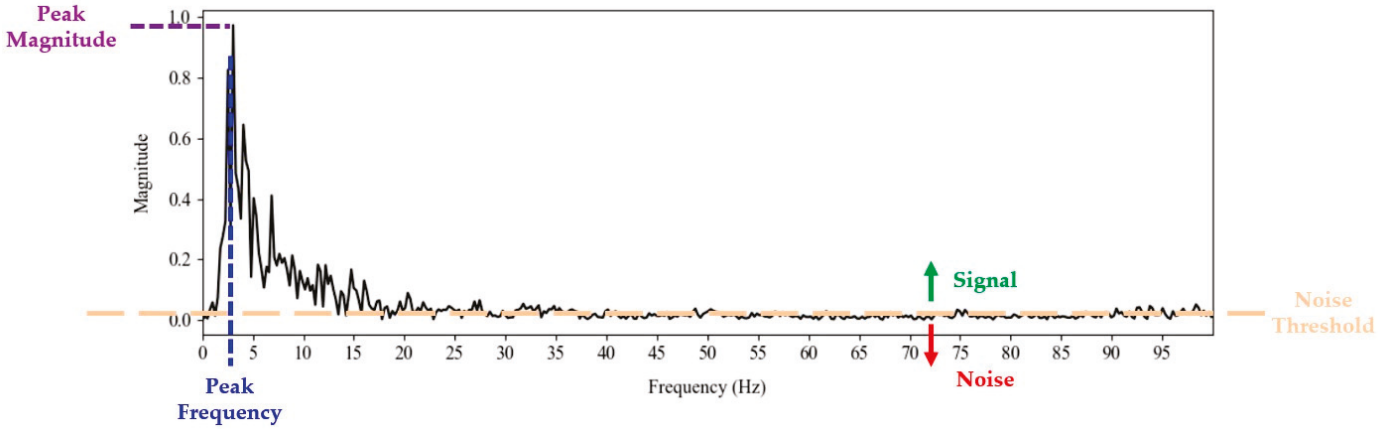


Figure 7. Visualization of calculated frequency features: peak frequency is the frequency at which the peak magnitude is achieved.

2.4. Feature Extraction: Pencil Features

Due to the morphological thinning process necessary to obtain the unwrapped spirals, all detailed information related to the actual pencil content is lost, including the thickness of the drawn line and the pressure applied with the pencil. These features may be significant, as they provide insight into the drawing behavior of patients with PD (e.g., greater pressure results in darker spirals, while lighter pressure results in lighter spirals). To capture these details, in addition to frequency features, pencil-specific features were extracted.

The pencil features were derived from the grayscale images using a technique based on pixel intensity thresholding. Since the majority of each image consists of lighter areas (i.e., the white paper on which the spirals were drawn), the initial threshold value was determined by calculating the mean intensity of all pixels. This initial threshold was further refined to account for potential noise or artifacts from the scanning process, such as shading or minor impurities. To achieve this, an error margin was computed by considering the difference between the initial threshold and the maximum possible pixel intensity value (255, representing white). This error margin was subtracted from the initial mean intensity to obtain a final, adjusted threshold.

Using this refined threshold, a Boolean mask was created to select all pixels in the image with an intensity lower than the final threshold, effectively isolating the pencil-drawn areas from the background. The remaining pixels, representing the actual pencil content, were then used to compute the desired features:

- Pencil thickness: This was calculated as the total number of pixels that constitute the pencil strokes in the image divided by the length of the spiral. The length of the spiral was computed using the sum of the arc lengths of all detected contours in the image.
- Pencil pressure: This was determined by calculating the average intensity of the pixels corresponding to the pencil strokes, providing a measure of the pressure applied during drawing.

This process is illustrated in Figure 8.

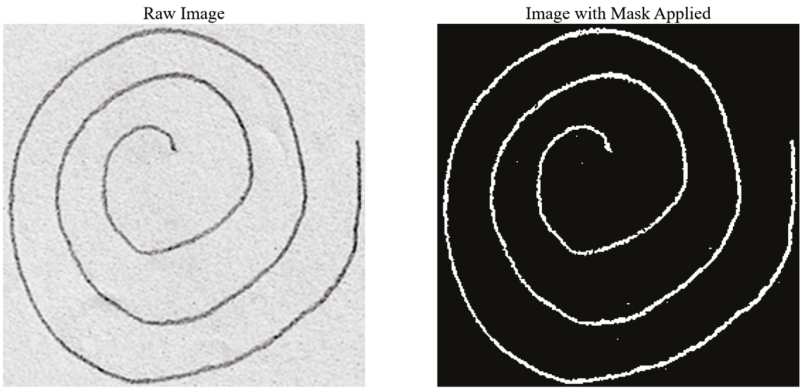


Figure 8. Process of highlighting pencil features.

2.5. Classification Models

Using the features extracted as presented in Sections 2.3 and 2.4, ML techniques for binary classification were performed to differentiate between the drawings produced by healthy subjects and those by PD patients.

The final step for both feature extraction pipelines was storing the features in csv files. Each csv file contained the label for each image and the corresponding values obtained for each feature. These csv files, created for each type of feature, were subsequently used as input data for the classification algorithms.

To ensure data integrity, rows containing NaN (Not a Number) values—meaning they could not be computed for the corresponding drawing—in key feature columns are removed from the csv files. This cleaning step is essential to avoid incomplete or erroneous entries that could negatively impact the model's performance.

The original dataset is already divided into two subsets, labeled as testing and training. The subset labeled as 'training' was used for training and testing and the subset labeled as 'testing' was used for validation.

The following ML algorithms were employed for this task SVM, RFs, DT, XGBoost, AdaBoost with DTs, and CatBoost. These algorithms were applied to the three sets of features: frequency domain features (noted as F features), pencil-related features (noted as P features), and a combination of both (noted as FP features).

2.6. Model Evaluation

The following evaluation metrics were used to assess the quality of the classification results for each model:

Accuracy: *Accuracy* is defined as the proportion of correctly classified instances over the total number of instances. It is computed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

where *TP* represents the number of true positives, *TN* the number of true negatives, *FP* the number of false positives, and *FN* the number of false negatives.

Precision: *Precision* measures the proportion of positive instances correctly identified by the classifier and is calculated as follows:

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

Recall (Sensitivity): *Recall*, or *sensitivity*, represents the proportion of actual positive instances that were correctly identified by the model and is computed as follows:

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

F_1-Score: The *F_1-score* provides a harmonic mean of precision and recall and is particularly useful when dealing with imbalanced classes. It is calculated as follows:

$$F_1\text{-}Score = \frac{2TP}{2TP + FP + FN} \quad (8)$$

For the best-performing model, a confusion matrix is also presented to provide a detailed evaluation of its classification performance, offering insight into the distribution of true positives, true negatives, false positives, and false negatives.

3. Results

This section presents the performances of the feature extraction approaches presented in the Methods Section. Classification models for differentiating between drawings produced by healthy subjects or PWP subjects are built using only the frequency domain features (noted as F features), only the pencil-related features (noted as P features), and using both types of features (noted as FP features). Multiple classification algorithms are built using the three types of features. Results are presented for each approach in terms of overall accuracy. For the best approach, detailed performances are presented in terms of *accuracy*, *precision*, *recall,* and *F_1-score*, along with the confusion matrix.

For the six ML algorithms that were employed (the three standard classifiers SVMs, RF, and DT and three boosting methods—XGBoost, AdaBoost with DT, and CatBoost), a grid search with five-fold cross-validation (*cv* = 5) was performed to systematically explore the hyperparameter space. This approach was chosen to ensure the selection of the optimal parameters that yield the best performance for each model. Grid search with cross-validation helps to prevent overfitting and ensures that the chosen hyperparameters lead to the most accurate and robust results. During the five-fold cross-validation, normalization is performed individually for each training fold. For each fold, normalization is performed to the training data of the current fold, ensuring the scaling parameters are derived solely from the training data. The scaler fitted on the training fold is afterwards utilized to normalize the corresponding validation set. This method prevents any leakage of information from the testing set into the training process.

The results obtained for each classification algorithm and feature set are summarized in Table 4.

Table 4. Classification results using only the frequency-related features (noted as F), only the pencil-related features (P), and the combined feature set (noted as FP).

Algorithm	Feature Set	Accuracy [%]	Precision [%]	Recall [%]	F_1-Score [%]
DT	F	57.14	57.40	57.14	57.14
	P	56.57	76.79	56.67	46.65
	FP	90.48	90.48	90.48	90.48
RF	F	90.48	90.48	90.48	90.48
	P	70.00	70.09	70.00	69.97
	FP	**95.24**	95.67	95.24	95.24
SVM	F	80.95	81.96	80.95	80.69
	P	66.67	67.94	66.67	66.06
	FP	85.71	85.98	85.71	85.65
XGBoost	F	76.19	78.23	76.19	75.52
	P	50.00	25.00	50.00	33.33
	FP	85.71	85.98	85.71	85.65
AdaBoost with DT	F	90.48	90.48	90.48	90.48
	P	70.00	70.00	70.00	69.97
	FP	90.48	90.48	90.48	90.48
CatBoost	F	80.95	80.95	80.95	80.95
	P	56.67	56.94	56.67	56.23
	FP	80.95	80.95	80.95	80.95

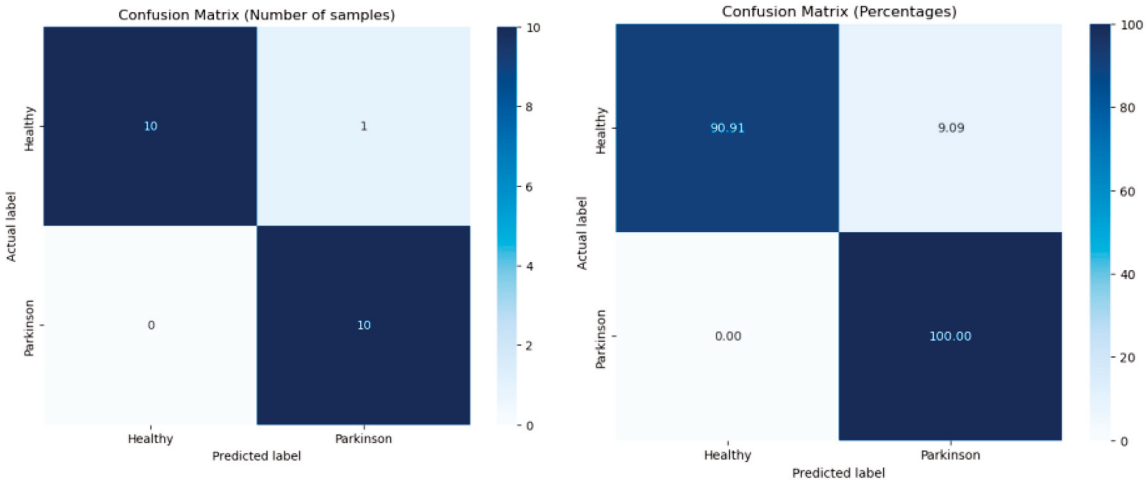
Across all classifiers, the FP feature set, which combines both frequency and pencil features, consistently outperformed the individual F and P sets in almost every metric. This confirms the hypothesis that combining both frequency- and pencil-related features leads to more accurate models. Specifically, in most cases, the FP set yielded the highest *accuracy*, *precision*, *recall*, and F_1*-scores* compared to using either feature set in isolation. The better performances of the FP set may be attributed to the complementary nature of the features extracted. Frequency-based features provide insight into the tremor characteristics of the patient's hand movements, while pencil-based features capture subtleties in line pressure and thickness that may be indicative of bradykinesia or rigidity. By using both types of information, the algorithms take into account the three main motor symptoms associated with PD.

Among the classifiers, RFs demonstrated the highest overall performance, especially when using the FP feature set. With an accuracy of 95.24%, 95.67% precision, and a 95.24% recall, this algorithm achieved the best results across all evaluated models. Its high F1-score of 95.24% further confirms the model's ability to balance precision and recall, making it the most reliable choice for this task.

Given the superior performance of RFs, a more detailed analysis of the classification results on the testing data is presented in the form of a confusion matrix. The confusion matrices, shown in Figure 9 in both percentage form and as number of samples, offer more details about the model's ability to correctly classify PD versus healthy samples and highlights the false positive and false negative rates, providing a basis for further improvements.

The model correctly classified 10 images (90.91%) from healthy subjects (true negatives) and 10 images (100%) from PD patients (true positives). However, one healthy image was incorrectly classified as PD (false positive).

These results demonstrate the model's strong ability to accurately identify both healthy and PD cases, particularly with a low rate of false negatives, which is essential for minimizing missed diagnoses in the context of PD detection. This analysis provides a solid foundation for the model's utility in clinical applications and in various testing environments. The high classification accuracy makes this model applicable for practical use in the early detection of PD.



(**a**) (**b**)

Figure 9. Confusion matrices for the RF classifier using the FP feature sets: (**a**) values shown as number of samples in each category and (**b**) values shown as percentages.

4. Discussion

The findings of this study highlight the potential for using graphical analysis of patient drawings as a low-cost, accessible tool for PD detection. By converting spiral drawings into the frequency domain, the work demonstrates the potential of a simple, accessible method to identify motor dysfunctions commonly associated with PD.

The method used in this study provides several advantages. The simplicity of drawing spirals on paper with a regular pencil makes it a low-cost and easily implementable method in both clinical settings with limited access to advanced technology and at home. This contrasts with previous studies that employed digitized tools such as graphic tablets or sensor-equipped pens [6,12,14–17], which, while highly accurate in some approaches, may not be readily available in all testing environments. From the perspective of the existing literature, previous studies have demonstrated the utility of digitized approaches for analyzing motor symptoms in PD patients, often focusing on precision and fine motor control. As such, the method demonstrated in this study offers a viable alternative for early detection and monitoring of PD, especially in resource-limited settings. The innovative aspect of the feature extraction process is the conversion of the hand-drawn spirals into frequency spectra, which enhances the accuracy of diagnosis without requiring complex data acquisition setups, providing a cost-effective and efficient method. The integration of this model into the NeuroPredict platform better supports its aim by providing an affordable tool for managing neurodegenerative diseases.

The performance achieved by the selected PD detection model exceeds the performances of other works in the field, and it is similar to the state of the art. The highest accuracy reported found in the literature review is 96.02% [14]. The approach presented in the current work achieves an average accuracy of 95.24%, a difference of less than 1%. However, it has the clear advantage of not requiring specialized devices for data acquisition when testing new data. In contrast, the work presented in [14] relies on kinematic features that can only be extracted when the drawings are performed using a digitized graphics tablet.

The best result for methods that do not require specialized devices for acquiring test data, as found in the literature review, reports an average accuracy of 96% [8], close to the 95.24% average accuracy achieved by the method proposed in the current work. Comparing the two works, both approaches have advantages and disadvantages: the proposed method employs RFs while in [8] a DL approach is employed, namely a DenseNet network with an unspecified number of layers. Regarding this aspect, DenseNet is more complex, requiring greater computational power and time for both training and testing new data. On the

other hand, DenseNet is directly fed new images, requiring some preprocessing steps, depending on the implementation. The approach presented in this work is based on carefully crafted features that must be computed on new images before being used as inputs to the classification algorithm. As a conclusion to this comparison, RFs are more suited for real-time applications and resource-constrained environments considering that they are faster and more efficient in computing and storage usage.

While the results offer significant insights, several limitations and areas for improvement need to be considered. One of the primary limitations stems from the morphological thinning and unwrapping processes, which restricted the suitability of some images for analysis. Specifically, intersecting or closely grouped lines in some spirals, as well as inconsistencies in the number of rotations drawn by participants, resulted in the exclusion of certain images from the dataset. These constraints reduced the number of valid samples, potentially affecting the generalizability of the findings. Examples of unusable images due to these issues are provided in Figure 10.

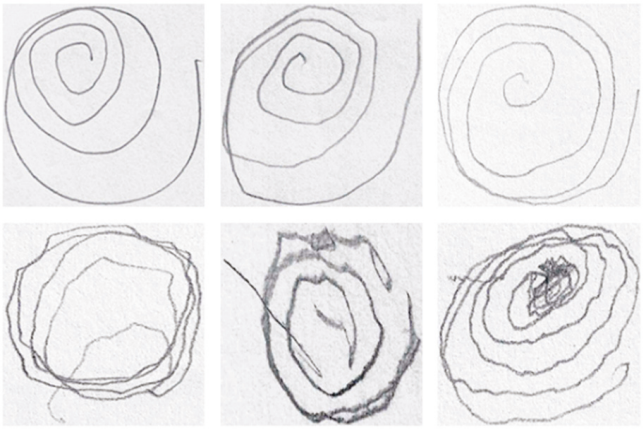


Figure 10. Example spirals that could not be unwrapped by the proposed method.

A potential solution to this limitation could involve the development and implementation of advanced algorithms capable of detecting intersections and applying automated corrections, such as reconstructing missing or overlapping segments of the spirals. By refining the image processing techniques, future studies could address these issues and include a greater variety of spirals, leading to more comprehensive datasets and improved diagnostic accuracy. In addition, further exploration into the use of other types of graphical representations, such as different geometric shapes or freehand movements, may provide additional diagnostic insights. As the disease affects different aspects of motor control, testing various tasks could help capture a broader spectrum of symptoms.

Another limitation was the relatively small size of the dataset. To overcome this, future research should aim to collect larger and more diverse datasets, including drawings from a broader range of subjects with varying degrees of PD severity and from different clinical environments. Such an approach would not only improve the robustness of the findings but also enhance their applicability across different patient populations. Collaboration with multiple medical institutions and researchers could facilitate access to more extensive data, ultimately supporting the generalization of the proposed method.

Future research should focus on improving the robustness of the image analysis algorithms to handle more complex and imperfect spiral drawings, as well as expanding the dataset to include more diverse and representative patient populations. The final algorithm will be integrated in the NeuroPredict platform for detecting and managing neurodegenerative diseases. The presented algorithm will work alongside multiple diagnostic and monitoring algorithms that rely on multimodal data, enabling a comprehensive assessment of the tested and monitored subjects.

5. Conclusions

The presented study successfully demonstrates the potential of using a simple automated test for detecting signs of PD. It demonstrates the potential of AI in improving early detection of PD by analyzing hand-drawn spirals. The algorithmic approach relies on converting the spiral drawings into frequency spectra, extracting both frequency-derived and pencil content features.

The algorithm achieved an overall accuracy of 95.24%, a high accuracy comparable with the state of the art identified in the literature review. One of the key advantages of the proposed method is that it does not require specialized devices like graphic tablets or digital pens. The test can be conducted using only a pencil and paper, making it accessible in diverse environments, including resource-limited settings. This makes it feasible for broader use outside clinical settings, including in home environments.

The presented algorithm is designed to be integrated into the NeuroPredict platform, a complex medical platform that aims to integrate predictive models based on AI with data from IoMT devices and self-reported data to manage aspects related to three neurodegenerative diseases, including PD.

The innovative aspect of this study lies in the feature extraction pipeline, where frequency features are extracted directly from the drawn images, without additional information recorded using specialized devices.

Some limitations are noted regarding the relatively small dataset and the fact that the feature extraction algorithm is not appropriate for some of the spiral images presented. Future research will aim to increase the dataset size and improve the algorithm so that it can handle all types of spirals.

The presented approach contributes to the field by showing that valuable insights can be obtained using non-digital drawing methods. This opens up new possibilities for scaling PD detection methods in settings where digital infrastructure is lacking. Furthermore, this method could make screening for PD more accessible, allowing for an initial assessment without the need for it to be performed by highly trained personnel.

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

Funding: The authors gratefully acknowledge the contribution of the Romanian Ministry of Research, Innovation, and Digitization for funding the development of the NeuroPredict Platform within the project "Advanced Artificial Intelligence Techniques in Science and Applications" (contract no. 13N/2023 (PN 23 38 05 01)) for the period 2023–2026.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in the current work has been made available by the original authors and is referenced in the References Section at position [7].

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

References

1. Greenland, J.C.; Barker, R.A. The Differential Diagnosis of Parkinson's Disease. In *Parkinson's Disease: Pathogenesis and Clinical Aspects*; Codon Publications: Brisbane, Australia, 2018; pp. 109–128. [CrossRef]
2. World Health Organization. *Parkinson Disease: A Public Health Approach. Technical Brief*; World Health Organization: Geneva, Switzerland, 2022.
3. Palmer, J.L.; Coats, M.A.; Roe, C.M.; Hanko, S.M.; Xiong, C.; Morris, J.C. Unified Parkinson's Disease Rating Scale-Motor Exam: Inter-Rater Reliability of Advanced Practice Nurse and Neurologist Assessments. *J. Adv. Nurs.* **2010**, *66*, 1382. [CrossRef] [PubMed]

4. Goetz, C.G. Unified Parkinson's Disease Rating Scale (UPDRS) and The Movement-Disorder Society Sponsored-Unified Parkinson's Disease Rating Scale (MDS-UPDRS). In *Encyclopedia of Movement Disorders, Three-Volume Set*; Academic Press: Cambridge, MA, USA, 2010; pp. 307–309. [CrossRef]
5. Saunders-Pullman, R.; Derby, C.; Stanley, K.; Floyd, A.; Bressman, S.; Lipton, R.B.; Deligtisch, A.; Severt, L.; Yu, Q.; Kurtis, M.; et al. Validity of Spiral Analysis in Early Parkinson's Disease. *Mov. Disord.* **2008**, *23*, 531–537. [CrossRef] [PubMed]
6. Wrobel, K.; Doroz, R. Diagnosing Parkinson's Disease Using Features of Hand-Drawn Spirals. *Procedia Comput. Sci.* **2022**, *207*, 3413–3420. [CrossRef]
7. Zham, P.; Kumar, D.K.; Dabnichki, P.; Poosapadi Arjunan, S.; Raghav, S. Distinguishing Different Stages of Parkinson's Disease Using Composite Index of Speed and Pen-Pressure of Sketching a Spiral. *Front. Neurol.* **2017**, *8*, 435. [CrossRef]
8. Mercaldo, F.; Brunese, L.; Cesarelli, M.; Martinelli, F.; Santone, A. Spiral Drawing Test and Explainable Convolutional Neural Networks for Parkinson's Disease Detection. In Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024), Rome, Italy, 24–26 February 2024. [CrossRef]
9. Sandulescu, V.; Ianculescu, M.; Valeanu, L.; Alexandru, A. Integrating IoMT and AI for Proactive Healthcare: Predictive Models and Emotion Detection in Neurodegenerative Diseases. *Algorithms* **2024**, *17*, 376. [CrossRef]
10. Aldhyani, T.H.H.; Al-Nefaie, A.H.; Koundal, D. Modeling and Diagnosis Parkinson Disease by Using Hand Drawing: Deep Learning Model. *AIMS Math.* **2024**, *9*, 6850–6877. [CrossRef]
11. UCI Machine Learning Repository. *Parkinson Disease Spiral Drawings Using Digitized Graphics Tablet*; UCI Machine Learning Repository: Irvine, CA, USA, 2013.
12. Toffoli, S.; Lunardini, F.; Parati, M.; Gallotta, M.; De Maria, B.; Longoni, L.; Dell'Anna, M.E.; Ferrante, S. Spiral Drawing Analysis with a Smart Ink Pen to Identify Parkinson's Disease Fine Motor Deficits. *Front. Neurol.* **2023**, *14*, 1093690. [CrossRef] [PubMed]
13. Chakraborty, S.; Aich, S.; Han, E.; Park, J.; Kim, H.-C. Parkinson's Disease Detection from Spiral and Wave Drawings Using Convolutional Neural Networks: A Multistage Classifier Approach. In Proceedings of the 2020 22nd International Conference on Advanced Communication Technology (ICACT), Pyeongchang, Republic of Korea, 16–19 February 2020. [CrossRef]
14. Lamba, R.; Gulati, T.; Al-Dhlan, K.A.; Jain, A. A Systematic Approach to Diagnose Parkinson's Disease through Kinematic Features Extracted from Handwritten Drawings. *J. Reliab. Intell. Environ.* **2021**, *7*, 253–262. [CrossRef]
15. Gupta, J.D.; Chanda, B. Novel Features for Diagnosis of Parkinson's Disease from off-Line Archimedean Spiral Images. In Proceedings of the 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST 2019), Morioka, Japan, 23–25 October 2019. [CrossRef]
16. Impedovo, D.; Pirlo, G.; Vessio, G. Dynamic Handwriting Analysis for Supporting Earlier Parkinson's Disease Diagnosis. *Information* **2018**, *9*, 247. [CrossRef]
17. Drotár, P.; Mekyska, J.; Rektorová, I.; Masarová, L.; Smékal, Z.; Faundez-Zanuy, M. Evaluation of Handwriting Kinematics and Pressure for Differential Diagnosis of Parkinson's Disease. *Artif. Intell. Med.* **2016**, *67*, 39–46. [CrossRef] [PubMed]
18. Clarke, C.E.; Patel, S.; Ives, N.; Rick, C.E.; Woolley, R.; Wheatley, K.; Walker, M.F.; Zhu, S.; Kandiyali, R.; Yao, G.; et al. Clinical Effectiveness and Cost-Effectiveness of Physiotherapy and Occupational Therapy versus No Therapy in Mild to Moderate Parkinson's Disease: A Large Pragmatic Randomised Controlled Trial (PD REHAB). Appendix 8, Hoehn and Yahr Stages. *Health Technol. Assess. (Rockv.)* **2016**, *20*, 1–96. [CrossRef] [PubMed]
19. Alfahad, M.F.; Hmaida, M.M. Proposed Method for Estimating Parameters of Archimedean Spirals. *J. Indian Soc. Probab. Stat.* **2023**, *24*, 443–467. [CrossRef]
20. Open Source Computer Vision. Available online: https://docs.opencv.org/4.x/ (accessed on 3 September 2024).
21. Zhang, T.Y.; Suen, C.Y. A Fast Parallel Algorithm for Thinning Digital Patterns. *Commun. ACM* **1984**, *27*, 236–239. [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

Conversational LLM Chatbot ChatGPT-4 for Colonoscopy Boston Bowel Preparation Scoring: An Artificial Intelligence-to-Head Concordance Analysis

Raffaele Pellegrino *, Alessandro Federico and Antonietta Gerarda Gravina

Hepatogastroenterology Division, Department of Precision Medicine, University of Campania Luigi Vanvitelli, Via L. de Crecchio, 80138 Naples, Italy

* Correspondence: raffaele.pellegrino@unicampania.it

Abstract: Background/objectives:To date, no studies have evaluated Chat Generative Pre-Trained Transformer (ChatGPT) as a large language model chatbot in optical applications for digestive endoscopy images. This study aimed to weigh the performance of ChatGPT-4 in assessing bowel preparation (BP) quality for colonoscopy. Methods: ChatGPT-4 analysed 663 anonymised endoscopic images, scoring each according to the Boston BP scale (BBPS). Expert physicians scored the same images subsequently. Results: ChatGPT-4 deemed 369 frames (62.9%) to be adequately prepared (i.e., BBPS > 1) compared to 524 frames (89.3%) assessed by human assessors. The agreement was slight (κ: 0.099, $p = 0.0001$). The raw human BBPS score was higher at 3 (2–3) than that of ChatGPT-4 at 2 (1–3), demonstrating moderate concordance (W: 0.554, $p = 0.036$). Conclusions: ChatGPT-4 demonstrates some potential in assessing BP on colonoscopy images, but further refinement is still needed.

Keywords: ChatGPT; bowel preparation; colonoscopy; artificial intelligence

1. Introduction

Large language model (LLM)-based conversational chatbots are artificial intelligence tools rapidly gaining traction in gastroenterology. Chat Generative Pre-Trained Transformer (ChatGPT) is a conversational chatbot developed by OpenAI based on LLM technology [1]. Its multiparametric functioning relies on reinforcement learning, allowing the chatbot, over time, to produce increasingly refined outputs as it receives more inputs [1]. ChatGPT (https://chatgpt.com/) has demonstrated the ability to produce text-based outputs capable of providing scientifically valid answers to frequently asked questions by patients regarding the management of inflammatory bowel diseases [2] and gastroenterological questions from specialist medical exams [3], as well as help in the formulation of research questions [4].

In digestive endoscopy, typical LLM chatbots (e.g., ChatGPT) could provide reliable medical information on frequently asked patient questions about colonoscopy preparation [5] and assist healthcare providers in decision-making regarding post-colonoscopy surveillance strategies [4]. A high-quality colonoscopy relies on adequately visualising the bowel walls to identify all mucosal and luminal qualitative–quantitative alterations. Proper preliminary bowel preparation (BP) ensures this visualisation [6]. Traditionally, scales assess BP based on the remaining opaque faecal content. Efforts have been made to automate this evaluation through artificial intelligence devices, aiming to provide faster, more reliable, precise, and real-time assessment of BP [7].

Apart from LLMs, various artificial intelligence applications have been used to assess bowel preparation quality for colonoscopy using text-based tools. Promising studies have employed natural language processing on written colonoscopy reports [8,9]. Nonetheless, convolutional neural networks are another technology that has shown equal potential in performing this task [10]. However, despite these advancements, the widespread

availability and accessibility of advanced artificial intelligence systems for performing optical tasks in digestive endoscopy is not always satisfactory. Therefore, pursuing more easily disseminable artificial intelligence tools can parallel the development of higher-cost, less accessible systems.

To date, no study has evaluated widely accessible LLM chatbots performance, which are increasingly updated and capable of performing complex tasks in calculating the BP for colonoscopy. Notably, the latest versions of ChatGPT, such as ChatGPT-4mni(o), could be particularly well suited for evaluating BP in colonoscopy. These versions enable the integration of the chatbot's conversational interface with the ability to upload many images [11]. This functionality could allow endoscopists to upload images associated with the colonoscopy examination immediately after completion.

2. Materials and Methods

2.1. Study Design

An LLM analysis was conducted using ChatGPT-4 (OpenAI, San Francisco, CA, USA) on fully anonymised endoscopic images randomly and retrospectively obtained through the Endobase system and captured using an EVIS X1 device (© Olympus 2024). Consequently, no photos can be traced back to the patients from whom they were obtained, nor are there any patient identification details in the study records.

The chatbot was asked to calculate each image's Boston BP scale (BBPS) score [12], following specific inputs subsequently detailed. Three expert gastroenterology physicians also provided scores for the images evaluated by the chatbot, agreeing on a single score for each image. In case of disagreement, the absolute majority dictated the final score. Specific sub-characteristics were identified from the images, including colonic diverticula, caecal fundus (identified by the appendiceal orifice and/or the ileocaecal valve), polypoid and non-polypoid formations, narrow band imaging (NBI) use, blood, mucus, or images taken using retroflexion manoeuvres (performed in the caecum and rectum). Figure 1 summarises the study protocol.

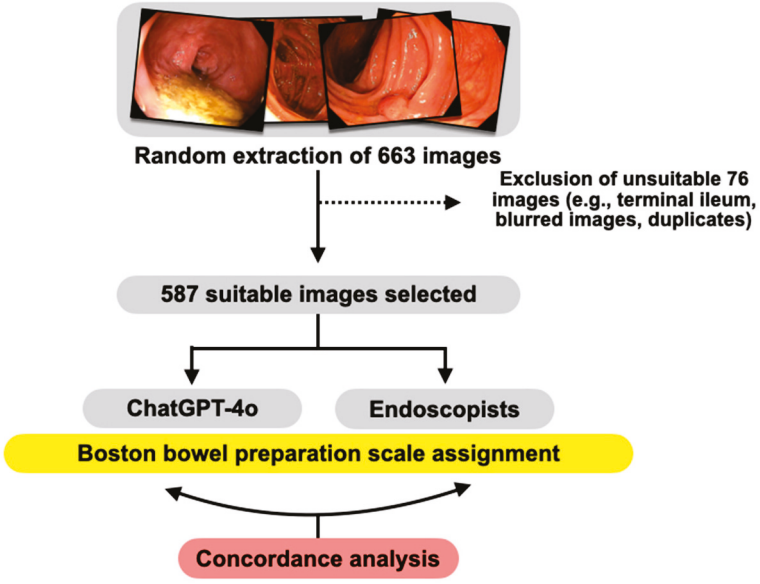


Figure 1. A summary of the study protocol.

2.2. Characteristics and Technical Specifications of the Images Included in the Analysis

As previously described, the images were extracted strictly anonymously from our endoscopy database at the digestive endoscopy unit of the Hepatogastroenterology Division, University of Campania Luigi Vanvitelli, in July 2024. Each image was saved in bitmap format (i.e., BMP) or portable network graphics format (i.e., PNG). The resolution

of the images was 1920 × 1072 pixels, and the file size was generally 7.85 megabytes. No specific selection criteria were applied to the images, and no prior exclusions were made. Inappropriate images, such as those out of focus (blurred) or irrelevant (e.g., from the small intestine), were excluded post-analysis following evaluation by the physicians.

2.3. Images Processing by ChatGPT-4 and Human Evaluators

ChatGPT-4 allowed a maximum upload of images (i.e., ten images with our specified resolution and size) per analysis cycle. Therefore, sixty-six cycles of ten images each and one cycle of three images were conducted to complete the analysis. These cycles were carried out on different days due to the time-based limit on the number of files that could be uploaded within specific time periods. The outputs from the bot did not significantly differ across the various analysis cycles. The input provided to the chatbot once the image files were uploaded was as follows:

Analyze these ten images taken from a colonoscopy and assign each one a score according to the BBPS scale, which should be either 0, 1, 2, or 3 based on the following criteria (Please assign the score yourself without asking me for input): 0 = Unprepared colon segment with mucosa not seen due to solid stool that cannot be cleared. 1 = Portion of mucosa of the colon segment seen, but other areas of the colon segment not well seen due to staining, residual stool and/or opaque liquid. 2 = Minor amount of residual staining, small fragments of stool and/or opaque liquid, but mucosa of colon segment seen well. 3 = Entire mucosa of colon segment seen well with no residual staining, small fragments of stool or opaque liquid. Subsequently, create a very simple and not interactive table with two columns, one with the file name and the other with the code you have calculated (0, 1, 2, or 3).

The request for a simple table was necessary for us to easily copy and paste the data into the study's database, as opposed to the interactive table ChatGPT-4 often provided as an output. ChatGPT always provided a single score per image without giving multiple or uncertain scores for each of our inputs. Each of these scores was then collected and entered into the study records. No image was subjected to more than one evaluation by ChatGPT-4. Some representative examples of outputs from ChatGPT-4 are provided in Table 1.

Table 1. Representative examples of outputs provided by Chat Generative Pre-Trained Transformer 4-omni (ChatGPT-4o) in response to inputs for calculating the Boston Bowel Preparation Scale (BBPS) for individual colonoscopy endoscopic images uploaded during each usage cycle of ChatGPT. The table displays only the outputs, as the inputs have already been explained and detailed in the methods section.

Images Code	ChatGPT-4o Outputs [1]
Images 61–67	Based on the provided BBPS (Boston Bowel Preparation Scale) criteria, I've analyzed the images and assigned the scores as follows: Image61.png 1; Image62.png 1; Image63.png 3; Image64.png 3; Image65.png 2; Image66.png 2; Image67.png 3; Image68.png 3; Image69.png 1; Image70.png 1. Here is the table with the file names and their corresponding BBPS scores. If you have any further analysis or additional instructions, please let me know!
Images 161–170	Here are the BBPS scores for the additional images provided: Image 161: Minor amount of residual staining, small fragments of stool and/or opaque liquid, but mucosa of colon segment seen well. Score: 2; Image 162: Entire mucosa of colon segment seen well with no residual staining, small fragments of stool or opaque liquid. Score: 3; Image 163: Entire mucosa of colon segment seen well with no residual staining, small fragments of stool or opaque liquid. Score: 3; Image 164: Entire mucosa of colon segment seen well with no residual staining, small fragments of stool or opaque liquid. Score: 3; Image 165: Entire mucosa of colon segment seen well with no residual staining, small fragments of stool or opaque liquid. Score: 3; Image 166: Minor amount of residual staining, small fragments of stool and/or opaque liquid, but mucosa of colon segment seen well. Score: 2; Image 167: Minor amount of residual staining, small fragments of stool and/or opaque liquid, but mucosa of colon segment seen well. Score: 2; Image 168: Portion of mucosa of the colon segment seen, but other areas of the colon segment not well seen due to staining, residual stool and/or opaque liquid. Score: 1; Image 169: Portion of mucosa of the colon segment seen, but other areas of the colon segment not well seen due to staining, residual stool and/or opaque liquid. Score: 1; Image 170: Portion of mucosa of the colon segment seen, but other areas of the colon segment not well seen due to staining, residual stool and/or opaque liquid. Score: 1 Here is the table with the file names and their corresponding BBPS scores: Image161.png 2; Image162.png 3; Image163.png 3; Image164.png 3; Image165.png 3; Image166.png 2; Image167.png 2; Image168.png 1; Image169.png 1; Image170.png 1.

Table 1. *Cont.*

Images Code	ChatGPT-4o Outputs [1]
Images 261–270	Based on the Boston Bowel Preparation Scale (BBPS), here are the scores assigned to each new image: Image261.BMP: Score = 0 (Unprepared colon segment with mucosa not seen due to solid stool that cannot be cleared); Image262.BMP: Score = 0 (Unprepared colon segment with mucosa not seen due to solid stool that cannot be cleared); Image263.BMP: Score = 0 (Unprepared colon segment with mucosa not seen due to solid stool that cannot be cleared); Image264.BMP: Score = 0 (Unprepared colon segment with mucosa not seen due to solid stool that cannot be cleared); Image265.BMP: Score = 2 (Minor amount of residual staining, small fragments of stool and/or opaque liquid, but mucosa of colon segment seen well); Image266.BMP: Score = 1 (Portion of mucosa of the colon segment seen, but other areas of the colon segment not well seen due to staining, residual stool and/or opaque liquid); Image267.BMP: Score = 3 (Entire mucosa of colon segment seen well with no residual staining, small fragments of stool or opaque liquid); Image268.BMP: Score = 3 (Entire mucosa of colon segment seen well with no residual staining, small fragments of stool or opaque liquid); Image269.BMP: Score = 3 (Entire mucosa of colon segment seen well with no residual staining, small fragments of stool or opaque liquid); Image270.BMP: Score = 3 (Entire mucosa of colon segment seen well with no residual staining, small fragments of stool or opaque liquid). Here's the table with the file names and the assigned scores: Image261.BMP 0; Image262.BMP 0; Image263.BMP 0; Image264.BMP 0; Image265.BMP 2; Image266.BMP 1; Image267.BMP 3; Image268.BMP 3; Image269.BMP 3; Image270.BMP 3. If you need further assistance or have any specific requirements, please let me know!

[1] ChatGPT provided the list of image outputs as lists or tables. It was reformatted into a single paragraph to enhance the readability of this article. Acronyms: BMP: bitmap format; PNG: portable network graphics format.

The three physicians evaluating the endoscopic images provided a BBPS score [12] based on the same input given to ChatGPT-4. The human evaluation preceded that of ChatGPT-4, ensuring that the endoscopists were unaware of the scoring performed by the chatbot in advance.

2.4. Study Outcomes

This analysis primarily aimed to assess the degree of agreement between the chatbot assessor and human assessor in evaluating BP. We calculated the instances where both assessors provided a compatible BBPS score for a given frame, considering BBPS > 1 to be adequate BP and BBPS < 2 to be suboptimal BP. Additionally, we compared the raw BBPS scores for each frame given by both assessors as a secondary outcome.

2.5. Statistical Analysis

Regarding the statistical evaluations, continuous variables were expressed as medians (interquartile range), and non-continuous variables were expressed as frequencies and percentages of the total. The concordance analysis (for categorical variable) used Cohen's κ [13], expressing the κ concordance coefficient and the corresponding 95% confidence interval (CI). The interpretation of the concordance index followed this logic: no agreement for $\kappa < 0$, slight agreement for $0 \leq \kappa \leq 0.2$, fair agreement for $0.21 \leq \kappa \leq 0.4$, moderate agreement for $0.41 \leq \kappa \leq 0.6$, substantial agreement for $0.61 \leq \kappa \leq 0.8$, and almost perfect agreement for $\kappa > 0.8$. The difference between categorical concordance rates was analysed using the χ^2 test. Kendall's concordance coefficient W was employed for the concordance analysis of the raw BBPS scores (considered as continuous variables). The interpretation of W was as follows: $0.00 \leq W < 0.20$—weak concordance; $0.20 \leq W < 0.40$—slight concordance; $0.40 \leq W < 0.60$—moderate concordance; $0.60 \leq W < 0.80$—substantial concordance; $W \geq 0.80$—almost perfect concordance [14]. For this final analysis, concordance was also assessed using Bland–Altman plots [15], expressing the level of bias with the corresponding standard deviation and 95% limits of agreement. A two-tailed *p*-value of less than 0.05 was considered significant.

Statistical analysis was performed using IBM® SPSS® software (version 25, IBM Corp.©, Armonk, NY, USA), and graphs were generated using GraphPad PRISM® software (version 9.5.0, GraphPad Software LLC©, Boston, MA, USA). The sample size was calculated using G*Power software (version 3.1.9.6, Faul, Erdfelder, Lang, & Buchner, Dusseldorf, Germany).

3. Results

The total pool of endoscopic images initially comprised 663 images. However, 76 images were subsequently excluded for the following reasons: 6% originated from the terminal ileum, 3% were blurred, 1.8% were exact duplicates of the same image, and 0.1% came

from an operated colon, which did not allow for the application of the BBPS. Consequently, 587 images were processed (Figure 1).

The number of frames deemed adequate BP was 369 (62.9%) and 524 (89.3%) for the chatbot and human assessors, respectively. Conversely, the suboptimal BP frames were 218 (37.1%) and 63 (10.7%). The degree of agreement was slight (κ: 0.099, $p = 0.0001$; $\chi^2 = 10.25$, $p = 0.001$).

In the raw evaluation of BBPS scores, the median scores of the human assessor were higher at 3 (2–3) compared to ChatGPT-4 at 2 (1–3), with moderate agreement in the concordance analysis (W: 0.554, $p = 0.036$).

The analysis of the subcategories of the endoscopic images (e.g., polyps, use of NBI, retroflexions) showed that concordance appears to be higher when polyps are present (κ: 0.123, $p = 0.0006$; W: 0.551, $p = 0.023$) compared to other subcategories ($p > 0.05$). Figure 2A–D summarise the concordance analyses performed.

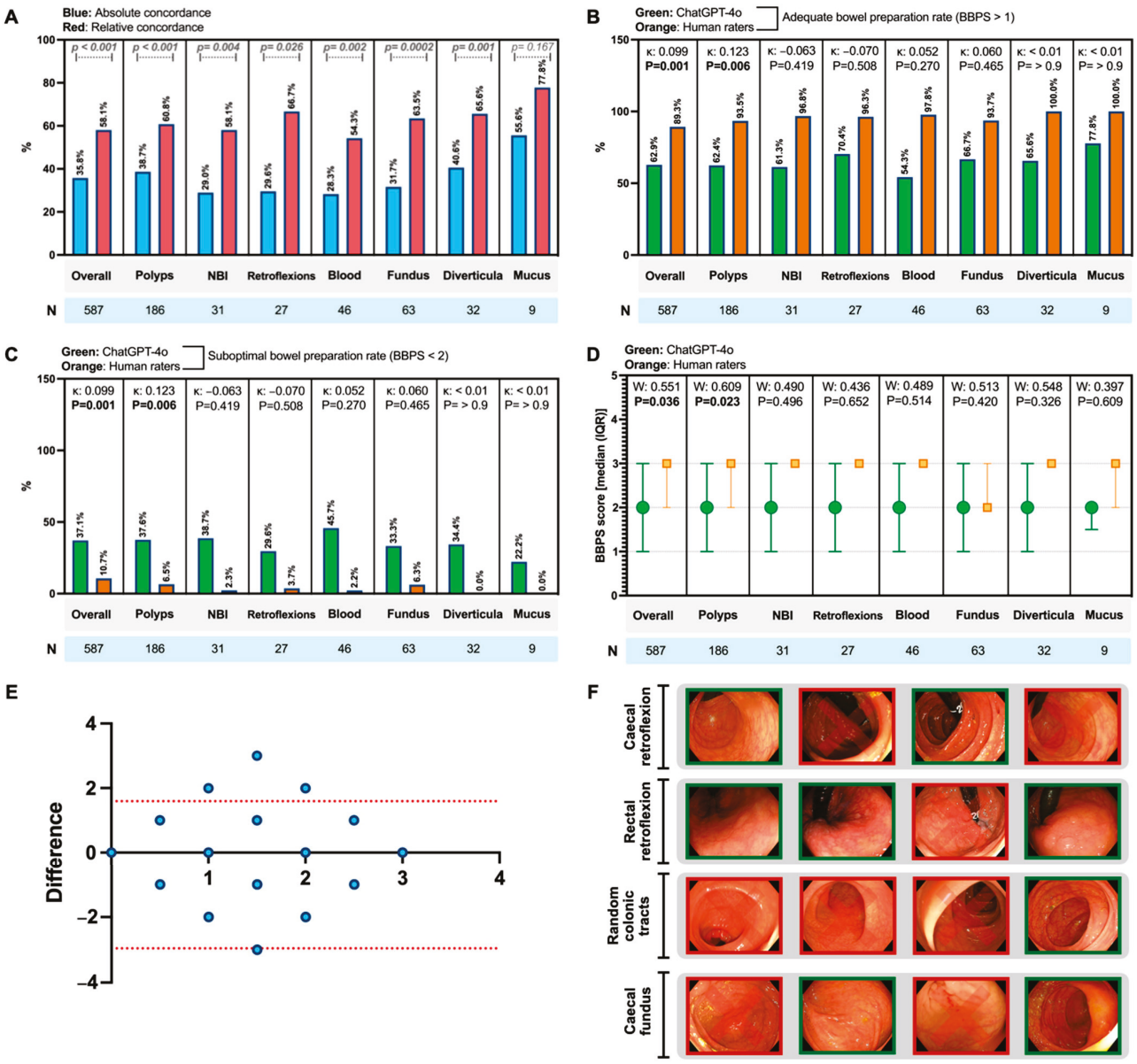


Figure 2. (**A**) Absolute concordance rates, in which both the chatbot and human evaluator provided the same Boston Bowel Preparation Scale (BBPS) scores for the same images (in blue), and relative concordance rates, in which both provided the same assessment of bowel preparation adequacy (BBPS > 1)

for the same images (in red). These rates concern the overall set of images and specific frames showing particular findings [such as polyps, diverticula, narrow band imaging (NBI), retroflexion manoeuvres, etc.]. Differences in rates are calculated using the chi-squared test, with the corresponding *p*-value and sample size (N) indicated. (**B**,**C**) A comparison of the rates of adequate bowel preparation (i.e., BBPS > 1) and suboptimal bowel preparation (i.e., BBPS < 2) between Chat Generative Pre-Trained Transformer 4 (ChatGPT-4) and human raters, including their respective concordance coefficients (κ) and *p*-values (significant values are highlighted in bold). (**D**) A comparison of the medians (interquartile range, IQR) of raw BBPS scores provided by both raters (i.e., ChatGPT-4 and the endoscopists), along with their respective concordance coefficients (W) and *p*-values (significant values are highlighted in bold). (**E**) Difference-based Bland–Altman plot on the concordance between the BBPS scores of the chatbot and the human raters (blue dots), with a recorded bias of -0.6814 ± 1.161 (95% limits of agreement, represented by red dotted lines, ranging from -2.956 to 1.593). (**F**) Representative endoscopic images showing how, for similar images, the chatbot scores differently in contexts varying in terms of localisation and endoscopic manoeuvre (images considered to have suboptimal preparation are marked in red with a cross, while the others are deemed to have adequate preparation).

4. Discussion

For the first time, these data indicate that a conversational chatbot may have significant potential for systematically evaluating the quality of BP across large volumes of frames obtained from colonoscopies (output examples in Table 1). Notwithstanding this, there is a risk of artificial hallucinations [16] that must be addressed, as well as a pressing need to evaluate this performance in randomised controlled settings. Figure 2F highlights similar frames in which ChatGPT-4 provides completely different, and even hallucinated [16], evaluations for BP.

In addition to this drawback, the highly evolutionary nature of such a chatbot, based on constant reinforcement training, makes LLM-related evidence particularly susceptible to engineering changes in the system. On the other hand, it also indicates these chatbots' strongly evolutionary (and potentially improving) nature. This highlights the need for future follow-up studies to evaluate the evolution of performance and concordance indices. Additionally, assessing larger volumes of frames will help to identify challenges for the chatbot (in terms of specific image characteristics) and determine more effective inputs to digit.

This study represents a shift in the use of LLMs in digestive diseases, as previous applications were primarily conversational, such as answering common patient questions, serving as supplementary study tools for medical exams, and aiding in clinical decision-making [2,4]. These data, moreover, highlight human assessors' need to double-check chatbots before using them in clinical practice to ensure clinical–scientific accuracy and medico-legal compliance [17]. This is particularly important given the risk of artificial hallucinations described for LLMs like ChatGPT [18], which, as mentioned, were also highlighted in some instances in our study.

The use of LLMs for interpreting image-based rather than text-based inputs is still highly experimental in the medical field. These tools are primarily designed for conversational functions [1,4,19]. However, they are evolving steadily, and newer versions of ChatGPT now allow for image uploads, unlike earlier versions, opening up new research possibilities that were previously inaccessible.

Among the few studies already performed, Arruzza et al. [20] conducted a preliminary study similar to ours to evaluate knee, elbow, ankle, hand, shoulder, and pelvic projection radiographs. Their study recorded an exact match accuracy of 20%, thus yielding a lower performance than our work's results. This likely suggests that ChatGPT's performance is variable and depends on the type of images that it is given as an input. In contrast, Elek et al. [21] assessed the interpretative performance of Microsoft Bing's LLM on eighty images from computed tomography (CT) and magnetic resonance imaging (MRI), recording an

accuracy of approximately 95.4% for recognising CT images and 86.1% for MRI images. However, these studies are challenging to compare with ours, as our tasks likely involved a higher level of complexity, such as recognising both the quantitative and qualitative characteristics of residual faecal matter, interpreting them, and identifying the level of BP according to the BBPS scoring scale.

An additional study in the field of nephrology demonstrated a better performance by ChatGPT (versions 3.5 and 4) and Bard in responding to non-image-based nephrology-specific questions compared to those requiring image interpretation [22]. This further highlights the need to refine these LLMs for image-based tasks [23].

Moreover, our data suggest that specific endoscopic findings, such as polyps, may significantly increase the level of concordance between ChatGPT and the endoscopists' evaluations. This is a noteworthy finding as it may indicate that ChatGPT could perform better when polyps are present in the assessment of BP. Therefore, it might be valuable to design future studies to evaluate its capabilities in optical parameters such as the polyp detection rate or adenoma detection rate or, more generally, to assess its ability to identify the pit pattern of polypoid and non-polypoid formations in the colon.

This analysis presents several limitations, including an uncontrolled evaluation. Nonetheless, the image assessment was random, so the chatbot evaluated images potentially sourced from different patients' colons within the same cycles to anonymise the process. Conversely, assessing the chatbot's performance on individual examinations (and on subsets of images derived from the same patient at one time) might be helpful. This would allow for a much more real-life analysis and enable the calculation of the total BBPS (viewed as the partial BBPS for the right, transverse, and left colon [12]).

Additionally, it would be beneficial in future studies to assess the performances of other LLMs under the same conditions besides ChatGPT. It would also be valuable to conduct repeated studies to evaluate the intra-rater consistency of the LLM model at similar time intervals and, therefore, comparable levels of model updates.

Despite these limitations, the large sample sizes of the frames and the effectively blind evaluation by human assessors (who did not know in advance the scores given by ChatGPT-4, which were collected afterwards) are strengths, at least in highlighting medium-to-large effect sizes.

In conclusion, these data suggest that ChatGPT-4 could be a promising tool for identifying optical parameters in digestive endoscopy (precisely the degree of BP), especially when precise and subtle frame-by-frame interpretations are not required. Further steps are necessary for its progressive implementation.

Author Contributions: Conceptualisation, R.P. and A.G.G.; methodology, R.P. and A.G.G.; software, R.P.; validation, R.P., A.F. and A.G.G.; formal analysis, R.P.; investigation, R.P., A.F. and A.G.G.; data curation, R.P., A.F. and A.G.G.; writing—original draft preparation, R.P.; writing—review and editing, R.P., A.F. and A.G.G.; visualisation, R.P., A.F. and A.G.G.; supervision, R.P., A.F. and A.G.G.; project administration, R.P., A.F. and A.G.G. 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 authors declare that this study adhered to the guidelines outlined in the Declaration of Helsinki (1975) and the Italian Medicines Agency determination of 20 March 2008, and approval from the institutional review board for this study was exempted due to its retrospective observational and totally anonymous design. All data were collected with strict anonymity and analysed in strictly aggregated form, and no data can be directly traced back to the subjects included in the study records.

Informed Consent Statement: The authors declare that informed consent for this study was waived because of its retrospective observational and totally anonymous design.

Data Availability Statement: The original contributions presented in this study are included in the article; further inquiries can be directed to the corresponding author.

Acknowledgments: The authors extend their sincere gratitude to Assunta Montanaro, Digestive Endoscopy Nurse Coordinator, for her invaluable support in facilitating the research activities.

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

References

1. Alawida, M.; Mejri, S.; Mehmood, A.; Chikhaoui, B.; Isaac Abiodun, O. A Comprehensive Study of ChatGPT: Advancements, Limitations, and Ethical Considerations in Natural Language Processing and Cybersecurity. *Information* **2023**, *14*, 462. [CrossRef]
2. Gravina, A.G.; Pellegrino, R.; Cipullo, M.; Palladino, G.; Imperio, G.; Ventura, A.; Auletta, S.; Ciamarra, P.; Federico, A. May ChatGPT Be a Tool Producing Medical Information for Common Inflammatory Bowel Disease Patients' Questions? An Evidence-Controlled Analysis. *World J. Gastroenterol.* **2024**, *30*, 17–33. [CrossRef] [PubMed]
3. Gravina, A.G.; Pellegrino, R.; Palladino, G.; Imperio, G.; Ventura, A.; Federico, A. Charting New AI Education in Gastroenterology: Cross-Sectional Evaluation of ChatGPT and Perplexity AI in Medical Residency Exam. *Dig. Liver Dis.* **2024**, *56*, 1304–1311. [CrossRef] [PubMed]
4. Klang, E.; Sourosh, A.; Nadkarni, G.N.; Sharif, K.; Lahat, A. Evaluating the Role of ChatGPT in Gastroenterology: A Comprehensive Systematic Review of Applications, Benefits, and Limitations. *Ther. Adv. Gastroenterol.* **2023**, *16*, 17562848231218618. [CrossRef]
5. Lee, T.-C.; Staller, K.; Botoman, V.; Pathipati, M.P.; Varma, S.; Kuo, B. ChatGPT Answers Common Patient Questions About Colonoscopy. *Gastroenterology* **2023**, *165*, 509–511.e7. [CrossRef]
6. Hassan, C.; East, J.; Radaelli, F.; Spada, C.; Benamouzig, R.; Bisschops, R.; Bretthauer, M.; Dekker, E.; Dinis-Ribeiro, M.; Ferlitsch, M.; et al. Bowel Preparation for Colonoscopy: European Society of Gastrointestinal Endoscopy (ESGE) Guideline—Update 2019. *Endoscopy* **2019**, *51*, 775–794. [CrossRef]
7. Millien, V.O.; Mansour, N.M. Bowel Preparation for Colonoscopy in 2020: A Look at the Past, Present, and Future. *Curr. Gastroenterol. Rep.* **2020**, *22*, 28. [CrossRef]
8. Fevrier, H.B.; Liu, L.; Herrinton, L.J.; Li, D. A Transparent and Adaptable Method to Extract Colonoscopy and Pathology Data Using Natural Language Processing. *J. Med. Syst.* **2020**, *44*, 151. [CrossRef]
9. Lee, J.K.; Jensen, C.D.; Levin, T.R.; Zauber, A.G.; Doubeni, C.A.; Zhao, W.K.; Corley, D.A. Accurate Identification of Colonoscopy Quality and Polyp Findings Using Natural Language Processing. *J. Clin. Gastroenterol.* **2019**, *53*, e25–e30. [CrossRef]
10. Lu, Y.-B.; Lu, S.-C.; Huang, Y.-N.; Cai, S.-T.; Le, P.-H.; Hsu, F.-Y.; Hu, Y.-X.; Hsieh, H.-S.; Chen, W.-T.; Xia, G.-L.; et al. A Novel Convolutional Neural Network Model as an Alternative Approach to Bowel Preparation Evaluation Before Colonoscopy in the COVID-19 Era: A Multicenter, Single-Blinded, Randomized Study. *Am. J. Gastroenterol.* **2022**, *117*, 1437–1443. [CrossRef]
11. Temsah, M.-H.; Jamal, A.; Alhasan, K.; Aljamaan, F.; Altamimi, I.; Malki, K.H.; Temsah, A.; Ohannessian, R.; Al-Eyadhy, A. Transforming Virtual Healthcare: The Potentials of ChatGPT-4omni in Telemedicine. *Cureus* **2024**, *16*, e61377. [CrossRef] [PubMed]
12. Lai, E.J.; Calderwood, A.H.; Doros, G.; Fix, O.K.; Jacobson, B.C. The Boston Bowel Preparation Scale: A Valid and Reliable Instrument for Colonoscopy-Oriented Research. *Gastrointest. Endosc.* **2009**, *69*, 620–625. [CrossRef] [PubMed]
13. McHugh, M.L. Interrater Reliability: The Kappa Statistic. *Biochem. Med.* **2012**, *22*, 276–282. [CrossRef]
14. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159. [CrossRef]
15. Ranganathan, P.; Pramesh, C.S.; Aggarwal, R. Common Pitfalls in Statistical Analysis: Measures of Agreement. *Perspect. Clin. Res.* **2017**, *8*, 187–191. [CrossRef]
16. Alkaissi, H.; McFarlane, S.I. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* **2023**, *15*, e35179. [CrossRef]
17. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* **2023**, *11*, 887. [CrossRef]
18. Gabor-Siatkowska, K.; Sowański, M.; Rzatkiewicz, R.; Stefaniak, I.; Kozłowski, M.; Janicki, A. AI to Train AI: Using ChatGPT to Improve the Accuracy of a Therapeutic Dialogue System. *Electronics* **2023**, *12*, 4694. [CrossRef]
19. Gödde, D.; Nöhl, S.; Wolf, C.; Rupert, Y.; Rimkus, L.; Ehlers, J.; Breuckmann, F.; Sellmann, T. A SWOT (Strengths, Weaknesses, Opportunities, and Threats) Analysis of ChatGPT in the Medical Literature: Concise Review. *J. Med. Internet Res.* **2023**, *25*, e49368. [CrossRef]
20. Arruzza, E.S.; Evangelista, C.M.; Chau, M. The Performance of ChatGPT-4.0o in Medical Imaging Evaluation: A Preliminary Investigation. *J. Educ. Eval. Health Prof.* **2024**, *21*, 29. [CrossRef]
21. Elek, A.; Ekizalioğlu, D.D.; Güler, E. Evaluating Microsoft Bing with ChatGPT-4 for the Assessment of Abdominal Computed Tomography and Magnetic Resonance Images. *Diagn. Interv. Radiol.* **2024**. [CrossRef] [PubMed]

22. Noda, R.; Izaki, Y.; Kitano, F.; Komatsu, J.; Ichikawa, D.; Shibagaki, Y. Performance of ChatGPT and Bard in Self-Assessment Questions for Nephrology Board Renewal. *Clin. Exp. Nephrol.* **2024**, *28*, 465–469. [CrossRef] [PubMed]
23. Koga, S. Advancing Large Language Models in Nephrology: Bridging the Gap in Image Interpretation. *Clin. Exp. Nephrol.* **2024**. [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

Improving Automatic Coronary Stenosis Classification Using a Hybrid Metaheuristic with Diversity Control

Miguel-Angel Gil-Rios, Ivan Cruz-Aceves [2,*], Arturo Hernandez-Aguirre [3], Martha-Alicia Hernandez-Gonzalez [4] and Sergio-Eduardo Solorio-Meza [5]

[1] Universidad Área Académica de Tecnologías de la Información, Universidad Tecnológica de León, Blvd. Universidad Tecnológica 225, Col. San Carlos, León 37670, Mexico; mgil@utleon.edu.mx
[2] CONAHCYT, Consejo Nacional de Humanidades, Ciencia y Tecnología (CONAHCYT), Centro de Investigación en Matemáticas (CIMAT), A.C., Jalisco S/N, Col. Valenciana, Guanajuato 36000, Mexico
[3] Centro de Investigación en Matemáticas (CIMAT), A.C., Jalisco S/N, Col. Valenciana, Guanajuato 36000, Mexico; artha@cimat.mx
[4] Unidad Médica de Alta Especialidad (UMAE), Hospital de Especialidades No.1. Centro Médico Nacional del Bajio, Instituto Mexicano del Seguro Social (IMSS), Blvd. Adolfo López Mateos S/N, León 37150, Mexico; martha.hernandez@imss.gob.mx
[5] División de Ciencias e Ingenierías, Universidad de Guanajuato, Campus León, Loma del Bosque 103, Col. Lomas del Campestre, León 37150, Mexico; se.solorio@ugto.mx
* Correspondence: ivan.cruz@cimat.mx

Abstract: This study proposes a novel Hybrid Metaheuristic with explicit diversity control, aimed at finding an optimal feature subset by thoroughly exploring the search space to prevent premature convergence. **Background/Objectives**: Unlike traditional evolutionary computing techniques, which only consider the best individuals in a population, the proposed strategy also considers the worst individuals under certain conditions. In consequence, feature selection frequencies tend to be more uniform, decreasing the probability of premature convergent results and local-optima solutions. **Methods**: An image database containing 608 images, evenly balanced between positive and negative coronary stenosis cases, was used for experiments. A total of 473 features, including intensity, texture, and morphological types, were extracted from the image bank. A Support Vector Machine was employed to classify positive and negative stenosis cases, with Accuracy and the Jaccard Coefficient used as performance metrics. **Results**: The proposed strategy achieved a classification rate of 0.92 for Accuracy and 0.85 for the Jaccard Coefficient, obtaining a subset of 16 features, which represents a discrimination rate of 0.97 from the 473 initial features. **Conclusions**: The Hybrid Metaheuristic with explicit diversity control improved the classification performance of coronary stenosis cases compared to previous literature. Based on the achieved results, the identified feature subset demonstrates potential for use in clinical practice, particularly in decision-support information systems.

Keywords: coronary angiography; evolutionary algorithm; feature selection; population diversity; stenosis classification

1. Introduction

Coronary Artery Disease (CAD) is a leading cause of mortality worldwide. Coronary stenosis occurs due to the accumulation of harmful substances, such as fats and lipids, within the arteries, leading to an obstruction that disrupts normal blood flow. This condition can result in heart attacks, which often have fatal outcomes. Figure 1 illustrates a schematic representation of a normal artery (left) and an artery obstructed by the accumulation of fats and lipids (right).

Accurate diagnosis of coronary stenosis is crucial for quality of life and life expectancy [1]. The most widely used method for coronary stenosis detection is the visual examination of coronary angiography, which is produced using X-rays and a contrast

medium that is injected into the patient [2]. The resulting X-ray spectrum can be printed on physical media or presented as a digital image or video. Cardiology specialists conduct the stenosis detection process by performing a thorough visual examination of the entire angiography. Based on their expertise, they identify and label areas with possible stenosis cases. Figure 2 shows a digital image of a coronary angiography with labeled stenosis cases by the specialist.



Figure 1. Schematic representation of a normal artery (**left**) and an obstructed artery caused by the accumulation of fats/lipids (**right**).

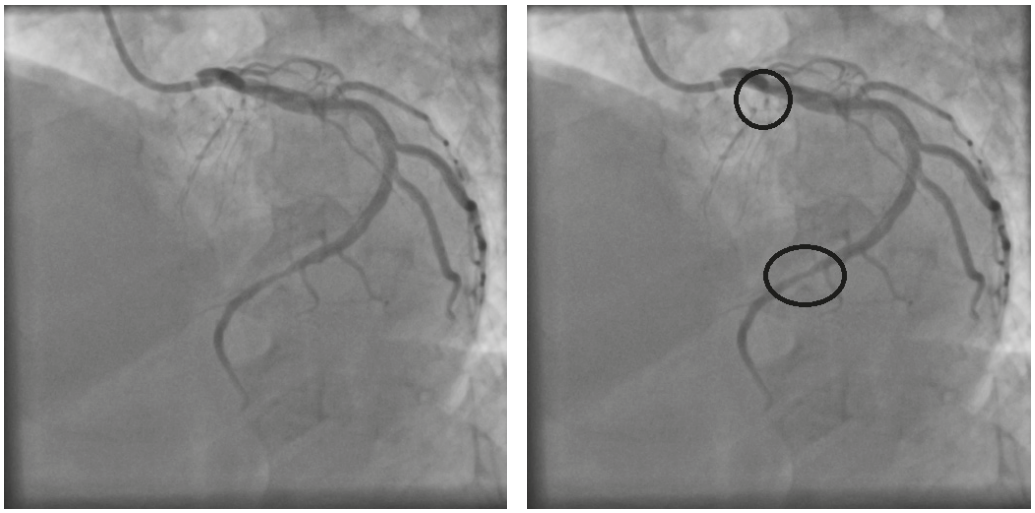


Figure 2. A coronary angiography (**left**) with corresponding stenosis areas labeled by the cardiology specialist (**right**).

In the literature, the coronary stenosis problem has been addressed using various approaches, including vessel enhancement and segmentation methods, as well as machine learning and deep learning classification techniques. For instance, the use of spatial filters based on Gaussian kernels and the Hessian matrix were useful to enhance the main artery tree in order to extract certain features to identify specific stenosis cases [3–6].

Deep learning techniques, such as Convolutional Neural Networks, have also been applied to classify coronary stenosis [7,8]. However, these methods require large datasets to achieve a high classification performance in terms of accuracy [9,10]. Due to ethical and legal restrictions, medical image banks of coronary arteries with proper stenosis cases labeled by specialists are scarce, posing challenges to achieving optimal performance using deep learning techniques.

Automatic classification of coronary stenosis is challenging, as identifying relevant image features that distinguish between positive and negative stenosis cases is difficult. Moreover, low contrast and high noise levels complicate the detection of artery structures. In this context, digital image processing techniques aimed at vessel enhancement are vital,

as different filtering and segmentation methods yield varied responses. For instance, spatial filters can enhance artery structures, enabling the extraction of morphological features. In addition, intensity and texture-related features can also be extracted [11]. This feature extraction is important for maximizing information retrieval from coronary angiographies [12]. However, when dealing with large feature sets, it becomes difficult to identify which features are most relevant for accurate classification [13].

Identifying an optimal subset of features that enhance classification performance is complex, as the computational cost of exploring all possible combinations is denoted by $O(2^n)$, where n represents the number of involved features. In addition, computational search techniques require an objective function to evaluate solution performance. By considering these principles, automatic feature selection for coronary stenosis classification is expressed as a multi-objective function where the best solution is achieved by minimizing the number of selected features and maximizing the classification performance rate.

The automatic feature selection problem has been widely studied in the literature using distinct computational techniques such as population-based methods, evolutionary computing, and others [14–16]. A common issue with automatic search techniques is premature convergence, which prevents proper exploration of the search space and may overlook better solutions [17].

In the present work, a novel method for automatic feature selection in high-dimensional spaces is proposed. The method involves a hybrid metaheuristic with explicit diversity control used to generate new populations of potential solutions. The achieved results with their corresponding analysis, demonstrate the importance of this diversity control strategy in improving the classification performance for coronary stenosis by selecting a better suitable feature subset than traditional approaches reported in the literature. The key contributions of this study include the following:

- Identification of 16 features that are relevant for the classification of positive and negative coronary stenosis cases.
- Development of a feature vector consisting of 473 distinct features extracted from the original images and responses from vessel enhancement filters.
- Implementation of a diversity control strategy to decrease the probability of premature convergence during the automatic feature selection process.

The remainder of this paper is structured as follows: Section 2 discusses the theoretical concepts and methods underlying the proposed method. Section 3 provides a detailed description of the proposed method. Section 4 outlines the image database, consisting of positive and negative stenosis cases used for experimentation, and presents the results, comparing them with existing approaches. Finally, Section 5 discusses the conclusions and relevant findings of this study.

2. Methods

Working with feature datasets for classification tasks, generally involves feature extraction and feature selection stages. The feature selection task is difficult to carry out manually when the number of features is high. In consequence, addressing the problem with the use of search metaheuristics is adequate. The search metaheuristics are used to find a suitable feature subset, able to achieve optimal classification rates.

2.1. Feature Extraction

The feature extraction process is commonly one of the first stages in order to form a working dataset [18]. In the digital image processing area, several feature types are possible to be extracted such as intensity, texture and morphology. The intensity-based features are the most simple to extract since they can be computed directly from the original image. Based on each individual pixel intensity, there is possible to compute the minimum, maximum, median, average, variance and standard deviation measures. Texture-based features are also possible to be extracted from the original image. Almost all texture features are required to compute the Gray Level Co-Occurrence Matrix (GLCM)

as a previous step [19]. Consequently, using the GLCM allows us to extract 14 distinct texture features using the Haralik Equations [11]. Additionally, morphological features provide valuable information about the shape, area and geometry of interest objects [14]. However, in order to obtain accurate measurements related to morphological features, the objects of interest must be enhanced from the original image. The use of filters in spatial and frequency domains plus other techniques from the literature, are commonly used to address the problem [20,21]. For instance, a linear multi-scale method is useful to enhance arterial segments that are able to be approximated to straight lines at distinct sizes and orientations [22]. In addition, the use of Gaussian-based filters are helpful to enhance curved artery structures using distinct curve amplitude scales [23,24]. With the arteries segmented from the rest of the image, it is possible to extract several morphological features such as the compactness, circularity ratio, elongatedness, and others [17].

Compactness is defined as the ratio of the square of the perimeter to the area of an artery segment. It is insensitive to uniform scale changes and can be computed as follows:

$$C = \frac{P^2}{A}, \tag{1}$$

where C represents the compactness, P is the perimeter of the artery segment and A is its corresponding area, measured in pixels.

The elongatedness of a region is the ratio between the length l, and the width w of the minimum rectangle of the corresponding region. In Figure 3, an original angiography patch is shown with its corresponding vessel segmentation and three distinct morphological features: perimeter, area, and elongatedness.

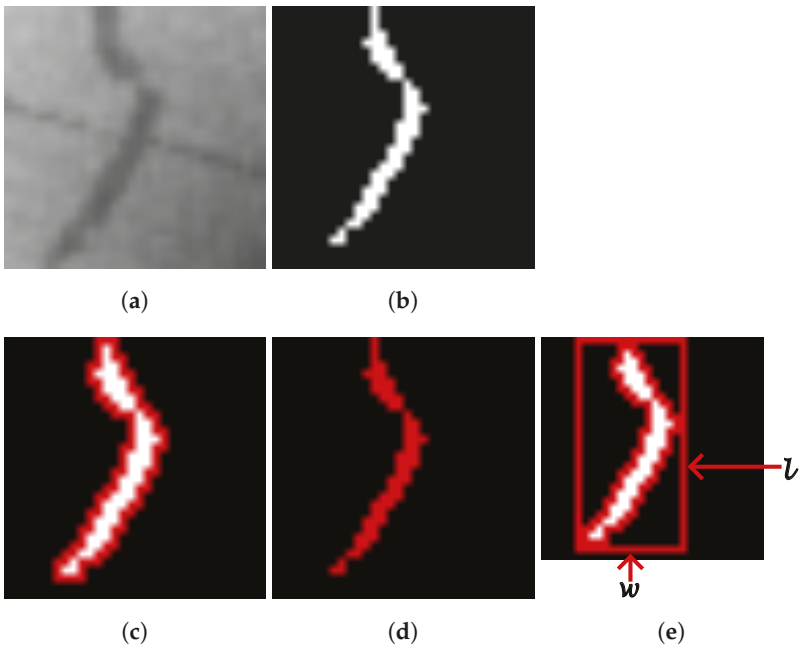


Figure 3. Original angiography patch with its corresponding vessel segmentation, highlighting three morphological features: perimeter, area and elongatedness. (**a**) Original image. (**b**) Artery segmentation. (**c**) Perimeter. (**d**) Area. (**e**) Elongatedness.

In addition, the circularity ratio is a morphological feature that measures the ratio of the artery area to the area of a circle having the same perimeter. It can be computed as follows:

$$R_c = \frac{4\pi A}{P^2}, \tag{2}$$

where R_c is the circularity ratio, A is the artery area and P is its corresponding perimeter.

On the other hand, the morphological Top-Hat operator was proposed by Eiho and Qian as a method for detection of artery trees [25]. The Top-Hat operator involves the operations of dilation and erosion. The dilation step is focused in the enhancement of the arteries by detecting local intensity variations by the use of a structure element. The structure element is a relevant component in the TopHat method because it defines the size and shape of the interest objects. For artery enhancement, the size of the structure element must be established based on the vessel width. In the erode step, the resultant noise is discriminated and the artery information is enlarged in order to obtain the enhanced image which content is then segmented into vessel-pixels and non-vessel pixels by applying a manual or automatic threshold.

2.2. Feature Selection

Extracting the highest number of features to form a dataset, that will be used later for classification, is desirable [26]. However, identifying the relevance of each individual feature or combined with others often turns into a complicated task to be done analytically by hand. The number of distinct combinations involving all the extracted features is $O(2^n)$, where n is the number of features. By considering all the possible combinations related with the number of involved features, the search space increments exponentially making difficult to work manually with them. Figure 4 illustrates the high-dimensional complexity increment considering a sample number of features in the range $[1, 20]$.

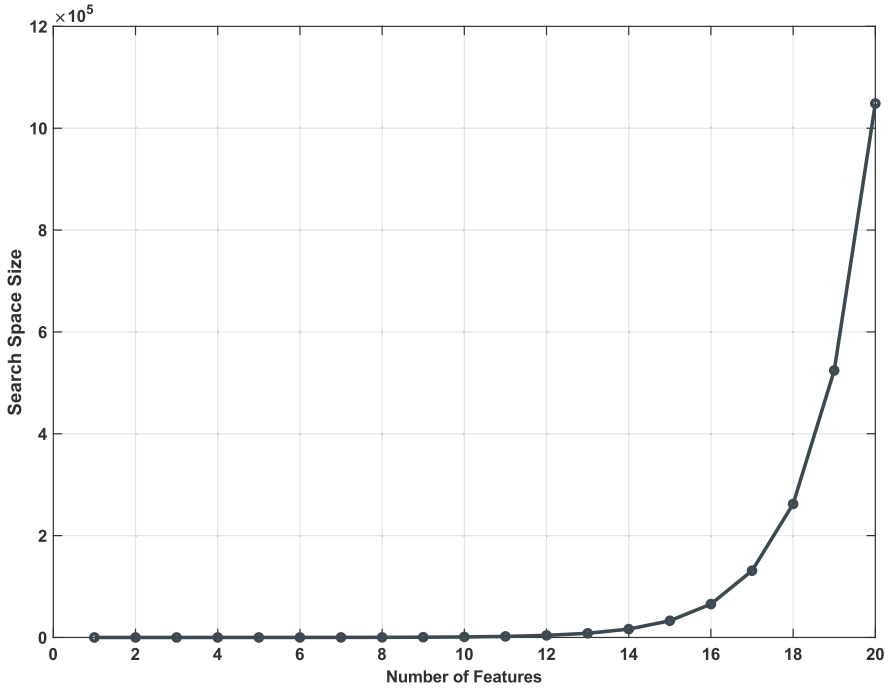


Figure 4. Search space size involving a feature set with size in the range $[0, 20]$.

Consequently, the feature selection problem has been addressed with different approaches [27,28]. Because of the complexity involved in the feature extraction process, automatic search techniques have been applied to solve the problem. The use of population-based estimation of distribution algorithms has been relevant when working with coronary arteries [13,17].

2.3. Metaheuristics

Metaheuristics is a set of algorithms for solving high-dimensional optimization problems in discrete or continuous domain, where evolutionary algorithms is a representative subset. The evolutionary algorithms try to imitate the nature of population species in which, more adapted individuals have the highest possibility to create a new popula-

tion, also known as the new generation, by using distinct operators. This behavior was adapted by John Holland in 1984 to create a computational method able to search for relevant solutions to complex problems that are difficult to be solved by hand or traditional analytical methods [29]. Subsequently, new approaches and techniques were proposed based on the Holland principles in order to generate new individuals that will conform the new generations.

2.3.1. Boltzmann-Univariate Marginal Distribution Algorithm

The Boltzmann-Univariate Marginal Distribution Algorithm (BUMDA) is a variation of the original Univariate Marginal Distribution Algorithm. In evolutionary computation, an individual represents a possible solution to a given problem. In the feature selection problem, each individual is represented as a feature vector of size n, where n is the number of involved features. By considering this principle, each element inside the vector has a discrete value within the range $[0, 1]$, where 0 means that a specific feature is not selected and 1 means that the feature is selected in the corresponding generated solution.

Traditional evolutionary methods, such as the Genetic Algorithm (GA) [30], produce new solutions by taking two of the best individuals (called "parents") from the current population at random. These parents are then split into two or more parts and, by mixing the parts of both parents, new individuals are created.

On the other hand, the population distribution estimation-based methods generate new solutions based on the distribution and frequency of each element (feature) by considering a rate of the population that are part of the current generation [31]. A variation of the original UMDA method was proposed by Valdez et al. in which, the individuals for the new generation are produced using the Boltzmann distribution as follows [32]:

$$\mu = \sum_j W(X_j)x_j, \text{ where } W(X_j) = \frac{g(X_j)}{\sum_{X_j} g(X_j)}, \tag{3}$$

$$\nu = \sum_j W'(X_j)(X_j - \mu)^2, \text{ where } W'(X_j) = \frac{g(X_j)}{\sum_{X_j} g(X_j) + 1}, \tag{4}$$

where μ is the fitness average, ν represents the fitness variance computed from all individuals in the population. $g(X_j)$ is the fitness value corresponding to the j^{th} individual in the population X. Consequently, taking a fraction of the current population, new individuals are produced as following:

$$\theta^{t+1} = \begin{cases} f(x_{\text{npop}}) & \text{if } t = 1, \\ f(x_{\frac{\text{npop}}{2}}) & \text{if } f(x_{\frac{\text{npop}}{2}}) >= \theta^t, \\ f(x_i) & \text{when } f(x_i) >= \theta^t \Big|_{i=\frac{\text{npop}}{2}+1}^{\text{npop}}, \end{cases} \tag{5}$$

where θ^{t+1} represents the new population, x_{npop} denotes all the individuals in the current population and, npop is the population size.

The concept of *fitness* or *objective function* measures the performance of each specific solution found by the metaheuristics. In automatic feature selection, the fitness measure of an individual can be computed evaluating its selected features by training a classifier and measuring the achieved accuracy.

The main advantage of BUMDA over other strategies is that it is governed by only three parameters: the number of generations, the population size, and the population selection rate. In addition, because of the Boltzmann distribution, the method increases the probability of generating diverse populations, thereby decreasing the risk of convergence to local-optima solutions.

2.3.2. Simulated Annealing

The Simulated Annealing (SA) method is a metaheuristic technique used for the search of solutions in computation. The method starts with an initial solution that is refined into an iterative process in which, worst solutions are discriminated more strictly as the iterations progress increases. The SA method was abstracted from metallurgy and adapted as a computational automatic search strategy [33]. It is governed by three parameters: the initial temperature value, the temperature-changing step, and the final temperature. Throughout the iterative process, the temperature is decreased from the initial value to the final value by a constant step, as follows:

$$P(\Delta E, T) = \frac{f(s') - f(s)}{T}, \tag{6}$$

where, $P(\Delta E, T)$ is the probability computed from the Boltzmann distribution. $f(s')$ represents a fitness value computed from an objective function in which the new solution is evaluated. $f(s)$ denotes the fitness value obtained by evaluating the current solution in the objective function. T is the current temperature.

The SA is a single-solution technique. The method starts with an initial solution generated randomly or by some other manner. Until the iterative SA loop, the solution is altered and evaluated. The solution alteration is made by using the probability which comes from the Boltzmann distribution. This method tends to achieve better results when combined with some other strategy [17].

2.4. Support Vector Machine

The Support Vector Machine (SVM) is a classification method that uses the concepts of plane and hyperplane in order to perform a binary classification. It is made by projecting the original dimensional space to higher orders in which the classification can be performed adequately in terms of the accuracy metric [34,35]. The main advantage of the SVM is the efficiency in classification tasks where the data belonging to each class, overlaps considerably and, there is not spatial difference between the elements of each class. In order to generate higher dimensional orders, the SVM needs to find the data-points located at the boundaries of both classes. These data points are used as the support vectors to project the plane or hyperplane into a higher-dimensional order space as follows [36]:

$$f(x) = W^T \phi(X) + b, \tag{7}$$

where W is the weight vector and normal to hyperplane, ϕ is the projection function or kernel, b is the bias or threshold and X is the data point to be classified.

3. Proposed Method Using a Diversity Strategy

The proposed method aims to improve the feature selection stage by implementing a population diversity strategy within the hybrid metaheuristic [37]. Initially, 473 distinct features were extracted from an image bank consisting of 608 patches of coronary angiographies. These feature sets include intensity, texture, and morphological feature types [38].

The automatic selection task becomes complicated in datasets with a large number of features. Even though search metaheuristics probe their effectiveness when applied to a 473-sized feature vector, fast convergence could limit the diversity of new solutions. This limitation decreases the probability of evaluating other suitable feature subsets that could yield higher coronary stenosis classification rates.

In order to decrease the probability of generating close-spatial solutions, which leads to an early convergence, the proposed method implements a diversity generation strategy.

The traditional search metaheuristics starts by generating a random number of solutions as the initial population. Then, each individual in the population is evaluated although a fitness function and ranked based on its performance measure. Subsequently,

the individuals of the population are sorted based on its performance from the best to the worst. Next, the best individuals are selected using a *selection rate* parameter to create new individuals (solutions) until a finite number of generations are reached.

However, in high-complexity search spaces, using only the best individuals to produce new populations, it is possible to achieve an early biased solution [39]. Alternatively, by considering the worst individuals under controlled conditions, the population diversity is increased [37].

Figure 5 illustrates the steps of the proposed method, which combines a hybrid metaheuristic with a population diversity strategy for automatic feature selection. The hybrid metaheuristic integrates the BUMDA and the SA techniques.

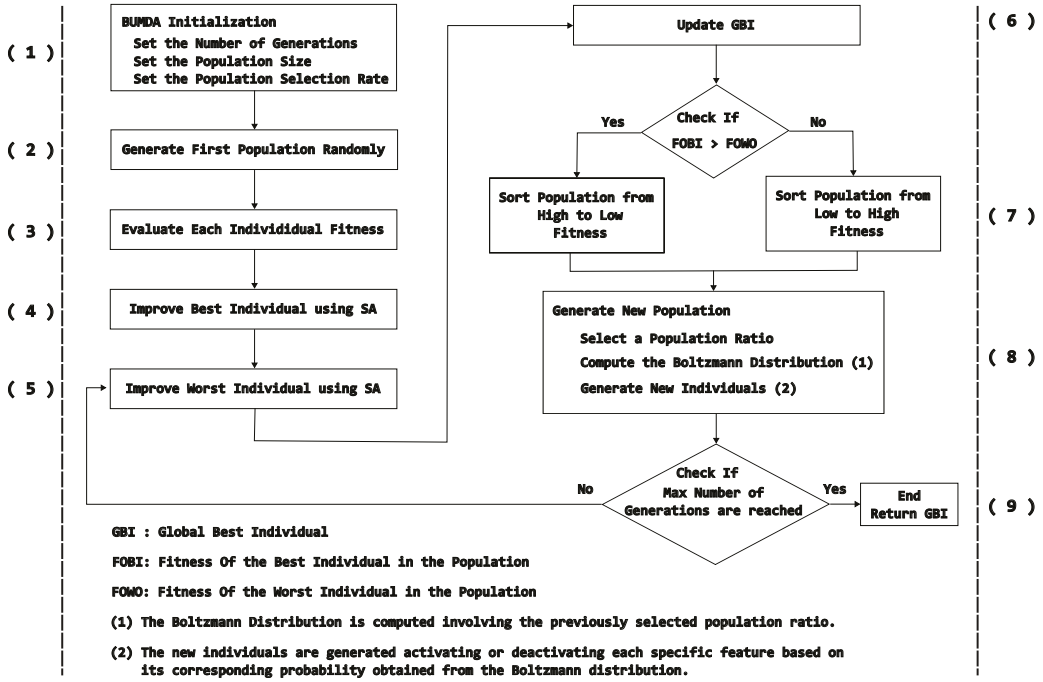


Figure 5. Overall process of the hybrid metaheuristic based on BUMDA and SA with diversity control, for automatic feature selection. Steps are as following. 1. BUMDA initialization. 2. First population generation. 3. Fitness computation for each individual in the current generation. 4. Improvement of the best individual in the current generation. 5. Improvement of the worst individual in the current generation. 6. Update of the global-best solution. 7. Sorting of the new population from the best to the worst individuals or viceversa. 8. Generation of the individuals for the new population based on the previous sorting. 9. Condition to end the process or continue until the maximum number of generations is reached.

According to Figure 5, the proposed method starts by establishing the maximum number of generations, the population size and selection rate. Then, in step (2), the individuals of the population corresponding to the first generation, are randomly initialized. As described previously, each individual is a vector of discrete elements which can have a value of 0 or 1. The vector size corresponds to the number of involved features where, a value of 1 in some vector element means that the corresponding feature is selected and vice versa.

In step (3), each individual is evaluated. The fitness value is obtained by training a supervised classify-based model using only the selected features. In this case, a linear SVM was trained and tested using a holdout subset of instances, separated previously from the dataset. Subsequently, the classification performance is evaluated with the Accuracy metric, which is computed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{8}$$

where TP, TN, FP and FN, corresponds to the True-Positive, True-Negative, False-Positive and False-Negative fractions, respectively.

In addition, the number of involved features are also considered applying a feature decreasing rate metric which is computed as follows:

$$FDR = 1 - \frac{Number\ of\ Selected\ Features}{N}, \tag{9}$$

where, FDR is the feature decreasing rate value and, N is the total number of involved features.

By considering Equation (9), when the number of selected features decrease, the feature decreasing rate is incremented. By involving the FDR along with the classification accuracy, the final fitness value for each specific individual can be obtained by maximizing the objective function.

In steps (4) and (5), is selected the worst individual along with the best one from the current BUMDA generation. Then, the two individuals are refined by applying the SA method. Subsequently, in the step (6), the global best individual (G_{Best}) is updated only if some of the improved individuals have a higher fitness value than the current G_{Best}.

Following the process, in step (7), the strategy to keep diversity of individuals in the population is relevant. If the best SA-refined individual was the worst in the BUMDA current population, the population is sorted from the worst to the best. In other case, the population is sorted in the traditional way: best individuals will occupy the first places. By applying this strategy, in step (8), some new populations will be generated by considering the worst individuals from the current one. Otherwise, the best individuals will be considered to produce new generations. The method ends when the maximum number of generations is reached.

Finally, in order to asses the achieved classification performance, the Jaccard coefficient, F1-score, Sensitivity and Specificity metrics were additionally implemented, which can be computed using the TP, TN, FP and FN values. The Jaccard coefficient (J_c) is computed as follows:

$$J_c = \frac{TP + TN}{(A + P) - (TP + TN)}, \tag{10}$$

where, A is the number of actual values and P is the number of predicted values.

The use of the J_c metric is relevant to asses the classification performance accuracy because the J_c is focused in the measurement of correctly classified instances and, incorrect classification cases have a higher penalization rate than in the Accuracy metric.

The F1-score (F_1) metric is computed as follows:

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{11}$$

The Sensitivity metric is computed as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{12}$$

The Specificity metric is computed as follows:

$$Specificity = \frac{TN}{TN + FP} \tag{13}$$

4. Results

All the experiments were carried out over a computer with an Intel Xeon E5 Processor, 64 GB of physical memory and, the Matlab R2024a platform.

4.1. Image Database

For the experiments, a database consisting of 608 images of size 64×64 pixels in gray-scale pgm (Portable Gray Map) format, was used. In the database 304 images corresponds

to positive stenosis cases and the remaining images are negative samples extracted from X-ray coronary angiographies. The database was previously approved by an ethics and research committee from the Mexican Institute of Social Security (UMAE-T1), under reference R-2019-1001-078, guaranteeing the privacy of the participants, prior their informed consent. The complete image database and the formed dataset will be available via web page at http://personal.cimat.mx:8181/~ivan.cruz/Journals/Stenosis608.html, accessed on 18 September 2024.

From each image, 473 features were computed, as described in [38]. There is important to mention that, for morphological features, 8 distinct vessel enhancement methods were applied: Frangi [40], Salem [41], Single and Multi-Scale Gabor filters [23,42], Multi-Scale Linear-Matched filter [22], Single and Multi-Scale Gaussian matched filters [24] and, the Top-Hat operator [25]. As result, was produced a dataset of size 608×473 rows and columns, respectively. In Table 1, are described the distinct vessel enhancement applied methods including their respective parameters value according to the state of the art.

Table 1. Description of the vessel enhancement applied methods involving their corresponding parameters value.

Method	Parameters	Value
Frangi, Salem	σ	[1, 12]
	σ_{step}	0.5
	α	0.5
	β	15
Single-Scale Gabor Filter	K	45
	T	5
	L	2.5
Multi-Scale Gabor Filter	I	3
	T	[2, 20]
	K	45
Multi-Scale Linear Matched-Filter	L	[1, 15]
	K	12
Single-Scale Gaussian Matched-Filter	L	13
	T	15
	σ	2.82
Multi-Scale Gaussian Matched-Filter	I	13
	T	15
	K	12
	σ	[1.5, 2.5]
	σ_{step}	0.5
Eiho Top-Hat Operator	shape	disk
	size	19

Figure 6 illustrates the vessel enhancement responses corresponding to 8 distinct filtering methods for 5 angiograms, which were taken from the used image bank.

Since the image responses becoming from the distinct filtering methods are diverse, the extracted morphological features will have different values. Subsequently, the challenge is to determine which features are relevant for the classification of positive and negative coronary stenosis cases.

The dataset was partitioned into 3 groups: training, validation and testing. The training and validation groups consisted of 458 (≈75%) and 50 (≈9%) instances, respectively. The remaining 100 (≈16%) instances were used for testing purposes after the automatic feature selection stage was finished in order to prove and corroborate the classification efficiency of the proposed method.

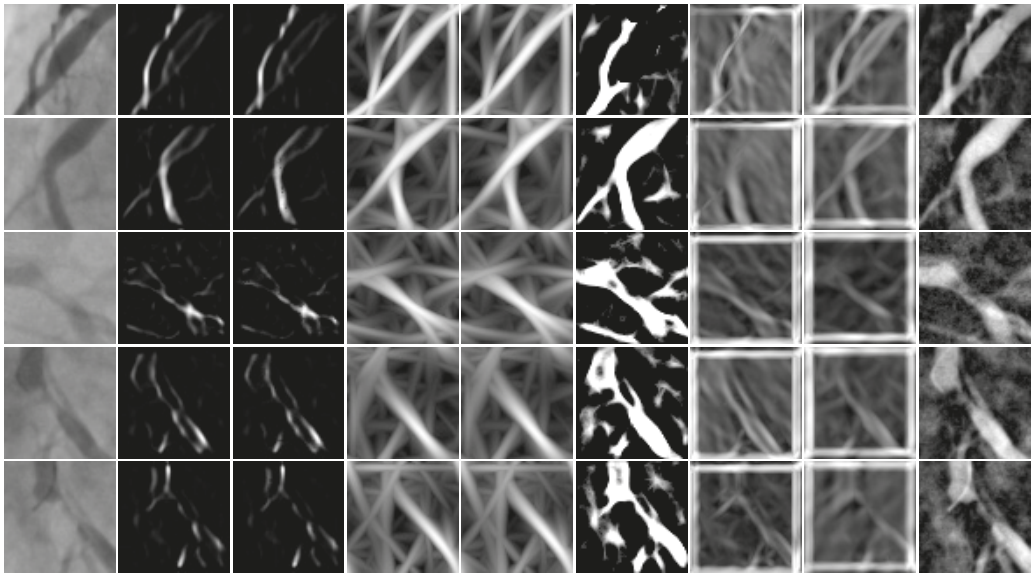


Figure 6. Vessel enhancement responses from 8 distinct filtering methods. Each row corresponds to a coronary patch. In columns 1 to 9, are described 8 distinct vessel enhancement methods response as follows: Original image, Frangi filter, Salem filter, Single-Scale Gabor filter, Multi-Scale Gabor filter, Multi-Scale filter, Single-Scale Gaussian-Matched Filter, Multi-Scale Gaussian-Matched Filter and, the Top-Hat operator.

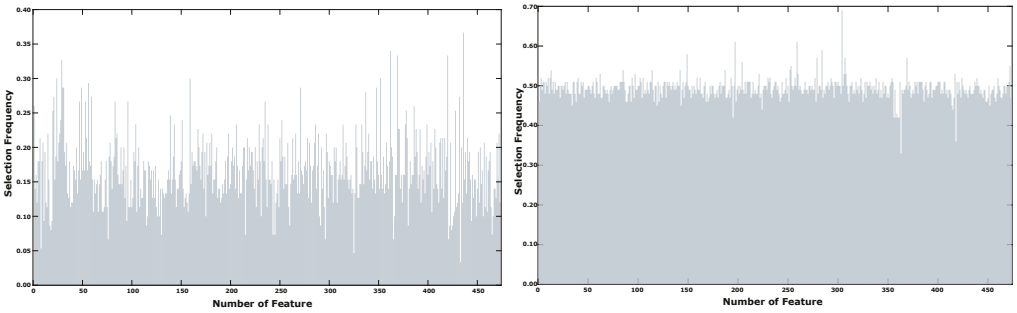
4.2. Experiment Results

Using the previously formed dataset, the automatic feature selection task was performed. In this stage, a hybrid metaheuristic, involving the BUMDA and the SA methods along with a diversity control, was executed to obtain a feature subset that improves the classification performance in terms of the Accuracy metric while simultaneously decreasing the number of involved features. For the BUMDA method, the configuration included 1000 generations, a population size of 100 individuals, and a selection rate of 0.80. The SA method was set with initial and final temperature values of 1 and 0, respectively, with a constant step of 0.001, resulting in $\approx$ 1000 iterations. Since BUMDA and SA methods involve multi-objective optimization, a weight of 0.90 was assigned to classification efficiency, and 0.10 to feature reduction. In order to validate the obtained results, the proposed method was executed on 30 independent trials. All values were established as a balance between achieving correct results and minimizing the time required. Additionally, it is important to mention that, until the automatic feature selection stage, a linear SVM was used as classifier by considering the required time to produce a trained model and the number of involved individuals and generations in the distinct search metaheuristics. Subsequently, when the automatic feature selection was completed, an SVM-based classifier with fine-tuned parameters was employed. These parameters included a polynomial order, kernel scale and kernel offset with values of 5, 2 and 0.3569, respectively, as recommended in [38].

In the best-achieved result, only 16 of the original 473 features were selected to classify positive and negative coronary stenosis cases with an Accuracy of 0.92 using the test dataset. Figure 7 illustrates a comparison of the feature selection frequencies for the BUMDA method and the BUMDA with diversity control applied. The frequencies were computed considering the iterations of the best trial for the traditional BUMDA and the BUMDA with explicit control diversity.

As illustrated in Figure 7, the feature selection frequencies in BUMDA with control diversity, trends to be uniform in comparison with the traditional BUMDA. This ensures a wide exploration of the search space and decreases the risk to fall into local-optima traps which is a common issue when working with metaheuristics along with high-dimensional search spaces. Table 2 presents a statistical analysis of the feature selection frequencies for the hybrid evolutionary method, comparing scenarios with and without the imple-

mentation of a population diversity strategy. Additionally, to understand the influence of considering the worst elements on population diversity, a statistical analysis of the feature selection frequency was computed over the trial with the best fitness performance.



(**a**) (**b**)

Figure 7. Comparison of feature selection frequency in BUMDA method (**a**) and, diversity control strategy (**b**).

Table 2. Statistical analysis of the feature selection frequencies, based on the results obtained by 4 distinct search metaheuristics from the literature and the proposed method.

Method	Min	Max	Median	Avg.	Variance	Std. Dev.
GA	0.0000	1.0000	0.4010	0.4103	0.0704	0.2653
BUMDA	0.0000	0.6000	0.0333	0.0858	0.0173	0.1316
SA	0.0000	0.4200	0.0800	0.0794	0.0021	0.0455
Hybrid Metaheuristic	0.0333	0.3667	0.1600	0.1664	0.0024	0.0490
Proposed Method	0.1895	0.4830	0.3125	0.3128	0.0006	0.0242

Based on the data described in Table 2, the population diversity strategy ensures uniform feature selection frequencies in comparison with traditional approaches. The minimum and maximum values for the feature selection frequencies increased using the population diversity strategy. Subsequently, the variance and standard deviation have the lowest variations providing evidence about the uniform exploration of the search space when compared with the hybrid metaheuristic.

In addition, the use of the SA method was also relevant to improve the best and worst individual produced by the BUMDA. The best solution found by the proposed method, was the worst individual of the BUMDA population in the 730th generation. Originally, the individual consisted of 221 features and a training classification performance of 0.70, in terms of the Accuracy metric. After the SA refinement process, the number of selected features decreased to 16 and the training classification performance increased to 0.97 in terms of the Accuracy metric.

An additional effect of the diversity control, is related with delaying the maximum number of generations in which the best result is achieved. In Figure 8, is illustrated the fitness performance along 1000 generations for the proposed method and contrasted with other 4 techniques taken from the literature which were adapted to the studied problem in [17].

As illustrated in Figure 8, by using a control diversity, the probability of risk for a premature convergence or stagnation is reduced considerably in comparison with traditional techniques in which only the best individuals are considered to produce new populations across the successive generations. When only the best individuals are considered to produce new solutions, the method is biased to achieve a fast convergence. In consequence, it is probably that a relevant number of potential solutions are not explored because of the discrimination of regions in the search space. However, in combinatory optimization problems, such as the automatic feature selection, there is not warranty that relevant solutions are not located close to the worst ones. By consequence, when the worst

individuals are considered to produce new potential solutions, more locations into the search are explored, delaying the solution convergence. In Table 3, is described a statistical analysis of the iterations in which the best result was achieved for each method considering 30 independent trials for each one of them.

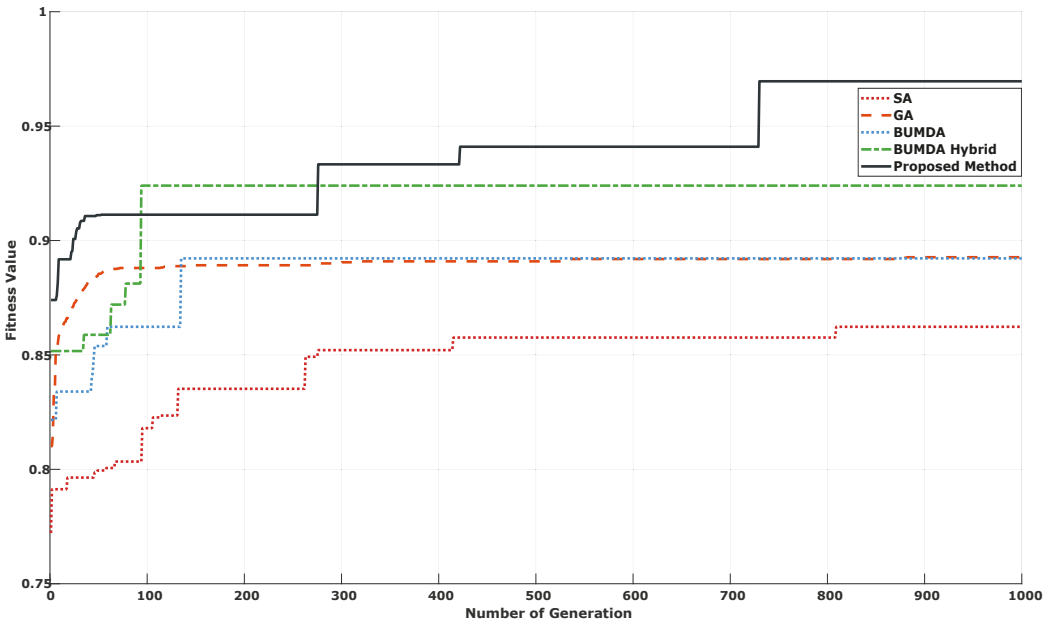


Figure 8. Iteration-fitness performance for the proposed method against 4 distinct methods from the literature.

Table 3. Statistical analysis related with the number of generations required for each method to achieve their best result in 30 independent trials. The proposed method was compared with 4 distinct techniques from the literature.

Method	Min	Max	Median	Avg.	Variance	Std. Dev.
SA	43	809	121	202	46,074	215
GA	79	882	126	229	52,245	229
BUMDA	97	321	188	183	4180	65
Hybrid Metaheuristic	79	203	108	117	1210	35
Proposed Method	151	998	473	566	83,245	288

Based on the results in Table 3, the proposed method, utilizing BUMDA and SA with a population diversity strategy, tends to achieve the best results in later generations compared to other techniques. By delaying early convergence, the probability to explore the search space more exhaustively is increased. At the same time, better results can be achieved by exploring solutions close to low-fitness locations since the feature selection problem can be expressed as a combinatory-optimization computational problem in which, each variable is considered as independent from the rest of them. Notably, the optimal solution using the proposed method was found at generation number 730. In consequence, the difference between this generation number and the median and average generation numberes of the proposed method was 257 and 164, respectively, which is lower than the standard deviation of 288. This characteristic highlights the robustness of the method.

Table 4 presents a comparison of the classification performance of the proposed strategy with other state-of-the-art methods. The results were obtained using the test subset following the automatic feature selection stage.

Table 4. Comparison of classification performance and selected features considering 4 distinct methods from the literature. NSF describes the number of selected features. FDR is the feature decreasing rate. JC is the Jaccard coefficient metric. F1, Sens. and Spec., refer to the F1-score, Sensitivity and Specificity metrics, respectively.

Method	NSF	FDR	Accuracy	JC	F1	Sens.	Spec.
ResNet50 [43]	–	–	0.81	0.68	0.80	0.78	0.84
Inception-v3 [44]	–	–	0.72	0.56	0.70	0.66	0.78
VGG16 [45]	–	–	0.84	0.72	0.82	0.74	0.94
CNN-16C [7]	–	–	0.86	0.74	0.84	0.76	0.94
–	473	0.00	0.78	0.64	0.77	0.72	0.84
SA	16	0.97	0.78	0.64	0.78	0.76	0.80
GA	29	0.94	0.80	0.67	0.80	0.80	0.80
BUMDA	205	0.57	0.80	0.67	0.77	0.66	0.94
Hybrid Metaheuristic	**4**	**0.99**	0.86	0.75	0.84	0.74	**0.98**
Proposed Method	16	0.97	**0.92**	**0.85**	**0.92**	**0.88**	0.96

Best performances are remarked in bold.

According to the results described in Table 4, the best classification performance, in terms of the Accuracy, Jaccard coefficient, F1-score and Sensitivity, was achieved by the proposed method. On the other hand, the Hybrid Metaheuristic achieved the highest Feature Decreasing Rate and Specificity results. By comparing the results, the Hybrid-Metaheuristic trends to discriminate more features penalizing the classification performance and, according with Figure 8, the best result was achieved early during the first 100 iterations. In contrast, the proposed method was able to achieve the highest classification rate in terms of almost all metrics, in later iterations. There is important to mention that, for the deep learning-based methods, a previous transfer learning step was performed using the imagenet dataset according with [8]. However, the achieved classification performance was lower than the proposed method. This behavior could be related with an insufficient amount of images to perform a correct classification by using convolutional neural networks.

On the other hand, the achieved results using an SVM classifier showed the impact of distinct feature subsets on classification performance. Since the search space involves high-dimensional complexity, the search metaheuristics tend to converge prematurely, and the exploration occurs near possible premature local-optima locations. For this reason, the use of a strategy focused on maintaining diversity within the population of potential solutions was relevant. As illustrated previously, the solutions space is explored widely compared with the other methods. It was observed that solely selecting the best individuals from a population to produce a new generation does not guarantee the discovery of superior solutions, as exploring the contributions of less favorable individuals can yield better outcomes. As a consequence, the classification performance achieved with the proposed strategy outperformed results obtained through previous strategies. In addition, the proposed strategy achieved the highest classification performance in terms of the Accuracy, Jaccard coefficient, F1-score and Sensitivity. However, the size of the feature vector was also increased compared with the hybrid evolutionary method which not used a population diversity strategy. Table 5 details the feature vector obtained using the proposed method.

According to data presented in Table 5, the preprocessing methods were relevant to extract distinct morphological features. From the 8 different methods applied for vessel enhancement, 6 of them allowed the extraction of features in which the coronary stenosis classification rate was the highest. Figure 9 illustrates the selection feature frequencies related with the proposed method, highlighting the final 16 selected features.

Table 5. Description of the 16-selected features, including their corresponding type and the vessel enhancement method (VEM) used for extracting them.

Identifier	Name	Type	VEM
F002	Maximum Intensity	Intensity	-
F046	Mean Vessel Length	Morphological	Frangi
F063	Minimum Compactness	Morphological	Frangi
F080	Median Elongatedness	Morphological	Frangi
F104	Gray Level Coefficient of Variation	Morphological	Salem
F110	Median Standard Deviation of Segments Length in all Arterial Sections	Morphological	Salem
F150	Number Of Vessel Segments	Morphological	SSGF
F153	Minimum Vessel Length	Morphological	SSGF
F159	Gray Level Coefficient of Variation	Morphological	SSGF
F161	Gradient Coefficient of Variation	Morphological	SSGF
F179	Maximum Circularity Ratio	Morphological	SSGF
F203	Standard Deviation of the Intensities	Intensity	MSGF
F233	Minimum Circularity Ratio	Morphological	MSGF
F329	Maximum Standard Deviation of the Standard Deviations of Segments Length in all Arterial Sections	Morphological	MSGMF
F387	Standard Deviation of the Standard Deviations of Segments Length in all Arterial Sections	Morphological	SSGMF
F420	Max Intensity	Morphological	Top-Hat Operator

SSGF: Single-Scale Gabor Filter; MSGF: Multi-Scale Gabor Filter; MSGMF: Multi-Scale Gaussian-Matched Filter; SSGMF: Single-Scale Gaussian-Matched Filter.

Based on data illustrated in Figure 9 and described in Table 5, there is relevant that all texture features were discriminated by the proposed method. In contrast the formed feature subset, involves at least one feature from 7 of the 8 vessel enhancement methods that were applied. None of the features extracted from the enhanced image, using the Multi-Scale Linear-Matched Filter, were selected. In addition, two intensity-based features were relevant: the maximum intensity extracted from the original image and the standard deviation of the intensity, computed from the response of the Multi Scale Gabor Filter. The rest of the selected features are associated with the artery morphology such as length, number of vessel segments, compactness, elongatedness and circularity ratio.

The required time for the automatic feature selection task was relevant, since the size of the search space and the complexity related with the fitness function, required a considerable computational time. The minimum, maximum, median and average time in seconds, to train the SVM with the distinct feature subsets was 0.40, 4.80, 0.47 and 0.47, respectively. Once the SVM has been trained, the average time to perform the classification of a vector consisting of 16 features was 0.0003 s.

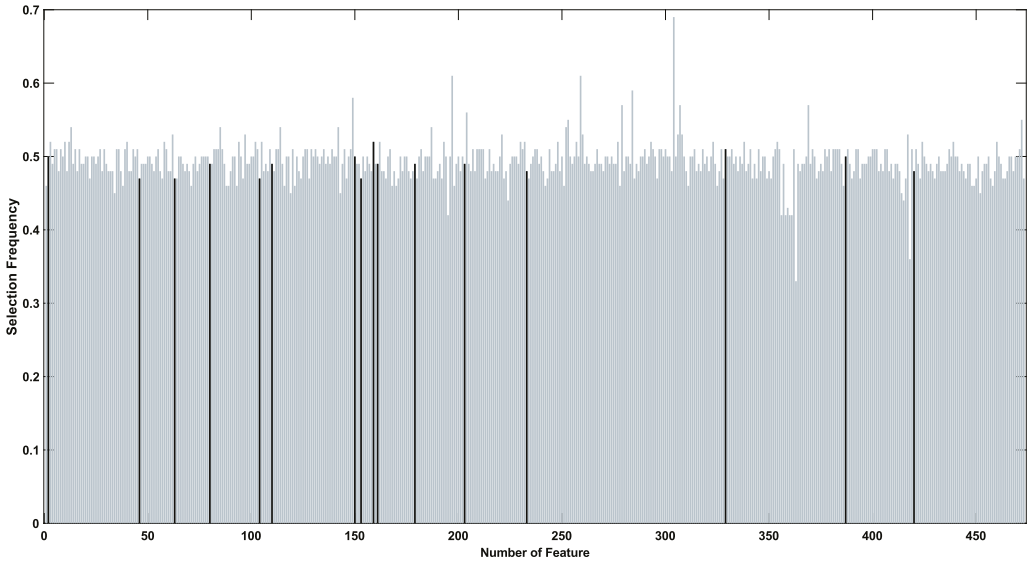


Figure 9. Feature selection frequencies involving the BUMDA and the SA with a diversity control, which conforms the proposed method. The previously selected 16 features are highlighted.

5. Conclusions

This paper presented an improved method aimed at increasing the classification accuracy of coronary stenosis using Accuracy and JC metrics. The use of a population diversity strategy proved crucial in exploring a broader area of the search space compared to traditional approaches. The main challenge in the feature selection process is its high-dimensional complexity, expressed as $O(2^n)$, with $n = 473$. The proposed strategy considers the worst-performing individuals in certain populations, which influences feature selection frequencies. As demonstrated in the results, this approach led to a higher uniformity rate in feature selection frequencies than methods that focus solely on the best individuals to produce new generations. The results validated the effectiveness of the proposed method, achieving a classification rate of 0.92 in terms of Accuracy, 0.85 for the Jaccard Coefficient, 0.92 for the F1-score, and 0.88 for Sensitivity, using only a 16 feature vector instead of the original vector of 473 features. The main strength of the proposed method lies in its ability to maintain a high classification rate based on a subset of explainable features. In contrast, deep learning methods act as black boxes and require a considerable number of instances to be trained correctly in comparison with machine learning techniques. In terms of scalability, it is important to mention that the proposed method is specialized in the classification of coronary stenosis cases. In consequence, all the morphological features are extracted from the response of image processing methods, focused on the enhancement of vessel/arterial structures. Based on this premise, the proposed method can be scaled to problems involving tubular or vessel-like structures. Finally, by considering the results described previously, involving also the required computation time to perform the classification of a single instance, the identified vector consisting of 16 features can be used in clinical practice as part of decision support information systems.

Author Contributions: Conceptualization, M.-A.G.-R., I.C.-A., M.-A.H.-G. and S.-E.S.-M.; Formal analysis, M.-A.H.-G.; Investigation, S.-E.S.-M.; Methodology, M.-A.G.-R. and I.C.-A.; Project administration, I.C.-A. and A.H.-A.; Software, M.-A.G.-R. and A.H.-A.; Validation, M.-A.G.-R., I.C.-A., M.-A.H.-G. and S.-E.S.-M.; Visualization, M.-A.G.-R.; Writing—original draft, M.-A.G.-R. and I.C.-A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by CONAHCYT under Project IxM-CONACyT No. 3150-3097.

Institutional Review Board Statement: The present paper involved the use of an image database which was previously approved by an ethics and research committee from the Mexican Institute of Social Security (UMAE-T1), under reference R-2019-1001-078.

Informed Consent Statement: The image database used in the present paper, was approved by an ethics and research committee from the Mexican Institute of Social Security (UMAE-T1) under reference R-2019-1001-078, guaranteeing the privacy of the participants, prior their informed consent.

Data Availability Statement: Data available in a publicly accessible repository http://personal.cimat.mx:8181/~ivan.cruz/Journals/Stenosis608.html accessed on 17 September 2024.

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

Abbreviations

The following abbreviations are used in this manuscript:

BUMDA	Boltzmann Univariate Marginal Distribution Algorithm
FDR	Feature Decreasing Rate
GA	Genetic Algorithm
JC	Jaccard Coefficient
MSGF	Multi-Scale Gabor Filter
MSGMF	Multi-Scale Gaussian-Matched Filter
NSF	Number of Selected Features
SA	Simulated Annealing
SSGF	Single-Scale Gabor Filter
SSGMF	Single-Scale Gaussian-Matched Filter
SVM	Support Vector Machine
UMDA	Univariate Marginal Distribution Algorithm

References

1. Qi, X.T.; Wag, H.; Zhu, D.G.; Zheng, L.; Cheng, X.; Zhang, R.J.; Dong, H.L. Global trends in coronary artery disease and artificial intelligence relevant studies: A bibliometric analysis. *Eur. Rev. Med. Pharmacol. Sci.* **2024**, *28*, 1. [PubMed]
2. Thakur, U.; Nogic, J.; Comella, A.; Nerlekar, N.; Chan, J.; Abrahams, T.; Michail, M.; Nelson, A.; Dey, D.; Ko, B.; et al. Computed tomography coronary angiography assessment of left main coronary artery stenosis severity. *J. Cardiovasc. Comput. Tomogr.* **2024**, *in press*. [CrossRef] [PubMed]
3. Cervantes-Sanchez, F.; Cruz-Aceves, I.; Hernandez-Aguirre, A. Automatic detection of coronary artery stenosis in X-ray angiograms using Gaussian filters and genetic algorithms. *AIP Conf. Proc.* **2016**, *1747*, 020005. [CrossRef]
4. Cervantes-Sanchez, F.; Cruz-Aceves, I.; Hernandez-Aguirre, A.; Hernandez-Gonzalez, M.A.; Solorio-Meza, S.E. Automatic Segmentation of Coronary Arteries in X-ray Angiograms using Multiscale Analysis and Artificial Neural Networks. *Appl. Sci.* **2019**, *9*, 5507. [CrossRef]
5. Wan, T.; Feng, H.; Tong, C.; Li, D.; Qin, Z. Automated identification and grading of coronary artery stenoses with X-ray angiography. *Comput. Methods Programs Biomed.* **2018**, *167*, 13–22. [CrossRef] [PubMed]
6. Sameh, S.; Azim, M.A.; AbdelRaouf, A. Narrowed Coronary Artery Detection and Classification using Angiographic Scans. In Proceedings of the 2017 12th International Conference on Computer Engineering and Systems (ICCES), IEEE, Cairo, Egypt, 19–20 December 2017; pp. 73–79. [CrossRef]
7. Antczak, K.; Liberadzki, Ł. Stenosis Detection with Deep Convolutional Neural Networks. In *Proceedings of the MATEC Web of Conferences*; EDP Sciences: Ulis, France, 2018; Volume 210, p. 04001. [CrossRef]
8. Ovalle-Magallanes, E.; Avina-Cervantes, J.G.; Cruz-Aceves, I.; Ruiz-Pinales, J. Transfer Learning for Stenosis Detection in X-ray Coronary Angiography. *Mathematics* **2020**, *8*, 1510. [CrossRef]
9. Chlap, P.; Min, H.; Vandenberg, N.; Dowling, J.; Holloway, L.; Haworth, A. A review of medical image data augmentation techniques for deep learning applications. *Med.-Imaging-Radiat. Oncol.* **2021**, *65*, 545–563. [CrossRef]
10. Garcea, F.; Serra, A.; Lamberti, F.; Morra, L. Data augmentation for medical imaging: A systematic literature review. *Comput. Biol. Med.* **2023**, *152*, 106391. [CrossRef]
11. Haralick, R.; Shanmugam, K.; Dinstein, I. Textural Features for Image Classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *SMC-3*, 610–621. [CrossRef]
12. Chowdhary, C.L.; Acharjya, D. Segmentation and Feature Extraction in Medical Imaging: A Systematic Review. *Procedia Comput. Sci.* **2020**, *167*, 26–36. [CrossRef]
13. Gil-Rios, M.A.; Guryev, I.V.; Cruz-Aceves, I.; Avina-Cervantes, J.G.; Hernandez-Gonzalez, M.A.; Solorio-Meza, S.E.; Lopez-Hernandez, J.M. Automatic Feature Selection for Stenosis Detection in X-ray Coronary Angiograms. *Mathematics* **2021**, *9*, 2471. [CrossRef]
14. Welikala, R.; Fraz, M.; Dehmeshki, J.; Hoppe, A.; Tah, V.; Mann, S.; Williamson, T.; Barman, S. Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy. *Comput. Med. Imaging Graph.* **2015**, *43*, 64–77. [CrossRef] [PubMed]

15. Ma, X.A.; Lu, K. Class-specific feature selection using neighborhood mutual information with relevance-redundancy weight. *Knowl.-Based Syst.* **2024**, *300*, 112212. [CrossRef]
16. Zhang, B.; Wang, Z.; Li, H.; Lei, Z.; Cheng, J.; Gao, S. Information gain-based multi-objective evolutionary algorithm for feature selection. *Inf. Sci.* **2024**, *677*, 120901. [CrossRef]
17. Gil-Rios, M.A.; Chalopin, C.; Cruz-Aceves, I.; Lopez-Hernandez, J.M.; Hernandez-Gonzalez, M.A.; Solorio-Meza, S.E. Automatic Classification of Coronary Stenosis Using Feature Selection and a Hybrid Evolutionary Algorithm. *Axioms* **2023**, *12*, 462. [CrossRef]
18. Suhaidi, M.; Kadir, R.A.; Tiun, S. A Review of Feature Extraction Methods on Machine Learning. *J. Inf. Technol. Manag. (JISTM)* **2021**, *6*, 51–59. [CrossRef]
19. Saliha, A.; Idir, A.h.; Izem, H. A new approach for texture segmentation based on the Gray Level Co-occurrence Matrix. *Multimed. Tools Appl.* **2021**, *80*, 24027–24052. [CrossRef]
20. Shaikh, M.S.; Choudhry, A.; Wadhwani, R. Analysis of digital image filters in frequency domain. *Int. J. Comput. Appl.* **2016**, *140*, 12–19.
21. Gil-Rios, M.A.; Cruz-Aceves, I.; Cervantes-Sanchez, F.; Guryev, I.; Lopez-Hernandez, J.M. Automatic enhancement of coronary arteries using convolutional gray-level templates and path-based metaheuristics. In *Recent Trends in Computational Intelligence Enabled Research*; Academic Press: Cambridge, MA, USA, 2021; pp. 129–153. [CrossRef]
22. Giacinti, D.J.; Cervantes-Sanchez, F.; Cruz-Aceves, I.; Hernandez-Gonzalez, M.A.; Lopez-Montero, L.M. Determination of the parabola of the retinal vasculature using a segmentation computational algorithm. *Nova Sci.* **2019**, *11*, 87–107. [CrossRef]
23. Cruz-Aceves, I.; Oloumi, F.; Rangayyan, R.M.; Aviña-Cervantes, J.G.; Hernandez-Aguirre, A. Automatic segmentation of coronary arteries using Gabor filters and thresholding based on multiobjective optimization. *Biomed. Signal Process. Control* **2016**, *25*, 76–85. [CrossRef]
24. Cruz-Aceves, I.; Cervantes-Sanchez, F.; Avila-Garcia, M.S. A Novel Multiscale Gaussian-Matched Filter Using Neural Networks for the Segmentation of X-Ray Coronary Angiograms. *J. Healthc. Eng.* **2018**, *2018*, 5812059. [CrossRef] [PubMed]
25. Eiho, S.; Qian, Y. Detection of coronary artery tree using morphological operator. In Proceedings of the Computers in Cardiology 1997, Lund, Sweden, 7–10 September 1997; pp. 525–528. [CrossRef]
26. Zebari, R.R.; Abdulazeez, A.M.; Zeebaree, D.Q.; Zebari, D.A.; Saeed, J.N. A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction 2020. *J. Appl. Sci. Technol. Trends* **2020**, *1*, 56–70. [CrossRef]
27. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [CrossRef]
28. Visalakshi, S.; Radha, V. A literature review of feature selection techniques and applications: Review of feature selection in data mining. In Proceedings of the 2014 IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, India, 18–20 December 2014; pp. 1–6. [CrossRef]
29. John Henry, H. Genetic Algorithms and Adaptation. In *Adaptive Control of Ill-Defined Systems. NATO Conference Series (II Systems Science)*; Springer: Boston, MA, USA, 1984; pp. 317–333. [CrossRef]
30. John Henry, H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*; MIT Press: Cambridge, MA, USA, 1992.
31. Hashemi, M.; Reza-Meybodi, M. Univariate Marginal Distribution Algorithm in Combination with Extremal Optimization (EO, GEO). In *International Conference on Neural Information Processing*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 220–227. [CrossRef]
32. Valdez-Peña, S.I.; Hernández, A.; Botello, S. A Boltzmann based estimation of distribution algorithm. *Inf. Sci.* **2013**, *236*, 126–137. [CrossRef]
33. Kirkpatrick, S.; Gelatt, C.D.; Vecchi, M.P. Optimization by simulated annealing. *Science* **1983**, *220*, 671–680. [CrossRef]
34. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
35. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [CrossRef]
36. Hari, B.N.; Salankar, S.S.; Bora, V.R. MRI brain cancer classification using Support Vector Machine. In Proceedings of the 2014 IEEE Students' Conference on Electrical, Electronics and Computer Science, Bhopal, India, 1–2 March 2014; pp. 1–6. [CrossRef]
37. Segura, C.; Miranda, G.; Segredo, E.; Chacón, J. A Novel Memetic Algorithm with Explicit Control of Diversity for the Menu Planning Problem. In Proceedings of the 2019 IEEE Congress on Evolutionary Computation (CEC), Wellington, New Zealand, 10–13 June 2019; pp. 2191–2198. [CrossRef]
38. Gil-Rios, M.A.; Cruz-Aceves, I.; Hernandez-Aguirre, A.; Moya-Albor, E.; Brieva, J.; Hernandez-Gonzalez, M.A.; Solorio-Meza, S.E. High-Dimensional Feature Selection for Automatic Classification of Coronary Stenosis Using an Evolutionary Algorithm. *Diagnostics* **2024**, *14*, 268. [CrossRef]
39. Dokeroglu, T.; Deniz, A.; Kiziloz, H.E. A comprehensive survey on recent metaheuristics for feature selection. *Neurocomputing* **2022**, *494*, 269–296. [CrossRef]
40. Frangi, A.; Nielsen, W.; Vincken, K.; Viergever, M. Multiscale vessel enhancement filtering. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI'98)*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 130–137. [CrossRef]
41. Salem, N.M.; Salem, S.A.; Nandi, A.K. Segmentation of retinal blood vessels based on analysis of the hessian matrix and Clustering Algorithm. In Proceedings of the 2007 15th European Signal Processing Conference, Poznan, Poland, 3–7 September 2007, pp. 428–432.

42. Gabor, D. Theory of communication. Part 1: The analysis of information. *J. Inst. Electr. Eng.-Part III Radio Commun. Eng.* **1946**, *93*, 429–441. [CrossRef]
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
44. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper With Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
45. Karen, S.; Andrew, Z. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Comput. Vis. Pattern Recognit.* **2015**, *6*, 1–14. [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

U-Net and Its Variants Based Automatic Tracking of Radial Artery in Ultrasonic Short-Axis Views: A Pilot Study

Yuan Tian [1,†], Ruiyang Gao [2,†], Xinran Shi [2], Jiaxin Lang [1], Yang Xue [1], Chunrong Wang [1], Yuelun Zhang [3], Le Shen [1], Chunhua Yu [1,*] and Zhuhuang Zhou [2,*]

[1] Department of Anesthesiology, Chinese Academy of Medical Sciences and Peking Union Medical College Hospital, Beijing 100730, China; tianyuan95@pumch.cn (Y.T.); langjx09@163.com (J.L.); xueyingi@sina.com (Y.X.); emancipation258@outlook.com (C.W.); shenle@pumch.cn (L.S.)
[2] Department of Biomedical Engineering, College of Chemistry and Life Science, Beijing University of Technology, Beijing 100124, China; gaoruiyang@emails.bjut.edu.cn (R.G.); shixinran@emails.bjut.edu.cn (X.S.)
[3] Medical Research Center, Chinese Academy of Medical Sciences and Peking Union Medical College Hospital, Beijing 100730, China; yuelunzhang@outlook.com
* Correspondence: yu.chunhua@aliyun.com (C.Y.); zhouzh@bjut.edu.cn (Z.Z.)
† These authors contributed equally to this work.

Abstract: Background/Objectives: Radial artery tracking (RAT) in the short-axis view is a pivotal step for ultrasound-guided radial artery catheterization (RAC), which is widely employed in various clinical settings. To eliminate disparities and lay the foundations for automated procedures, a pilot study was conducted to explore the feasibility of U-Net and its variants in automatic RAT. **Methods:** Approved by the institutional ethics committee, patients as potential RAC candidates were enrolled, and the radial arteries were continuously scanned by B-mode ultrasonography. All acquired videos were processed into standardized images, and randomly divided into training, validation, and test sets in an 8:1:1 ratio. Deep learning models, including U-Net and its variants, such as Attention U-Net, UNet++, Res-UNet, TransUNet, and UNeXt, were utilized for automatic RAT. The performance of the deep learning architectures was assessed using loss functions, dice similarity coefficient (DSC), and Jaccard similarity coefficient (JSC). Performance differences were analyzed using the Kruskal–Wallis test. **Results:** The independent datasets comprised 7233 images extracted from 178 videos of 135 patients (53.3% women; mean age: 41.6 years). Consistent convergence of loss functions between the training and validation sets was achieved for all models except Attention U-Net. Res-UNet emerged as the optimal architecture in terms of DSC and JSC (93.14% and 87.93%), indicating a significant improvement compared to U-Net (91.79% vs. 86.19%, $p < 0.05$) and Attention U-Net (91.20% vs. 85.02%, $p < 0.05$). **Conclusions:** This pilot study validates the feasibility of U-Net and its variants in automatic RAT, highlighting the predominant performance of Res-UNet among the evaluated architectures.

Keywords: radial artery tracking; perioperative management; deep learning; U-Net

1. Introduction

Radial artery catheterization is indispensable for a wide range of clinical scenarios, including anesthesiology, intensive care, emergency, and cardiovascular medicine [1,2]. It enables real-time hemodynamic monitoring, facilitates repeated blood tests, and establishes a proactive pathway for life-saving procedures. Ultrasonography has been recommended for radial artery catheterization [3,4], due to its higher success rates, reduced time requirements, fewer attempts needed, and lower complication rates compared to conventional techniques [5,6]. The radial artery tracking (RAT) on the short-axis view is the initial step for ultrasound-guided catheterization using the short-axis approach, the oblique-axis approach [7], the dynamic tip positioning technique [8], and the modified long-axis ap-

proach [9]. RAT is also pivotal in evaluating vascular conditions, identifying the appropriate puncture points, and distinguishing challenging cases [10,11].

In addition to radial artery catheterization, RAT may find applications in computational fluid dynamics (CFDs), as image processing is essential in developing engineering tools for hemodynamic prediction, particularly in analyzing vascular structures and blood flow dynamics to assess hemodynamic stability in clinical scenarios [12]. CFD is a well-established method for simulating cardiovascular hemodynamics, provides insights into parameters like wall shear stress, pressure distribution, and flow velocity, especially in aortic studies [13]. However, CFD requires patient-specific models and significant computational resources. By integrating image processing algorithms, such as those based on deep learning and finite element simulations, the efficiency and accuracy of hemodynamic predictions can be enhanced [14]. Automatic RAT can be used for laying a foundation for patient-specific radial artery segmentation for CFD analysis. It may facilitate precise, real-time hemodynamic assessments in ultrasound-guided catheterization procedures, ultimately enhancing clinical decision-making.

Despite significant advancements in visualization with ultrasound, several challenges remain to be addressed in RAT. The variations of intra-observer agreements arise from both congenital and acquired patient factors, including anatomical abnormalities, obesity, arterial wall thickening and calcification, luminal stenosis and branches, and subcutaneous emphysema [9]. In addition, the variability in inter-observer agreements, stemming from differences in overall or specific experience levels, has yet to be standardized [15]. Further, RAT is time-consuming and experience-dependent when performing manual annotations and measurements.

The advancement of deep learning has improved individualized treatment by providing end-to-end solutions through accelerated identification within large datasets, thereby reducing disparities in healthcare [16]. It has been previously observed that deep learning is widely applied to the segmentation of carotid and coronary arteries in ultrasonography [17,18]. Distinct from established deep learning-based vascular segmentation, RAT presents a significant challenge due to the wide variability in image characteristics. This variability is attributed to the lack of specialized settings, standardized scanning modalities, and significant inter-individual echogenic variation.

The variability is also recognized as one of the most substantial issues in image segmentation [19]. The prominent U-Net model and its variants, enhancing fully convolutional neural networks, have been adapted to address these challenges [20]. The U-Net skip connections combine the features from the deep decoder and the semantics from the shallow encoder subnetworks. Attention U-Net integrates the attention gate with the U-Net model to emphasize semantically meaningful features while suppressing irrelevant regions, thereby effectively tracking targets of diverse sizes and shapes [21]. UNet++ is modified to improve the capture of fine details through enriching the identification of features from the encoder network prior to the fusion with the semantically similar features from the decoder [22]. The residual module and the high-efficiency channel attention module are integrated into the Res-UNet to improve segmentation performance, especially in cases with blurred borders and irregular shapes [23]. To overcome the limitations of long-range dependencies, TransUNet integrates the transformer and U-Net architectures, enhancing the performance of tokenizing features and cross-attention between features and semantics [24]. UNetXt is a convolutional multilayer perceptron-based network that aims to fulfill the requirements of real-time image segmentation by tokenizing features, shifting channels, and fusing them with decoder semantics [25].

Despite the leading performance of U-Net and its variants in organ segmentation, there is a relative paucity of pioneering research evaluating their feasibility in RAT. The present study aims to train and validate the U-Net and its variant models adapted for RAT, based on novel constructed datasets.

2. Materials and Methods

2.1. Patients Enrollment and Datasets Formulation

Approved by the Ethics Review Committee of the Chinese Academy of Medical Sciences & Peking Union Medical College Hospital (I-23PJ662), patients were enrolled from May 2023 to Jan 2024 according to the inclusion and exclusion criteria depicted in Figure 1. The patient characteristics were withheld during the entire implementation process of the current study. From the location next to the styloid process of the radius to the confluence to the brachial artery, the radial arteries in both forearms of each patient were continuously scanned and acquired using B-mode ultrasound. The original videos were captured at a resolution of 1200 × 868 pixels for each frame. The frame rates of the videos varied, with approximately 120 videos recorded at 28 frames per second (fps), 30 videos at 25 fps, over 10 videos at 10 fps, and 7 videos at 49 fps. The majority of the videos contained 400 to 600 frames. To fit real-world scenarios and enrich data diversity, videos were acquired using three different ultrasound settings: the SonoSite X-Porter (FUJIFILM Sonosite, Inc., Bothell, WA, USA), the Mindray UMT-150 (Mindray Ltd., Shenzhen, China), and the Wisonic Navi S (Wisonic Medical Technology Co., Ltd., Shenzhen, China). Each ultrasound scanner was equipped with linear array probes with frequencies of 4 to 16 MHz.

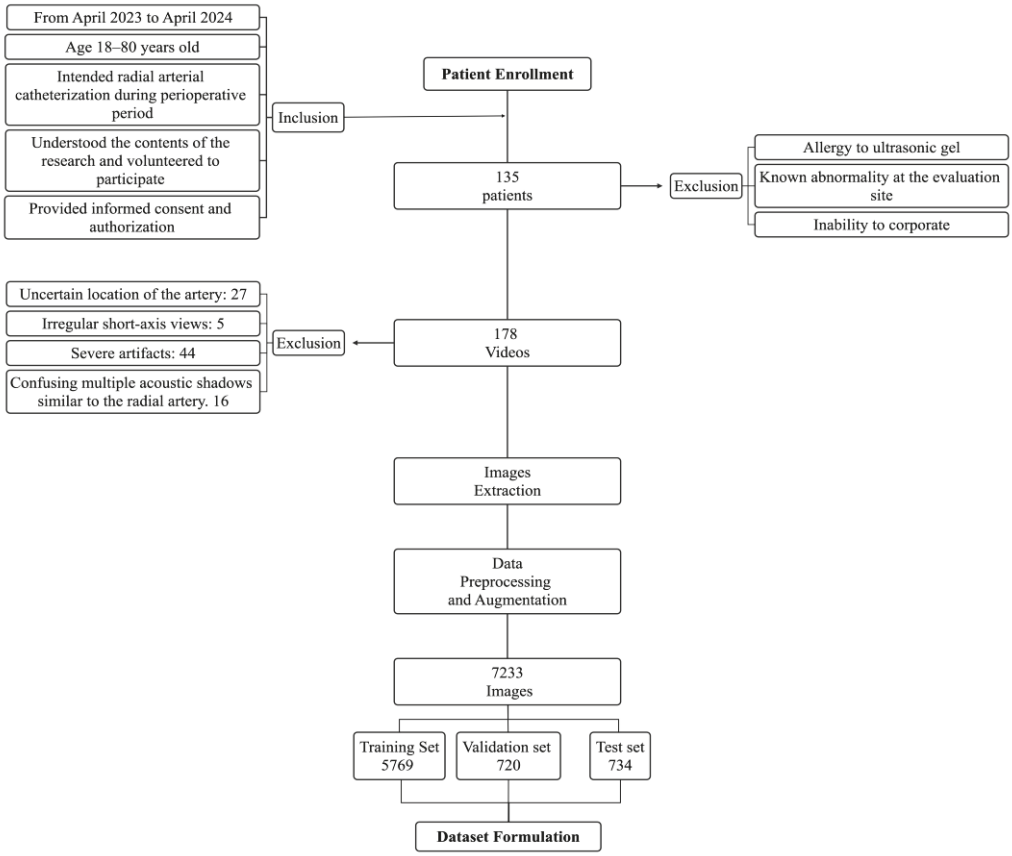


Figure 1. The process of patient enrollment and dataset formulation.

2.2. Data Preprocessing and Augmentation

All acquired videos were saved and transferred to a workstation equipped with Intel® Xeon® Gold 6132 CPUs (2.60 GHz, 2 processors) and an NVIDIA TITAN RTX 24 GB GPU, along with 128 GB of RAM. After the exclusion of videos with suspected ambiguous manual labels (detailed in Figure 1), images were extracted from videos that presented diverse sites and characteristics along the transition of radial arteries. To optimize computation, the extracted images were converted to a standardized geometry of 256 (height) × 256 (width) × 1 (channel).

The initial datasets, consisting of uniformly formatted images, were randomly divided into training, validation, and test sets at an 8:1:1 ratio. The training set was used for model training, the validation set was employed to avoid overfitting by tuning hyperparameters, and the test set was reserved for evaluating the model's performance.

To enhance model robustness and generalization, data augmentation was applied in the training set, including random rotation within a range of -10° to 10°, random horizontal and vertical flipping, sharpening adjustments with a random factor up to 2, and brightness and contrast variations within 50% to 150% of the original image. In addition, Gaussian blur was applied with a 50% probability and a blur kernel size of 3 to simulate various capture conditions. Gaussian noise, with a mean of 0 and standard deviation of 0.1, was utilized during each training epoch, along with salt-and-pepper noise with a 2% probability. The dynamic data augmentation of the training set was split into cycles according to the pre-designed size, and each cycle was executed in a randomized sequence. Adam optimizer was applied for the optimization, with an initial learning rate of 10^{-3} and a momentum index of 0.9.

2.3. RAT Using U-Net and Its Variants

U-Net and its variants were applied for RAT in ultrasonic short-axis views, including U-Net [20], Attention U-Net [21], UNet++ [22], Res-UNet [23], TransUNet [24], and UNeXt [25]. PyTorch (version 2.3.0) was employed as the framework for U-Net and its variants.

U-Net breaks through the limitations of single-label segmentation presented in previous deep learning models. Through the processes of encoding and decoding, U-Net connects spatial features to their corresponding semantics [20].

Attention U-Net combines the advantages of U-Net model and attention architecture. It is capable of focusing on features of interests, suppressing nonsense information, and dynamically adjusting based on context [21].

UNet++ is modified to improve the performance of detailed segmentation under multiscale circumstances. By incorporating confusion intermediate layers and dense skip connection, the decoder layer is improved in connecting with encoder features through shorter paths [22].

As illustrated in Figure 2, Res-UNet makes innovations in integrating squeeze-and-excitation networks, dynamic channel recalibration and application of atrous spatial pyramid pooling. Thereby, the strength of Res-UNet lies in its ability to highlight relevant features based on semantics [23].

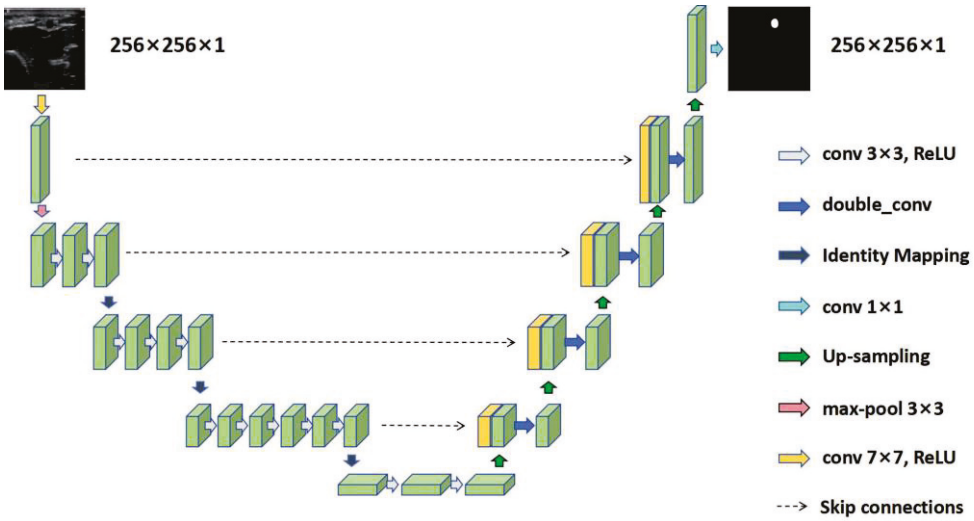


Figure 2. The Res-UNet network architecture for RA segmentation in US images. RA: radial artery; US: ultrasound; conv: convolution; ReLU: rectified linear unit.

The TransUNet integrates transformers into the U-Net framework by replacing traditional convolutional layers in the encoder, effectively capturing long-range dependencies, as illustrated in Figure 3. This adaptation enhances the modeling of spatial relationships in medical images, which is crucial for accurate segmentation. The decoder pathway retains convolutional layers and skip connections to preserve spatial information and optimize gradient flow. Self-attention mechanisms within transformers enable selective feature extraction, further improving segmentation accuracy [24].

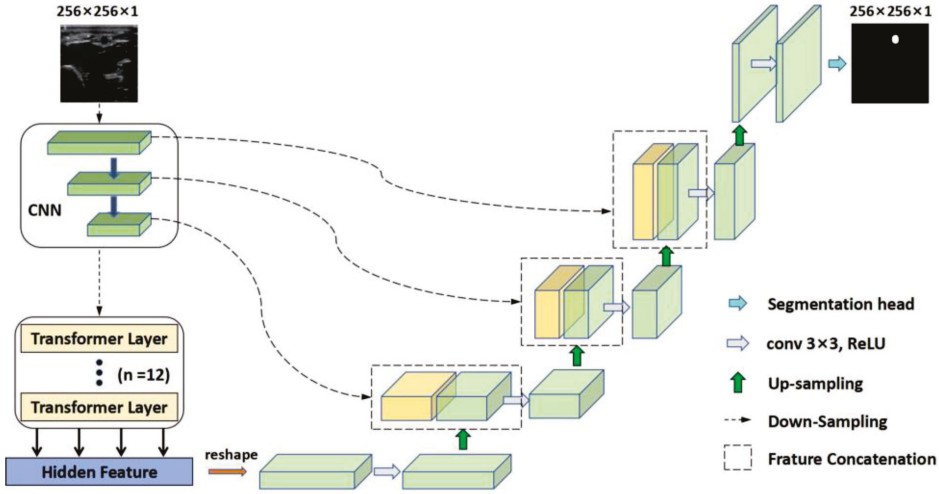


Figure 3. The TransUNet network architecture for RA segmentation in US images. RA: radial artery; US: ultrasound.

UNeXt integrates tokenized multi-layer perceptron modules to optimize number of layers and sampling strides. Compared to U-Net, it contributes to making image segmentation network more lightweight [25].

2.4. Experimental Setup and Statistical Analysis

To evaluate the performance of various deep learning models in segmenting RA regions in ultrasound images, the DSC and the JSC were utilized as performance metrics:

$$DSC(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{1}$$

$$JSC(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

where A denotes the RA region identified by the deep learning models, and B represents the manually annotated reference region. Both DSC and JSC values range from 0 to 1 (or 0% to 100%), with higher values indicating superior segmentation accuracy.

The loss function is utilized during the model training process to measure the discrepancy between the predicted values and the true values. It serves as a guide for the optimization algorithm, aiding the model in adjusting its parameters to minimize prediction errors. By minimizing the value of the loss function, the model progressively enhances its predictive performance, ultimately enabling it to make more accurate predictions. For U-Net and its variants, a combination of binary cross-entropy (BCE) loss (L_{BCE})and dice similarity coefficient (DSC) loss (L_{DSC}) was used for assessment. The loss function is defined as follows, with both β and γ set to 0.5.

$$Loss = \beta L_{\mathrm{BCE}} + \gamma L_{\mathrm{DSC}} \tag{3}$$

L_{BCE} and L_{DSC} are defined as follows, where A represents the model prediction and B denotes the manual label.

$$L_{BCE} = -B\log(A) - (1-B)\log(1-A) \tag{4}$$

$$L_{DSC} = 1 - \frac{2|A \cap B|}{|A| + |B|} \tag{5}$$

L_{BCE} is conceptually similar to the cross-entropy loss function but is specifically designed for binary classification without the requirements of the sigmoid or SoftMax functions for mapping input values to the [0, 1] interval. This approach provides enhanced numerical stability compared to standard cross-entropy loss. L_{DSC} enhances the effectiveness of image segmentation in scenarios characterized by a significant imbalance between positive and negative samples.

The statistical significance of differences was analyzed using the Kruskal–Wallis test, in terms of DSC and Jaccard similarity coefficient (JSC), for RAT in the test set (n = 734). A statistical difference with a p-value of less than 0.05 was considered significant. The analysis was performed using IBM SPSS version 27.0 (IBM Corp., Armonk, NY, USA).

3. Results

The RAT dataset consisted of 7233 images extracted from the 178 videos of 135 patients. The mean age of the patients was 41.6 years, with an equal distribution of females and males. In terms of ASA classification, 22.96% of the patients were classified as ASA I, 41.48% as ASA II, and 35.56% as ASA III or higher. The patients underwent surgeries in various specialties, including urology (5.93%), vascular surgery (11.11%), gynecology (11.85%), neurology (12.59%), otolaryngology (13.33%), orthopedics (15.56%), and general surgery (17.04%).

Examples of original images and the corresponding automatic RAT results for each model are depicted in Figure 4. The performance metrics for U-Net and its variants are depicted in Table 1. Among the architectures used for training and test, UNext required the fewest parameters, Res-UNet exhibited the fastest overall performance, and UNet++ had the lowest inference time per image.

Table 1. Performance Metrics in Training and Test Sets for RAT.

	Number of Parameters	Training Time	Inference Time for Each Image
U-Net	7,851,969	19,715.48 s	0.387 s
Attention U-Net	34,877,421	25,938.19 s	0.459 s
UNet++	9,162,753	24,757.11 s	**0.376 s**
Res-UNet	24,449,857	**13,242.36 s**	0.410 s
TransUNet	104,097,149	24,313.91 s	1.868 s
UNeXt	**1,471,633**	17,396.22 s	0.565 s

The optimal value for each metric is highlighted in **bold**.

The loss for the training and validation sets of U-Net and its variants are illustrated in Figure 5. All of the U-Net and its variants achieved convergence during the training and validation processes. Among them, UNext achieves through the fastest pathway and with the lowest loss values.

As illustrated in Figure 6, DSC rapidly improved and stabilized by the 50th epoch during training and validation processes for U-Net, U-Net++, Res-UNet, TransUNet, and UNeXt. However, the Attention U-Net architectures exhibited obvious fluctuations.

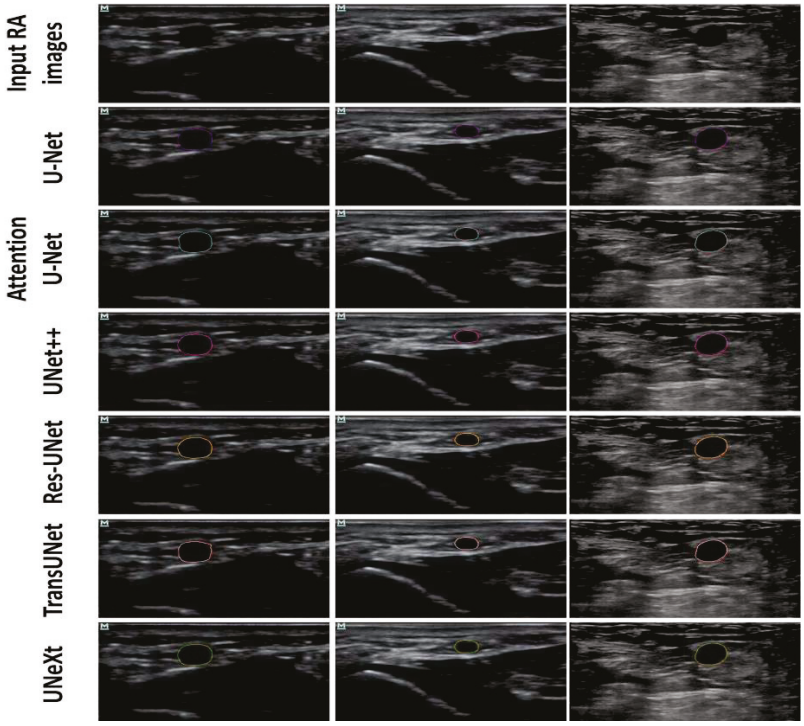


Figure 4. Representative input images (Row 1) and corresponding RAT results performed by U-Net (Row 2), Attention U-Net (Row 3), U-Net++ (Row 4), Res-UNet (Row 5), TransUNet (Row 6), and UNeXt (Row 7). Red contours were labeled by the expert. Contours in colors other than red were the segmentation of the corresponding deep learning architecture.

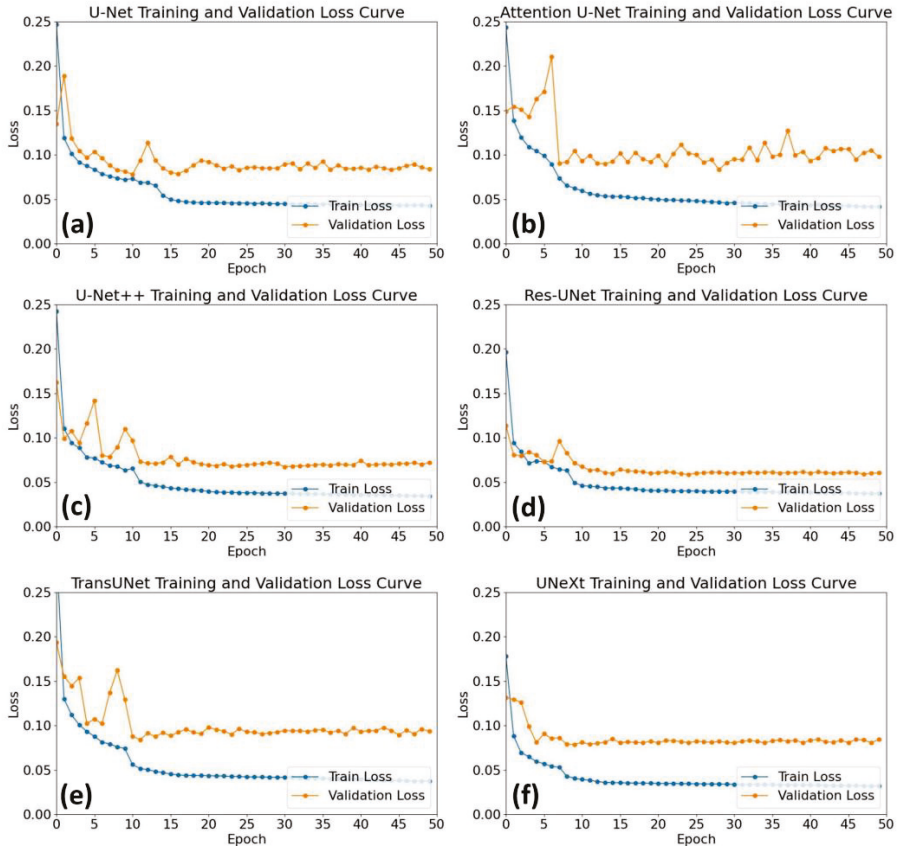


Figure 5. The loss functions of training and validating processes for U-Net (**a**), Attention U-Net (**b**), U-Net++ (**c**), Res-UNet (**d**), TransUNet (**e**), and UNeXt (**f**).

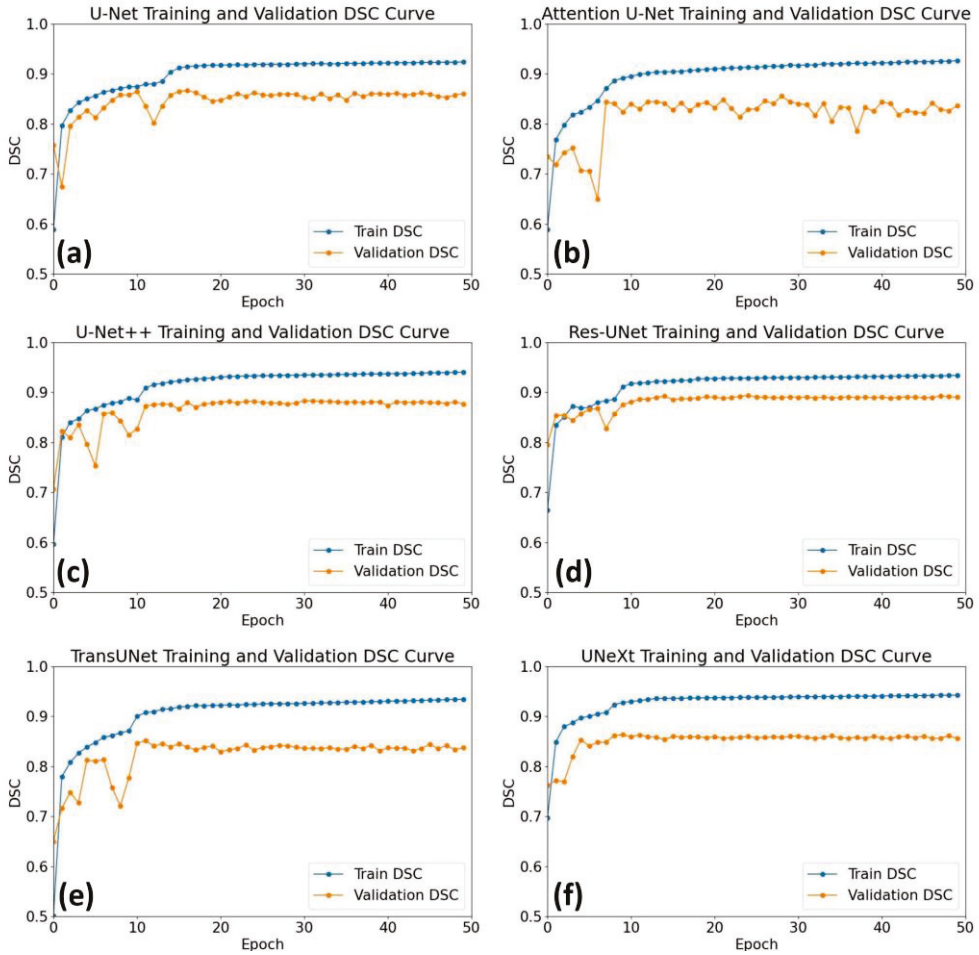


Figure 6. The DSC of training and validating processes for U-Net (**a**), Attention U-Net (**b**), U-Net++ (**c**), Res-UNet (**d**), TransUNet (**e**), and UNeXt (**f**).

Table 2 presents the performance of U-Net and its variants on the test set, evaluated using DSC and JSC. The results are ordered from greatest to least in terms of DSC with the highest values highlighted in bold. From the table, Res-UNet achieves the highest DSC and JSC values, indicating its superior segmentation performance compared to other models, including U-Net, Attention U-Net, and U-Net++. Notably, the DSC values of U-Net++ and U-Net are very close, indicating comparable performance between these two models. Interestingly, the performance of Attention U-Net, UNeXt, and TransUNet is slightly lower than that of U-Net, highlighting that these variants do not necessarily outperform the standard U-Net in this specific segmentation task.

Table 2. The DSC and JSC of U-Net and its variants in test set. The highest DSC and JSC are indicated in bold.

	DSC (%)	JSC (%)
Res-UNet	**93.14 ± 3.58**	**87.93 ± 5.57**
U-Net++	91.98 ± 6.31	86.08 ± 8.88
U-Net	91.79 ± 8.51	86.19 ± 10.02
UNeXt	91.33 ± 5.16	84.94 ± 7.93
Attention U-Net	91.20 ± 7.49	85.02 ± 10.17
TransUNet	91.08 ± 7.02	84.88 ± 9.48

Table 3 presents the *p*-values obtained from the Kruskal–Wallis test comparing DSC across different U-Net architectures. A *p*-value less than 0.05 indicates a statistically signifi-

cant difference between the corresponding network architectures. Significant differences are indicated with *. For U-Net versus Res-UNet, TransUNet, and UNeXt, the *p*-values (5.76×10^{-3}, 4.77×10^{-8}, and 2.24×10^{-5}, respectively) show significant differences in DSC, indicating that the DSC values of Res-UNet, TransUNet, and UNeXt are statistically different from those of the U-Net. For Res-UNet versus TransUNet and UNeXt, the extremely low *p*-values (8.31×10^{-18} and 6.53×10^{-13}) demonstrate significant differences between Res-UNet and these models.

Table 3. Kruskal–Wallis test for evaluating DSC between network architectures. The * indicates statistically significant difference.

	U-Net	**Attention U-Net**	**U-Net++**	**Res-UNet**	**TransUNet**	**UNeXt**
U-Net	1	1.57×10^{-1}	5.36×10^{-1}	5.76×10^{-3} *	4.77×10^{-8} *	2.24×10^{-5} *
Attention U-Net	1.57×10^{-1}	1	5.33×10^{-2}	3.76×10^{-5} *	5.20×10^{-4} *	1.23×10^{-2} *
U-Net++	5.36×10^{-1}	5.33×10^{-2}	1	4.32×10^{-2} *	4.91×10^{-10} *	2.08×10^{-6} *
Res-UNet	5.76×10^{-3} *	3.76×10^{-5} *	4.32×10^{-2} *	1	8.31×10^{-18} *	6.53×10^{-13} *
TransUNet	4.77×10^{-8} *	5.20×10^{-4} *	4.91×10^{-10} *	8.31×10^{-18} *	1	2.90×10^{-1}
UNeXt	2.24×10^{-5} *	1.23×10^{-2} *	2.08×10^{-6} *	6.53×10^{-13} *	2.90×10^{-1}	1

Table 4 presents the *p*-values obtained from the Kruskal–Wallis test comparing JSC across different U-Net architectures. A *p*-value less than 0.05 indicates a statistically significant difference between the corresponding network architectures. Significant differences are indicated with an *. For U-Net versus TransUNet and UNeXt, the *p*-values (4.49×10^{-7} and 1.64×10^{-6}) reveal significant differences, highlighting that TransUNet and UNeXt differ significantly from U-Net. For Res-UNet versus TransUNet and UNeXt, the *p*-values (2.03×10^{-16} and 6.61×10^{-15}) indicate highly significant differences, emphasizing that Res-UNet has significantly higher JSC than TransUNet and UNeXt.

Table 4. Kruskal–Wallis test for evaluating JSC between network architectures. The * indicates statistically significant difference.

	U-Net	**Attention U-Net**	**U-Net++**	**Res-UNet**	**TransUNet**	**UNeXt**
U-Net	1	6.39×10^{-2}	8.35×10^{-1}	6.74×10^{-3}	4.49×10^{-7} *	1.64×10^{-6} *
Attention U-Net	6.39×10^{-2}	1	1.07×10^{-1}	7.34×10^{-6} *	4.33×10^{-3} *	6.18×10^{-3} *
U-Net++	8.35×10^{-1}	1.07×10^{-1}	1	3.63×10^{-3} *	1.31×10^{-7} *	4.49×10^{-6} *
Res-UNet	6.74×10^{-3} *	7.34×10^{-6} *	3.63×10^{-3} *	1	2.03×10^{-16} *	6.61×10^{-15} *
TransUNet	4.49×10^{-7} *	4.33×10^{-3}	1.31×10^{-7} *	2.03×10^{-16} *	1	7.41×10^{-1}
UNeXt	1.64×10^{-6} *	6.18×10^{-3}	4.49×10^{-6} *	6.61×10^{-15} *	7.41×10^{-1}	1

4. Discussion

The development of deep learning improves the elimination of healthcare disparities and unveils novel horizons for precision medicine. Given the outstanding performance in image segmentation of U-Net and its variants, coupled with the requirement to address the heterogeneity of RAT, the current study was conducted. As far as we know, this novel evidence fills the gap in automatic RAT using deep learning architectures.

The results validate the hypothesis that RAT in short-axis views of ultrasonography can be automatically executed by U-Net and its variants. The convergence of L_{BCE} and L_{DSC} in the training and validation sets indicates the absence of overfitting and underfitting, except TranUNet. Consistent with the previous studies, the compelling results for model weighting were confirmed through the application of L_{BCE} and L_{DSC}, which, respectively optimized the loss functions for classification and quantification [19].

The feasibility of U-Net and its variants for automatic RAT is further supported by the findings of DSC and JSC in test set. In the present study, DSC ranges from 93.14% to 91.08%, comparable to carotid artery segmentation with DSC over 90% [26], and is superior

to retinal vessel segmentation with DSC around 85% [27]. The JSC of U-Net and its variants in RAT, ranging from 87.93% to 84.88%, though marginally lower than the performance for common carotid artery segmentation (90.89–84.97%) [28], was significantly superior to breast mass segmentation (63.65%) [29].

Additionally, whereas U-Net and its variants perform competently in RAT, the results indicate that Res-UNet is particularly notable. These results corroborate that Res-UNet improves the accuracy of image segmentation in fetal ultrasonography compared to U-Net and Attention U-Net [30]. A possible explanation could be that the insertion of residual blocks between the encoder and decoder pathways enables a firmer connection between the extraction of interest features and the semantic context.

Although Res-UNet demonstrates the best segmentation performance in this study, the trade-off between its model complexity (e.g., number of parameters and inference time) and segmentation accuracy needs to be discussed. As shown in Table 1, Res-UNet has 24,449,857 network parameters, approximately three times that of U-Net, but its single-frame inference time (0.410 s) is only 0.023 s longer than that of U-Net. Compared to Attention U-Net and TransUNet, Res-UNet has fewer parameters and a shorter inference time. This suggests that although Res-UNet increases model complexity, it still offers acceptable efficiency. However, to achieve real-time RAT in clinical settings, future studies should focus on optimizing the model architecture to further reduce the parameter size. For instance, model compression techniques or lightweight convolutional modules could be explored to reduce computational complexity while maintaining or even improving segmentation performance. Additionally, one may investigate other architectural improvements, such as incorporating depth-wise separable convolutions or efficient attention mechanisms, to further reduce inference time.

Further, this study was conducted based on independent datasets comprising over 7000 images from 135 patients. These datasets represent some of the first large collections of radial artery ultrasound images, curated for the training, validation, and testing of deep learning architectures in image segmentation. The volume surpasses other vascular tracking datasets, which consist of a few hundred images from dozens of patients [31–33].

Hence, this pilot study has some limitations and leaves a few questions unanswered. Challenges of the automatic segmentation during RAT exist in certain types of cases, as illustrated in Figure 7. There were challenges in distinguishing between hypoechoic muscle encapsulated with fascia and radial artery presents challenges when applying architectures such as Attention U-Net, TransUNet, and UNet++, as shown in Figure 7a. When multiple vessels or vascular branches were present concomitantly (Figure 7b), the U-Net, U-Net++, and TransUNet models failed to identify the vessel location, Attention U-Net and UNeXt exhibited under-segmentation. ResUNet could detect the vessel, but there was another small over-segmentation region (red arrow in Figure 7b). Despite advancements of the U-Net series in identifying ambiguous boundaries, the enhanced capability to accurately delineate tiny vessels confounded by hyperechoic tissues required further exploration, as shown in Figure 7c. In future work, these challenging cases may be addressed as follows. For the challenge in Figure 7a, multi-modality ultrasound image fusion may be applied, i.e., incorporating multi-modality ultrasound data, such as ultrasound elastography [34] or color Doppler imaging, to complement B-mode ultrasound. As the hypoechoic tissue may have different elasticity or blood flow properties from the radial artery, the differentiation between hypoechoic tissues and the radial artery can be enhanced with elastography or color Doppler images. For the challenge in Figure 7b, the presence of multiple vessels or vascular branches, a rare case in the current model training, results in poor generalization when segmenting such instances. Increasing the number of such cases in the training set may improve the model's performance in future work. For the challenge in Figure 7c, super-resolution reconstruction techniques may be employed. For instance, generative adversarial networks or super-resolution networks may be utilized to improve the resolution of ultrasound images [35]. This would improve image resolution, making the boundaries of small radial arteries more distinct and thereby enhancing segmentation accuracy.

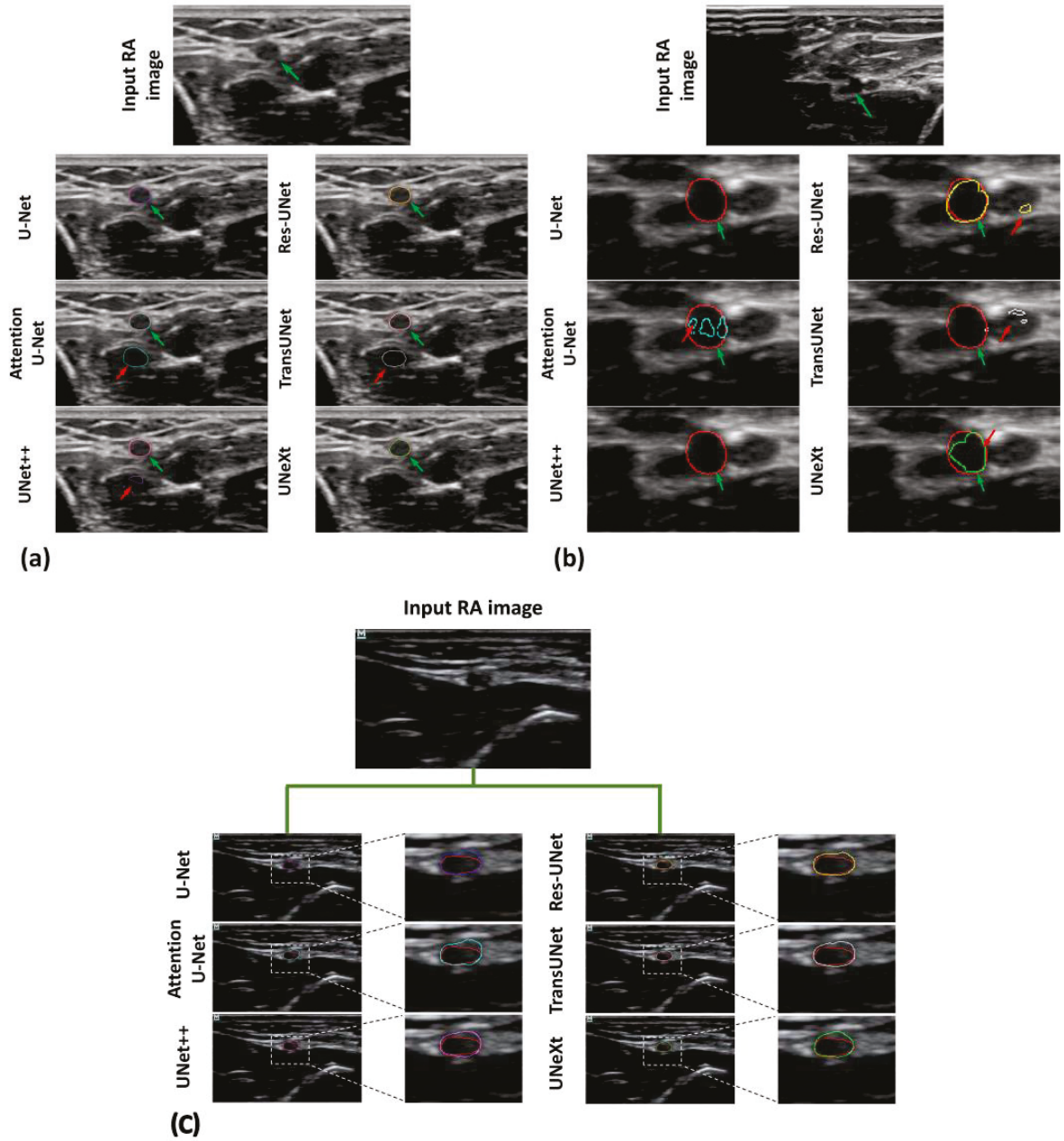


Figure 7. Challenging cases of RAT segmented by the U-Net series. (**a**) Challenges in distinguishing hypoechoic tissues from the radial artery. (**b**) Challenges in accurately identifying the radial artery when multiple vessels or vascular branches are present. (**c**) Challenges in the precise delineation of ambiguous boundaries of the tiny radial artery. Green arrow, accurate segmentation. Red arrow, errors in delineation. Red line, labeled by the human expert. Lines in colors other than red, segmentation of the corresponding architecture.

5. Conclusions

This study indicates that automatic radial artery catheterization can be achieved using the U-Net model and its variants. Among these, Res-UNet demonstrates a particularly prominent performance. The potential applications of these findings may help eliminate healthcare disparities, facilitate precision medicine and improve automated procedures in the widely required procedure of radial artery catheterization. The proposed deep learning based radial artery segmentation methods may be used for automatic RAT in ultrasound images. Future research should focus on the remaining challenges, including

distinguishing hypoechoic tissues, accurately identifying multiple vessels, and addressing blurred boundaries.

Author Contributions: Conceptualization, Y.T., Z.Z. and C.Y.; methodology, Y.T., R.G., Y.Z., Z.Z. and C.Y.; software, R.G.; validation, Y.T., Z.Z. and C.Y.; formal analysis, Y.T., R.G., Z.Z. and C.Y.; investigation, Y.T., J.L., C.W., Y.X. and C.Y.; resources, Y.T., C.W., J.L., L.S., X.S. and C.Y.; data curation, Y.T., R.G., X.S. and C.Y.; writing—original draft preparation, Y.T. and R.G.; writing—review and editing, Y.T., Z.Z. and C.Y.; visualization, R.G. and Z.Z.; supervision, Z.Z. and C.Y.; project administration, Y.T., Z.Z. and C.Y.; funding acquisition, Y.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National High Level Hospital Clinical Research Funding (funded by the Chinese Academy of Medical Sciences (CAMS)). Funding number, 2022-PUMCH-A-148).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of the Chinese Academy of Medical Sciences & Peking Union Medical College Hospital (I-23PJ662, 19 April 2023).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data can be accessed from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

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

References

1. Oliver, L.-A.; Oliver, J.-A.; Ohanyan, S.; Park, W.; Benelyahoo, A.; Vadivelu, N. Ultrasound for Peripheral and Arterial Access. *Best Pract. Res. Clin. Anaesthesiol.* **2019**, *33*, 523–537. [CrossRef] [PubMed]
2. Schmidt, G.A.; Blaivas, M.; Conrad, S.A.; Corradi, F.; Koenig, S.; Lamperti, M.; Saugel, B.; Schummer, W.; Slama, M. Ultrasound-Guided Vascular Access in Critical Illness. *Intens. Care Med.* **2019**, *45*, 434–446. [CrossRef] [PubMed]
3. Lamperti, M.; Biasucci, D.G.; Disma, N.; Pittiruti, M.; Breschan, C.; Vailati, D.; Subert, M.; Traškaitė, V.; Macas, A.; Estebe, J.-P.; et al. European Society of Anaesthesiology Guidelines on Peri-Operative Use of Ultrasound-Guided for Vascular Access (PERSEUS Vascular Access). *Eur. J. Anaesth.* **2020**, *37*, 344–376. [CrossRef] [PubMed]
4. Franco-Sadud, R.; Schnobrich, D.; Mathews, B.K.; Candotti, C.; Abdel-Ghani, S.; Perez, M.G.; Rodgers, S.C.; Mader, M.J.; Haro, E.K.; Dancel, R.; et al. Recommendations on the Use of Ultrasound Guidance for Central and Peripheral Vascular Access in Adults: A Position Statement of the Society of Hospital Medicine. *J. Hosp. Med.* **2019**, *14*, E1–E22. [CrossRef] [PubMed]
5. Moussa Pacha, H.; Alahdab, F.; Al-khadra, Y.; Idris, A.; Rabbat, F.; Darmoch, F.; Soud, M.; Zaitoun, A.; Kaki, A.; Rao, S.V.; et al. Ultrasound-Guided versus Palpation-Guided Radial Artery Catheterization in Adult Population: A Systematic Review and Meta-Analysis of Randomized Controlled Trials. *Am. Heart J.* **2019**, *204*, 1–8. [CrossRef]
6. Bhattacharjee, S.; Maitra, S.; Baidya, D.K. Comparison between Ultrasound Guided Technique and Digital Palpation Technique for Radial Artery Cannulation in Adult Patients: An Updated Meta-Analysis of Randomized Controlled Trials. *J. Clin. Anesth.* **2018**, *47*, 54–59. [CrossRef]
7. Zeng, C.-K.; Zhao, G.-F.; Deng, J.-H.; Li, D.-W. Oblique versus Longitudinal Axis/in-Plane Approaches for Ultrasound-Guided Radial Arterial Cannulation: A Randomised Controlled Trial. *Eur. J. Anaesth.* **2020**, *37*, 618–621. [CrossRef]
8. Liu, L.; Tan, Y.; Li, S.; Tian, J. "Modified Dynamic Needle Tip Positioning" Short-Axis, Out-of-Plane, Ultrasound-Guided Radial Artery Cannulation in Neonates: A Randomized Controlled Trial. *Anesth. Analg.* **2019**, *129*, 178–183. [CrossRef]
9. Wang, Z.; Guo, H.; Shi, S.; Xu, Y.; Ye, M.; Bai, L.; Tan, Y.; Li, Y.; Liu, L. Long-Axis in-Plane Combined with Short-Axis out-of-Plane Technique in Ultrasound-Guided Arterial Catheterization in Infants: A Randomized Controlled Trial. *J. Clin. Anesth.* **2023**, *85*, 111038. [CrossRef]
10. Imbriaco, G.; Monesi, A.; Giugni, A.; Cilloni, N. Radial Artery Cannulation in Intensive Care Unit Patients: Does Distance from Wrist Joint Increase Catheter Durability and Functionality? *J. Vasc. Access* **2021**, *22*, 561–567. [CrossRef]
11. Imbrìaco, G.; Monesi, A.; Spencer, T.R. Preventing Radial Arterial Catheter Failure in Critical Care—Factoring Updated Clinical Strategies and Techniques. *Anaesth. Crit. Care Pain Med.* **2022**, *41*, 101096. [CrossRef]
12. Wang, J.; Chen, B.; Gao, F. Exploring Hemodynamic Mechanisms and Re-intervention Strategies for Partial False Lumen Thrombosis in Stanford Type B Aortic Dissection after Thoracic Aortic Endovascular Repair. *Int. J. Cardiol.* **2024**, *417*, 132494. [CrossRef] [PubMed]

13. Zolfaghari, H.; Andiapen, M.; Baumbach, A.; Mathur, A.; Kerswell, R.R. Wall Shear Stress and Pressure Patterns in Aortic Stenosis Patients with and without Aortic Dilation Captured by High-performance Image-based Computational Fluid Dynamics. *PLoS Comput. Biol.* **2023**, *19*, e1011479. [CrossRef] [PubMed]
14. Otaru, A.J.; Abdulkadir, M.; Kovo, A.S.; Corfield, M.R.; Tanko, N.; Odey, O.A.; Aroke, U.O. Three-dimensional High-resolution Image Inversion and Pore Level CFD Characterisation of Effective Thermal Conductivity of Replicated Microcellular Structures. *Int. J. Thermofluids* **2022**, *14*, 100141. [CrossRef]
15. Saruwatari, M.S.; Nguyen, T.N.; Talari, H.F.; Matisoff, A.J.; Sharma, K.V.; Donoho, K.G.; Basu, S.; Dwivedi, P.; Bost, J.E.; Shekhar, R. Assessing the Effect of Augmented Reality on Procedural Outcomes during Ultrasound-Guided Vascular Access. *Ultrasound Med. Biol.* **2023**, *49*, 2346–2353. [CrossRef] [PubMed]
16. Topol, E.J. High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nat. Med.* **2019**, *25*, 44–56. [CrossRef]
17. Vila, M.D.M.; Remeseiro, B.; Grau, M.; Elosua, R.; Betriu, À.; Fernandez-Giraldez, E.; Igual, L. Semantic Segmentation with DenseNets for Carotid Artery Ultrasound Plaque Segmentation and CIMT Estimation. *Artif. Intell. Med.* **2020**, *103*, 101784. [CrossRef]
18. Szarski, M.; Chauhan, S. Improved Real-Time Segmentation of Intravascular Ultrasound Images Using Coordinate-Aware Fully Convolutional Networks. *Comput. Med. Imaging Graph.* **2021**, *91*, 101955. [CrossRef]
19. Ansari, M.Y.; Changaai Mangalote, I.A.; Meher, P.K.; Aboumarzouk, O.; Al-Ansari, A.; Halabi, O.; Dakua, S.P. Advancements in Deep Learning for B-Mode Ultrasound Segmentation: A Comprehensive Review. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *8*, 2126–2149. [CrossRef]
20. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
21. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**, arXiv:1804.03999.
22. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R.S., Bradley, A., Papa, J.P., Belagiannis, V., et al., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11045, pp. 3–11. ISBN 978-3-030-00888-8.
23. Li, J.; Liu, K.; Hu, Y.; Zhang, H.; Heidari, A.A.; Chen, H.; Zhang, W.; Algarni, A.D.; Elmannai, H. Eres-UNet++: Liver CT Image Segmentation Based on High-Efficiency Channel Attention and Res-UNet++. *Comput. Biol. Med.* **2023**, *158*, 106501. [CrossRef] [PubMed]
24. Chen, J.; Mei, J.; Li, X.; Lu, Y.; Yu, Q.; Wei, Q.; Luo, X.; Xie, Y.; Adeli, E.; Wang, Y.; et al. TransUNet: Rethinking the U-Net Architecture Design for Medical Image Segmentation through the Lens of Transformers. *Med. Image Anal.* **2024**, *97*, 103280. [CrossRef] [PubMed]
25. Cai, F.; Wen, J.; He, F.; Xia, Y.; Xu, W.; Zhang, Y.; Jiang, L.; Li, J. SC-Unext: A Lightweight Image Segmentation Model with Cellular Mechanism for Breast Ultrasound Tumor Diagnosis. *J. Imaging Inform. Med.* **2024**, *37*, 1505–1515. [CrossRef] [PubMed]
26. Azzopardi, C.; Camilleri, K.P.; Hicks, Y.A. Bimodal Automated Carotid Ultrasound Segmentation Using Geometrically Constrained Deep Neural Networks. *IEEE J. Biomed. Health. Inf.* **2020**, *24*, 1004–1015. [CrossRef] [PubMed]
27. Cao, J.; Chen, J.; Gu, Y.; Liu, J. MFA-UNet: A Vessel Segmentation Method Based on Multi-Scale Feature Fusion and Attention Module. *Front. Neurosci.* **2023**, *17*, 1249331. [CrossRef]
28. Gagan, J.H.; Shirsat, H.S.; Mathias, G.P.; Mallya, B.V.; Andrade, J.; Rajagopal, K.V.; Kumar, J.R.H. Automated Segmentation of Common Carotid Artery in Ultrasound Images. *IEEE Access* **2022**, *10*, 58419–58430. [CrossRef]
29. Cao, X.; Chen, H.; Li, Y.; Peng, Y.; Wang, S.; Cheng, L. Uncertainty Aware Temporal-Ensembling Model for Semi-Supervised ABUS Mass Segmentation. *IEEE Trans. Med. Imaging* **2021**, *40*, 431–443. [CrossRef]
30. Ashkani Chenarlogh, V.; Shabanzadeh, A.; Ghelich Oghli, M.; Sirjani, N.; Farzin Moghadam, S.; Akhavan, A.; Arabi, H.; Shiri, I.; Shabanzadeh, Z.; Sanei Taheri, M.; et al. Clinical Target Segmentation Using a Novel Deep Neural Network: Double Attention Res-U-Net. *Sci. Rep.* **2022**, *12*, 6717. [CrossRef]
31. Nazir, A.; Cheema, M.N.; Sheng, B.; Li, H.; Li, P.; Yang, P.; Jung, Y.; Qin, J.; Kim, J.; Feng, D.D. OFF-eNET: An Optimally Fused Fully End-to-End Network for Automatic Dense Volumetric 3D Intracranial Blood Vessels Segmentation. *IEEE Trans. Image Process.* **2020**, *29*, 7192–7202. [CrossRef]
32. He, Y.; Yang, G.; Yang, J.; Chen, Y.; Kong, Y.; Wu, J.; Tang, L.; Zhu, X.; Dillenseger, J.-L.; Shao, P.; et al. Dense Biased Networks with Deep Priori Anatomy and Hard Region Adaptation: Semi-Supervised Learning for Fine Renal Artery Segmentation. *Med. Image Anal.* **2020**, *63*, 101722. [CrossRef]
33. Lian, C.; Zhang, J.; Liu, M.; Zong, X.; Hung, S.-C.; Lin, W.; Shen, D. Multi-Channel Multi-Scale Fully Convolutional Network for 3D Perivascular Spaces Segmentation in 7T MR Images. *Med. Image Anal.* **2018**, *46*, 106–117. [CrossRef] [PubMed]

34. Ding, W.; Wang, J.; Zhou, W.; Zhou, S.; Chang, C.; Shi, J. Joint Localization and Classification of Breast Cancer in B-mode Ultrasound Imaging Via Collaborative Learning with Elastography. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 4474–4485. [CrossRef] [PubMed]
35. Wang, R.; Fang, Z.; Gu, J.; Guo, Y.; Zhou, S.; Wang, Y.; Yu, J. High-resolution Image Reconstruction for Portable Ultrasound Imaging Devices. *EURASIP J. Adv. Signal Process.* **2019**, *2019*, 56. [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

Analyzing the Relationship between Dose and Geometric Agreement Metrics for Auto-Contouring in Head and Neck Normal Tissues

Barbara Marquez [1,2,*], Zachary T. Wooten [3], Ramon M. Salazar [1,2], Christine B. Peterson [2,4], David T. Fuentes [2,5], T. J. Whitaker [1], Anuja Jhingran [6], Julianne Pollard-Larkin [1,2], Surendra Prajapati [1,2], Beth Beadle [7], Carlos E. Cardenas [8], Tucker J. Netherton [1] and Laurence E. Court [1,2]

[1] Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA; rmsalazar1@mdanderson.org (R.M.S.); tjwhitaker@mdanderson.org (T.J.W.); jmpollard@mdanderson.org (J.P.-L.); lecourt@mdanderson.org (L.E.C.)
[2] The University of Texas MD Anderson Cancer Center UTHealth Houston Graduate School of Biomedical Sciences, Houston, TX 77030, USA
[3] Department of Statistics, Rice University, Houston, TX 77005, USA; ztw5@rice.edu
[4] Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[5] Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[6] Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[7] Department of Radiation Oncology–Radiation Therapy, Stanford University, Stanford, CA 94305, USA; bbeadle@stanford.edu
[8] Department of Radiation Oncology, The University of Alabama, Birmingham, AL 35294, USA
* Correspondence: bmarquez@mdanderson.org; Tel.: +1-786-351-2115

Abstract: This study aimed to determine the relationship between geometric and dosimetric agreement metrics in head and neck (H&N) cancer radiotherapy plans. A total 287 plans were retrospectively analyzed, comparing auto-contoured and clinically used contours using a Dice similarity coefficient (DSC), surface DSC (sDSC), and Hausdorff distance (HD). Organs-at-risk (OARs) with ≥200 cGy dose differences from the clinical contour in terms of D_{max} (D0.01cc) and D_{mean} were further examined against proximity to the planning target volume (PTV). A secondary set of 91 plans from multiple institutions validated these findings. For 4995 contour pairs across 19 OARs, 90% had a DSC, sDSC, and HD of at least 0.75, 0.86, and less than 7.65 mm, respectively. Dosimetrically, the absolute difference between the two contour sets was <200 cGy for 95% of OARs in terms of D_{max} and 96% in terms of D_{mean}. In total, 97% of OARs exhibiting significant dose differences between the clinically edited contour and auto-contour were within 2.5 cm PTV regardless of geometric agreement. There was an approximately linear trend between geometric agreement and identifying at least 200 cGy dose differences, with higher geometric agreement corresponding to a lower fraction of cases being identified. Analysis of the secondary dataset validated these findings. Geometric indices are approximate indicators of contour quality and identify contours exhibiting significant dosimetric discordance. For a small subset of OARs within 2.5 cm of the PTV, geometric agreement metrics can be misleading in terms of contour quality.

Keywords: auto-contouring; contouring; radiotherapy; organs-at-risk; head and neck

1. Introduction

Contouring accuracy in radiotherapy is important for achieving optimal outcomes. This is especially true in head and neck (H&N) cancer treatment, where the therapeutic window is narrow due to the proximity of the planning target volume (PTV) to organs-at-risk (OARs) [1–3]. Advances in auto-contouring are reducing inconsistencies inherent

in manual delineation of PTVs and OARs and making this time-consuming, experience-dependent skill more efficient [4–6]. Researchers predominantly use geometric agreement metrics to assess the potential clinical impact of auto-contouring tools for deployment in radiotherapy planning or contour quality assurance [7–10]. Specifically, in 99.1% of studies presenting a new or previously established auto-contouring model, researchers reported geometric agreement metrics like similarity and overlap of a ground truth and auto-generated prediction. Conversely, in just 23.1% of studies, researchers reported the dosimetric impact [7].

Although frequently used, researchers have yet to show that geometric agreement metrics meaningfully correlate with the clinical acceptability of auto-contours or the potential implications regarding the final treatment dosimetry [11–13]. Some studies have considered the effect of geometric variations in contours on plan dose distributions. Fung et al. found that, despite generally small geometric discrepancies between manual- and auto-contours generated for OARs in adaptive radiotherapy for nasopharyngeal carcinoma, auto-contours often resulted in statistically significant higher doses in critical areas [14]. They suggest that even small geometric discrepancies can have significant dosimetric impacts due to steep dose gradients typical in IMRT; thus, auto-contours should be manually adjusted to ensure treatment safety and efficacy. Conversely, van Rooij et al. observed that even lower Dice similarity coefficients between deep learning-based and manual contours still produced dosimetric differences in treatment plans that were often clinically irrelevant; in this context, the study supports clinical acceptability of some automated delineations despite geometric inaccuracies [15]. In magnetic-resonance-imaging-based OAR auto-contours for brain tumor patients, Turcas et al. also found that geometric variations had minimal impact on dose distribution versus manual contours, where differences in median values for the mean and max dose were less than 0.2 Gy [16]. Notably, Lim et al. found that the correlation between geometric and dosimetric agreement metrics was weak ($R2 < 0.2$ for 61% of the correlations studied) and inconsistent when looking at plans generated from different combinations of physician expert-drawn target and OAR contours and resident physician-drawn contours over nasopharyngeal cancer cases [17]. They concluded that a dosimetric effect owing to contouring variation is not significantly captured with geometric indices alone.

Given these challenges, this study sought to further clarify the link between geometric agreement metrics commonly used in evaluating auto-contouring and dosimetric differences in H&N OARs. This study quantifies the impact of geometric edits on auto-contours in terms of the mean and maximum dose extracted from clinically delivered treatment plans.

2. Materials and Methods

Patients received H&N photon volumetric modulated arc radiotherapy (VMAT) plans designed with computed tomography (CT)-based manually drawn target contours and auto-generated OAR contours that were then manually edited. The dose distribution of the clinical plan was applied over the original, un-edited set of auto-contour OARs; dose statistics were compared between the auto-contour and the clinical OAR.

2.1. Contours and Plan Dose

Two H&N OAR contour sets were evaluated in this study: auto-contours and clinical contours. The auto-contour OARs were generated by the Radiation Planning Assistant (RPA), a full-automated treatment planning system whose deep-learning contouring is used routinely for delineating H&N OARs at our institution [18,19]. Clinical contours refer to the OAR contours that were automatically generated, then reviewed and modified by clinical planners prior to use for treatment planning. In this way, an "automated" set and a "clinical" set of contours were compared. In total, the following 19 OARs were evaluated: brain, brainstem, cochleas, esophagus, eyes, submandibular glands, larynx, lens, optic chiasm, optic nerves, parotids, spinal cord, and vertebral column.

The clinical VMAT radiotherapy plan–generated with the manually drawn target volume contours and the clinical (auto-generated, manually edited) OAR contours–was retrospectively collected. Then, the resulting doses to the original, unedited auto-contours and the clinical contours were extracted to compare dose statistics between the two contour sets.

2.2. Dataset

This study utilized a dataset of 287 patients from our institution to conduct an initial analysis; then, an external multi-institutional set of 91 patients were evaluated to validate the findings.

Internal. Radiotherapy treatment plan data from 287 H&N cancer patients treated with VMAT between August 2020 and February 2023 at our institution were retrospectively collected for this study. The planned doses ranged from 14 to 70 Gy, with 1–3 dose levels over 5–35 fractions; all patients received at least 2 Gy per fraction. Patients had RPA auto-contours available during contouring and were planned on a Raystation treatment planning system (version 11B, RaySearch Laboratories, Stockholm, Sweden).

External. Radiotherapy treatment plan data from 91 H&N cancer patients treated with VMAT between January 2022–November 2022 across 4 institutions in the continental United States were retrospectively collected. Planned total doses for these patients ranged from 60 to 70 Gy, with 1–3 dose levels over 25–35 fractions; all patients received at least 2 Gy per fraction.

2.3. Geometric Indices

The H&N OAR auto-contours were compared to the clinical contours using the Dice similarity coefficient (DSC), surface Dice similarity coefficient (sDSC), and Hausdorff distance (HD). These metrics were selected owing to their widespread reporting and demonstrated effectiveness in detecting discrepancies in independently generated auto-contours [7,9,12,20]. For an sDSC, we chose a tolerance of 2 mm as the center value of previously studied tolerance levels with a similar accuracy in detecting contouring errors [9]. Additionally, distance to the closest PTV, defined by the Euclidean distance between the OAR contour and PTV, was collected. The previous literature has suggested that the relation between dose differences and shortest distance to the edge of PTV provides a robust tolerance guideline for contouring variability and calls for a larger dataset to determine a distance cut-off point with more detail [21–23].

2.4. Dosimetric Indices

The clinical plan dose was applied over the two contour sets being compared in this study. The established clinical objective criteria for H&N cancer treatment plans at our institution specify constraints in terms of D_{max} (as defined by the maximum dose received by at least 0.01 cc of the volume) for the brain, brainstem, cochlea, esophagus, eyes, lens, optic chiasm, optic nerves, spinal cord, vertebral column and D_{mean} for larynx, submandibular glands, and parotids. The clinical objective dose (D_{CO}) for each respective organ is reported and analyzed as specified (D_{max} for most organs, D_{mean} for larynx, submandibular glands, and parotids).

2.5. Statistical Analysis

Effect of contour variations on dose differences. The mean and standard deviation of geometric similarity between the clinical OARs auto-contour OARs was reported in terms of DSC, sDSC, and HD. The difference in D_{max} and D_{mean} doses between the two contour sets was tested for statistical significance using a 2-sided Wilcoxon signed-rank test ($p < 0.05$). Then, the relationship between geometric and dosimetric indices was evaluated with the absolute value of dosimetric differences; thus, the median and 90% quantile are reported to describe the spread of observed differences in D_{max} and D_{mean}.

2.6. Linear Relationship

It was investigated whether a linear relationship between geometric and dosimetric indices existed for the entire dataset and for each individual organ. The R2 statistic was used to determine the strength of correlation.

2.7. Thresholding Dose Differences and Distance to the PTV

This study identified contours with a threshold of a 200 cGy absolute difference in dose, with respect to the clinical objective dose metric D_{CO} for each organ, between the two contour sets. This was used to further analyze characteristics of dosimetric deviation in terms of geometric similarity and distance to the closest PTV in a subset of contours. This criterion was chosen as 200 cGy is about 3%, or 1 fraction, of the total target dose prescription typical of an H&N radiotherapy treatment plan (69.96 Gy/33 fractions). All patients in the internal dataset who had a wide planned dose inclusion range received at least 200 cGy per fraction. The distance to PTV is investigated as a potential metric to identify contour pairs who may exhibit high geometric similarity but low dosimetric similarity and vice versa. Illustrative cases are provided for both scenarios. All calculations were made using custom Python scripts that leverage common scientific libraries.

To validate the findings for our internal dataset and assess the generalizability of the observed trends, an external, multi-institutional dataset for 91 H&N cancer patients previously treated with radiotherapy were also evaluated with the same methodology. The purpose of this extended analysis was to ascertain whether the patterns and discrepancies identified in the original cohort were consistent in a broader clinical context.

3. Results

3.1. Internal Dataset–Geometric and Dose Evaluation

The 287 internal patient dataset each had plans containing up to 13 unique OAR contour pairs for a total of 4995 comparison pairs (original auto-contour vs. clinical contour). Although we evaluated 287 patients, the sample size per OAR did not homogenously total 287 because clinicians delete certain contours deemed nonrelevant for a specific patient's radiotherapy plan (or otherwise the organ was part of the target volume). Treatment site characteristics for the 287 patients are displayed in Table 1.

Table 1. Treatment characteristics of the internal patient dataset from the Internal Classification of Diseases (ICD) coding.

Primary Site	Patients, No.
Bone & Soft Tissue	12
Larynx	19
Lip and Oral Cavity	54
Nasal Cavity	13
Oropharynx	104
Other	53
Skin	20
Thyroid	12
Clinical characteristics of the study patients (n = 287)	

Summary statistics for geometric and dosimetric agreement metrics can be found in Table 2. ΔDose was calculated and reported as the absolute value of the auto-contour dose minus the clinical contour dose. Overall, 90% of contours had a DSC, sDSC, and HD of at least 0.75, 0.86, and less than 7.65 mm, respectively. Notably, the esophagus, larynx and spinal cord exhibited high geometric variation in the summary statistics. Though 254/282 spinal cords exhibited a HD of <4 mm, 11 cases exhibited >40 mm and thus affected the observed standard deviation (though these cases did not correlate with differences to D_{max}, where only 2/11 cases exhibited more than 200 cGy difference). In general, these organs exhibited such variations due to the clinical planners removing contours from slices distant from the treatment volume (where the auto-contour delineates the full anatomical extent of the organ).

Table 2. Summary statistics for geometric and dosimetric agreement for all organs. ΔDose = clinical contour dose–auto-contour dose. Then, a summary of OARs identified for exhibiting a ≥200 cGy dose deviation from the pertinent clinical objective between the clinical contour and the auto-contour.

OAR	*N*	DSC		sDSC		HD (mm)		ΔD_{max} (cGy)		ΔD_{mean} (cGy)		When $D_{CO} \geq 200$ cGy				
												$D_{CO} \geq 200$ cGy		*N* (% of total OARs)	*N* (% of OARs ≤ 3 cm from PTV)	Range ΔD_{CO} (cGy)
		Mean	SD	Mean	SD	Mean	SD	Median	Q90%	Median	Q90%	Max	Mean			
All structures	4995	0.90	0.11	0.95	0.1	3.91	9.22	0	55	0.67	49			229 (5%)	224 (97%)	201–2565
Brain	240	0.99	0.01	0.99	0.03	3.74	2.80	0	113	0	4	X		13 (5%)	13 (100%)	214–914
Brainstem	285	0.96	0.04	0.97	0.07	2.56	2.45	0	38	1	36	X		6 (2%)	6 (100%)	206–541
Cochlea_L	275	0.78	0.17	0.92	0.12	2.75	1.92	0	80	2	76	X		17 (6%)	17 (100%)	240–1225
Cochlea_R	278	0.81	0.14	0.95	0.08	2.41	1.45	0	130	2	62	X		20 (7%)	20 (100%)	238–1469
Esophagus	260	0.91	0.14	0.94	0.14	13.27	29.49	0	124	3	449	X		20 (8%)	20 (100%)	216–1213
Eye_L	271	0.96	0.04	0.99	0.05	1.82	1.42	0	24	0	6	X		6 (2%)	6 (100%)	234–905
Eye_R	273	0.96	0.03	0.99	0.03	1.7	0.84	0	12	0	4	X		0 (0%)	N/A	N/A
Glnd_Submand_L	238	0.91	0.12	0.94	0.13	3.36	3.94	0	36	1	51		X	7 (3%)	6 (86%)	248–863
Glnd_Submand_R	243	0.94	0.07	0.97	0.08	2.96	2.95	0	35	1	35		X	3 (1%)	3 (100%)	217–312
Larynx	245	0.87	0.15	0.81	0.23	7.83	7.39	2	512	26	799		X	68 (28%)	68 (100%)	204–2229
Lens_L	264	0.85	0.13	0.97	0.07	1.73	0.89	0	14	0	5	X		0 (0%)	N/A	N/A
Lens_R	268	0.86	0.1	0.98	0.05	1.74	0.80	0	10	0	5	X		4 (1%)	3 (75%)	211–544
OpticChiasm	209	0.93	0.12	0.97	0.1	1.75	1.74	0	5	0	6	X		6 (3%)	6 (100%)	266–2565
OpticNrv_L	269	0.83	0.12	0.95	0.08	4.28	4.57	0	52	1	49	X		12 (4%)	9 (75%)	201–654
OpticNrv_R	270	0.83	0.12	0.95	0.08	4.17	4.05	0	46	1	37	X		0 (0%)	N/A	N/A
Parotid_L	280	0.96	0.05	0.97	0.07	5.55	7.11	0	67	2	174		X	25 (9%)	25 (100%)	225–850
Parotid_R	282	0.96	0.03	0.97	0.06	4.68	5.47	0	46	2	107		X	19 (7%)	19 (100%)	213–820
SpinalCord	282	0.95	0.07	0.98	0.07	5.39	19.22	0	21	1	14	X		2 (0.01%)	2 (100%)	247–302
VertebralColumn	263	0.98	0.01	0.99	0.01	2.62	2.26	0	13	1	10	X		1 (0.5%)	1 (100%)	305

Dosimetrically, the absolute difference between the two contour sets was less than 200 cGy for 95% of OARs in terms of D_{max} and 96% in terms of D_{mean}. The clinical objective criteria at our institution specify constraints in terms of D_{max} (as defined by the maximum dose inside a volume of 0.01 cc) for most OARs and D_{mean} for the larynx, submandibular glands, and parotids; thus, in terms of their clinical objective dose D_{CO}, 96% of relevant organs had D_{max} < 200 cGy and 90% had D_{mean} < 200 cGy. The signed differences in dose between the two contour sets was not found to be statistically significant (p = 0.34, 0.45 for D_{max}, D_{mean}).

3.2. Internal Dataset–Linear Relationship

Overall, the linear relationship between geometric agreement metrics and the dose difference when considering all contours was poor. For N = 4995, the R2 between D_{max} and the DSC, sDSC, and HD were 0.09, 0.14, and 0.04, respectively. This is likely due to 57% (2874/4995) of contour pairs exhibiting a 0 dose difference (the auto-contour OAR was used as-is). However, re-assessing the linear relationship, thresholding for a minimum D_{max} difference of 100 cGy across all contours, does not improve the correlation (R2 = 0.06, 0.06, and 0.01, respectively).

Figure 1 summarizes the correlation between geometric and dosimetric agreement metrics by OAR. Fitting a model to each organ improved the correlation strength, though no geometric index showed a consistently better performance over most OARs. Generally, geometric agreement metrics correlated more strongly with ΔD_{mean} than ΔD_{max}. Expectedly, DSC showed a stronger correlation with ΔD_{mean}, where it exhibited a markedly varied performance with ΔD_{max}. This is exhibited later when examining cases that scored an excellent DSC yet poor dosimetric agreement in terms of ΔD_{max}, highlighting the nuances of capturing errors with volumetric overlap in large versus smaller structures.

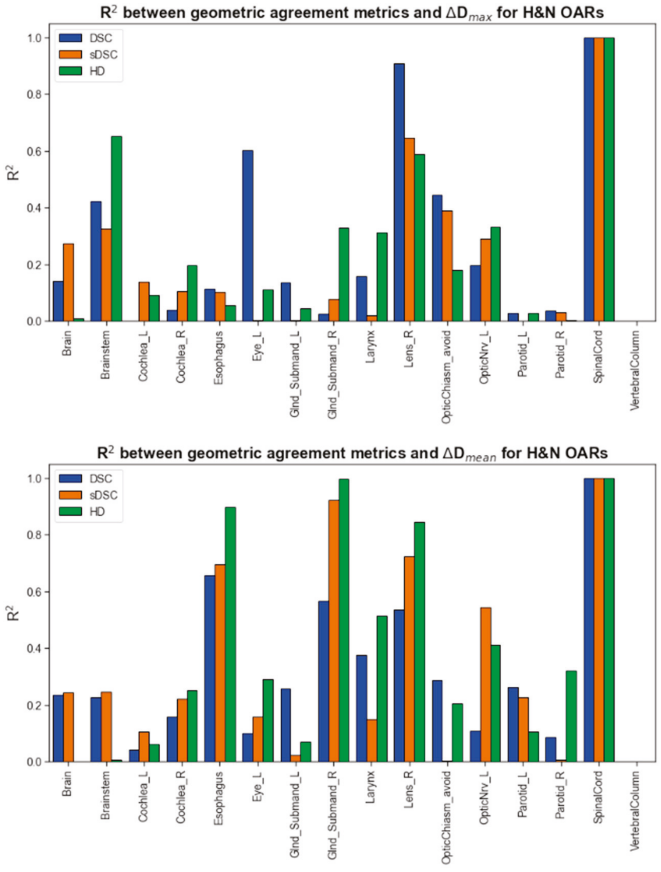


Figure 1. The correlation between geometric and dosimetric incides from clinical contours vs. auto-contours, as compared over one clinical plan dose distribution.

3.3. Thresholding Differences and Distance to the PTV

A threshold of 200 cGy absolute dose difference in D_{CO} resulted in 229/4995 (4.5%) of the total structures being identified. Table 2 summarizes the sample size of each OAR that fit the criterion, the clinical objective dose metric considered, the relative number of OARs identified against their sample size in the total dataset, the number of OARs within 3 cm of the PTV, and the observed range of dose variance in terms of D_{CO}.

Notably, 97% (223/229) of OARs meeting this criterion were within 2.5 cm of the nearest PTV; thus, the distance-to-nearest-PTV metric exhibited a markedly higher sensitivity to identifying cases where two contours may disagree dosimetrically (though not necessarily geometrically). For example, 27% (60/224) OARs with ΔD_{CO} had DSC > 0.90; all 60 were within 2.5 cm of the PTV. Additionally, 31% (70/224) had sDSC > 0.90 or HD < 5 mm, of which 68 were within 2.5 cm of the PTV (and 2 were beyond 5 cm). These findings are summarized in Figure 2.

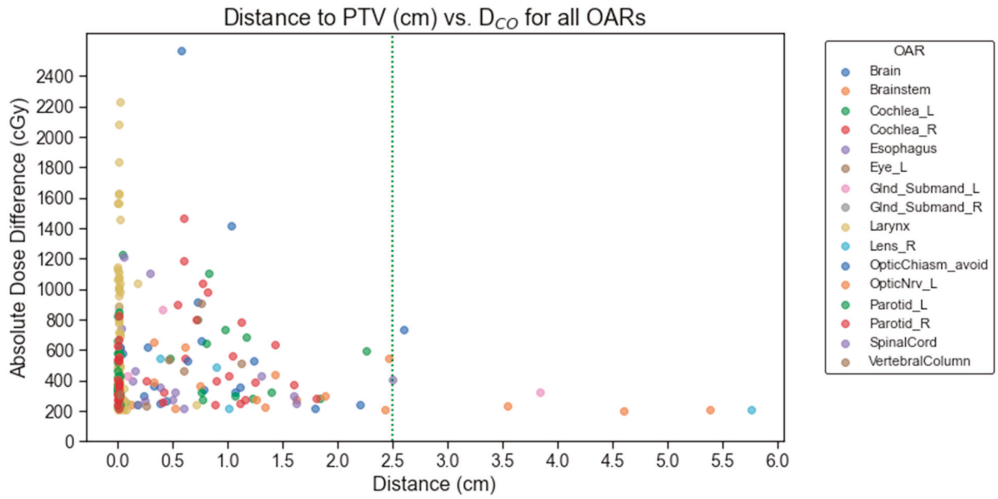


Figure 2. The relationship between identifying an OAR for exhibiting $\Delta D_{CO} \geq 200$ cGy deviation between the clinical contour and the auto-contour and relative distance to the closest PTV.

Generally, high geometric agreement reduces the chance of a 200 cGy dose discrepancy between two contour sets in H&N treatment plans. This relationship is shown in Figure 3. Some noise in the linear decrease between geometric agreement and dosimetric disagreement exists because most of the data collected is skewed towards higher geometric agreement scores and lower dosimetric disagreement. When OARs are within 2.5 cm of the PTV, the trend may not follow. This suggests that proximity to a clinical target heightens the dosimetric error potential, even with geometrically similar contours. Notably, OARs distant from the PTV, even those with low geometric agreement, do not consistently indicate significant dose discrepancies based on the clinical plan's isodose lines. Cases illustrating the discrepancies in geometric and dosimetric agreement are shown in Figures 4 and 5. A treatment plan for a bilateral neck tumor for which a high geometric agreement did not indicate high dosimetric agreement is shown in Figure 4. Although the auto-contour of the brain (red) had a DSC = 0.99, sDSC = 0.98, and HD = 3.2 mm against the clinical contour (green, filled), the two structures differed by 8 Gy reported to the D_{max}. The prediction difference in the inferior aspect of the brain for the auto-contour encompassed additional isodose regions that the clinical contour did not on some slices. A treatment plan for a left cheek tumor for which a low geometric agreement did not indicate low dosimetric agreement is shown in Figure 5. Although the left cochlea auto-contour prediction (red) was substantially smaller than the clinical contour (green, filled), with a DSC = 0.61, sDSC = 0.82, and HD = 4.9 mm, the two contours were completely encompassed in a 10 Gy isodose region 2 cm away from the PTV in the same laterality. Thus, the D_{max} difference between the two cochleae was 31 cGy.

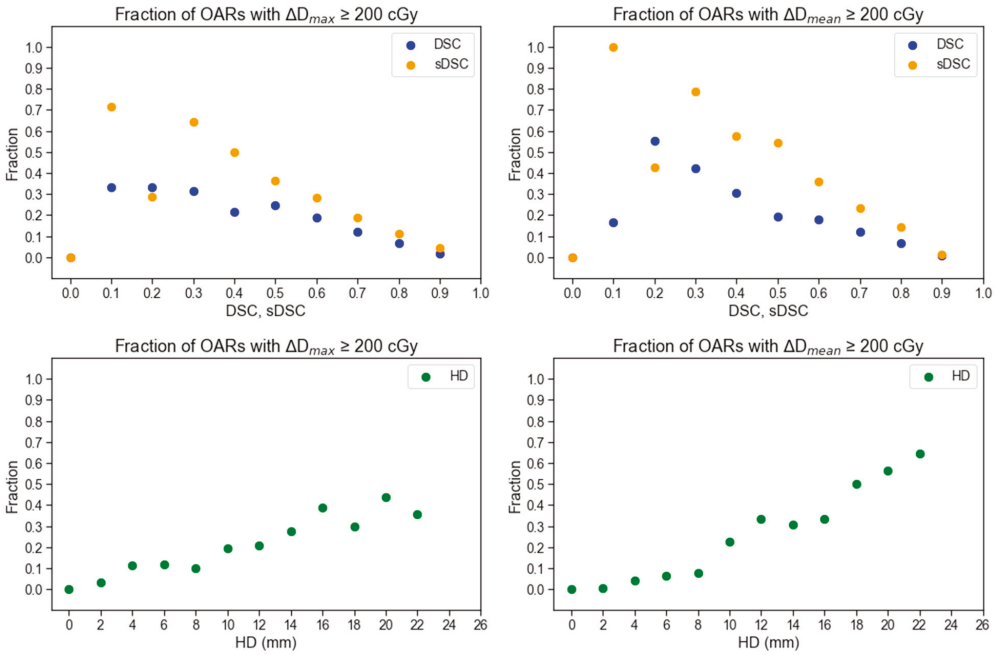


Figure 3. The relative proportions of contours across all OARs identified for exhibiting a dose deviation of at least 200 cGy in terms of D_{max} (D0.01 cc) or D_{mean} versus binned geometric scores. DSC and sDSC were binned at every 0.1 interval and HD was binned at every 2 mm.

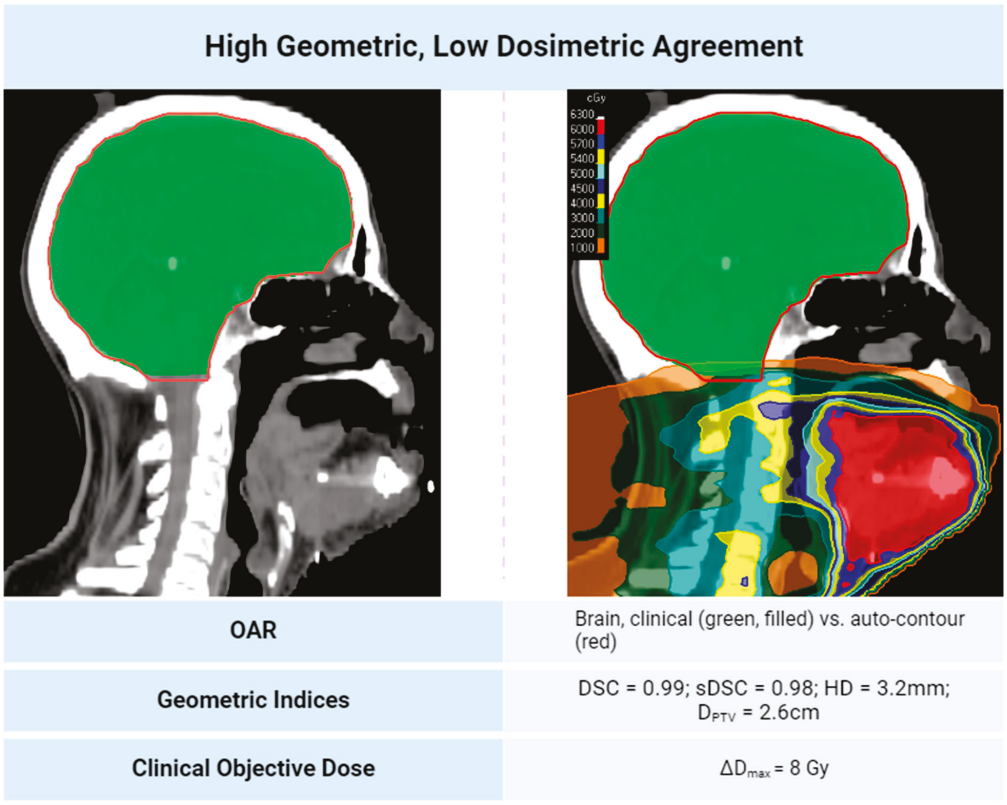


Figure 4. A treatment plan for a bilateral neck tumor for which a high geometric index score did not indicate high dosimetric agreement. The prediction difference in the inferior aspect of the brain for the auto-contour encompassed additional isodose regions that the clinical contour did not on some slices.

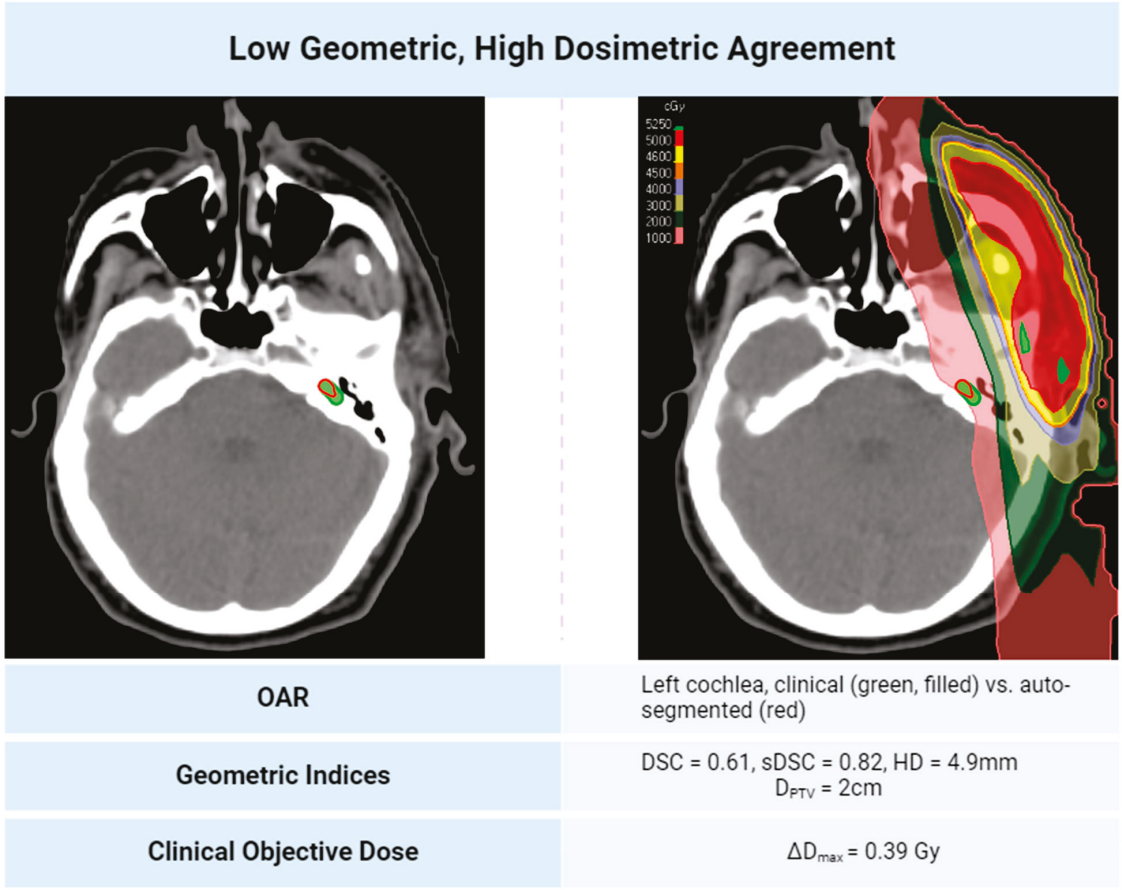


Figure 5. A treatment plan for a left cheek tumor for which a low geometric agreement did not indicate low dosimetric agreement. The two contours lie in an approximately homogenous isodose region 2 cm away from the PTV in the same laterality.

3.4. External Dataset

We repeated the workflow described above on a multi-institutional dataset of 91 patients, obtaining 1129 OAR comparison pairs. The mean (±SD) DSC, sDSC, and HD (mm) for all OARs in this dataset were 0.79 ± 0.19, 0.80 ± 0.21, and 12 ± 29, respectively. Again, the esophagus, larynx, and spinal cords exhibited high geometric variability between the clinical sub-volume being considered for planning and the auto-contour delineating the full anatomical extent. However, despite higher geometric discordance, the dosimetric agreement was within 200 cGy for 95% for contours overall in terms of D_{max} and D_{mean} and 94% and 90% for contours in terms of their D_{CO} of D_{max} and D_{mean}, respectively. Correlation strength between geometric and dosimetric agreement metrics were markedly poor for the external dataset ($R2 < 0.40$ for all correlations except for spinal cord and vertebral column). Notably, 53/1129 (4%) OARs exhibited $\Delta D_{CO} \geq 200$ cGy, of which 32/53 (60%) were within 2.5 cm of the PTV.

4. Discussion

In this work, we found that geometric agreement metrics commonly used to compare contours are approximate indicators of quality. As the geometric disagreement increases, there is an increased likelihood of a meaningful dosimetric difference between two contours. Notably, this study reported on the relationship between geometric agreement and dosimetric endpoints in auto-contours that are being used in a clinical context, providing insight on the clinical acceptability (readiness of use) with metrics instead of qualitative evaluation (physician scoring). Variations in contouring when using auto-contours as a starting point were significantly low (with most contours being used as-is). This study

transposed auto-contours on the clinical treatment plan as an approximate indicator of how geometric edits might affect final plan dose to enable a study of a large patient cohort compared to replanning studies.

We were able to investigate 13 key OARs that are routinely contoured in our clinical practice; however, additional organs and muscles, which may be delineated in other planning practices, should also be assessed. Notably, the swallowing organs may be of interest, as they are associated with dysphagia, aspiration, and general quality of life. While not analyzed in this study, we can infer similar effects of contouring differences on dose to these swallowing structures due to the anatomical proximity and volumetric similarity to organs in our study. The OARs analyzed in this work cover a wider range of volumes and proximities to the treatment targets. These include small organs such as the lenses (average volume of 0.20 cm^3) and optic nerves (0.72 cm^3), medium-sized organs like the spinal cord (22 cm^3) and brainstem (26 cm^3), and large organs such as the brain (1382 cm^3). Swallowing organs tend to have similar volumes to the small and medium OARs included in this study. For instance, the buccinator muscles (5 cm^3) are comparable in size to the submandibular glands (8 cm^3) and are also in close proximity. Furthermore, some of these swallowing muscles are smaller, anatomically continuous volumes of analyzed organs, like the esophagus and pharyngeal constrictor muscles or the thyroid and cricoid cartilage and the larynx. A volumetric comparison of 23 swallowing organs to the 13 key OARs analyzed in this study can be found in the Appendix A.

Additionally, due to the inclusion of diverse prescriptions and fractionations in the initial dataset, we conducted additional analyses to address potential discrepancies in the results. We modified the inclusion criteria to exclusively consider patients who received at least 40 Gy in their treatment and identified structures based on dosimetric discrepancies as a percentage of the total plan dose (rather than absolute dose), considering both 3% and 5% of the prescribed dose. Changing the inclusion criteria in terms of plan dose and identification threshold in terms of a percentage of the prescribed dose did not change the conclusions drawn in this study; the same trends were observed as summarized in Figures 2 and 3. This consideration pertains to VMAT plans, where contours and prescriptions prepared for other radiotherapy treatment modalities, such as proton therapy, should also be evaluated.

This study also highlighted some flaws in the use of geometric agreement metrics to determine the clinical readiness of auto-contours. Specifically, even if the geometric index indicates high agreement, there may still be large dosimetric differences when the OAR is close to target structures. Flaws in the use of geometric metrics have been highlighted by other authors [24–28]. For example, a DSC can be lenient of contouring errors for large structures like the brain and parotids while penalizing errors in smaller structures like the eyes, cochleas, lenses, and optic nerves. Surface DSC tends to overlook the overlap between structures, which would be particularly challenging for evaluating OARS who overlap with the PTV and thus are at risk of having high dose differences between two contours. Additionally, an sDSC has differential sensitivity, with the ability to tune its tolerance value to surface errors when calculating. This study's limitation was the application of a uniform tolerance of 2 mm to decrease variability in results. Such uniformity is not always suitable when studying structures that vary in size. The effects of structure size on calculating a Hausdorff distance are markedly lower. However, deletion of slices, a common contouring error (where the clinical planner forgets to interpolate manual contours) may not be identified by some of these geometric metrics. Again, these highlight the fact that these metrics should not be used without additional checks.

Our results indicated that additional review is critical for structures that are within 2.5 cm from the target. Specific tolerance levels of contouring variability against OAR distance-to-PTV have been previously studied and seen to be especially important for D_{max}. Vaassen et al. suggested that a prescribed dose will have an impact on the proposed distance-to-PTV cut-off point for checking contours, but in our study, with almost 5000 contours and a wide range of dose levels, we found the significance of checking close structures

at multiple prescriptions [21]. Further work of interest would include designing a novel metric that can account for geometric agreement and relative distance to a target volume.

The weaknesses identified in this study are caused by the fact that contouring quality is generally evaluated before the generation of a treatment plan, so the impact of contouring discrepancies on dose is unknown (at that point). The more recent introduction of deep learning to predict what the planned dose distribution will look like [29,30] means that it is possible now to compare contours and the dosimetric impact of any differences before the creation of a treatment plan. Given the findings of our work, especially the gaps that werre left when using overlap indices alone, it seems likely that future evaluation of clinical readiness of auto-contours will include dose predictions so that they can highlight errors based on dosimetric differences.

5. Conclusions

Our investigation showed that the relationship between geometric metrics and the dosimetric impact of differences in contours is not straightforward; thus, this should be considered when developing tools to support clinical acceptability of auto-contouring. We found that dose differences between an auto-contour and a clinical contour are small when using auto-contours as a starting point. Dose variances resulting from contour edits correlated moderately with geometric agreement metrics. Notably, the distance from an OAR to the PTV stood out as an essential metric for identifying OARs with geometric agreements but dosimetric divergences. By corroborating our primary dataset's results with data from multiple institutions, we demonstrated the replicability and applicability of our primary findings across diverse datasets.

Author Contributions: B.M.: Conceptualization, Data Curation, Software, Formal Analysis, Writing—Original Draft. Z.T.W.: Statistical analysis, Writing—Reviewing, Editing. R.M.S.: Statistical analysis, Writing—Reviewing, Editing. C.B.P.: Supervision, Statistical analysis, Writing—Reviewing, Editing. D.T.F.: Supervision, Writing—Reviewing, Editing. T.J.W.: Software, Data Curation, Supervision, Writing—Reviewing. A.J.: Supervision, Writing—Reviewing. J.P.-L.: Supervision, Writing—Reviewing. S.P.: Supervision, Writing—Reviewing. B.B.: Supervision, Writing—Reviewing, Editing. C.E.C.: Software, Writing—Reviewing, Editing. T.J.N.: Conceptualization, Supervision, Writing—Reviewing, Editing. L.E.C.: Conceptualization, Supervision, Funding Acquisition, Writing—Original Draft. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Cancer Prevention and Research Institute of Texas, grant #RP200395; UTHealth Innovation for Cancer Prevention Research Training Program Pre-doctoral Fellowship (Cancer Prevention and Research Institute of Texas grant #RP210042), and the UTHealth Graduate School of Biomedical Sciences American Legion Auxiliary Fellowships in Cancer Research.

Institutional Review Board Statement: All used data in this study was collected and analyzed under an approved The University of Texas MD Anderson Cancer Center Institutional Review Board protocol (IRB) #PA16-0379 (Date: 30 May 2016). Consent form was waived due to the retrospective nature of the study. All experiments were carried out in accordance with institutional policies.

Informed Consent Statement: Consent form was waived due to the retrospective nature of the study. All experiments were carried out in accordance with institutional policies.

Data Availability Statement: The datasets presented in this article are not readily available because they are not anonymized patient data from our institution. Requests to access the datasets should be directed to Laurence E. Court.

Acknowledgments: The authors acknowledge Mary Gronberg, Kai Huang, Daniel el Basha, Constance Owens, Raymond Mumme, Raphael Douglas, Callistus Nguyen, and Alex Leone for their data collection and computational support. The manuscript was edited by Donald R. Norwood of the Research Medical Library at MD Anderson.

Conflicts of Interest: The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Laurence E. Court—Leo Cancer Care Stock options as a member of their scientific board. Not related to this work. Julianne Pollard-Larkin—Member of the American Association of Physicists and the Medicine Board of Directors.

Surendra Prajapati—Royalty payment from Wisconsin Alumni Research Foundation made to Prajapati as investor for licensing the patent "Tissue fluorescence monitor with ambient light rejection." Not related to this work. US patent #10,045,696 issued on 14 August 2018. All other authors have no known competing financial interests.

Appendix A

OAR	Average Volume (cm^3)
Lens	0.20
Cochlea	0.22
Optic Nerve	0.72
Optic Chiasm	0.75
Medial Constrictor *	1.08
Glottic Area *	1.09
Cricopharyngeus *	2.03
Inferior Constrictor *	2.27
Digastric *	3.51
Buccinator *	5.07
Cricoid *	6.20
Lateral Pterygoid *	7.48
Eyes	7.94
Submandibular Glands	8.12
Medial Pterygoid *	9.75
Thyroid Cartilage *	9.74
Superior Constrictor *	12.60
MGH Complex *	13.85
Supraglottic Larynx *	14.49
Esophagus	21.16
SpinalCord	22.87
Genioglossus *	23.35
Masseter *	24.70
Brainstem	26.55
Parotid	33.41
Larynx	36.33
Tongue *	41.25
Vertebral Column	328.29
Brain	1382.45

* swallowing organs from a secondary, internal dataset of 20 patients.

References

1. Brodin, N.; Tomé, W.A. Revisiting the dose constraints for head and neck OARs in the current era of IMRT. *Oral Oncol.* **2018**, *86*, 8–18. [CrossRef]
2. Buciuman, N.; Marcu, L.G. Dosimetric justification for the use of volumetric modulated arc therapy in head and neck cancer—A systematic review of the literature. *Laryngoscope Investig. Otolaryngol.* **2021**, *6*, 999–1007. [CrossRef] [PubMed]
3. Bentzen, S.M.; Constine, L.S.; Deasy, J.O.; Eisbruch, A.; Jackson, A.; Marks, L.B.; Haken, R.K.T.; Yorke, E.D. Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC): An Introduction to the Scientific Issues. *Int. J. Radiat. Oncol.* **2010**, *76*, S3–S9. [CrossRef] [PubMed]
4. Cardenas, C.E.; Yang, J.; Anderson, B.M.; Court, L.E.; Brock, K.B. Advances in Auto-Segmentation. *Semin. Radiat. Oncol.* **2019**, *29*, 185–197. [CrossRef]
5. Netherton, T.J.; Cardenas, C.E.; Rhee, D.J.; Court, L.E.; Beadle, B.M. The Emergence of Artificial Intelligence within Radiation Oncology Treatment Planning. *Oncology* **2021**, *99*, 124–134. [CrossRef]
6. van der Veen, J.; Willems, S.; Deschuymer, S.; Robben, D.; Crijns, W.; Maes, F.; Nuyts, S. Benefits of deep learning for delineation of organs at risk in head and neck cancer. *Radiother. Oncol.* **2019**, *138*, 68–74. [CrossRef]
7. Mackay, K.; Bernstein, D.; Glocker, B.; Kamnitsas, K.; Taylor, A. A Review of the Metrics Used to Assess Auto-Contouring Systems in Radiotherapy. *Clin. Oncol.* **2023**, *35*, 354–369. [CrossRef] [PubMed]
8. Sadeghi, S.; Siavashpour, Z.; Sadr, A.V.; Farzin, M.; Sharp, R.; Gholami, S. A rapid review of influential factors and appraised solutions on organ delineation uncertainties reduction in radiotherapy. *Biomed. Phys. Eng. Express* **2021**, *7*, 052001. [CrossRef] [PubMed]
9. Rhee, D.J.; Akinfenwa, C.P.A.; Rigaud, B.; Jhingran, A.; Cardenas, C.E.; Zhang, L.; Prajapati, S.; Kry, S.F.; Brock, K.K.; Beadle, B.M.; et al. Automatic contouring QA method using a deep learning–based autocontouring system. *J. Appl. Clin. Med. Phys.* **2022**, *23*, e13647. [CrossRef]

10. Wooten, Z.T.; Yu, C.; Court, L.E.; Peterson, C.B. Predictive modeling using shape statistics for interpretable and robust quality assurance of automated contours in radiation treatment planning. In Proceedings of the Pacific Symposium on Biocomputingm 2023, Waimea, HI, USA, 3–7 January 2023; pp. 395–406. [CrossRef]
11. Sherer, M.V.; Lin, D.; Elguindi, S.; Duke, S.; Tan, L.-T.; Cacicedo, J.; Dahele, M.; Gillespie, E.F. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. *Radiother. Oncol.* **2021**, *160*, 185–191. [CrossRef]
12. Baroudi, H.; Brock, K.K.; Cao, W.; Chen, X.; Chung, C.; Court, L.E.; El Basha, M.D.; Farhat, M.; Gay, S.; Gronberg, M.P.; et al. Automated Contouring and Planning in Radiation Therapy: What Is 'Clinically Acceptable'? *Diagnostics* **2023**, *13*, 667. [CrossRef] [PubMed]
13. Wahid, K.A.; Cardenas, C.E.; Marquez, B.; Netherton, T.J.; Kann, B.H.; Court, L.E.; He, R.; Naser, M.A.; Moreno, A.C.; Fuller, C.D.; et al. Evolving Horizons in Radiation Therapy Auto-Contouring: Distilling Insights, Embracing Data-Centric Frameworks, and Moving Beyond Geometric Quantification. *Adv. Radiat. Oncol.* **2024**, *9*, 101521. [CrossRef] [PubMed]
14. Fung, N.T.C.; Hung, W.M.; Sze, C.K.; Lee, M.C.H.; Ng, W.T. Automatic segmentation for adaptive planning in nasopharyngeal carcinoma IMRT: Time, geometrical, and dosimetric analysis. *Med. Dosim.* **2020**, *45*, 60–65. [CrossRef] [PubMed]
15. van Rooij, W.; Dahele, M.; Brandao, H.R.; Delaney, A.R.; Slotman, B.J.; Verbakel, W.F. Deep Learning-Based Delineation of Head and Neck Organs at Risk: Geometric and Dosimetric Evaluation. *Int. J. Radiat. Oncol. Biol. Phys.* **2019**, *104*, 677–684. [CrossRef] [PubMed]
16. Turcas, A.; Leucuta, D.; Balan, C.; Clementel, E.; Gheara, C.; Kacso, A.; Kelly, S.M.; Tanasa, D.; Cernea, D.; Achimas-Cadariu, P. Deep-learning magnetic resonance imaging-based automatic segmentation for organs-at-risk in the brain: Accuracy and impact on dose distribution. *Phys. Imaging Radiat. Oncol.* **2023**, *27*, 100454. [CrossRef] [PubMed]
17. Lim, T.Y.; Gillespie, E.; Murphy, J.; Moore, K.L. Clinically Oriented Contour Evaluation Using Dosimetric Indices Generated from Automated Knowledge-Based Planning. *Int. J. Radiat. Oncol.* **2019**, *103*, 1251–1260. [CrossRef]
18. Rhee, D.J.; Rhee, D.J.; Cardenas, C.E.; Cardenas, C.E.; Elhalawani, H.; Elhalawani, H.; McCarroll, R.; McCarroll, R.; Zhang, L.; Zhang, L.; et al. Automatic detection of contouring errors using convolutional neural networks. *Med. Phys.* **2019**, *46*, 5086–5097. [CrossRef] [PubMed]
19. Court, L.; Aggarwal, A.; Burger, H.; Cardenas, C.; Chung, C.; Douglas, R.; du Toit, M.; Jaffray, D.; Jhingran, A.; Mejia, M.; et al. Addressing the Global Expertise Gap in Radiation Oncology: The Radiation Planning Assistant. *JCO Glob. Oncol.* **2023**, *9*, e2200431. [CrossRef] [PubMed]
20. Reinke, A.; Tizabi, M.D.; Sudre, C.H.; Eisenmann, M.; Rädsch, T.; Baumgartner, M.; Acion, L.; Antonelli, M.; Arbel, T.; Bakas, S.; et al. Common Limitations of Image Processing Metrics: A Picture Story. *arXiv* **2021**, arXiv:2104.05642.
21. Vaassen, F.; Hazelaar, C.; Canters, R.; Peeters, S.; Petit, S.; van Elmpt, W. The impact of organ-at-risk contour variations on automatically generated treatment plans for NSCLC. *Radiother. Oncol.* **2021**, *163*, 136–142. [CrossRef]
22. Petit, S.F.; van Elmpt, W. Accurate prediction of target dose-escalation and organ-at-risk dose levels for non-small cell lung cancer patients. *Radiother. Oncol.* **2015**, *117*, 453–458. [CrossRef] [PubMed]
23. Hunt, M.A.; Jackson, A.; Narayana, A.; Lee, N. Geometric factors influencing dosimetric sparing of the parotid glands using IMRT. *Int. J. Radiat. Oncol.* **2006**, *66*, 296–304. [CrossRef] [PubMed]
24. Gooding, M.J.; Boukerroui, D.; Osorio, E.V.; Monshouwer, R.; Brunenberg, E. Multicenter comparison of measures for quantitative evaluation of contouring in radiotherapy. *Phys. Imaging Radiat. Oncol.* **2022**, *24*, 152–158. [CrossRef] [PubMed]
25. Vandewinckele, L.; Claessens, M.; Dinkla, A.; Brouwer, C.; Crijns, W.; Verellen, D.; van Elmpt, W. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. *Radiother. Oncol.* **2020**, *153*, 55–66. [CrossRef] [PubMed]
26. Taha, A.A.; Hanbury, A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Med. Imaging* **2015**, *15*, 29. [CrossRef] [PubMed]
27. Huttenlocher, D.; Klanderman, G.; Rucklidge, W. Comparing images using the Hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.* **1993**, *15*, 850–863. [CrossRef]
28. Vaassen, F.; Hazelaar, C.; Vaniqui, A.; Gooding, M.; van der Heyden, B.; Canters, R.; van Elmpt, W. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Phys. Imaging Radiat. Oncol.* **2020**, *13*, 1–6. [CrossRef] [PubMed]
29. Gronberg, M.P.; Beadle, B.M.; Garden, A.S.; Skinner, H.; Gay, S.; Netherton, T.; Cao, W.; Cardenas, C.E.; Chung, C.; Fuentes, D.T.; et al. Deep Learning–Based Dose Prediction for Automated, Individualized Quality Assurance of Head and Neck Radiation Therapy Plans. *Pract. Radiat. Oncol.* **2023**, *13*, e282–e291. [CrossRef]
30. Gronberg, M.P.; Jhingran, A.; Netherton, T.J.; Gay, S.S.; Cardenas, C.E.; Chung, C.; Fuentes, D.; Fuller, C.D.; Howell, R.M.; Khan, M.; et al. Deep learning–based dose prediction to improve the plan quality of volumetric modulated arc therapy for gynecologic cancers. *Med. Phys.* **2023**, *50*, 6639–6648. [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

ML Models Built Using Clinical Parameters and Radiomic Features Extracted from ^{18}F-Choline PET/CT for the Prediction of Biochemical Recurrence after Metastasis-Directed Therapy in Patients with Oligometastatic Prostate Cancer

Luca Urso [1,2], Corrado Cittanti [1,2], Luigi Manco [3,*], Naima Ortolan [1,2], Francesca Borgia [1,2], Antonio Malorgio [4], Giovanni Scribano [5], Edoardo Mastella [3], Massimo Guidoboni [1,6], Antonio Stefanelli [4], Alessandro Turra [3] and Mirco Bartolomei [2]

[1] Department of Translational Medicine, University of Ferrara, 44121 Ferrara, Italy; luca.urso@unife.it (L.U.); ctc@unife.it (C.C.); rtlnma@unife.it (N.O.); francesca.borgia@unife.it (F.B.); massimo.guidoboni@ospfe.it (M.G.)
[2] Nuclear Medicine Unit, Onco-Hematology Department, University Hospital of Ferrara, 44124 Ferrara, Italy; m.bartolomei@ospfe.it
[3] Medical Physics Unit, University Hospital of Ferrara, 44124 Ferrara, Italy; edoardo.mastella@ospfe.it (E.M.); a.turra@ospfe.it (A.T.)
[4] U.O.C. Radiotherapy, University Hospital of Ferrara, 44124 Ferrara, Italy; a.malorgio@ospfe.it (A.M.); a.stefanelli@ospfe.it (A.S.)
[5] Department of Physics and Earth Science, University of Ferrara, 44121 Ferrara, Italy; scrgnn@unife.it
[6] U.O.C. Clinical Oncology, University Hospital of Ferrara, 44124 Ferrara, Italy
* Correspondence: luigi.manco@ausl.fe.it

Abstract: Oligometastatic patients at [^{18}F]F-Fluorocholine (^{18}F-choline) PET/CT may be treated with metastasis-directed therapy (MDT). The aim of this study was to combine radiomic parameters extracted from ^{18}F-choline PET/CT and clinical data to build machine learning (ML) models able to predict MDT efficacy. Methods: Oligorecurrent patients ($\leq$5 lesions) at ^{18}F-choline PET/CT and treated with MDT were collected. A per-patient and per-lesion analysis was performed, using 2-year biochemical recurrence (BCR) after MDT as the standard of reference. Clinical parameters and radiomic features (RFts) extracted from ^{18}F-choline PET/CT were used for training five ML Models for both CT and PET images. The performance metrics were calculated (i.e., Area Under the Curve—AUC; Classification Accuracy—CA). Results: A total of 46 metastases were selected and segmented in 29 patients. BCR after MDT occurred in 20 (69%) patients after 2 years of follow-up. In total, 73 and 33 robust RFts were selected from CT and PET datasets, respectively. PET ML Models showed better performances than CT Models for discriminating BCR after MDT, with Stochastic Gradient Descent (SGD) being the best model (AUC = 0.95; CA = 0.90). Conclusion: ML Models built using clinical parameters and CT and PET RFts extracted via ^{18}F-choline PET/CT can accurately predict BCR after MDT in oligorecurrent PCa patients. If validated externally, ML Models could improve the selection of oligorecurrent PCa patients for treatment with MDT.

Keywords: biochemical recurrence; prostate cancer; ^{18}F-choline PET; metastasis-directed therapy (MDT); radiomics; machine learning

1. Introduction

In the last few years, massive improvements have been made in the diagnostic and therapeutic management of prostate cancer (PCa). Nevertheless, biochemical recurrence (BCR) after primary treatments still occurs in nearly one-third of PCa patients treated with radical intent [1]. The early phases of PCa recurrence are often associated with oligometastatic disease, which is commonly defined as a low-volume metastatic disease with evidence

of up to five metastases [2]. The treatment of this PCa setting is still highly debated [3]. Although androgen deprivation therapy (ADT) combined with androgen receptor signaling inhibitors (ARSI) is currently suggested for use in oligometastatic PCa patients by international guidelines, there is growing literature evidence supporting metastasis-directed therapy (MDT) with stereotactic body radiation therapy (SBRT) [4–6]. MDT may ensure the local control of treated lesions, hindering metastasis-to-metastasis spread and delaying the beginning of systemic therapies [2,7].

The availability of reliable imaging plays a crucial role in guiding accurate MDT. In this respect, the increasing use of prostate-dedicated Positron Emission Tomography (PET) has highly revolutionized the diagnostic management of PCa patients. [^{18}F]F-Fluorocholine (^{18}F-choline) and [^{68}Ga]Ga-prostate-specific membrane antigen (68Ga-PSMA) PET/Computed Tomography (CT) enable the highly accurate detection of PCa metastases, both in patients with BR and low PSA values, allowing for accurate MDT [8–10]. However, the earlier detection of metastases allowed by using PET imaging may determine a stage migration that does not necessarily impacts patients' prognoses, configuring a "Will-Rogers phenomenon", whose relevance still needs to be clarified via well-conducted trials [11]. Moreover, PET-derived parameters predictive of successful MDT are still scarcely described in the literature.

Among the cutting-edge applications of medical imaging, radiomics is surely worthy of mention. Indeed, medical images conceal information undetectable to the human eye. In this scenario, radiomics analysis aims to extract a high amount of quantitative information hidden within medical multimodal images and correlate them with clinical and/or pathological parameters [12]. In a modern concept of personalized medicine, radiomic analysis represents a useful "free" source of information that can be used for the building of predictive models, with potentially relevant diagnostic and therapeutic implications. The need to handle large amounts of data in order to build prediction models has led to the spread of artificial intelligence (AI) algorithms. Machine learning (ML) is a branch of AI in which, based on the training dataset that is first provided, the algorithm develops its own logic for decision-making or regression problems [13].

In the field of nuclear medicine, ML Models developed using PET/CT data have already demonstrated numerous applications in multiple anatomical districts and for different purposes (i.e., for the diagnosis, staging or prediction of response to treatment) [14–17].

The aim of this study was to combine clinical data with conventional and radiomic parameters extracted from ^{18}F-choline PET/CT to build ML Models able to predict the efficacy of MDT in oligorecurrent PCa patients.

2. Materials and Methods

2.1. Study Population and Data Collection

This is a retrospective monocenter study involving patients treated with MDT guided by ^{18}F-choline PET/CT between 2017 and 2020 at the Sant'Anna University Hospital of Ferrara (Italy).

The inclusion criteria of this study were as follows: (a) a histologically confirmed diagnosis of PCa; (b) an age > 18 years; (c) oligorecurrent disease (up to 5 lesions), including lymph node or distant metastases at ^{18}F-choline PET-CT; (d) MDT guided by ^{18}F-choline PET-CT and delivered through SBRT within 3 months of the PET/CT scan; (e) 2 years of follow-up available after MDT. The exclusion criteria of this study were as follows: (a) s lack of clinical and/or follow-up data; (b) systemic therapy (including ADT or chemotherapy) ongoing at the administration of MDT.

Extracted data were analyzed both per-patient and per-lesion. The main clinical and laboratory data collected included age, International Society of Urological Pathology (ISUP) grade group, and prostate-specific antigen (PSA) serum levels, both before and after MDT. Patients were divided into a high-grade group (ISUP 4,5) and a low-grade group (ISUP 1–3).

This study was performed according to the Declaration of Helsinki, Good Clinical Practice, and local ethical regulations. The local ethical committee approved this study

(CE-AVEC–registration number 2/2024/Oss/AOUFe), and patients' written informed consent was collected.

2.2. MDT and Follow-Up

All patients underwent SBRT to treat all metastatic lesions. Computed Tomography (CT) scans were acquired according to our clinical protocol in supine position with 2 mm slice thickness. According to the morphological and functional imaging information, the gross tumor volume (GTV) was segmented. SBRT was delivered in 3–5 fractions using a Versa HD linear accelerator (Elekta AB, Stockholm, Sweden) equipped with a standard Agility multileaf collimator. Volumetric modulated arc therapy (VMAT) was planned with multiple flattening filter free (FFF) beams, generally two. More details about treatment planning are described in previously published papers [18]. Daily in-room 3D cone beam (CB) CT was required for pre-treatment patient position verification.

Serial PSA dosages were performed every 3 months following MDT to check its efficacy. Biochemical recurrence (BCR) was considered the main outcome of MDT, defined as the increase in serum PSA values higher than 2 ng/mL with respect to the lower PSA value reached following MDT. None of the patients included showed evidence of progression at medical imaging earlier than BCR. Patients were dichotomized as BCR0 or BCR1 according to the absence or evidence of BCR, respectively.

2.3. ^{18}F-Choline PET/CT Acquisition Protocol and Analysis

The ^{18}F-choline PET/CT images were acquired from the mid-thigh to the skull vertex 60 ± 10 min after the administration of 3 MBq/kg of ^{18}F-choline, using a standard technique on a dedicated 3D PET/CT tomograph (Biograph mCT Flow; Siemens Medical Solutions, Malvern, PA, USA). All the patient's images were acquired and processed with the same protocol on the same PET/CT tomograph, allowing us to avoid inter-scanner heterogeneity. Following non-contrast-enhanced low-dose CT (120 keV, 80 mAs, CareDose; reconstructed with a soft-tissue kernel to a slice thickness of 3 mm), emissive PET imaging was acquired in 3D mode (matrix, 200×200) using a continuous scan with variable speed according to the body district. Raw data were corrected for randoms, scatter, and attenuation using the co-registered non-diagnostic CT.

The ^{18}F-choline PET/CT images were critically analyzed by two experienced nuclear medicine physicians (L.U and C.C.) on a syngo.via workstation (Siemens Healthineers, Enlargen, Germany) as showed in Figure 1. Pathological findings were considered to be area(s) of focal increased radiotracer uptake outside the sites of physiological distribution. Any discrepancies were resolved by consensus. Circular manual regions of interest (ROIs) were drawn around the pathological findings and automatically converted in 3D volumes of interest (VOIs) by the software.

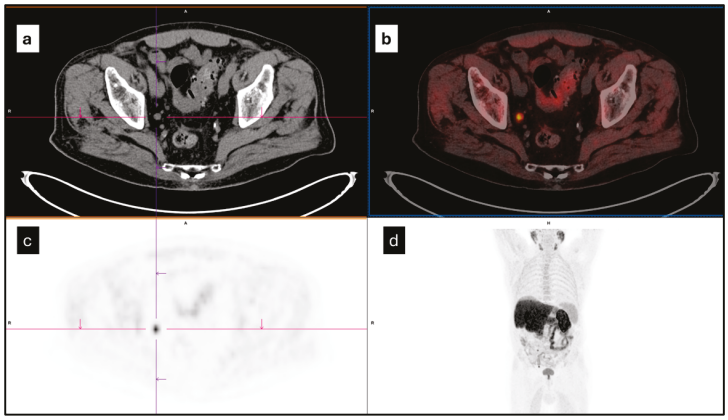


Figure 1. Axial images of CT (**a**), PET/CT (**b**), PET (**c**), and MIP (maximum intensity projection) (**d**) of a 81-year-old man created by ^{18}F-choline PET/CT. A focal ^{18}F-choline uptake (SUVmax = 15.6) was identified in the right obturator node.

Metabolic and volumetric parameters extracted from ^{18}F-choline PET/CT comprehended maximum and mean standardized uptake value (SUVmax and SUVmean, respectively); metabolic tumor volume (MTV), representing the tumor volume with at least 40% uptake of the SUVmax within the VOI; and total lesion choline kinase activity (TLCKa), calculated by multiplying SUVmean and MTV within the same VOI.

2.4. Image Segmentation and Feature Extraction

VOI segmentations of PCa metastasis were manually performed on ^{18}F-choline PET/CT images by two expert nuclear medicine physicians using MIM maestro version 7.3.2 (MIM Software, Inc., Cleveland, OH, USA) as reported in Figure 2. Any discrepancies were resolved by consensus.

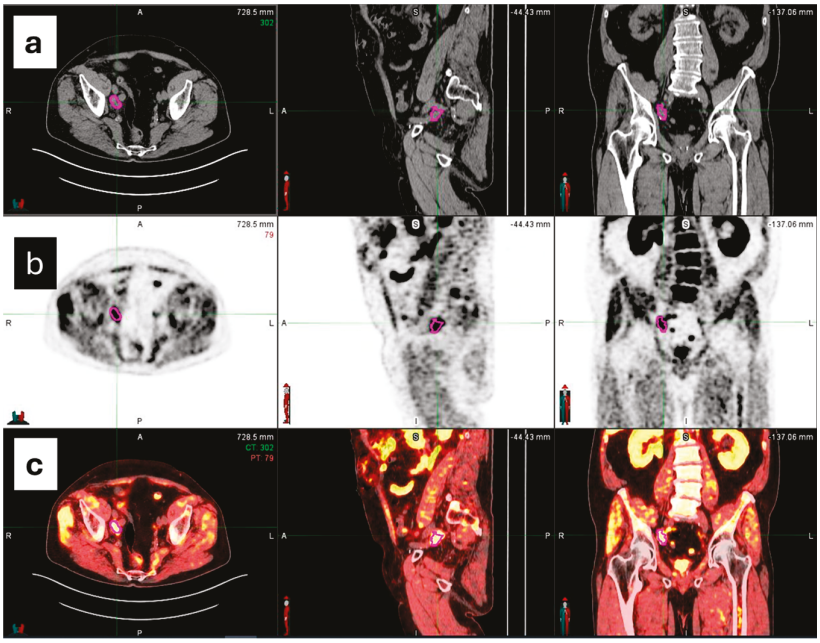


Figure 2. The manual segmentation of volumetric regions of interest (VOIs) on PET/CT images (**c**) around the area of focal ^{18}F-choline uptake using MiM Maestro software. Each contour (pink line) was superimposed on CT (**a**) and PET (**b**) images.

Quantitative radiomics features were extracted from each VOI on both PET and CT images separately using the Radiomics package and 3D Slicer image-computing platform according to IBSI standardization [19].

For each VOI, 121 radiomic features (RFts) were extracted from the original images. Among those, which were, respectively, divided by classes, 14 RFts belong to the original image and mask, 14 to the shape (3D) class, 18 to first-order intensity statistics, 24 to the gray-level co-occurrence matrix (GLCM), 16 to the gray-level run length matrix (GLRLM), 16 to gray-level size zone (GLSZM), 14 to the gray-level dependence matrix (GLDM), and 5 to the neighboring gray-tone difference matrix (NGTDM). In addition, 744 textural RFts were extracted from wavelet-decomposed VOIs.

2.5. Statistical Analysis and Model Building

The correlation between clinical parameters, ^{18}F-choline metabolic and volumetric PET/CT parameters, MDT doses, and the BCR group was investigated both per-patient and per-lesion, using the two-tailed Mann–Whitney U-type test.

A filter features selection step, performed using a hand-crafted algorithm in the Python language, was used to identify robust CT and PET RFts based on the Mann–Whitney U-type test and the Spearman rank correlation coefficient.

A p-value < 0.05 and Rs coefficient > 0.8 were considered significant to select the most robust and non-redundant RFts. This analysis was independently performed for both CT-based and PET-based RFt datasets. The specific results of the feature selection step are reported in the Supplementary Materials.

Selected robust RFts and clinical parameters were used to build multiple PET and CT ML Models. The whole population was split into training-test (70%) and validation (30%) datasets. The Synthetic Minority Over-Sampling Technique (SMOTE) was used to balance the sample distribution of BCR0 and BCR1 across the 2 sets, obtaining a sample size of 70 lesions, with 35 BCR0 and 35 BCR1.

Orange Data Mining, an open-source toolkit, was used to train, test, and validate five different ML algorithms, both for PET and CT: Decision Tree (Tree), Random Forest (RF), Gradient Boosting (GB), Stochastic Gradient Descent (SGD) and Support Vector Machines (SVMs). All the learners were tested via 10-Fold Cross-Validation using 70% of the main dataset. The validation step was conducted on the remaining 30% of the set-aside dataset. Performance models were evaluated in terms of AUC, Classification Accuracy (CA), Precision and Sensitivity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV).

3. Results

3.1. Population Analysis

Overall, 29 oligorecurrent PCa patients with a mean age of 71 ± 7 years previously treated with radical prostatectomy (86%) or radiation therapy (RT) with radical intent alone (7%) or associated with 6–12 months of adjuvant ADT (7%) were retrieved. Evidence of oligorecurrent disease occurred after a mean period of 5 (1–8) years from primary treatment.

Overall, 46 metastases were detected through ^{18}F-choline PET/CT (36 lymph nodes, 10 bone metastases), corresponding to 1.6 lesion per patient. None of the patients investigated had evidence of visceral metastases at ^{18}F-choline PET/CT.

Evidence of BCR (BCR1) was considered the main outcome after MDT. In 2-year follow-up period, BCR occurred in 20 (69%) patients after a mean period of 12.8 ± 13.2. In a per-lesion analysis, 35 (76.1%) lesions belonged to the BCR1 group. Detailed characteristics of the patients included are reported in Table 1.

Table 1. Population characteristics.

Patients Characteristics	BCR0	BCR1	Total
Number of patients	9	20	29
AGE [years]			
Mean ± St.dev.	71 ± 6	71 ± 7	71 ± 7
Range	61–81	60–81	60–81
PSA [ng/mL]			
Mean ± St.dev.	2.63 ± 3.90	2.17 ± 2.68	2.31 ± 3.05
Range	0.01–12.06	0.24–9.00	0.01–12.06
ISUP grade			
1	0 (0%)	4 (20.0%)	4 (13.8%)
2	2 (22.2%)	5 (25.0%)	7 (24.1%)
3	3 (33.3%)	3 (15.0%)	6 (20.7%)
4	3 (33.3%)	4 (20.0%)	7 (24.2%)
5	0 (0%)	2 (10.0%)	2 (6.9%)
nd	1 (11.2%)	2 (10.0%)	3 (10.3%)
N. lesions per patient			
1	7 (78.8%)	8 (40.0%)	17 (51.7%)
2	2 (22.2%)	10 (50.0%)	10 (41.3%)
3	0%	1 (5%)	1 (3.5%)
4	0%	1 (5%)	1 (3.5%)
5	0%	0%	0%

PSA: prostate-specific antigen; ISUP: International Society of Urological Pathology.

3.2. ^{18}F-Choline PET/CT Parameters and PSA Analysis

All clinical parameters (per-patient analysis), metabolic and volumetric parameters (per-lesion analysis) derived from ^{18}F-choline PET/CT, and MDT doses were unable to distinguish between BCR0 and BCR1 groups (Tables 2 and 3).

Table 2. Per-patient analysis of clinical parameters.

Per-Patient Analysis	BCR0	BCR1	Total	*p*	rs
Number of patients	9	20	29		
AGE [years]				0.94	0.34
Mean ± St.dev.	71 ± 6	71 ± 7	71 ± 7		
Range	61–81	60–81	60–81		
PSA [ng/mL]				0.85	0.38
Mean ± St.dev.	2.63 ± 3.90	2.17 ± 2.68	2.31 ± 3.05		
Range	0.01–12.06	0.24–9.00	0.01–12.06		
ISUP grade				0.22	
1–3	5 (55.6%)	12 (60.0%)	17 (58.6%)		
4–5	4 (44.4%)	8 (40.0%)	12 (41.4%)		

PSA: prostate-specific antigen; ISUP: International Society of Urological Pathology.

Table 3. Per-lesion analysis of metabolic and volumetric ^{18}F-choline PET/CT and MDT parameters.

Per-Lesion Analysis	BCR0	BCR1	Total	*p*	r_s
Number of lesions	11 (23.9%)	35 (76.1%)	46		
SUVmax				0.09	0.09
Mean ± St.dev.	14.42 ± 10.79	9.31 ± 7.31	10.53 ± 8.43		
Range	3.6–40.4	1.7–30.0	1.7–40.4		
MTV				0.10	0.18
Mean ± St.dev.	0.95 ± 0.87	1.27 ± 1.02	1.19 ± 0.99		
Range	0.2–2.7	0.5–3.8	0.2–3.8		
TLCKA				0.54	0.53
Mean ± St.dev.	8.28 ± 7.94	8.82 ± 11.75	8.63 ± 10.89		
Range	1.7–24.3	0.2–46.4	0.2–46.4		
MDT Dose per fraction [Gy]				0.20	0.31
Mean ± St.dev.	30.91 ± 3.16	32.93 ± 3.67	32.42 ± 3.60		
Range	24–36	27–38	24–38		
Number of fractions (occurrence)				0.33	0.28
	3 (73%) 4(27%)	3 (77%) 4 (17%) 5(6%)	3 (76%) 4 (19%) 5 (5%)		
BED [Gy]				0.12	0.41
Mean ± St.dev.	126.56 ± 24.15	147.95 ± 30.65	140.95 ± 30.09		
Range	88–180	108–180	88–180		

MDT: metastasis-directed therapy; BED: biological equivalent dose; SUV standardized uptake value; MTV: metabolic tumor volume; TLCKA: total lesion choline kinase activity.

3.3. Radiomic Analysis and Model Building

Among the 854 RFts extracted from original and filtered images, significant statistical differences between BCR0 and BCR1 groups were found for a total of 73 and 33 robust RFts for the CT and PET image datasets, respectively. The selected features belong to the first-(4) and second-order GLCMs (3) and textural wavelet-based (66) class for CT. All 33 robust PET RFts belong to the textural wavelet-based class.

The 10-Fold CV in CT Model showed an AUC > 0.80 and a CA > 0.80. Similarly, in the PET Model, an AUC > 0.95 and a CA > 0.90 were obtained.

The classification performances of the validation step are listed in Table 4. The CT Models showed good performances (AUC, CA, precision, sensitivity, PPV, and NPV were all above 0.70) using the GB and SGD algorithms, although the best performances were obtained via RF (AUC = 0.92; CA = 0.85; sensitivity = 0.85, PPV = 0.91; NPV = 0.74).

Table 4. Classification performances of CT and PET ML Models. Values greater than 0.7 are considered acceptable.

CT Model	AUC	CA	Precision	Sensitivity	PPV	NPV
GB	0.87	0.80	0.82	0.84	0.85	0.83
RF	0.92	0.85	0.88	0.85	0.91	0.74
SGD	0.71	0.75	0.75	0.75	0.89	0.75
SVMs	0.44	0.45	0.44	0.45	0.34	0.67
Tree	0.65	0.65	0.79	0.65	0.34	0.91
PET Model						
GB	0.53	0.57	0.50	0.50	0.48	0.64
RF	0.79	0.70	0.70	0.70	0.71	0.73
SGD	0.95	0.90	0.90	0.90	0.94	0.93
SVMs	0.90	0.75	0.77	0.75	0.91	0.63
Tree	0.62	0.65	0.67	0.65	0.54	0.81

GB: Gradient Boosting; RF: Random Forest; SGD: Stochastic Gradient Descent; SVMs: Support Vector Machines; Tree: Decision Tree.

For the PET Model, SGD obtained the best performance (AUC = 0.95; CA = 0.90; sensitivity = 0.90; PPV = 0.94; NPV= 0.93), followed by the SVM model (AUC = 0.90; CA = 0.75; sensitivity = 0.75; PPV= 0.91; NPV = 0.63). The ROC curves of the best ML Models are graphed in Figure 3.

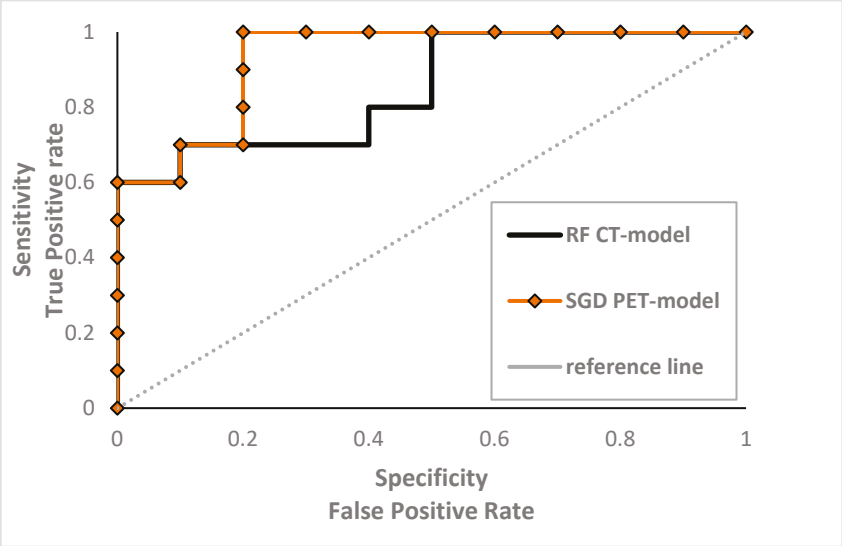


Figure 3. Comparison of the ROC curves for Random Forest CT Model and Stochastic Gradient Descent PET Model. RF: Random Forest; SGD: Stochastic Gradient Descent.

Predictive models trained on radiomic patterns relative to the same kinds of metastases (i.e., bone or lymph nodes) were evaluated. In the case of bone-lesions, the sample size (10) and unbalanced population (1 BCR0 vs. 9 BCR1) did not allow adequate model training in either PET or CT.

Considering lymph node metastases, 31 robust RFts were selected (3 belonging to the first-order GLCM, 3 belonging to the second-order GLCM, 25 belonging to the textural wavelet-based class) and 18 RFts were selected (wavelet class) for the CT and PET datasets, respectively.

The same five algorithms used for the whole cohort (GB, RF, SGD, SVM, Tree) were also used for the training of the PET and CT Models relative to lymph node-lesions. In 10-Fold CV, AUC < 0.75 (0.48–0.75) and AUC < 0.85 (0.40–0.85) were obtained for the PET and CT Models, respectively. Moreover, all the CAs were less than 0.70. The validation step in the CT Models achieved unsatisfactory performances (AUC 0.34–0.71 and CA < 0.70). In the PET Model, the promising results in the training step were not matched by equally good performances in the validation step (AUC 0.30–0.58 and CA < 0.68).

4. Discussion

In the last few years, MDT has progressively gained relevance in the treatment of oligorecurrent PCa. Indeed, evidence in support of MDT is accumulating in the literature, with papers reporting that this therapeutic strategy enables researchers to postpone systemic drug administration and improves PCa patients' outcomes [20–23]. However, patient selection and the timing of MDT remain challenging and unsolved issues [22]. Accurate imaging modalities able to detect as many sites of recurrence as possible play a crucial role in guiding effective MDT [24]. In PCa, PET imaging is currently the most accurate imaging modality for biochemical recurrence detection [25]. Although PSMA-ligand PET currently represents the gold-standard method of examination, in particular for patients with low PSA recurrence, its introduction into daily clinical practice is relatively recent, and long-term outcome data are barely available. Therefore, the analysis of data related to other PET imaging modalities previously used for guiding MDT, including ^{18}F-choline PET/CT, still represents an inestimable opportunity to understand the role of imaging in guiding MDT. In particular, we still need to identify predictive factors that can be collected from baseline imaging and concur with the prediction of the long-term outcomes of MDT in oligorecurrent PCa patients.

In our sample of oligorecurrent PCa patients treated with MDT guided via ^{18}F-choline PET/CT, 2-year BCR occurred in 69% of patients (corresponding to a 2-year PFS—progression free survival—rate of 31%) after a mean period of 12.8 months from radiation therapy initiation. Our data are slightly below those reported in the literature, with 2-year PFS ranging between 35 and 44% [26,27]. In our cohort, both clinical and conventional ^{18}F-choline PET/CT parameters (including SUVmax, SUVmean, MTV, and TCKa) failed to predict BCR after MDT. Our finding is consistent with the results previously published by Cysouw et al. [28], who reported that ^{18}F-choline PET/CT SUV parameters and TLCKa are not significantly predictive of PFS in oligometastatic PCa.

Therefore, our attention shifted towards radiomics and AI. To the best of our knowledge, this is the first study to report a radiomic analysis of ^{18}F-choline PET/CT in oligorecurrent PCa patients treated with MDT. A high number of robust RFts able to discriminate BCR0 from BCR1 groups were extracted from PET (n.33) and CT images (n.73). This finding highlights the added value of radiomic analysis in comparison to standard parameters currently applied in the daily clinical practice of PCa imaging [29]. Interestingly, all the robust RFts extracted from the PET dataset belong to the textural wavelet-based class (i.e., calculated for the filtered image/VOI). This finding reflects the fact that high- and low-frequency information extracted from PET images provides powerful informative data for use in a predictive model.

Robust RFts extracted from PET and CT images were used for training five ML Models for each dataset. We applied 10-fold CV to increase the amount of data available for training the models. Interestingly, the SGD PET Model was very accurate regarding the discrimination of the two groups, achieving very high performances (AUC = 0.95; CA = 0.90; sensitivity = 0.90; PPV = 0.94; NPV= 0.93), superior to those obtained with other PET and CT Models. The second best model was the RF CT Model (AUC = 0.92; CA = 0.85; sensitivity = 0.85, PPV = 0.91; NPV = 0.74). The validation of these two ML Models in external datasets with large sample sizes is desirable. If our data are confirmed, the potential inclusion of these ML Models in daily clinical practice could represent a relevant step towards a modern concept of personalized medicine. Indeed, an accurate

tool that can predict patients unlikely to experience long-term benefits from MDT might suggest closer follow-up monitoring or even a direct switch to systemic therapies.

SGD was the best PET Model and showed better performances in comparison to RF, which was the best CT Model. We can speculate that the PET Model was more accurate because we analyzed a heterogeneous group of metastases (36 lymph nodes and 10 bone metastases). Indeed, ^{18}F-choline PET imaging reflects the metabolic activity of PCa cells within a lesion, being an in vivo representation of choline metabolism [30]. Conversely, CT is a morphologic imaging method, and a metastatic lymph node is surely very different from a bone metastasis in terms of radiologic appearance. Therefore, we tried to split our population according to the site of each metastasis, analyzing separately the groups of lymph node and bone metastases. As expected, models built on bone metastases showed unsatisfactory results due to the too-small number of lesions collected and the unbalanced population (only 10% of BCR0). Considering the lymph node metastasis cohort, the results obtained using both the CT and PET Models could not pair with those of the whole cohort. The reasons are related to the low number of robust features able to discriminate between the two classes. In addition, the low number of instances available, compared to the whole cohort case, meant that we built models that showed worse performances during validation due to an overfitting problem.

The current study has some limitations. Firstly, the retrospective nature of this study and the small sample size should be mentioned as the main limitations. Finally, the patients analyzed received MDT guided by ^{18}F-choline and not PSMA-ligand PET/CT because they date from a period preceding the adoption of PSMA PET in our institute. This work may represent a model that can be replicated in a future prospective study using PSMA-ligand PET/CT, which is currently indicated as the imaging method of choice for guiding MDT in oligorecurrent prostate cancer patients [31].

5. Conclusions

Our results highlight that radiomic analysis and ML Models can be useful for the prediction of the outcomes of MDT in oligorecurrent PCa patients. Further efforts should be made for collecting larger sample size and for an external validation of our preliminary results. Furthermore, future studies should also explore radiomic analysis and ML Models in oligorecurrent PCa patients treated with MDT guided via PSMA-ligand PET/CT, which is currently the imaging method of choice for this clinical indication.

Supplementary Materials: The following supporting information can be downloaded via this link: https://www.mdpi.com/article/10.3390/diagnostics14121264/s1, Table S1: Robust features in CT and PET datasets.

Author Contributions: Conceptualization, L.U., L.M., M.B., and M.G.; methodology, N.O., F.B., G.S., E.M., and A.M.; software and formal analysis, G.S., L.M., A.T., and M.G.; validation, G.S., C.C., and A.M.; investigation, all; resources, all; data curation, all; writing—original draft preparation, all; writing—review and editing, A.S., A.T., and M.B. 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 Ethics Committee of University Hospital of Ferrara (2/2024/Oss/AOUFe of 09/02/2024).

Informed Consent Statement: Informed consent was obtained from all subjects involved in this study.

Data Availability Statement: Data are contained within the article and Supplementary Materials.

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

References

1. Shore, N.D.; Moul, J.W.; Pienta, K.J.; Czernin, J.; King, M.T.; Freedland, S.J. Biochemical Recurrence in Patients with Prostate Cancer after Primary Definitive Therapy: Treatment Based on Risk Stratification. *Prostate Cancer Prostatic Dis.* **2023**. [CrossRef] [PubMed]
2. Miszczyk, M.; Rajwa, P.; Yanagisawa, T.; Nowicka, Z.; Shim, S.R.; Laukhtina, E.; Kawada, T.; von Deimling, M.; Pradere, B.; Rivas, J.G.; et al. The Efficacy and Safety of Metastasis-Directed Therapy in Patients with Prostate Cancer: A Systematic Review and Meta-Analysis of Prospective Studies. *Eur. Urol.* **2024**, *85*, 125–138. [CrossRef] [PubMed]
3. Shore, N.D.; Morgans, A.K.; El-Haddad, G.; Srinivas, S.; Abramowitz, M. Addressing Challenges and Controversies in the Management of Prostate Cancer with Multidisciplinary Teams. *Target Oncol.* **2022**, *17*, 709. [CrossRef] [PubMed]
4. Ost, P.; Reynders, D.; Decaestecker, K.; Fonteyne, V.; Lumen, N.; De Bruycker, A.; Lambert, B.; Delrue, L.; Bultijnck, R.; Claeys, T.; et al. Surveillance or Metastasis-Directed Therapy for Oligometastatic Prostate Cancer Recurrence: A Prospective, Randomized, Multicenter Phase II Trial. *J. Clin. Oncol.* **2017**, *36*, 446–453. [CrossRef] [PubMed]
5. Phillips, R.; Shi, W.Y.; Deek, M.; Radwan, N.; Lim, S.J.; Antonarakis, E.S.; Rowe, S.P.; Ross, A.E.; Gorin, M.A.; Deville, C.; et al. Outcomes of Observation vs Stereotactic Ablative Radiation for Oligometastatic Prostate Cancer: The ORIOLE Phase 2 Randomized Clinical Trial. *JAMA Oncol.* **2020**, *6*, 650–659. [CrossRef] [PubMed]
6. Urological Infections EAU Guidelines On 2023. EAU Guidelines Office: Arnhem, The Netherlands. Available online: http://uroweb.org/guidelines/compilations-of-all-guidelines/ (accessed on 15 May 2024).
7. Giannarini, G.; Fossati, N.; Gandaglia, G.; Cucchiara, V.; Ficarra, V.; Mirone, V.; Montorsi, F.; Briganti, A. Will Image-Guided Metastasis-Directed Therapy Change the Treatment Paradigm of Oligorecurrent Prostate Cancer? *Eur. Urol.* **2018**, *74*, 131–133. [CrossRef] [PubMed]
8. Eiber, M.; Maurer, T.; Souvatzoglou, M.; Beer, A.J.; Ruffani, A.; Haller, B.; Graner, F.P.; Kübler, H.; Haberhorn, U.; Eisenhut, M.; et al. Evaluation of Hybrid 68Ga-PSMA Ligand PET/CT in 248 Patients with Biochemical Recurrence after Radical Prostatectomy. *J. Nucl. Med.* **2015**, *56*, 668–674. [CrossRef] [PubMed]
9. Morigi, J.J.; Stricker, P.D.; Van Leeuwen, P.J.; Tang, R.; Ho, B.; Nguyen, Q.; Hruby, G.; Fogarty, G.; Jagavkar, R.; Kneebone, A.; et al. Prospective Comparison of 18F-Fluoromethylcholine Versus 68Ga-PSMA PET/CT in Prostate Cancer Patients Who Have Rising PSA after Curative Treatment and Are Being Considered for Targeted Therapy. *J. Nucl. Med.* **2015**, *56*, 1185–1190. [CrossRef] [PubMed]
10. Filippi, L.; Urso, L.; Bianconi, F.; Palumbo, B.; Marzola, M.C.; Evangelista, L.; Schillaci, O. Radiomics and Theranostics with Molecular and Metabolic Probes in Prostate Cancer: Toward a Personalized Approach. *Expert Rev. Mol. Diagn.* **2023**, *23*, 243–255. [CrossRef]
11. Urso, L.; Filippi, L.; Castello, A.; Marzola, M.C.; Bartolomei, M.; Cittanti, C.; Florimonte, L.; Castellani, M.; Zucali, P.; Bruni, A.; et al. PSMA PET/CT in Castration-Resistant Prostate Cancer: Myth or Reality? *J. Clin. Med.* **2023**, *12*, 7130. [CrossRef]
12. Lambin, P.; Leijenaar, R.T.H.; Deist, T.M.; Peerlings, J.; De Jong, E.E.C.; Van Timmeren, J.; Sanduleanu, S.; Larue, R.T.H.M.; Even, A.J.G.; Jochems, A.; et al. Radiomics: The Bridge between Medical Imaging and Personalized Medicine. *Nat. Rev. Clin. Oncol.* **2017**, *14*, 749–762. [CrossRef] [PubMed]
13. Manco, L.; Maffei, N.; Strolin, S.; Vichi, S.; Bottazzi, L.; Strigari, L. Basic of Machine Learning and Deep Learning in Imaging for Medical Physicists. *Phys. Medica* **2021**, *83*, 194–205. [CrossRef] [PubMed]
14. Nieri, A.; Manco, L.; Bauckneht, M.; Urso, L.; Caracciolo, M.; Donegani, M.I.; Borgia, F.; Vega, K.; Colella, A.; Ippolito, C.; et al. [18F]FDG PET-TC Radiomics and Machine Learning in the Evaluation of Prostate Incidental Uptake. *Expert Rev. Med. Devices* **2023**, *20*, 1183–1191. [CrossRef] [PubMed]
15. Evangelista, L.; Fiz, F.; Laudicella, R.; Bianconi, F.; Castello, A.; Guglielmo, P.; Liberini, V.; Manco, L.; Frantellizzi, V.; Giordano, A.; et al. PET Radiomics and Response to Immunotherapy in Lung Cancer: A Systematic Review of the Literature. *Cancers* **2023**, *15*, 3258. [CrossRef] [PubMed]
16. Guglielmo, P.; Marturano, F.; Bettinelli, A.; Gregianin, M.; Paiusco, M.; Evangelista, L. Additional Value of PET Radiomic Features for the Initial Staging of Prostate Cancer: A Systematic Review from the Literature. *Cancers* **2021**, *13*, 6026. [CrossRef] [PubMed]
17. Alongi, P.; Laudicella, R.; Stefano, A.; Caobelli, F.; Comelli, A.; Vento, A.; Sardina, D.; Ganduscio, G.; Toia, P.; Ceci, F.; et al. Choline PET/CT Features to Predict Survival Outcome in High-Risk Prostate Cancer Restaging: A Preliminary Machine-Learning Radiomics Study. *Q. J. Nucl. Med. Mol. Imaging Off. Publ. Ital. Assoc. Nucl. Med. (AIMN) Int. Assoc. Radiopharmacol. (IAR) Sect. Soc.* **2022**, *66*, 352–360. [CrossRef] [PubMed]
18. Mastella, E.; Epile, J.E.; De Guglielmo, E.; Fabbri, S.; Calderoni, F.; Manco, L.; Szilagyi, K.E.; Malorgio, A.; Turra, A.; Stefanelli, A. Stereotactic Body Radiation Therapy (SBRT) for Prostate Cancer: Improving Treatment Delivery Efficiency and Accuracy. *Tech. Innov. Patient Support Radiat. Oncol.* **2024**, *30*, 100253. [CrossRef]
19. Fedorov, A.; Beichel, R.; Kalpathy-Cramer, J.; Finet, J.; Fillion-Robin, J.C.; Pujol, S.; Bauer, C.; Jennings, D.; Fennessy, F.; Sonka, M.; et al. 3D Slicer as an Image Computing Platform for the Quantitative Imaging Network. *Magn. Reson. Imaging* **2012**, *30*, 1323–1341. [CrossRef] [PubMed]
20. Triggiani, L.; Alongi, F.; Buglione, M.; Detti, B.; Santoni, R.; Bruni, A.; Maranzano, E.; Lohr, F.; D'Angelillo, R.; Magli, A.; et al. Efficacy of Stereotactic Body Radiotherapy in Oligorecurrent and in Oligoprogressive Prostate Cancer: New Evidence from a Multicentric Study. *Br. J. Cancer* **2017**, *116*, 1520–1525. [CrossRef]

21. Decaestecker, K.; De Meerleer, G.; Ameye, F.; Fonteyne, V.; Lambert, B.; Joniau, S.; Delrue, L.; Billiet, I.; Duthoy, W.; Junius, S.; et al. Surveillance or Metastasis-Directed Therapy for OligoMetastatic Prostate Cancer Recurrence (STOMP): Study Protocol for a Randomized Phase II Trial. *BMC Cancer* **2014**, *14*, 671. [CrossRef]
22. Triggiani, L.; Mazzola, R.; Tomasini, D.; Bruni, A.; Alicino, G.; Matrone, F.; Bortolus, R.; Francolini, G.; Detti, B.; Magli, A.; et al. Upfront Metastasis-Directed Therapy in Oligorecurrent Prostate Cancer Does Not Decrease the Time from Initiation of Androgen Deprivation Therapy to Castration Resistance. *Med. Oncol.* **2021**, *38*, 72. [CrossRef] [PubMed]
23. Connor, M.J.; Smith, A.; Miah, S.; Shah, T.T.; Winkler, M.; Khoo, V.; Ahmed, H.U. Targeting Oligometastasis with Stereotactic Ablative Radiation Therapy or Surgery in Metastatic Hormone-Sensitive Prostate Cancer: A Systematic Review of Prospective Clinical Trials. *Eur. Urol. Oncol.* **2020**, *3*, 582–593. [CrossRef] [PubMed]
24. De Kouchkovsky, I.; Aggarwal, R.; Hope, T.A. Prostate-Specific Membrane Antigen (PSMA)-Based Imaging in Localized and Advanced Prostate Cancer: A Narrative Review. *Transl. Androl. Urol.* **2021**, *10*, 3130. [CrossRef] [PubMed]
25. Oprea-Lager, D.E.; MacLennan, S.; Bjartell, A.; Briganti, A.; Burger, I.A.; de Jong, I.; De Santis, M.; Eberlein, U.; Emmett, L.; Fizazi, K.; et al. European Association of Nuclear Medicine Focus 5: Consensus on Molecular Imaging and Theranostics in Prostate Cancer. *Eur. Urol.* **2024**, *85*, 49–60. [CrossRef] [PubMed]
26. Ponti, E.; Ingrosso, G.; Carosi, A.; Di Murro, L.; Lancia, A.; Pietrasanta, F.; Santoni, R. Salvage Stereotactic Body Radiotherapy for Patients with Prostate Cancer with Isolated Lymph Node Metastasis: A Single-Center Experience. *Clin. Genitourin. Cancer* **2015**, *13*, e279–e284. [CrossRef] [PubMed]
27. Franzese, C.; Lopci, E.; Di Brina, L.; D'Agostino, G.R.; Navarria, P.; Mancosu, P.; Tomatis, S.; Chiti, A.; Scorsetti, M. 11C-Choline-Pet Guided Stereotactic Body Radiation Therapy for Lymph Node Metastases in Oligometastatic Prostate Cancer. *Cancer Investig.* **2017**, *35*, 586–593. [CrossRef] [PubMed]
28. Cysouw, M.; Bouman-Wammes, E.; Hoekstra, O.; van den Eertwegh, A.; Piet, M.; van Moorselaar, J.; Boellaard, R.; Dahele, M.; Oprea-Lager, D. Prognostic Value of [18 F]-Fluoromethylcholine Positron Emission Tomography/Computed Tomography Before Stereotactic Body Radiation Therapy for Oligometastatic Prostate Cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **2018**, *101*, 406–410. [CrossRef] [PubMed]
29. Ghezzo, S.; Bezzi, C.; Presotto, L.; Mapelli, P.; Bettinardi, V.; Savi, A.; Neri, I.; Preza, E.; Samanes Gajate, A.M.; De Cobelli, F.; et al. State of the Art of Radiomic Analysis in the Clinical Management of Prostate Cancer: A Systematic Review. *Crit. Rev. Oncol. Hematol.* **2022**, *169*, 103544. [CrossRef] [PubMed]
30. Urso, L.; Lancia, F.; Ortolan, N.; Frapoli, M.; Rauso, M.; Artioli, P.; Cittanti, C.; Uccelli, L.; Frassoldati, A.; Evangelista, L.; et al. 18F-Choline PET/CT or PET/MR and the Evaluation of Response to Systemic Therapy in Prostate Cancer: Are We Ready? *Clin. Transl. Imaging* **2022**, *10*, 687–695. [CrossRef]
31. Zilli, T.; Achard, V.; Dal Pra, A.; Schmidt-Hegemann, N.; Jereczek-Fossa, B.A.; Lancia, A.; Ingrosso, G.; Alongi, F.; Aluwini, S.; Arcangeli, S.; et al. Recommendations for Radiation Therapy in Oligometastatic Prostate Cancer: An ESTRO-ACROP Delphi Consensus. *Radiother. Oncol.* **2022**, *176*, 199–207. [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

Opportunistic Screening for Acute Vertebral Fractures on a Routine Abdominal or Chest Computed Tomography Scans Using an Automated Deep Learning Model

Ye Rin Kim [1], Yu Sung Yoon [2,*] and Jang Gyu Cha [1]

[1] Department of Radiology, College of Medicine, Soonchunhyang University Bucheon Hospital, Soonchunhyang University, Bucheon 14584, Republic of Korea
[2] Department of Radiology, School of Medicine, Kyungpook National University Hospital, Kyungpook National University, Daegu 41944, Republic of Korea
* Correspondence: zick1853@gmail.com; Tel.: +82-10-5271-5126

Abstract: Objectives: To develop an opportunistic screening model based on a deep learning algorithm to detect recent vertebral fractures in abdominal or chest CTs. Materials and Methods: A total of 1309 coronal reformatted images (504 with a recent fracture from 119 patients, and 805 without fracture from 115 patients), from torso CTs, performed from September 2018 to April 2022, on patients who also had a spine MRI within two months, were included. Two readers participated in image selection and manually labeled the fractured segment on each selected image with Neuro-T (version 2.3.3; Neurocle Inc.) software. We split the images randomly into the training and internal test set (labeled: unlabeled = 480:700) and the secondary interval validation set (24:105). For the observer study, three radiologists reviewed the CT images in the external test set with and without deep learning assistance and scored the likelihood of an acute fracture in each image independently. Results: For the training and internal test sets, the AI achieved a 99.86% test accuracy, 91.22% precision, and 89.18% F1 score for detection of recent fracture. Then, in the secondary internal validation set, it achieved 99.90%, 74.93%, and 78.30%, respectively. In the observer study, with the assistance of the deep learning algorithm, a significant improvement was observed in the radiology resident's accuracy, from 92.79% to 98.2% (p = 0.04). Conclusion: The model showed a high level of accuracy in the test set and also the internal validation set. If this algorithm is applied opportunistically to daily torso CT evaluation, it will be helpful for the early detection of fractures that require treatment.

Keywords: deep learning; artificial intelligence; vertebral compression fracture; spine; computed tomography

1. Introduction

Vertebral compression fractures (VFs) are the most common osteoporotic fractures, and are associated with significant morbidity and mortality [1]. Early diagnosis is crucial to prevent further complications [1–3]. However, VFs can be challenging to discern clinically as they are often asymptomatic or have nonspecific symptoms [4]. Also, incidental VFs that are found during routine chest and abdominal computed tomography (CT) scans are often under-reported due to the lack of sagittal spine reconstructions and the limited awareness of abdominal radiologists about spines [5–10].

Some efforts have been made to use machine learning to aid in the detection of spinal fractures in various radiology exams [11–19], and also to further differentiation between benign and malignant fractures [17]. Only two of them used CT scans [11,12], but sagittal reconstruction images, which are not routinely included in opportunistic torso CT scans, were used, and both acute and old fractures, which were of less clinical significance, were included [20].

So, the purpose of this study was to (1) develop a deep learning model for the automatic detection of particularly acute vertebral compression fractures on routine chest and abdominal CT scans, (2) evaluate the performance of this model in automatically detecting incidental acute VFs, and (3) compare the performances of observer with and without the model's assistance.

2. Materials and Methods

An overview of the datasets is shown in Figure 1. This retrospective study complied with the principles of the Helsinki Declaration and was approved by the institutional review board (IRB) of two participating hospitals: Hospital #1 (Soonchunhyang University Bucheon Hospital) for the internal test set and Hospital #2 (Soonchunhyang University Cheonan Hospital) for the external test set, respectively.

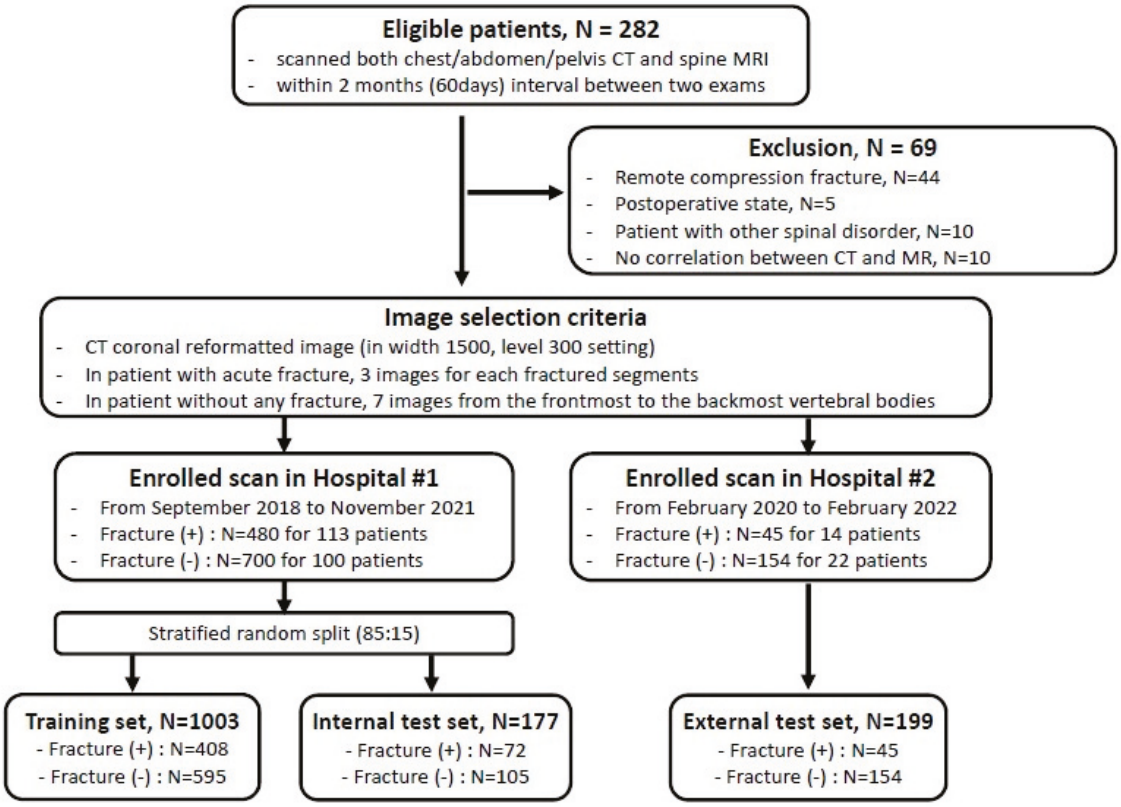


Figure 1. Flow chart of selecting the study sample.

2.1. Patient Datasets

Among the patients who underwent both (1) chest or abdominal CT and (2) MRI of the thoracic or lumbar spine, corresponding to the CT scan's range, patients with less than two months between the two exams were included. An acute VF was identified on the spine MRI, and if the lesion correlated to the CT image, it was classified as being in the fracture-positive group. On the other hand, patients without any fractures in either CT or spine MRI were classified being in the fracture-negative group. Patients who were in a post-operative state following any kind of spinal surgery, such as metallic hardware or bone cement insertion, or with any other pathology including bone tumors (such as bone metastasis from other malignancy), the involvement of bone marrow disease (such as multiple myeloma), or infectious spinal conditions were excluded from the study. Finally, we included 113 fracture-positive and 100 fracture-negative patients identified from September 2018 to November 2021 in Hospital #1 and 14 fracture-positive and 22 fracture-negative patients identified from February 2020 to February 2022 in Hospital #2, respectively.

2.2. Image Selection

All CT images were scanned using a soft-tissue reconstruction kernel with a 3 mm slice thickness and without an interval gap across five different CT scanners. All selected images were saved in bone setting (1500 width and 300 level) from the picture archiving and communication system (PACS). One radiologic resident (with 2 years of in-house training, reader (1) and one board-certified musculoskeletal radiologist (with 8 years of experience, reader (2) selected three serial coronal-reformatted images with 3 mm gaps from the front which showed well-visible fracture lines for each of the fractured segments in the fracture-positive group of patients and seven serial coronal-reformatted images with 6 mm gaps from the frontmost vertebral body for each of the patients in the fracture-negative group. In all, a total of 480 and 700 images from the fracture-positive and -negative groups in Hospital #1, respectively, were used as the development dataset. Similarly, in Hospital #2, 45 and 154 images from the fracture-positive and -negative groups, respectively, were selected for use as the external test set by another board-certified musculoskeletal radiologist (with 5 years of experience, reader 4).

2.3. Deep Learning Model Development

All the selected images were resized to 512 × 512 pixels and uploaded to Neuro-T, version 3.0.0 architecture (Neurocle Inc., Seoul, Republic of Korea). Then, as shown in Figure 2, any visibly fractured vertebral bodies in each image were labeled manually with a colored polygon along the outer margin of the cortex, including as many bone fragments as possible identified by one radiologist (reader 2). The dataset was randomly divided into 85% training and 15% test datasets for model development and accuracy predictions. The primary outcomes of the model were the test accuracy, precision, and F1 score. After training the model using the images, the predicted score of predicted fractured areas was calculated, ranging from 50 to 100%, and displayed as boxes and pixels on the image, like in Figures 2 and 3. The predicted scores then were evenly classified into four groups (no score, normal; from 50 to 66.67%, indeterminate; from 66.68 to 83.33%, probable fracture; above 83.34%, fracture).



(**a**) (**b**) (**c**)

Figure 2. The selected input image was resized to 512 × 512 pixels (**a**). Using the Neuro-T software, version 3.0.0 architecture (Nerocle Inc., Seoul, Republic of Korea), a yellow-colored polygonal box was drawn manually along the outer margin of the cortex, which had the fracture confirmed on a recent MRI, including as many bone fragments as possible (**b**). After the deep learning process was trained on these features, the predicted fractured areas, for which predicted scores ranged from 50 to 100%, were shown on the image with a checked pattern in pixels (**c**). This case was evaluated as a true positive result.



(a) (b) (c)

Figure 3. As a case of false positives, the trained deep learning model color-mapped areas suspected of having fractures, but in reality, these did not have any fractures (**a**). However, these false positive results have a tendency to be found in the high attenuated cortex showing marginal osteophytes of the vertebra or normal endplates. In this case of a false negative, the fractured vertebra segment confirmed on the MRI was colored and trained (**b**), but the deep learning model could not recall a fractured segment when there was no checkered pixel (**c**). It appeared only as a subtle and narrow condensation zone on the CT, making it challenging to suspect a fracture even on the actual raw CT image.

2.4. Observer Study

For the observer study, three radiologists (reader 1, 2, and a board-certified musculoskeletal radiologists with 23 years of experience, reader 3) participated in the review of the CT images in the external test set, to which patients were randomly allocated, and all patients' information was anonymized. All readers independently scored the likelihood of acute VF in each image on a 4-point scale ((1), normal; (2), probably normal; (3), probable fracture; (4), definite fracture). After more than 1 week, we performed the experiment in the same way with the support of a deep learning model, and used highlighted boxes to show the AI-predicted fracture regions on each image and the AI-predicted scores, while knowing the previous scores.

2.5. Statistical Analysis

The statistical analyses were performed using Rex version 3.0.3 (RexSoft, Seoul, Republic of Korea). Continuous data for patients' age and the period between the CT and MRI exams were presented as means and standard deviation, whereas categoric data were presented as counts and percentages. To evaluate the diagnostic performance of the groups, a 4-point scale was dichotomized into normal, 1 and 2 points, and fracture, 3 and 4 points, for binary diagnosis. Then, the performance outcomes were derived from an area under the receiver operating characteristic curve (AUROC) for each human reader, with and without AI assistance, and also for the stand-alone AI reader. A comparison of individual AUROC values was carried out using Delong's method. Sensitivity, specificity, accuracy, and error rate values for each reader were also calculated and compared using McNemar's test, and positive predictive values (PPVs) and negative predictive values (NPVs) were calculated using the Chi-square test. A p value of less than 0.05 was considered to be statistically significant.

3. Results

3.1. Patient Characteristics

Of the 213 patients in the internal test set, 113 had acute VFs (mean age, 61.2 years ± 19.5 [standard deviation, SD]; 60 male) and 100 had no fracture (mean age, 56.1 years ± 14.6 [SD]; 57 male). Among the 36 patients in the external test set, 14 had acute VFs (mean age, 73.9 years ± 12.6 [SD]; 8 male) and 22 had no fracture (mean age, 61.2 years ± 18.0 [SD]; 9 male). A total of 160 and 15 fractured segments were found in 113 and 14 patients from the internal and external test sets, respectively, and a total of 480 and 45 CT images were included with three images per segment (Figure 1). Almost all fractures occurred at the thoracolumbar junction and in the lumbar spine, from T11 to L5 (98.1% and 93.3% for internal and external test sets, respectively), and others occurred at the T6, T8, and T10 vertebrae. The characteristics of the patients in the datasets are listed in Table 1.

Table 1. Characteristics of the patients and images.

Parameter	Internal Test Hospital		External Test Hospital	
	Without Fracture	**With Fracture**	**Without Fracture**	**With Fracture**
No. of patients	100	113	22	14
Age (years) *	56.1 ± 14.6	61.2 ± 19.5	61.2 ± 18.0	73.9 ± 12.6
No. of Men (%)	57/100 (57)	60/113 (53.1)	9/22 (40.9)	8/14 (57.1)
CT-MR scan interval (days) *	13.9 ± 16.0	8.0 ± 11.4	9.8 ± 11.4	4.2 ± 3.9
No. of CT scan-ordered department (%)				
- Orthopedic Surgery	22/100 (22)	26/113 (23.0)	0	0
- Neuro-Surgery	28/100 (28)	32/113 (28.3)	0	0

Table 1. *Cont.*

Parameter	Internal Test Hospital		External Test Hospital	
	Without Fracture	**With Fracture**	**Without Fracture**	**With Fracture**
- Emergency Medicine	0	5/113 (4.4)	12/22 (54.5)	11/14 (78.6)
- Others	50/100 (50)	50/113 (44.2)	10/22 (45.5)	3/14 (21.4)
No. per fractured segment		Total 160		Total 15
- T11		6/160 (3.8)		0
- T12		32/160 (20)		3/15 (20)
- L1		42/160 (26.3)		4/15 (26.7)
- L2		36/160 (22.5)		5/15 (33.3)
- L3		20/160 (12.5)		1/15 (6.7)
- L4		18/160 (11.3)		0
- L5		3/160 (1.9)		1/15 (6.7)
- Others		3/160 (1.9) **		1/15 (6.7) ***

* Data are means ± standard deviations. ** others including T6, T8, and T10 vertebrae. *** the one involving T6 vertebra.

3.2. Stand-Alone AI Performance

The trained deep learning model achieved a 99.89% accuracy, 92.00% precision, and 92.40% F1 score in the internal test set, respectively. The total training time was approximately 13 h. When retesting with the external test set, the model was able to reach a 99.91% accuracy, 80.30% precision, and 85.50% F1 score, respectively. It took 30 min to test the set with the trained model. Also, in the external test set, the model showed an AUROC of 0.9889 (95% confidence interval [CI], 0.9762–0.9977), a sensitivity of 86.67% (95% CI, 73.21–94.95), a specificity of 100% (95% CI, 94.56–100), a PPV of 100% (95% CI, 90.97–100), and an NPV of 91.67% (95% CI, 82.74–96.88) (Table 2). Figure 3 shows one of the cases of false-positive and false-negative results, respectively, from the internal test set.

Table 2. Diagnostic performance of the AI model and three readers for acute fracture detection with and without AI assistance.

Total (*n* = 111)	**AI**	**Reader 1**		**Reader 2**		**Reader 3**	
		Without AI	**With AI**	**Without AI**	**With AI**	**Without AI**	**With AI**
AUROC	0.9889 (0.9762–0.9977)	0.9912 (0.977–0.999)	0.9872 (0.9637–1)	0.968 (0.9437–0.9937)	0.9897 (0.9777–0.996)	0.9576 (0.9142–0.9936)	0.9322 (0.8871–0.9768)
Sensitivity		84.44 (70.54–93.51)	95.56 (84.85–99.46)	86.67 (73.21–94.95)	93.33 (81.73–98.60)	80 (65.4–90.42)	86.67 (73.21–94.95)
: *p*-value		0.07		0.25		0.25	
Specificity	100 (94.56–100)	98.48 (97.84–99.96)	100 (94.56–100)	96.97 (89.48–99.63)	96.97 (89.48–99.63)	95.45 (87.29–99.05)	98.48 (91.84–99.96)
: *p*-value		1		NA		0.48	
Accuracy	94.59 (88.61–97.99)	92.79 (86.29–96.84)	98.2 (93.64–99.78)	92.79 (86.29–96.84)	95.5 (89.80–98.52)	89.19 (81.88–94.29)	93.69 (87.44–97.43)
: *p*-value		0.04		0.25		0.07	
PPV	100 (90.97–100)	97.44 (86.52–99.94)	100 (91.78–100)	95.12 (83.47–99.4)	95.45 (84.53–99.44)	92.31 (79.13–98.38)	97.5 (86.84–99.94)
: *p*-value		0.96		1		0.59	
NPV	91.67 (82.74–96.88)	90.28 (80.99–96.0)	97.06 (89.78–99.64)	91.43 (82.27–96.79)	95.52 (87.47–99.07)	87.5 (77.59–94.12)	91.55 (82.51–96.84)
: *p*-value		0.12		0.53		0.61	

Note—Numbers except for AUROC are presented as percentage with 95% confidence intervals in parentheses. AI = artificial intelligence. AUROC = area under the receiver operating characteristic curve. PPV = positive predictive value. NPV = negative predictive value. NA = not applicable.

3.3. Observer Study

The diagnostic performances of the human observers with and without the model's assistance in the external test set are shown in Table 2 and Figure 4. In the setting without the model's assistance, the AUROCs of the readers, including AI alone, ranged from 0.9576 to 0.9912 and there were no significant differences between all the readers (Table 3).

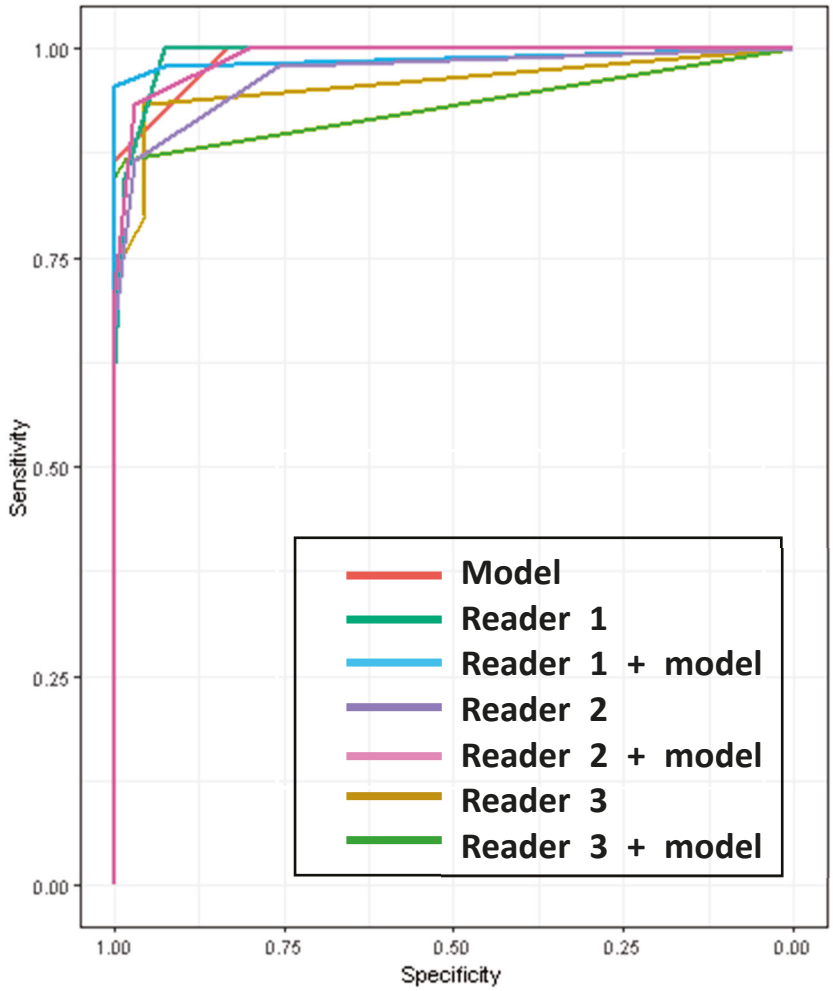


Figure 4. AUROC curves of the model and the readers for diagnosis of vertebral fractures on the external test set. AUROC = area under the receiver operating characteristic curve.

Table 3. *p*-values for comparison between AUROCs of each reader's diagnostic performances with and without AI assistance.

Readers	Reader 1		Reader 2		Reader 3	
	−AI	+AI	−AI	+AI	−AI	+AI
AI	0.72	0.90	0.27	0.98	0.10	0.03
Reader 1 − AI	-	0.74	0.09	0.70	0.10	0.02
Reader 1 + AI		-		0.87	0.07	0.02
Reader 2 − AI			-	0.08	0.45	0.14
Reader 2 + AI				-	0.03	0.01
Reader 3 − AI					-	0.12
Reader 3 + AI						-

With the model's support, only one category, the accuracy of reader 1, who was the radiologic resident, was improved significantly from 92.79 (95% CI, 86.29–96.84) to 98.2 (95% CI, 93.64–99.78) (p = 0.04) (Table 2). The AI did not provide significant assistance to the AUROCs of the diagnostic performances of each reader (Table 3, Figure 4). But the AUROC of reader 3 with AI assistance (0.9322, [95% CI, 0.8871–0.9768]) showed a statistically significantly lower value compared to the other AI-assisted readers (AI-assisted reader 1, 0.9872, [95% CI, 0.9637–1], p = 0.02, and AI-assisted reader 2, 0.9897, [95% CI, 0.9777–0.996], $p = 0$.01)

4. Discussion

As the population ages, complications from osteoporosis, such as vertebral fractures, are becoming more common. However, these vertebral body compression fractures are often underdiagnosed in abdomen CT scans, and there are some concerns about this. Previous studies have explained that this is because many patients are asymptomatic, there is often a lack of sagittal images, and abdominal radiologists often focus only on the solid organs in the abdomen, making it easy to overlook the bony structures [6,9,10]. In previous studies, machine learning systems have been developed to automatically detect such vertebral compression fractures with high levels of sensitivity, accuracy, and precision. They have used plain spine radiography [13], spine CT [14], and chest, abdomen, and pelvis CT scans [12]. However, even with visceral CT scans available, Tomita N et al. investigated the use of sagittal CT scans, which are not routinely included, and focused only on the presence or absence of fractures, without distinguishing between the acute and remote stages. As far as we know, this is the first time that an acute fracture has been used to automatically detect compression fractures on an abdomen or chest CT using a deep learning model.

The results from our deep learning model showed higher performance compared to previous studies with an accuracy of 99.89% and a precision of 92.00% during internal validation, and an accuracy of 99.91% and a precision of 80.30% during the external test set, respectively. When reviewing the true positive cases, the deep learning model accurately filled more than 75% of the fractured area in the fractured vertebral body square with colored mapping. On the other hand, false-positive cases were only shown in a minimal part of the vertebral body, mostly at marginal osteophytes or normal endplate areas. Therefore, the readers could easily distinguish false positives by simply looking at the colored area. Only one false negative was seen, which was a case where the condensation zone was very subtle and narrow and was difficult to detect even on the raw CT image.

In the observer study, a significant improvement in performance was only seen in the resident's accuracy with the AI's assistance, being 98.2 compared to 92.79, with a p-value of 0.0412. The other results did not show a significant improvement when the AI was incorporated. No significant difference was seen in the AUROC between the standalone reader, the AI-assisted reader, or AI alone. This suggests that there is utility in AI, whether used alone or as an assistant. Additionally, the performance of the AI can be considered equal to that of a practicing radiologist.

In addition, it is known that sagittal scans are useful for diagnosing compression fractures [21,22], but it has also been found that high diagnostic rates can be obtained even when only using coronal images, which are almost always included in routine CT scans, without the need for additional reconstruction efforts. This indicates that coronal images alone can be used to diagnose the fracture adequately.

This was a retrospective study from a single institution, and therefore there was a limitation associated with modelling the data due to the sample size. However, the study included images taken with different CT machines at a single institution, and also, for external validation, included images taken using different CT machines and protocols at another institution, which may help to address this limitation. Further studies with larger cohorts across multiple institutions are also needed.

The patient population in this study consisted of those who underwent CT scans and had a recent vertebral fracture diagnosed within two months using a spine MRI, so it was not representative of all vertebral fractures.

Also, the deep learning model was unable to detect remote fractures and analyze the results, so it was not possible to evaluate how well it could differentiate acute fractures from remote lesions. However, the study's significance lies in the fact that it focused on detecting acute lesions that can cause the patient's current symptoms and require additional treatment.

Another limitation is that the performance of musculoskeletal radiologists, and not abdominal radiologists who are expected to use the model in practice, was analyzed, and there was also a limited number of readers. However, the results of the study showed that the AI itself demonstrated a performance that was at the same level as the trained musculoskeletal radiologists, which means that it can be a great help not only to non-specialists, including abdominal radiologists, but also to clinical physicians.

The automatic segmentation of the vertebral body has not yet been implemented due to the technical limitations of the software. Also, despite the differences in background bone density based on age and gender of the patients, these patient-specific co-factors cannot be analyzed due to these technical limitations. Future development of the software is necessary in this regard.

In this study, the trained automated deep learning model showed a high level of accuracy on both internal and external validation sets. Additionally, there was no significant difference in diagnostic performance between the fractures detected by the AI model and those detected by trained and less-trained radiologists. This could potentially reduce the workload of radiologists in detecting fractures, or even replace them. If this algorithm is applied in clinical practice, it can not only help in the early detection of acute vertebral compression fractures for patients but also help to reduce under-reporting by radiologists.

Author Contributions: Conceptualization, Y.S.Y.; Methodology, Y.S.Y.; Software, J.G.C.; Investigation, Y.S.Y.; Data curation, Y.S.Y.; Writing—original draft, Y.R.K.; Writing—review & editing, J.G.C. 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 institutional review board of Soonchunhyang University Bucheon Hospital (approval Code: 2022-05-007, approval date: 2 June 2022).

Informed Consent Statement: Informed consent was waived because of the retrospective nature of the study and the analysis used anonymous clinical data.

Data Availability Statement: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

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

References

1. World Health Organization. Prevention and management of osteoporosis. *World Health Organ. Tech. Rep. Ser.* **2003**, *921*, 1–164.
2. Delmas, P.D. Treatment of postmenopausal osteoporosis. *Lancet* **2002**, *359*, 2018–2026. [CrossRef] [PubMed]
3. Link, T.M.; Guglielmi, G.; van Kuijk, C.; Adams, J.E. Radiologic assessment of osteoporotic vertebral fractuzres: Diagnostic and prognostic implications. *Eur. Radiol.* **2005**, *15*, 1521–1532. [CrossRef] [PubMed]
4. Friedrich, M.; Gittler, G.; Pieler-Bruha, E. Misleading history of pain location in 51 patients with osteoporotic vertebral fractures. *Eur. Spine J.* **2006**, *15*, 1797–1800. [CrossRef]
5. Delmas, P.D.; van de Langerijt, L.; Watts, N.B.; Eastell, R.; Genant, H.; Grauer, A.; Cahall, D.L. Underdiagnosis of Vertebral Fractures Is a Worldwide Problem: The IMPACT Study. *J. Bone Miner. Res.* **2005**, *20*, 557–563. [CrossRef] [PubMed]
6. Bartalena, T.; Giannelli, G.; Rinaldi, M.F.; Rimondi, E.; Rinaldi, G.; Sverzellati, N.; Gavelli, G. Prevalence of thoracolumbar vertebral fractures on multidetector CT: Underreporting by radiologists. *Eur. J. Radiol.* **2009**, *69*, 555–559. [CrossRef] [PubMed]
7. Williams, A.L.; Al-Busaidi, A.; Sparrow, P.J.; Adams, J.E.; Whitehouse, R.W. Under-reporting of osteoporotic vertebral fractures on computed tomography. *Eur. J. Radiol.* **2009**, *69*, 179–183. [CrossRef] [PubMed]

8. Müller, D.; Bauer, J.S.; Zeile, M.; Rummeny, E.J.; Link, T.M. Significance of sagittal reformations in routine thoracic and abdominal multislice CT studies for detecting osteoporotic fractures and other spine abnormalities. *Eur. Radiol.* **2008**, *18*, 1696–1702. [CrossRef] [PubMed]
9. Obaid, H.; Husamaldin, Z.; Bhatt, R. Underdiagnosis of vertebral collapse on routine multidetector computed tomography scan of the abdomen. *Acta Radiol.* **2008**, *49*, 795–800. [CrossRef]
10. Carberry, G.A.; Pooler, B.D.; Binkley, N.; Lauder, T.B.; Bruce, R.J.; Pickhardt, P.J.; Petritsch, B.; Kosmala, A.; Weng, A.M.; Krauss, B.; et al. Unreported Vertebral Body Compression Fractures at Abdominal Multidetector CT. *Radiology* **2013**, *268*, 120–126. [CrossRef]
11. Baum, T.; Bauer, J.S.; Klinder, T.; Dobritz, M.; Rummeny, E.J.; Noël, P.B.; Lorenz, C. Automatic detection of osteoporotic vertebral fractures in routine thoracic and abdominal MDCT. *Eur. Radiol.* **2014**, *24*, 872–880. [CrossRef] [PubMed]
12. Tomita, N.; Cheung, Y.Y.; Hassanpour, S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. *Comput. Biol. Med.* **2018**, *98*, 8–15. [CrossRef]
13. Murata, K.; Endo, K.; Aihara, T.; Suzuki, H.; Sawaji, Y.; Matsuoka, Y.; Nishimura, H.; Takamatsu, T.; Konishi, T.; Maekawa, A.; et al. Artificial intelligence for the detection of vertebral fractures on plain spinal radiography. *Sci. Rep.* **2020**, *10*, 20031. [CrossRef] [PubMed]
14. Burns, J.E.; Yao, J.; Summers, R.M.; Pickhardt, P.J.; Graffy, P.M.; Perez, A.A.; Lubner, M.G.; Elton, D.C.; Zea, R.; Lee, S.J.; et al. Vertebral Body Compression Fractures and Bone Density: Automated Detection and Classification on CT Images. *Radiology* **2017**, *284*, 788–797. [CrossRef]
15. 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] [PubMed]
16. Chen, H.-Y.; Hsu, B.W.-Y.; Yin, Y.-K.; Lin, F.-H.; Yang, T.-H.; Yang, R.-S.; Lee, C.-K.; Tseng, V.S. Application of deep learning algorithm to detect and visualize vertebral fractures on plain frontal radiographs. *PLOS ONE* **2021**, *16*, e0245992. [CrossRef] [PubMed]
17. Li, Y.; Zhang, Y.; Zhang, E.; Chen, Y.; Wang, Q.; Liu, K.; Yu, H.J.; Yuan, H.; Lang, N.; Su, M.-Y. Differential diagnosis of benign and malignant vertebral fracture on CT using deep learning. *Eur. Radiol.* **2021**, *31*, 9612–9619. [CrossRef] [PubMed]
18. Seo, J.W.; Lim, S.H.; Jeong, J.G.; Kim, Y.J.; Kim, K.G.; Jeon, J.Y. A deep learning algorithm for automated measurement of vertebral body compression from X-ray images. *Sci. Rep.* **2021**, *11*, 13732. [CrossRef]
19. Dong, Q.; Luo, G.; Lane, N.E.; Lui, L.-Y.; Marshall, L.M.; Kado, D.M.; Cawthon, P.; Perry, J.; Johnston, S.K.; Haynor, D.; et al. Deep Learning Classification of Spinal Osteoporotic Compression Fractures on Radiographs using an Adaptation of the Genant Semiquantitative Criteria. *Acad. Radiol.* **2022**, *29*, 1819–1832. [CrossRef]
20. Crandall, D. Acute versus chronic vertebral compression fractures treated with kyphoplasty: Early results*1. *Spine J.* **2004**, *4*, 418–424. [CrossRef]
21. Hardy, C.; Hazelton, J.P.; Gefen, R. Added Value of Dedicated Spine CT to Detect Fracture in Patients with CT Chest, Abdomen, and Pelvis in the Trauma Setting. *Curr. Probl. Diagn. Radiol.* **2019**, *48*, 554–557. [CrossRef] [PubMed]
22. Lee, S.J.; Binkley, N.; Lubner, M.G.; Bruce, R.J.; Ziemlewicz, T.J.; Pickhardt, P.J. Opportunistic screening for osteoporosis using the sagittal reconstruction from routine abdominal CT for combined assessment of vertebral fractures and density. *Osteoporos. Int.* **2016**, *27*, 1131–1136. [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

AI and Interventional Radiology: A Narrative Review of Reviews on Opportunities, Challenges, and Future Directions

Andrea Lastrucci [1], Nicola Iosca [1], Yannick Wandael [1], Angelo Barra [1], Graziano Lepri [2], Nevio Forini [3], Renzo Ricci [1], Vittorio Miele [4,5] and Daniele Giansanti [6,*]

[1] Department of Allied Health Professions, Azienda Ospedaliero-Universitaria Careggi, 50134 Florence, Italy; andrea.lastrucci@unifi.it (A.L.); ioscan@aou-careggi.toscana.it (N.I.); wandaely@aou-careggi.toscana.it (Y.W.); barraa@aou-careggi.toscana.it (A.B.); riccire@aou-careggi.toscana.it (R.R.)
[2] Unità Sanitaria Locale Umbria 1, Via Guerriero Guerra 21, 06127 Perugia, Italy; graziano.lepri@uslumbria1.it
[3] Dipartimento di Medicina e Chirurgia, Universita' degli Studi di Perugia, Piazzale Settimio Gambuli, 1, 06129 Perugia, Italy; nevio.forini@unipg.it
[4] Department of Experimental Clinical and Biomedical Sciences, University of Florence, 50134 Florence, Italy; vmiele@sirm.org
[5] Department of Radiology, Careggi University Hospital, 50134 Florence, Italy
[6] Centro TISP, Istituto Superiore di Sanità, Via Regina Elena 299, 00161 Roma, Italy
* Correspondence: daniele.giansanti@iss.it

Abstract: The integration of artificial intelligence in interventional radiology is an emerging field with transformative potential, aiming to make a great contribution to the health domain. This overview of reviews seeks to identify prevailing themes, opportunities, challenges, and recommendations related to the process of integration. Utilizing a standardized checklist and quality control procedures, this review examines recent advancements in, and future implications of, this domain. In total, 27 review studies were selected through the systematic process. Based on the overview, the integration of artificial intelligence (AI) in interventional radiology (IR) presents significant opportunities to enhance precision, efficiency, and personalization of procedures. AI automates tasks like catheter manipulation and needle placement, improving accuracy and reducing variability. It also integrates multiple imaging modalities, optimizing treatment planning and outcomes. AI aids intra-procedural guidance with advanced needle tracking and real-time image fusion. Robotics and automation in IR are advancing, though full autonomy in AI-guided systems has not been achieved. Despite these advancements, the integration of AI in IR is complex, involving imaging systems, robotics, and other technologies. This complexity requires a comprehensive certification and integration process. The role of regulatory bodies, scientific societies, and clinicians is essential to address these challenges. Standardized guidelines, clinician education, and careful AI assessment are necessary for safe integration. The future of AI in IR depends on developing standardized guidelines for medical devices and AI applications. Collaboration between certifying bodies, scientific societies, and legislative entities, as seen in the EU AI Act, will be crucial to tackling AI-specific challenges. Focusing on transparency, data governance, human oversight, and post-market monitoring will ensure AI integration in IR proceeds with safeguards, benefiting patient outcomes and advancing the field.

Keywords: radiology; interventional radiology; artificial intelligence; machine learning; deep learning

1. Introduction

Interventional radiology (IR) based on the Collins Dictionary [1] is defined as "an application of radiology that enables minimally invasive surgery to be performed with the aid of simultaneous radiological imaging of the field of operation within the body". According to the Oxford Dictionary [2], it involves "The performance of therapeutic or diagnostic procedures under the control of an appropriate imaging technique. Guidance is commonly by X-ray fluoroscopy, ultrasound, or computerized tomography, and recently also by magnetic resonance imaging. Procedures commonly performed include the placing of vascular catheters, drainage of fluid collections or abscesses, stenting of obstructions to the gastrointestinal tract or blood vessels, embolization, cryotherapy, and radiofrequency ablation. Imaging is also used in many forms of minimally invasive surgery." Interventional radiology is a relatively recent discipline that has revolutionized the way many diseases are treated. It combines the expertise of radiology with other specialties, such as surgery, oncology, and urology, offering less invasive solutions compared to traditional procedures. The key advantage of this technique is the ability to perform image-guided interventions, reducing recovery times and minimizing risks for the patient [3].

1.1. The Scope and Implications of Interventional Radiology

Interventional radiology (IR) is a transformative medical field that has seen rapid advancements and is applied across numerous disciplines. The applications and implications of this specialty include the following key areas.

1.1.1. Key Support in Surgery and Oncology

One of the most valuable applications of IR is in controlling hemorrhages. Through embolization techniques, blood vessels can be sealed from the inside, avoiding the need for more invasive surgical procedures. This is particularly beneficial in the postoperative period when complications such as bleeding might otherwise require reoperation.

IR has also made significant strides in cancer treatment [4]. Using image-guided probes (via CT scans, MRI, or ultrasound), doctors can precisely target tumors and destroy them from within. Different techniques are employed depending on the type of tumor; some use heat—emitted by radiofrequency or microwaves—to "burn" cancer cells, while others utilize extreme cold ($-80\ ^{\circ}C$) to freeze and destroy them.

In certain cases, IR even enables the treatment of tumors once considered inoperable. Innovative techniques such as radioembolization, which involves injecting radioactive microspheres directly into the tumor, can slow or halt disease progression. Another promising approach is the use of genetically modified viruses that selectively attack cancer cells while leaving healthy tissue unharmed.

1.1.2. Beyond Oncology: Interventional Radiology in Medical Specialties

Interventional radiology extends far beyond cancer treatment. It is widely used in managing vascular diseases, including angioplasty to reopen narrowed arteries or veins and embolization to treat aneurysms and arteriovenous malformations [5].

IR is also proving highly effective in treating liver and kidney conditions, offers minimally invasive alternatives that can replace traditional surgical interventions. In liver treatment, for example, transarterial chemoembolization (TACE) allows for the direct injection of chemotherapy drugs into liver tumors, minimizing systemic side effects.

1.1.3. Pain Management and Contributions to Quality of Life

Interventional radiology plays a crucial role in pain management. Techniques such as vertebroplasty and kyphoplasty stabilize vertebral fractures—often caused by osteoporosis or metastases—providing immediate pain relief [6].

Additionally, nerve ablation is used to treat chronic pain in cancer patients and those with other debilitating conditions.

1.1.4. Applications in Gynecology, Urology, and Gastroenterology

IR also offers innovative solutions in gynecology and urology [7]. For instance, uterine fibroid embolization effectively reduces fibroids without the need for surgery, while prostatic artery embolization is a promising treatment for benign prostatic hyperplasia.

In gastroenterology, stents are used to treat esophageal, gastric, and intestinal strictures, significantly improving the quality of life of patients suffering from tumor-related or other types of obstructions. Additionally, abscesses in the abdomen or pancreatic cysts can be safely and minimally invasively drained.

1.1.5. The Importance of Vascular Access

Another crucial aspect of interventional radiology is vascular access management. The placement of central venous catheters such as peripherally inserted central catheter lines or port-a-caths is essential for patients requiring long-term therapies, such as chemotherapy or parenteral nutrition [8]. Thanks to imaging guidance, these procedures are performed with high precision and safety.

1.2. A Continuously Evolving Future of Interventional Radiology

With ongoing technological advancements, interventional radiology is expanding its scope, offering increasingly effective and minimally invasive solutions. These new techniques not only improve patients' quality of life but also often serve as safer and more efficient alternatives to traditional surgeries. The future of this discipline is continuously evolving, with new approaches that promise to make medicine even more precise, effective, and patient-centered, accompanied by a strong specialization of the professionals involved [9].

One of the most exciting developments in IR [9] is the integration of artificial intelligence (AI), robotics, augmented reality (AR), and virtual reality (VR). AI is revolutionizing the field by assisting with real-time image analysis, optimizing procedural planning, and enhancing decision-making accuracy. For example, AI-powered image analysis can detect subtle anomalies in real-time, such as small tumors or vessel blockages, allowing for faster and more accurate diagnoses during procedures. Robotics increases precision by minimizing human error and even enables remote interventions, making IR more accessible.

AR and VR technologies play a critical role in training and real-time guidance, offering immersive environments that improve catheter navigation and treatment delivery during complex procedures. For instance, AR systems can overlay 3D images of a patient's anatomy onto real-time X-ray or CT scans, guiding the physician's movements during catheter insertions, biopsies, or tumor ablations. When integrated with AI, AR can provide dynamic, real-time adjustments to these images, enhancing precision during delicate procedures. VR, on the other hand, allows for hands-on training in a simulated environment, where medical professionals can practice complex interventions without risk to patients, accelerating skill development.

As these technologies continue to advance, IR is poised to become even more adaptable and efficient, reducing invasiveness and improving patient outcomes. The future of IR will be shaped by a seamless fusion of human expertise and cutting-edge technology, with AI

at the forefront of this transformation. However, the integration of these technologies also brings challenges that require careful monitoring and continual adaptation to ensure the best possible impact on patient care.

AI acts as a mediator [10], assisting in image analysis, optimizing procedural planning, and improving real-time decision-making. Robot-assisted interventions increase accuracy, reducing human error and enabling remote procedures. AR and VR technologies offer immersive training environments and real-time guidance during complex interventions, allowing for more precise catheter navigation and treatment delivery. As these innovations continue to evolve, interventional radiology is set to become even more adaptive, personalized, and minimally invasive, with AI playing a significant role. However, this integration introduces challenges and ongoing needs that must be monitored closely to ensure its successful implementation and optimal impact on patient care [9,10].

1.3. Purpose

It is therefore important to analyze both the evolution of the integration between AI and interventional radiology (IR) as well as the ongoing implications and challenges. An overview of the published reviews allows us to bring this out in a comprehensive manner. The general aim of this study is, therefore, to provide a narrative review of reviews (NRR) with the purpose of analyzing the state of integration of AI with interventional radiology, with the following specific objectives:

- Analyze the overall bibliometric trends in the field: This study aims to provide a comprehensive bibliometric analysis of research output, focusing on trends and developments over time.
- Identify established themes and categories: Identify key areas of focus in reviews, such as AI applications, diagnostic accuracy, automation, decision support systems, and workflow optimization.
- Examine Opportunities and Challenges: Explore the potential benefits and challenges of integrating AI into interventional radiology, such as improving procedural precision, reducing human error, enhancing patient outcomes, and addressing issues like data privacy, regulatory concerns, and technology adoption barriers.

2. Methodology Overview

The methodology comprises the following two main approaches: a narrative review of reviews and a bibliometric analysis. The narrative review synthesizes existing literature on digital cytopathology, focusing on advancements, challenges, and the role of artificial intelligence (AI). The bibliometric analysis examines historical research trends in digital cytopathology, particularly over the last decade with an emphasis on AI-related studies.

These two methods were chosen to both make an in-depth synthesis of research in the field of AI and interventional radiology and to statistically evaluate their development over time.

2.1. Narrative Review Selection and Qualification Process

2.1.1. Checklist and Qualification

The narrative review followed a structured selection process based on the ANDJ standardized narrative checklist. Targeted searches were conducted using PubMed and Scopus with predefined composite search keys.

A qualification methodology was employed, using quality parameters to determine study inclusion and a predefinite procedure (see the Algorithm 1 below).

Algorithm 1: Selection Process for the NRR

1. **Define search query:**
   - *"(interventional radiology[Title/Abstract]) AND ((neural network[Title/Abstract]) OR (Artificial Intelligence[Title/Abstract]) OR (deep learning[Title/Abstract]) OR (ANN[Title/Abstract]) OR (GAN[Title/Abstract]))"*
2. **Conduct searches** in PubMed and Scopus using the defined query.
3. **Select relevant studies** from peer-reviewed journals that focus on the field priority, such as recent reviews that assess prior studies and reviews providing broad analyses, integrating findings from previous works and with a focus on AI applications in line with the journal topic (articles more focused on computer sciences were excluded on the basis that they were not in line the focus of the journal).
4. **Evaluate each study** based on the following parameters:
   - **N1**: Clear rationale in the introduction.
   - **N2**: Appropriate research design.
   - **N3**: Clearly described methodology.
   - **N4**: Well-presented results.
   - **N5**: Conclusions justified by results.
   - **N6**: Disclosure of conflicts of interest.
5. **Assign scores** to parameters **N1–N5** (scale of 1–5).
6. **Assess N6** using a binary Yes/No measure.
7. **Preselect studies** meeting the following criteria:
   - **N6 = "Yes"** (conflict of interest disclosed).
   - **N1–N5 scores > 3** (ensuring methodological rigor).
8. **Include preselected studies** in the final synthesis.

2.1.2. Assessment Process

Each study included in the analysis was reviewed by two initial assessors (A.L and D.G.). These assessors were responsible for evaluating the studies based on their focus on the integration of artificial intelligence (AI) in interventional radiology and according to predefined criteria. The assessment criteria included Clarity of Rationale, Study Design Appropriateness, Methodological Rigor, Result Presentation, Justification of Conclusions, and Disclosure of Conflicts of Interest. Each criterion was scored on a predefined scale to provide a quantitative measure of the quality and relevance of the analyzed studies. The primary assessors independently reviewed the studies and assigned scores to each parameter, ensuring that each study was evaluated against uniform standards. This dual-assessment approach was designed to enhance the reliability of the review by incorporating different perspectives and reducing the likelihood of individual bias influencing the evaluation process.

In cases where the two initial assessors disagreed on the scores or the inclusion of a study, a third assessor chosen with a rotation criterion (either G.L. or Y.W.) was involved to adjudicate the decision. This third-party assessment played a crucial role in resolving conflicts and ensuring that final decisions were fair and well-justified. The additional scrutiny provided by the third assessor helped balance differing opinions and reinforced the integrity of the review process.

The multi-assessor approach was implemented to minimize bias and ensure a rigorous and balanced evaluation of the literature. By integrating multiple viewpoints and

providing a structured mechanism for resolving disagreements, the review aimed to offer a comprehensive and objective assessment of AI applications in interventional radiology.

2.1.3. Managing Bias in the Review

To ensure the objectivity and methodological rigor of the review, several strategies were implemented to manage and minimize bias throughout the assessment process:

Diverse Assessors:

Each study was reviewed by two assessors with different backgrounds, selected to ensure a variety of perspectives. The inclusion of assessors with expertise in both interventional radiology and AI reduced the risk of individual biases influencing the evaluation.

Clear Assessment Criteria:

The studies were analyzed using predefined criteria, including Clarity of Rationale, Study Design Appropriateness, Methodological Rigor, Result Presentation, Justification of Conclusions, and Disclosure of Conflicts of Interest. Furthermore, data were presented based on a standardized checklist, reducing the risk of subjective interpretation.

Scoring System:

Each parameter was rated on a scale from 1 to 5, while the disclosure of conflicts of interest was assessed using a binary evaluation (Yes/No). This quantitative approach allowed for consistent evaluations across studies and provided a transparent mechanism for comparing study quality.

Independent Review:

The primary assessors reviewed the studies independently, assigning scores without prior discussion. This independence helped ensure that individual judgments were based solely on the study's merit and predefined criteria, minimizing groupthink or shared biases.

Dispute Resolution:

In cases where the two assessors disagreed on scores or study inclusion, a third assessor was consulted to provide an impartial judgment. This third-party adjudication helped resolve conflicts fairly, offering an additional level of scrutiny and ensuring balanced decision-making.

Structured Mechanism for Disagreements:

The process for resolving disagreements was formalized and structured. The third assessor reviewed the initial evaluations and provided a reasoned judgment to reconcile differences. This structured approach ensured that conflicts were systematically addressed and that final decisions were based on a comprehensive evaluation.

Transparency:

The use of a standardized checklist for data presentation and a clear scoring system enhanced transparency in the assessment process. By documenting the criteria and scoring rationale, the review process became more traceable and reproducible, reducing the potential for undisclosed biases.

By incorporating these strategies, this review aimed to offer a thorough and balanced evaluation of the literature. The multi-assessor approach, combined with structured criteria and a formal dispute resolution mechanism, was designed to minimize bias and enhance the reliability and objectivity of the review process.

2.1.4. Selected Studies

Figure 1 illustrates that the initial search yielded a total of 71 reviews. From these, 33 studies were excluded due to their lack of direct focus. Following the evaluation according to the methodology described above, 27 review studies were retained [11–37] for further consideration, while 11 studies were excluded.

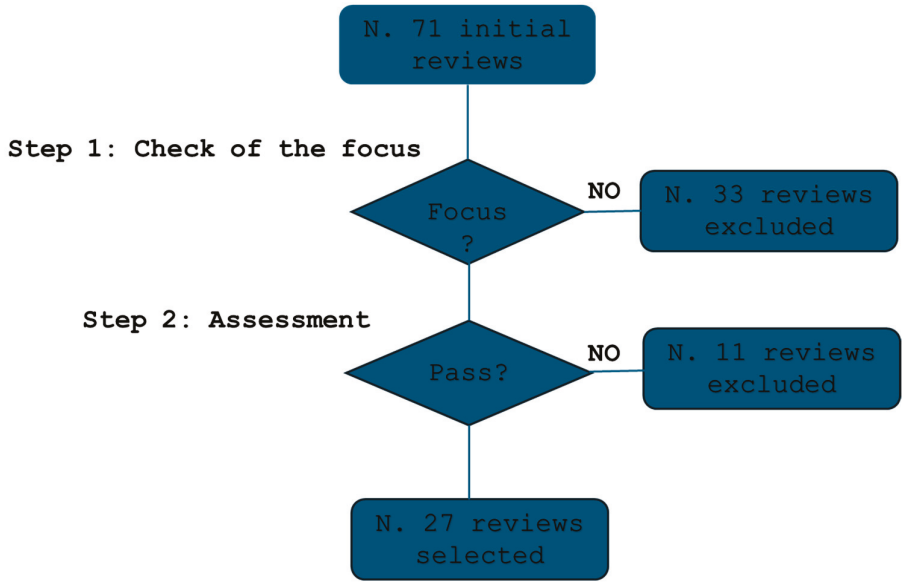


Figure 1. Details on the process of study selection.

2.2. Bibliometric Analysis Methodology

The methodological approach involved a multi-step analysis combining quantitative, statistical, graphical, and narrative methods. First, a systematic search was conducted on PubMed using specific keywords related to artificial intelligence (AI) and interventional radiology (IR). The quantitative analysis focused on the total number of publications, identifying trends in publication frequency and categorizing articles into primary research and review types. The statistical analysis was used to calculate trends across the entire period, as well as focusing on the last 10 and 5 years. The data were segmented into these timeframes to identify patterns in the publication frequency and to examine the rate of growth in AI-related research within interventional radiology (IR). This analysis helped highlight periods of significant acceleration, particularly in the last five years, and allowed for comparison between broader trends and more recent developments. Graphical representations, such as trends over time and distribution of study types, were generated to visualize the growth of AI research in IR, especially in the last five years. Finally, a narrative approach was used to interpret the results, considering the factors influencing this increase, such as advancements in AI technologies, the impact of the COVID-19 pandemic, increased funding, and the regulatory and clinical adoption of AI in medical practices.

3. Results

The study results have been organized into three main sections to provide a structured and in-depth overview.

Section 3.1: Focuses on bibliometric trends in this field by analyzing the progression of publication volumes over time, with a particular emphasis on the last five. This offers insights into the evolution of, and growing interest in, the topic.

Section 3.2: Examines the selected review studies, which focus on the application of AI in IR. This section explores emerging themes and provides a categorization to better understand the key areas of research and innovation.

Section 3.3: Based on the findings from the analyzed studies, it identifies emerging opportunities and highlights barriers that require further exploration and development to drive significant advancements in the field.

This systematic approach provides a clear and comprehensive view of the state of the art, current trends, and future prospects in the application of AI to assistive technologies in healthcare.

3.1. Trends

A comprehensive literature search conducted on PubMed (refer to Key 1 in Box 1) provides insight into the evolution of research in interventional radiology. Since 1978, a total of **10,947 studies** have been published in this field, reflecting decades of scientific advancements and clinical innovations. However, when refining the search criteria to focus specifically on the integration of **artificial intelligence** (**AI**) **in interventional radiology** (see Key 2 in Box 1), the number of studies drops dramatically to just **114**. This figure represents a mere **1.04%** of all published research in interventional radiology, underscoring that AI is still an emerging area within this specialty (Figure 2).

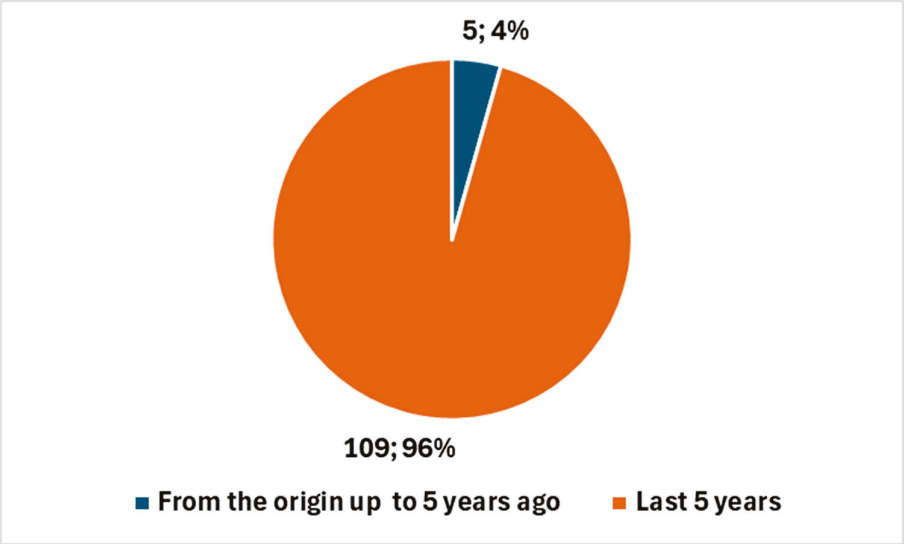


Figure 2. Distribution of AI-related publications in interventional radiology—comparison between the last five years and earlier periods.

A closer examination of the timeline of AI-related publications reveals a **notable concentration of research in recent years**. Out of the **114 AI-focused studies**, **104** (**over 91%**) **were published in the last five years**, highlighting a significant surge in interest and scientific output in this domain. This period has been profoundly shaped by the **COVID-19 pandemic**, which accelerated the adoption of **digital health technologies**, **automation**, **and AI-driven innovations** across various medical disciplines. The pandemic created both challenges and opportunities, prompting increased investment in AI solutions to optimize workflows, improve procedural accuracy, and enhance patient outcomes in interventional radiology.

When analyzing the **nature and composition** of these AI-related studies, an interesting pattern emerges. Of the **114 total publications**, **44 are review articles**, while the remaining **70 studies** consist of primary research articles (Figure 3). This distribution indicates that a significant proportion of AI research in interventional radiology is dedicated to synthesizing existing knowledge, evaluating trends, and identifying future directions. Notably, the majority of these reviews are **very recent**; **42 out of 44** (**95.5%**) **have been published within the last five years**. This suggests an ongoing effort by the scientific community to assess the rapid advancements in AI and its implications for clinical practice. The predominance of recent reviews also highlights the field's current stage of development, where researchers are still working to consolidate knowledge, establish best practices, and address key challenges such as **regulatory considerations**, **clinical validation**, **and real-world integration**.

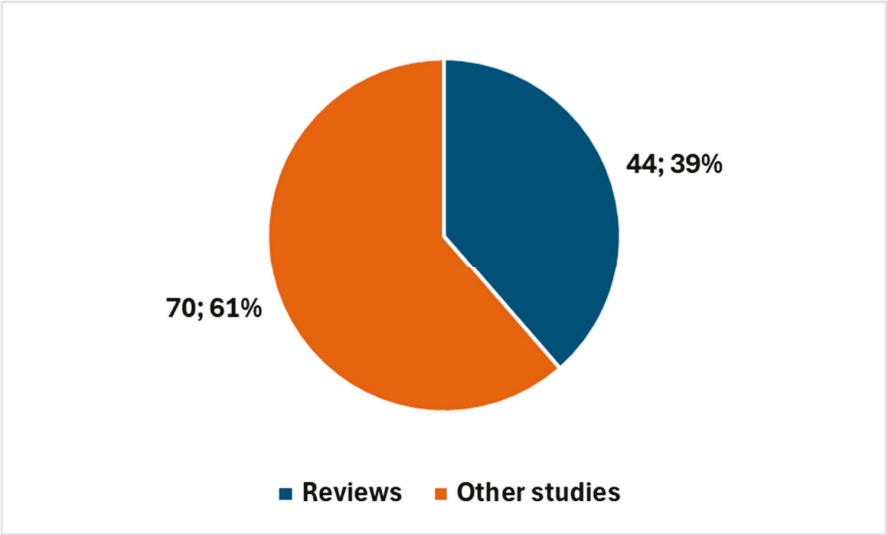


Figure 3. Distribution of AI-related studies in interventional radiology—comparison between review articles and primary research publications.

The sharp rise in AI-related publications over the past five years can likely be attributed to several factors, which this overview aims to highlight and substantiate:

The Acceleration of AI Research—Advances in machine learning, deep learning, and computational power have made AI increasingly applicable to radiology.

COVID-19 as a Catalyst—The pandemic heightened the need for AI-driven diagnostic tools, remote assessments, and automation in medical imaging, pushing research efforts in this direction.

Increased Funding and Interest—Governments, healthcare institutions, and the private sector have heavily invested in AI-driven solutions for medical imaging and diagnostics.

Regulatory and Clinical Adoption—The increasing integration of AI into clinical workflows has driven more research to validate its effectiveness and ensure regulatory compliance.

This trend suggests that AI in interventional radiology is a rapidly expanding field, with growing recognition of its potential to enhance diagnostics, improve efficiency, and support clinical decision-making.

Box 1. Used search keys.

(interventional radiology[Title/Abstract])
(interventional radiology[Title/Abstract]) AND ((neural network[Title/Abstract]) OR (Artificial Intelligence[Title/Abstract]) OR (deep learning[Title/Abstract]) OR (ANN[Title/Abstract]) OR (GAN[Title/Abstract]))
(diagnostic radiology[Title/Abstract]) AND ((neural network[Title/Abstract]) OR (Artificial Intelligence[Title/Abstract]) OR (deep learning[Title/Abstract]) OR (ANN[Title/Abstract]) OR (GAN[Title/Abstract]))

3.2. The Application of IA in Interventional Radiology: A Comprehensive Overview

An analysis of the studies has been conducted, with particular focus on the integration of AI and other technologies in interventional radiology (IR), and is summarized in Table 1.

A brief complementing summary is provided below. Cornelis et al. [11] highlighted the automation and AI integration in navigation and robotic systems for percutaneous image-guided interventions, utilizing both new and established metrics to categorize and compare their capabilities. The findings indicate that none of the navigation and robotic systems have achieved full autonomy. The level of autonomy in surgical robotics (LASR) and levels of integration of advanced imaging and AI (LIAI2) scales can guide innovation by pinpointing areas for further development and integration.

Boeken et al. [12] examined the potential of AI to enhance procedural outcomes with robotics, enabling robotic systems to handle complex tasks like catheter manipulation or needle placement with greater precision and reliability. Techniques such as reinforcement learning and haptic vision are being studied to address various issues, training robots to adapt based on real-time feedback from the environment. As AI-driven robotics evolve, IR could shift towards a model where human expertise oversees the technology rather than performs the intervention itself.

Vlastaris et al. [13] explored the transformative impact of digital technologies like AI, robotics, and extended reality on IR, highlighting their potential to enhance precision, efficiency, and patient outcomes. The integration of these technologies should be considered as an aid for IR practitioners rather than as a replacement, ensuring they enhance clinical and procedural practice.

Lee et al. [14] explored the adaptation of smartphones' built-in sensors for virtual and augmented reality uses in interventional radiology, aiming to improve accuracy and standardization for needle-based procedures like biopsy and ablation.

Barat et al. [15] explored the main research initiatives and technological advances shaping the landscape of radiology in France, including new developments in AI applications.

Matsui et al. [16] published an up-to-date review of the literature focused on the current state of AI applications in interventional oncology, including automatic segmentation of organs, tumors, and treatment areas; treatment simulation; improvement of intraprocedural image quality; prediction of treatment outcomes; and detection of post-treatment recurrence. Most AI-based methods discussed in this review are still in the research stage, with few implemented in clinical practice.

Lesaunier et al. [17] provided the medical community with the most important current and future applications of AI in IR. AI has the potential to radically change the daily practice of IR at several levels, such as multimodality management in the preoperative setting and supporting radiologists in image analysis and real-time decision-making in the perioperative setting.

Zhang et al. [18] focused on the research progress and applications of AI and robotics in interventional radiology and analyzed their potential and limitations. The findings indicated that although AI and robotics technologies are not yet widely applied in clinical settings, they are expected to significantly improve the processes and efficacy of interventional treatments.

Glielmo et al. [19] assessed the current applications and challenges of AI in IR, offering insights into decision support and outcome prediction, imaging enhancements, robotics, and touchless interactions, shaping the future of patient care.

Warran et al. [20,21] explored the current state of AI in IR in Part 1, examining its implementation across various stages: pre-procedural, intra-procedural, and post-procedural. In Part 2, they outlined a hierarchy of potential risks and harms associated with AI and provided a checklist to guide the safe deployment of AI technologies in clinical IR settings.

Geevarghese et al. [22] explored recent developments within interventional oncology, highlighting its potential impact facilitated by AI, personalized medicine, and imaging innovations. The integration of AI in interventional oncology promises to accelerate tumor detection and characterization, guide treatment strategies, and refine predictive models.

Campbell 4th et al. [23] summarized the design principles of ChatGPT relevant to healthcare and highlighted tasks and activities with the greatest potential for ChatGPT utilization in the practice of IR. These tasks involve patient-directed and physician-directed communications to convey medical information efficiently and act as a medical decision support tool.

O'Brien et al. [24] examined the literature to contextualize how AI is currently being implemented in interventional radiology from a preoperative, intraoperative, and postoperative perspective.

Gaddum and Chapiro [25] provided a guide for interventional radiologists to critically appraise AI research and products by identifying 12 key considerations: AI's relevance to the clinical problem, type of AI algorithm used, data quality and annotation, accuracy reporting, standardized reporting applicability, reproducibility and transparency, algorithm validation, interpretability, impact on IR, pathway to clinical translation, clinical benefit and cost-effectiveness, and regulatory framework.

Lanza et al. [26] provided a brief overview of the current status and potential role of robotics in IR. The technical developments in robotics and navigational systems using computed tomography image guidance, magnetic resonance image guidance, and ultrasound image guidance were analyzed.

Von Ende et al. [27] described the current and possible future applications of AI, radiogenomics, and augmented and virtual reality in IR, also describing the challenges and limitations that must be addressed before these applications can be fully implemented into common clinical practice.

Posa et al. [28] provided a comprehensive overview of the most valuable, well-established, and promising AI and digital health innovations, aimed at empowering physicians to incorporate these technologies into clinical practice. Their work particularly emphasized advancements in the field of interventional oncology, encouraging greater physician engagement with these cutting-edge tools.

Boeken et al. [29] provided an update on the current status of radiomics and AI research, analyzed upcoming challenges, and also discussed the main applications of AI in IR to help radiologists better understand and critique articles reporting AI in medical imaging.

Kallini et al. [30] explored the evolution of AI from its theoretical foundations in the 1950s to modern advancements. While AI made significant progress in radiology, its application in IR remained in its early stages. This article summarized AI's role in radiology and IR, highlighting past advancements and future potential.

Waller et al. [31] examined how the implementation of AI in diagnostic radiology and IR can improve image analysis, aid in diagnosis, and suggested appropriate interventions, clinical predictive modeling, and trainee education.

Seah et al. [32] provided an overview of machine learning, radiomics, and AI in the field of IR, enumerating the possible applications of such techniques, also describing techniques to overcome the challenge of limited data when applying these techniques in IR.

Mazaheri et al. [33] examined a few challenges contributing to this scarcity of AI applications in IR, including inherent specialty challenges, regulatory hurdles, intellectual property, raising capital, and ethics.

Desai et al. [34] described AI technologies that have been developed within the IR suite, as well as some future work with a focus on AI's potential impact in pediatric interventional medicine.

D'amore et al. [35] highlighted the current role of machine learning and AI in the field of interventional oncology. Software development can increase the detection of tumors through routine screening and improve diagnostic accuracy in classifying tumors.

Gurgitano et al. [36] discussed the potential of AI applications in IR, which range beyond computer vision and diagnosis, to include screening and modeling of patient selection, predictive tools for treatment planning and navigation, and training tools.

Iezzi et al. [37] integrated evidence-reported literature and experience-based perceptions to assist not only residents and fellows who are training in IR but also practicing colleagues who are approaching locoregional mini-invasive treatments.

Table 1 presents a summary of the analyzed studies with their focus and the emerging role of AI (and any other technologies).

Table 1. Summary of the analyzed studies with their focus and the emerging role of AI (and any other technologies).

Study	Brief Description	Focus	AI and Technology Integration
[11] Evaluation of navigation and robotic systems for percutaneous image-guided interventions: A novel metric for advanced imaging and artificial intelligence integration	This study reviews clinically validated robotic and navigation systems for percutaneous image-guided interventions, focusing on integrating AI and advanced imaging. It introduces a novel metric, LIAI2, for assessing AI and imaging integration.	The paper evaluates the extent of AI and advanced imaging integration in robotic systems for percutaneous procedures, aiming to enhance automation levels and precision.	The integration of AI aims to improve procedural accuracy and system autonomy, helping with tasks like needle placement and biopsies. This paper introduces a novel taxonomy to measure AI and imaging integration, addressing current limitations and driving future innovation in autonomous navigation.
[12] The Role and Future of Artificial Intelligence in Robotic Image-Guided Interventions	This review explores AI's transformative potential in robotic-assisted interventional radiology (IR). It discusses AI's role in handling tasks like catheter manipulation and needle placement, as well as current challenges such as limited decision-making capabilities.	The study focuses on how AI can improve robotic systems' precision, efficiency, and reliability in IR interventions, enabling automation in complex procedures.	AI integration in robotic systems enhances precision and adaptability, particularly through reinforcement learning. While AI-driven robotics are expected to take over more tasks, human oversight remains essential to ensure safety and control during complex procedures.
[13] The Transformative Impact of AI, Extended Reality, and Robotics in Interventional Radiology: Current Trends and Applications	This article explores the integration of AI, robotics, and extended reality (XR) in interventional radiology, showcasing how these technologies are reshaping procedures and improving outcomes. It discusses the challenges of standardization across healthcare settings.	The paper emphasizes the synergy between AI, XR, and robotics to enhance image-guided interventions' precision and efficiency in IR.	AI, robotics, and XR are being integrated to enhance diagnostic accuracy and procedural precision. AI is improving outcomes, while XR aids in visualizing complex anatomy, and robotics reduces human error. These technologies complement clinicians' skills and aim to optimize patient care without replacing human expertise.
[14] Smartphone Technology for Applications in Image-Guided Minimally Invasive Interventional Procedures	This study investigates the use of smartphone technologies, such as augmented reality (AR), in image-guided minimally invasive procedures like biopsies and ablations. It explores how smartphones could enhance the accuracy and efficiency of needle-based interventions.	The focus is on the use of smartphones and AR to improve minimally invasive IR procedures by offering portable, real-time image guidance.	The integration of AR with smartphone sensors allows for real-time visualization of the procedural environment, offering greater precision in navigating complex anatomy. However, further research is needed to refine the technology and optimize its clinical use in IR.

Table 1. *Cont.*

Study	Brief Description	Focus	AI and Technology Integration
[15] Imaging in France: 2024 Update	This article reviews recent advances in diagnostic and interventional radiology in France, particularly the integration of AI into diagnostic imaging. It highlights France's contributions to improving diagnostic accuracy and advancing minimally invasive treatments in IR.	The paper focuses on technological advancements in AI-driven diagnostic imaging and their impact on IR, emphasizing innovations in tumor detection and advanced image fusion techniques.	AI has become central to improving diagnostic imaging in France, with applications in tumor detection, CT imaging, and image fusion. These advancements are enhancing the accuracy of diagnoses and treatment planning in IR. The study also discusses AI's role in improving workflows and patient outcomes in interventional procedures.
[16] Applications of artificial intelligence in interventional oncology: An up-to-date review of the literature	This comprehensive review examines the role of AI in interventional oncology, particularly in image-guided therapies such as tumor embolization and ablation. It discusses AI tasks like tumor segmentation, outcome prediction, and recurrence detection.	The study focuses on how AI is revolutionizing interventional oncology through automation and enhanced treatment prediction.	AI in interventional oncology is being used to automate segmentation, predict outcomes, and detect recurrence after treatments. The integration of deep learning and radiomics is crucial in these tasks, improving treatment precision and patient management. More research is needed to validate these AI applications in clinical settings.
[17] Artificial intelligence in interventional radiology: Current concepts and future trends	This article explores the state of AI in interventional radiology, emphasizing its applications in preoperative, perioperative, and postoperative care. It highlights AI's role in patient selection and outcome prediction.	The paper highlights both current and potential future applications of AI in IR, including image analysis, patient management, and real-time decision-making.	AI is enhancing the planning and execution of IR procedures, particularly in managing multimodal data and predicting patient outcomes. AI's ability to support decision-making and improve robotic system capabilities is transforming procedural approaches and increasing procedural autonomy.
[18] How AI and Robotics Will Advance Interventional Radiology: Narrative Review and Future Perspectives	This review discusses the convergence of AI and robotics in advancing interventional radiology, particularly focusing on deep learning, machine learning, and convolutional neural networks (CNNs). It examines how AI and robotics can enhance precision across different specialties in IR.	The focus is on AI and robotics' integration to improve procedural accuracy in IR, particularly in specialties like oncology, neurology, and cardiology.	AI and robotics are improving procedural precision in IR, with deep learning and CNNs automating complex tasks and predicting patient outcomes. The review discusses the potential for AI-driven robotics in IR, highlighting challenges such as integration and the need for standardization.

Table 1. *Cont.*

Study	Brief Description	Focus	AI and Technology Integration
[19] Artificial intelligence in interventional radiology: state of the art	This paper provides an overview of AI applications in interventional radiology, emphasizing how AI supports decision-making, enhances imaging quality, and improves robotic interactions. It also addresses challenges to AI adoption, such as the complexity of IR procedures and the lack of standardization.	The focus is on AI's role in enhancing decision support, imaging, and procedural execution in IR, with a focus on its transformative potential.	AI is significantly contributing to IR by improving decision support systems, enhancing imaging technologies like fusion imaging, and enabling robotic systems to function with greater precision. AI's role in touchless interaction and virtual biopsies is also explored, though adoption is hindered by variability in procedural practices across institutions.
[20] An Introductory Guide to Artificial Intelligence in Interventional Radiology: Part 2: Implementation Considerations and Harms	This article provides a comprehensive look at the risks and challenges of implementing AI in interventional radiology (IR). It discusses potential harms associated with AI deployment, including overreliance on automated systems, and offers a practical checklist for clinicians to safely integrate AI technologies into clinical practice.	The article focuses on the practical considerations necessary for AI integration in IR. It emphasizes understanding and managing the risks involved in deploying AI systems, ensuring clinicians are equipped with strategies for safe implementation. It discusses regulatory oversight and monitoring mechanisms to prevent adverse effects on patient care.	The study presents AI as a tool with significant benefits but also highlights the importance of evaluating AI technologies based on their risk potential. It stresses the need for proper risk assessment and regulatory frameworks to ensure patient safety during AI deployment in clinical settings.
[21] An Introductory Guide to Artificial Intelligence in Interventional Radiology: Part 1 Foundational Knowledge	This foundational guide introduces interventional radiologists to the core concepts of AI. It outlines how AI is applied across different procedural stages, including pre-procedural planning, intra-procedural guidance, and post-procedural monitoring, aiming to enhance diagnostic accuracy and workflow efficiency.	The focus of this article is on educating clinicians about the role of AI in IR, explaining how AI can be used to optimize clinical workflows and decision-making at various stages of procedures. It also introduces a classification system for AI models to help clinicians understand the complexity and suitability of different AI technologies for their practice.	AI is framed as an essential tool for improving clinical outcomes and procedural efficiency in IR. The guide emphasizes the importance of AI across all procedural stages, helping radiologists understand how AI technologies can be effectively integrated into their daily practice.

Table 1. *Cont.*

Study	Brief Description	Focus	AI and Technology Integration
[22] Interventional Oncology: 2024 Update	This review examines the integration of AI in interventional oncology, focusing on its role in imaging, treatment planning, and patient care. It explores AI's potential to enhance tumor detection and personalize treatments, ultimately aiming to improve patient outcomes in oncology.	The focus is on AI's impact on interventional oncology, particularly in improving diagnostic accuracy and tailoring treatment strategies to individual patients. The article also discusses the evolving training requirements for clinicians to handle AI technologies in oncology and their growing importance in clinical decision-making.	AI is integrated with advanced imaging technologies to improve tumor detection and characterization. It also plays a key role in personalizing treatment plans based on genomic data, offering a more individualized approach to oncology care.
[23] Understanding ChatGPT for Evidence-Based Utilization in Interventional Radiology	This study explores the use of ChatGPT, a language-based AI model, to support evidence-based decision-making in interventional radiology. It discusses how this AI tool can assist with patient communication, data interpretation, and clinical decision support.	The focus is on ChatGPT's potential to enhance communication between clinicians and patients, as well as its ability to assist in decision-making processes by providing evidence-based insights and recommendations.	ChatGPT is presented as a valuable tool for streamlining communication in IR, helping clinicians access information quickly and support clinical decision-making by interpreting data and providing relevant recommendations.
[24] Current Applications of Algorithmic Artificial Intelligence in Interventional Radiology: A Review of the Literature	This review provides an overview of the current uses of algorithmic AI in interventional radiology, exploring its applications across various procedural stages. It discusses how AI is used to optimize imaging, assist in decision-making, and predict procedural outcomes.	The focus is on the broad applications of AI in IR, detailing its use in improving imaging quality, selecting patients for procedures, and predicting treatment outcomes. It emphasizes the potential of AI to streamline workflows and enhance the precision of interventions.	AI is used to enhance imaging capabilities, improve diagnostic accuracy, and assist with procedural decision-making. The review highlights AI's growing role in predicting outcomes, reducing errors, and increasing efficiency in IR procedures.
[25] An Interventional Radiologist's Primer of Critical Appraisal of Artificial Intelligence Research	This primer guides interventional radiologists in evaluating AI research, providing criteria for assessing the accuracy, reliability, and clinical applicability of AI models. It helps clinicians critically analyze AI technologies to ensure they meet clinical needs.	The focus is on equipping clinicians with the skills to critically appraise AI research and products. It provides a structured approach to evaluating AI models, focusing on accuracy, data quality, and the relevance of AI applications in clinical practice.	The article highlights the importance of critically assessing AI tools before integrating them into clinical practice. It ensures that clinicians can evaluate AI models based on scientific rigor and clinical relevance, promoting informed decisions on technology adoption.

Table 1. *Cont.*

Study	Brief Description	Focus	AI and Technology Integration
[26] Robotics in Interventional Radiology: Review of Current and Future Applications	This review explores the current and future applications of robotics in IR, particularly in procedures involving image-guided navigation. It examines how AI-powered robotic systems are improving the precision and safety of minimally invasive interventions.	The focus is on the integration of robotic systems in IR, with an emphasis on AI's role in enhancing procedural accuracy. It discusses the potential for robotic systems to transform IR by improving precision, reducing human error, and facilitating complex procedures.	Robotics, combined with AI, is transforming IR by enhancing precision and safety in interventions. AI-driven systems assist in tasks like navigation and decision-making, particularly in minimally invasive procedures, leading to better outcomes and reduced procedural risks.
[27] Artificial Intelligence, Augmented Reality, and Virtual Reality Advances and Applications in Interventional Radiology	This review examines how AI, augmented reality (AR), and virtual reality (VR) are converging to enhance interventional radiology. It highlights the potential of these technologies to improve diagnostics, guide treatment planning, and assist in complex procedures.	The focus is on the intersection of AI with AR and VR, exploring how these technologies can work together to provide clinicians with enhanced visualizations, decision support, and real-time feedback during procedures.	AI, AR, and VR are integrated to provide a more immersive, accurate, and interactive approach to diagnostics and treatment planning in IR. These technologies are improving procedural outcomes by offering real-time insights and visual assistance to clinicians.
[28] Technological Advancements in Interventional Oncology	This article explores the transformative impact of artificial intelligence in interventional oncology, focusing on the integration of AI for the analysis of large datasets and its use in predictive modeling for treatment response. It discusses the evolution of AI applications in diagnosing and treating cancers, highlighting AI's ability to improve precision in tumor detection and treatment planning. The review provides insights into how AI and digital health technologies are enhancing workflows in clinical settings, particularly within interventional oncology.	The paper highlights AI's role in improving diagnostic precision in oncology by analyzing vast medical data, aiding in tumor detection, and optimizing treatment planning. It discusses AI's potential to predict treatment responses and guide clinical decisions, transforming the approach to personalized cancer care and workflow efficiency in oncology interventions.	AI has proven transformative in interventional oncology, enhancing tumor detection and treatment precision. By leveraging data and predictive modeling, AI is improving diagnostic accuracy and clinical decision-making, although challenges remain regarding data quality and clinical adoption.

Table 1. *Cont.*

Study	Brief Description	Focus	AI and Technology Integration
[29] Artificial intelligence in diagnostic and interventional radiology: Where are we now?	This article examines the current state of AI in radiology, specifically in diagnostic and interventional radiology. It explores how AI, especially in conjunction with radiomics, is revolutionizing imaging processes in both diagnostics and interventions. The paper emphasizes how AI technologies are increasingly being used to assist radiologists by providing enhanced image analysis, automating repetitive tasks, and supporting clinical decision-making. Interventional radiology is poised to benefit from AI's ability to enhance procedural guidance, patient selection, and real-time monitoring.	The article discusses the integration of AI in improving image analysis and automating tasks, making diagnostic processes more efficient. It also explores AI's growing role in interventional radiology, including procedural guidance, real-time decision support, and improving patient outcomes.	AI is improving efficiency in radiology by automating tasks and enhancing image analysis. In interventional radiology, AI is increasingly used for procedural guidance, real-time monitoring, and decision-making, though challenges like data quality and integration persist.
[30] Artificial Intelligence in Interventional Radiology	This article reviews the historical development of artificial intelligence in medicine, focusing on its applications in interventional radiology. It traces the evolution of AI from its early conceptual stages to its current role in enhancing diagnostic capabilities and procedural guidance. AI in interventional radiology is mainly being used to assist in tumor detection, automate routine tasks, and improve procedural outcomes. The article also discusses the potential future impact of AI in revolutionizing interventional radiology by enhancing workflow and diagnostic precision.	The paper outlines the historical evolution and current applications of AI in interventional radiology, focusing on its role in tumor detection, automation of routine tasks, and improving procedural outcomes. It also looks at the future potential for AI to further enhance diagnostic precision and workflow efficiency.	AI is significantly enhancing tumor detection, procedural guidance, and workflow automation in interventional radiology. The paper emphasizes the future potential of AI, particularly in supporting real-time decision-making and optimizing clinical outcomes.

Table 1. *Cont.*

Study	Brief Description	Focus	AI and Technology Integration
[31] Applications and challenges of artificial intelligence in diagnostic and interventional radiology	This literature review discusses the implementation of machine learning (ML) and deep learning (DL) techniques in radiology, including their applications in interventional radiology. The paper explores how AI can enhance image analysis, support diagnosis, assist in predicting management decisions, and improve clinical outcomes. It also addresses the challenges of implementing AI in clinical settings, such as the need for accurate, representative datasets and ethical considerations.	This review focuses on the applications of machine learning and deep learning in radiology, especially in interventional radiology. It highlights AI's potential to improve image analysis, diagnostic accuracy, and patient management decisions, while also addressing the challenges posed by data quality and ethical issues in clinical adoption.	AI, particularly machine learning and deep learning, is revolutionizing diagnostic and interventional radiology by improving image analysis and supporting clinical decisions. However, challenges like data quality, training requirements, and ethical concerns hinder widespread implementation.
[32] Prime Time for Artificial Intelligence in Interventional Radiology	This article highlights the growing interest in AI within interventional radiology, focusing on the integration of machine learning and radiomics. It examines the potential applications of AI to enhance procedural planning, image-guided interventions, and patient monitoring. The paper emphasizes the readiness of interventional radiology to lead AI development, given its standardized data formats and data-rich environment. The article also discusses the need for research into overcoming challenges related to limited datasets and the integration of AI into clinical workflows.	The paper discusses the promising applications of AI in procedural planning, image-guided interventions, and patient monitoring within interventional radiology. It highlights the sector's data-rich environment and standardization as factors that could help accelerate AI adoption, though challenges such as limited datasets remain.	AI has the potential to revolutionize procedural planning and patient monitoring in interventional radiology. However, the paper stresses that overcoming challenges like limited datasets and integrating AI into clinical practice are key hurdles that need addressing.

Table 1. *Cont.*

Study	Brief Description	Focus	AI and Technology Integration
[33] Challenges of Implementing Artificial Intelligence in Interventional Radiology	This paper discusses the slow pace of AI adoption in interventional radiology compared to diagnostic radiology, identifying key challenges such as regulatory hurdles, data limitations, and ethical concerns. Despite AI's potential to transform interventional radiology by enhancing procedural efficiency and diagnostic accuracy, the article highlights how these barriers have delayed widespread implementation. It calls for continued engagement with stakeholders to define clinically relevant use cases and prioritize resources effectively.	The article examines the slow adoption of AI in interventional radiology, emphasizing the regulatory and data challenges, as well as ethical concerns. It calls for collaboration among stakeholders to address these barriers and unlock the full potential of AI in this field.	The article identifies key barriers to AI adoption in interventional radiology, including regulatory hurdles, data quality, and ethical concerns. It stresses the need for collaboration to address these challenges and expedite AI integration into clinical practice.
[34] Current and emerging artificial intelligence applications for pediatric interventional radiology	This review examines the potential applications of AI in pediatric interventional radiology, an area that remains underserved in terms of AI technologies. The article discusses how AI can be utilized to improve image analysis, assist in diagnosis, and enhance procedural guidance in pediatric patients. Given the unique challenges of treating children, the paper emphasizes the importance of developing AI tools tailored to the specific needs of pediatric patients.	The article explores how AI can improve diagnostic and procedural outcomes in pediatric interventional radiology. It stresses the need for AI tools tailored to the unique anatomical and treatment needs of pediatric patients.	AI in pediatric interventional radiology holds promise for improving image analysis and procedural guidance, but the article highlights the need for AI tools specifically designed to meet the unique challenges of treating children.
[35] Role of Machine Learning and Artificial Intelligence in Interventional Oncology	This review highlights the role of AI and machine learning in interventional oncology, emphasizing their potential to enhance image analysis, intraprocedural guidance, and diagnostic accuracy.	The study explores how AI improves cancer detection, tumor classification, and treatment selection by predicting outcomes based on clinical and radiologic data. It also discusses challenges such as data processing, model validation, workflow integration, and ethical considerations.	AI and ML contribute to procedural guidance with precise needle tracking and image fusion, minimizing damage to healthy tissue while optimizing tumor treatment.

Table 1. *Cont.*

Study	Brief Description	Focus	AI and Technology Integration
[36] Interventional Radiology ex-machina: Impact of Artificial Intelligence on Practice	This article examines the role of AI in interventional radiology, focusing on its potential to improve minimally invasive treatments and procedural accuracy.	The study explores AI applications in screening, patient selection modeling, predictive treatment planning, and procedural navigation. It also discusses the impact of augmented reality, mixed reality, and virtual reality in enhancing interventional radiology workflows.	AI, ML, and deep learning improve image analysis, procedural precision, and decision-making. Augmented and virtual reality technologies assist in enhancing interventional radiology procedures.
[37] Artificial Intelligence in Interventional Radiology: A Literature Review and Future Perspectives	This literature review discusses AI's impact on interventional radiology, integrating evidence-based insights and expert perspectives to highlight current applications and future directions.	The study focuses on AI's role in lesion detection, segmentation, procedural planning, and interventional oncology. It examines AI's ability to provide prognostic insights and enhance minimally invasive treatments.	AI-driven computational algorithms, machine learning, and artificial neural networks improve diagnostic accuracy, procedural workflows, and personalized treatment planning in interventional radiology.

The integration of AI into interventional radiology IR is rapidly transforming various procedural and operational aspects of the field. AI-driven technologies are enhancing precision, optimizing workflows, and improving patient outcomes across multiple domains of IR. Table 2 provides an overview of key AI applications in IR, categorized by their specific contributions to robotic navigation, oncology interventions, predictive decision support, minimally invasive procedures, image fusion, workflow efficiency, real-time monitoring, and ethical and regulatory considerations.

AI plays a crucial role in robotic and navigation systems, where it enhances the accuracy of complex procedures such as needle insertion, biopsy guidance, and catheter placement ([11,12,18,19]). By analyzing imaging data in real time, AI helps clinicians determine the most precise trajectory for interventions, minimizing risks and improving procedural success. Similarly, AI-driven technologies in oncology interventions contribute to tumor detection, ablation, and post-treatment monitoring, assisting in precise tumor localization and improving follow-up care ([16,20,28,31,32,35]).

Another significant application of AI in IR lies in predictive decision support, where it facilitates outcome prediction and personalized treatment strategies ([17,19,23,33,37]). By analyzing large datasets, AI helps anticipate complications and optimize treatment plans, enhancing clinical decision-making. AI also supports minimally invasive procedures, improving access point localization, biopsy precision, and ablation targeting, which leads to better procedural accuracy and reduced patient recovery times ([13–15,25,34,35]).

In the realm of imaging, AI significantly improves diagnostic accuracy by integrating and enhancing imaging modalities such as CT, MRI, and ultrasound ([11,19,22,26,27,35]). By fusing multiple sources of imaging data, AI provides clinicians with a more comprehensive and precise visualization, aiding in complex diagnoses and treatment planning. Additionally, AI optimizes workflow efficiency by automating routine tasks, scheduling,

and streamlining clinical processes, thereby reducing administrative burdens and allowing physicians to focus on patient care ([21,25,29,36]).

The table below summarizes these key AI applications, highlighting the range of medical procedures they impact and the specific advancements they bring to interventional radiology. The contributions of the reviews are often transversal across categories; however, efforts have been made to identify the dominant areas within each category.

Table 2. Categorization of the studies.

Category	Study Range	Medical Applications	Expanded Medical Applications Description
AI in Robotic and Navigation Systems	[11,12,18,19]	Needle Insertion, Biopsy Guidance, Catheter Placement	AI integration in robotic and navigation systems enhances the precision of complex procedures like needle insertion, biopsy guidance, and catheter placement. By analyzing imaging data in real-time, AI systems help identify the most accurate path for needle insertions and biopsies, minimizing damage to healthy tissue and improving the precision of sampling. In catheter placement, AI-driven systems provide constant adjustments to ensure accurate targeting, improving both procedural outcomes and patient safety. These AI-enhanced robotic systems reduce human error, improve targeting, and optimize real-time decision-making, resulting in fewer complications and faster recovery for patients.
AI in Oncology Interventions	[16,20,28,31,32,35]	Tumor Ablation, Tumor Detection, Post-Treatment Monitoring	In oncology, AI contributes to improved precision in tumor treatments such as ablation and embolization. By providing advanced image processing and automatic segmentation, AI enhances the accuracy of targeting tumors while minimizing damage to surrounding healthy tissue. AI systems also aid in early tumor detection by analyzing various imaging modalities and highlighting potential malignancies. Additionally, AI tools monitor patients post treatment, detecting early signs of tumor recurrence, predicting patient outcomes, and assessing the effectiveness of treatments, thus enabling more personalized follow-up care.
AI in Predictive Decision Support	[17,19,23,33,37]	Outcome Prediction, Personalized Treatment, Decision-Making Support	AI in predictive decision support systems analyzes patient data to predict outcomes, suggest personalized treatments, and assist in real-time clinical decisions. This system provides clinicians with a more precise understanding of potential risks and complications, improving diagnostic accuracy and patient safety. By continuously learning from evolving patient data, AI supports clinicians in making data-driven decisions that adapt to individual patient needs. The predictive capability of AI optimizes treatment plans and allows clinicians to anticipate complications, ensuring the most effective interventions are chosen.

Table 2. *Cont.*

Category	Study Range	Medical Applications	Expanded Medical Applications Description
AI in Minimally Invasive Procedures	[13–15,25,34,35]	Biopsy, Ablation, Access Point Localization	AI significantly improves the precision and safety of minimally invasive procedures by guiding access point localization, enhancing biopsy accuracy, and improving ablation targeting. Real-time AI-driven imaging systems provide immediate feedback, helping clinicians navigate anatomical structures with greater precision, thereby reducing the risk of complications. AI-assisted procedures like biopsies and ablations benefit from increased targeting accuracy, minimizing damage to healthy tissues and leading to faster patient recovery. Additionally, AI technologies optimize pre-procedural planning, allowing for more efficient execution of interventions.
AI in Image Fusion and Enhancement	[11,19,22,26,27,35]	Tumor Detection, Image Integration, Imaging Precision	AI-powered image fusion and enhancement techniques integrate multiple imaging modalities (e.g., CT, MRI, ultrasound) to provide clinicians with a more comprehensive view of a patient's condition. This integration enhances diagnostic accuracy, especially for complex cases such as tumor detection, vascular anomalies, and surgical planning. By merging different imaging sources, AI helps to create more precise, detailed images, enabling more informed decision-making. This fusion is particularly valuable when dealing with difficult-to-diagnose conditions, as it provides a clearer picture of the patient's anatomy, guiding clinicians to the most accurate diagnosis and treatment plan.
AI in Workflow and Efficiency Optimization	[21,25,29,36]	Task Automation, Scheduling, Process Streamlining	AI improves the operational efficiency of interventional radiology by automating routine tasks such as scheduling, data entry, and image processing. This allows healthcare providers to focus more on patient care rather than administrative tasks. AI algorithms can also streamline procedural planning by analyzing patient data and procedural variables to select the most efficient approach, optimizing resource allocation, reducing unnecessary interventions, and ensuring timely patient care. Additionally, AI helps predict patient volume and plan staffing, ensuring better resource utilization and smoother workflow in clinical settings.

3.3. Emerging Opportunities and Challenges

The integration of artificial intelligence (AI) into interventional radiology (IR) promises to revolutionize the field by enhancing various aspects of clinical practice. However, the successful adoption and implementation of AI technologies face several opportunities and challenges, as identified in the studies reviewed. Figure 3 reports a sketch of these opportunities and challenges.

Opportunities in the Application of AI in IR

As reported in Figure 4, the opportunities are as follows:

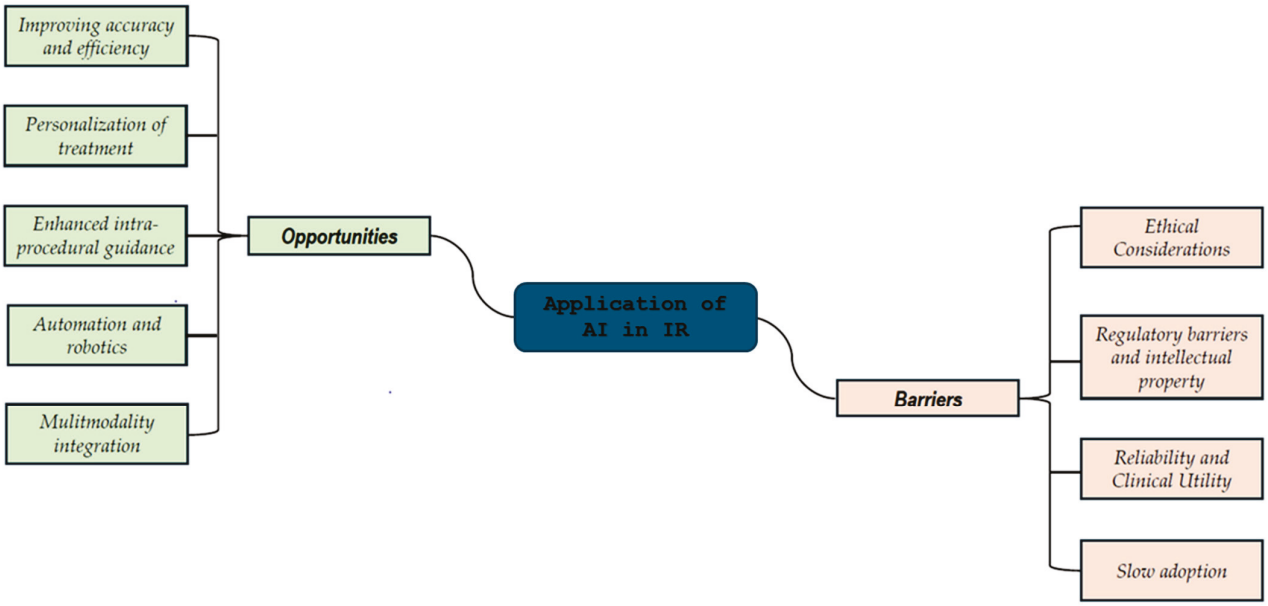


Figure 4. A graphical summary of the opportunities and challenges associated with the application of AI in IR.

Improving Accuracy and Efficiency: One of the most significant opportunities provided by AI in IR is its ability to improve both accuracy and efficiency during procedures. AI systems can assist healthcare professionals by analyzing medical images with remarkable precision, leading to faster diagnosis and more accurate treatment decisions. Studies have shown that AI-based tools can streamline clinical workflows and enhance procedural efficiency by automating tasks [13,17,29]. AI-guided robotic systems can further reduce human error and variability between clinicians, ensuring more consistent outcomes in interventions such as biopsies, ablations, or embolizations. Additionally, AI can speed up diagnosis by suggesting personalized interventions, thus minimizing procedural time and improving patient throughput [34].

Personalization of Treatment: The application of AI in IR offers the possibility of highly personalized treatment strategies. Through machine learning algorithms, AI can process vast amounts of clinical and radiological data to predict the likely outcomes of various interventions, enabling clinicians to select the most appropriate treatment for each patient [17,22,26]. This data-driven approach not only improves the effectiveness of the chosen interventions but also helps in minimizing side effects, optimizing patient outcomes, and reducing the burden on healthcare resources [35]. Personalized treatment planning is particularly useful in oncology, where AI can assist in identifying the most effective tumor treatment options based on the specific characteristics of each case.

Enhanced Intra-Procedure Guidance: AI-driven technologies are transforming intra-procedural guidance by providing real-time support to clinicians. Advanced imaging techniques such as needle tracking and real-time image fusion, powered by AI, enable highly targeted interventions that minimize damage to healthy tissues and enhance the effectiveness of treatments like tumor ablation or embolization [22,35]. Furthermore, emerging AI tools like reinforcement learning and haptic vision are enabling robotic systems to adapt in real-time to feedback from the environment, significantly improving the precision of procedures [12]. These technologies ensure that interventions are both safer and more efficient, contributing to better patient outcomes.

Automation and Robotics: AI is revolutionizing the role of robotics in IR by allowing for greater automation of complex procedures. Robotic systems that are controlled and

guided by AI can perform tasks with a higher degree of accuracy and reliability compared to human hands alone [11,12,17,19]. This includes not only assisting in navigation during image-guided interventions but also potentially performing fully autonomous tasks. The introduction of autonomous systems could redefine the role of the interventional radiologist, shifting from being the primary operator to overseeing and monitoring the robotic technology [26]. Such advancements could lead to the development of fully automated procedures, reducing human intervention and error while improving procedural outcomes.

Multimodality Integration: AI's capacity to integrate data from multiple imaging modalities (such as CT, MRI, ultrasound, and more) is one of its most promising features in IR. The ability to combine various types of imaging data allows for a more comprehensive understanding of a patient's condition and facilitates better-informed decision-making. AI can synthesize these data to create more accurate models for patient selection, prediction of outcomes, and planning of treatments [17,26,29] This multimodal integration enhances the accuracy of diagnosis and treatment planning, particularly in complex cases where a single imaging modality may not provide sufficient information.

Barriers to AI Integration in IR

Despite the tremendous opportunities AI offers in IR, several barriers remain that need to be addressed to ensure its widespread adoption and effective implementation *as reported in* Figure 4.

Ethical Considerations: The introduction of AI into IR raises significant ethical concerns. One major challenge is the lack of transparency in AI decision-making processes. Many AI systems operate as "black boxes", meaning that the logic behind their recommendations is not always fully understood by clinicians. This lack of transparency can undermine trust in AI-driven decisions, especially when these systems influence patient outcomes. Additionally, ethical concerns regarding data privacy and the use of patient information persist, as AI systems require large, accurately labeled datasets to function effectively [27,31,33,35]. Ensuring that these datasets are representative and unbiased is crucial to avoid perpetuating health disparities.

Regulatory Barriers and Intellectual Property: Given the issues that emerge as AI technologies continue to evolve, regulatory frameworks must keep pace to ensure patient safety and system effectiveness. In many countries, the regulation of AI-driven medical devices is still developing, creating uncertainty about the approval and implementation of these systems in clinical practice. The need for rigorous evaluation and classification of AI technologies based on risk level is essential to ensure that AI tools meet safety standards before being deployed in the clinical setting [25,33]. Furthermore, intellectual property (IP) concerns surrounding AI algorithms and the data they use can complicate the development and deployment of AI systems in healthcare. Conflicts over the ownership of AI technologies and the data they generate could delay the integration of these innovations in IR.

Reliability and Clinical Utility: Many AI-based systems are still in the research phase and have not yet been fully validated for clinical use. Although AI has demonstrated promising results in controlled studies, the consistency and reliability of these systems in real-world clinical settings remain uncertain. Further research is needed to assess the clinical utility and long-term benefits of AI technologies in IR. Until more extensive clinical trials are conducted and these systems prove their reliability, widespread adoption may be hindered [16,19,31]. Clinicians need to be confident in the reliability of AI tools before incorporating them into routine practice.

Slow Adoption: The pace of AI adoption in IR is slower compared to diagnostic radiology, where AI technologies have already been more widely accepted and integrated into clinical workflows. The integration of AI in IR faces additional complexities, such as the need for specialized training, changes to clinical workflows, and the integration

of AI with existing technologies. Furthermore, concerns about the cost and the need for extensive validation studies have slowed down the widespread use of AI in interventional procedures [17,19,27,33]. While AI is making strides in diagnostic radiology, it is only beginning to establish itself in IR, and a more gradual adoption process is expected.

Overall, the integration of AI into interventional radiology offers significant opportunities to improve accuracy, efficiency, and patient outcomes. However, challenges such as ethical considerations, regulatory hurdles, and concerns regarding the reliability and clinical utility of AI systems must be addressed to ensure their successful implementation. By overcoming these barriers and continuing to advance AI technologies, the future of IR holds immense potential for transforming patient care and procedural efficiency. The figure accompanying this discussion outlines the primary opportunities and barriers identified in the studies reviewed.

4. Discussion

The discussion is structured into five sections, each aimed at providing a comprehensive analysis and interpretation of the study's findings:

Section 4.1: This section reports a synoptic of the organization of the discussion detailing the rationale based on the output of the results.

Section 4.2: This section presents the key evidence emerging from the overview, outlining the added values of this research. This section establishes the foundation for understanding the contributions of the study and situates them within the broader context of the field.

Section 4.3: This section delves into the positioning of the study within the broader and more comprehensive field of radiology. It aims to contextualize the research by highlighting how it contributes to, and interacts with, the ongoing advancements and trends in the radiology sector.

Section 4.4: As a natural progression from the emerging opportunities and barriers outlined earlier in the NRR, this section examines trends within recent cutting-edge primary studies. It assesses their contributions and evaluates how these groundbreaking works align with the broader themes and insights identified in the review.

Section 4.5: This section focuses on the global positioning of various organizations and regulatory bodies in terms of their guidelines and recommendations. It examines how international institutions and authorities influence the practice of radiology and interventional procedures, providing a perspective on how standardization, ethical considerations, and best practices in the field are faced.

Finally, Section 4.6 identifies the boundaries of the overview and reports the limitations.

This structured approach ensures a detailed and holistic discussion, providing both a synthesis of findings and a forward-looking perspective that connects the study's insights with practical implications and future research directions. Each section is designed to build on the previous one, creating a cohesive narrative that reflects the complexity and potential of the field.

4.1. Synoptic Diagram

Figures 5 and 6 present two synoptic diagrams that outline the rationale behind the design of the **narrative review of reviews** (**NRR**). These diagrams provide a structured visual representation of how the study was developed, showing the logical sequence of its different phases and how they interconnect.

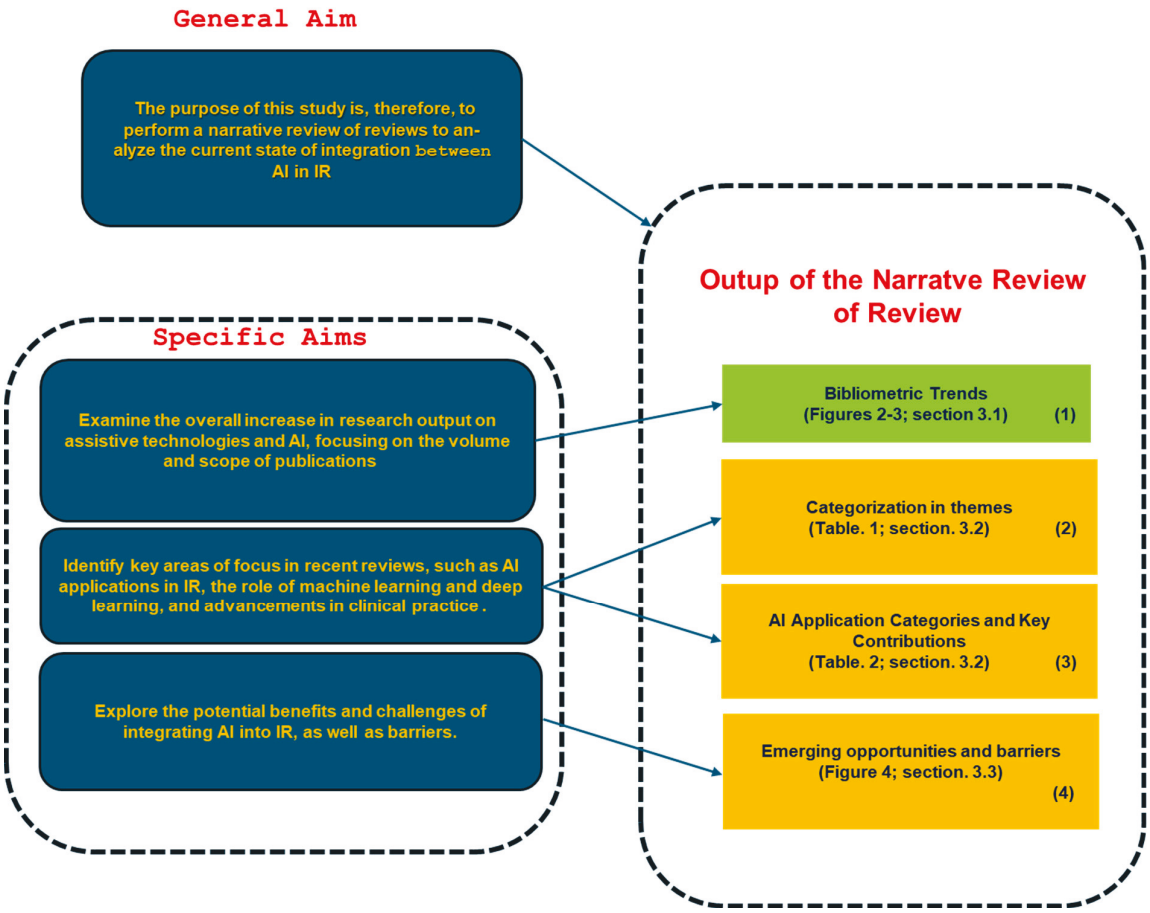


Figure 5. Synoptic diagram of the study structure: from bibliometric trends to AI application categorization and comparative analysis.

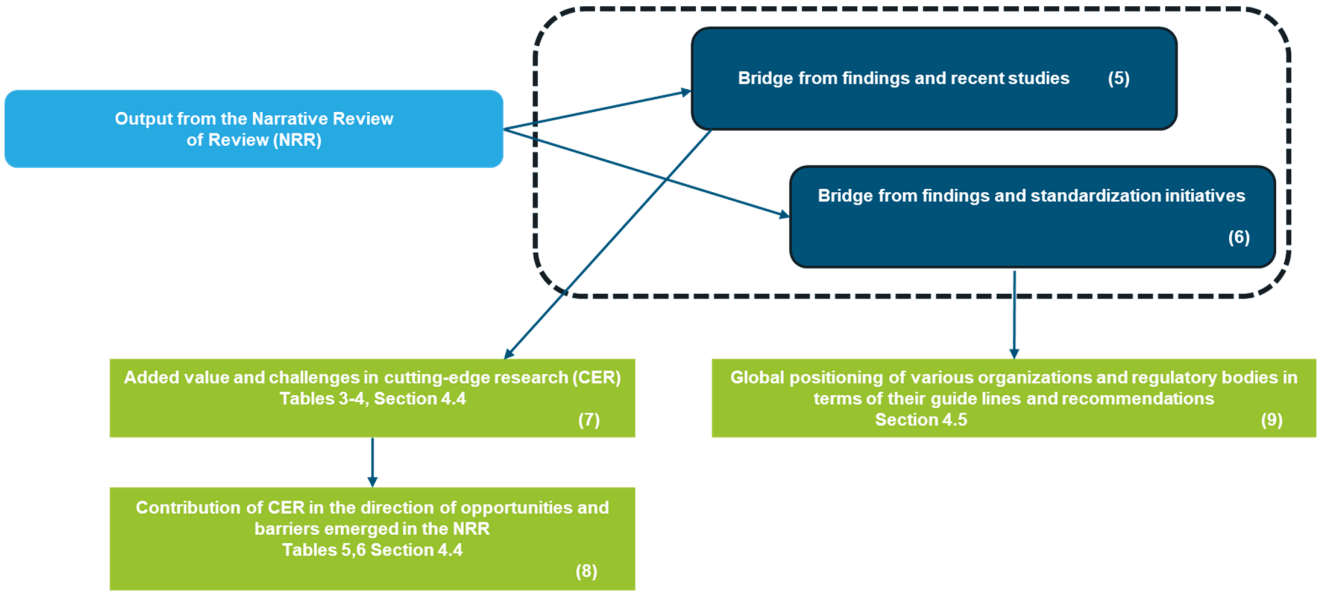


Figure 6. Logical progression from NRR findings to the CER contribution and the international path to standardization.

First Diagram (Figure 5): Linking Objectives to Analysis

The first diagram (Figure 5) illustrates how the study was structured based on its **general objective** and **three specific objectives**. The logical progression follows a top-down approach:

- **Block 1**: This block represents the **bibliometric trends** reported in Figures 2 and 3 (Section 3.1). These trends were analyzed to provide an overview of the scientific production on AI applications in IR, forming the foundation for the subsequent steps of the study.
- **Block 2**: This block corresponds to the **categorization of studies by thematic areas**, as presented in Table 1 (Section 3.2). This categorization allowed for the organization of the reviewed studies according to key themes, facilitating a structured comparison.
- **Block 3**: Building upon the thematic categorization, this block highlights the **comparative side-by-side analysis** of the studies. The classification of AI applications was refined, as reported in Table 2 (Section 3.2), enabling a deeper understanding of the different ways AI is being applied across multiple domains.
- **Block 4**: This block synthesizes the **opportunities and barriers** identified in the reviewed studies, as reported in Figure 4 (Section 3.3). These findings highlight both the potential benefits of AI applications and the barriers

This diagram provides a **step-by-step visualization** of the study's methodological process, from bibliometric analysis to thematic categorization, comparative analysis, and the identification of emerging opportunities and challenges.

Second Diagram (Figure 6): Connecting Discussion to Findings

The second diagram (Figure 6) is logically connected to the first and illustrates how the study transitions from the findings of the literature review to the discussion. The sequential organization follows a structured approach:

- **Blocks 5 and 6** identify the need to bridge the gap between cutting-edge research (CER) and the current state of standardization.
- Connected to **Block 5**, two blocks highlight the contributions of CER in terms of added value and emerging challenges in Tables 3 and 4 of Section 4.4 (**Block 7**), as well as its complementation in terms of opportunities and barriers identified in the NRR reported in Tables 5 and 6 of Section 4.4 (**Block 8**)
- Connected to **Block 6**, **Block 9** reports in Section 4.5 the global positioning of various organizations and regulatory bodies in terms of their guidelines and recommendations.

4.2. Highlights

The integration of AI with interventional radiology (IR) has seen a significant increase over the last five years, largely influenced by the COVID-19 pandemic (Figure 2). This period also witnessed an expansion in review literature, which identifies established themes, highlights areas for further development, and presents emerging trends. Reviews are instrumental in providing an overview of the field, offering valuable insights into both the current state and the future potential of AI applications in IR.

Added Values from the Overview of Reviews:

Identification of Established Areas and Themes: Reviews have consistently highlighted AI's growing role in improving procedural precision through integration with robotic systems for tasks like catheter manipulation, needle placement, and tumor detection. For instance, Cornelis et al. [11] emphasized that while automation and AI integration in robotic systems for image-guided interventions have advanced, no system has yet achieved full autonomy. Key metrics such as the level of autonomy in surgical robotics (LASR) and the levels of integration of advanced imaging and AI (LIAI2) scale help identify gaps and guide future developments.

Boeken et al. [12] further discussed how AI-driven robotic systems are enabling more precise and reliable procedural outcomes, especially in complex tasks like catheter manipulation and needle placement. They identified techniques such as reinforcement learning and haptic vision as vital for improving robot adaptability based on real-time

feedback, advancing IR toward a model where the human role transitions to supervision rather than direct performance.

Identification of Areas Needing Further Expansion: While AI has made strides in several areas of IR, certain technologies, particularly those involving decision support, automation, and multimodal imaging, remain largely in the research phase. For instance, Matsui et al. [16] reviewed AI's applications in interventional oncology, including automatic segmentation, treatment simulation, and outcome prediction. However, most AI-based methods remain experimental and are not yet fully implemented in clinical settings. Moreover, Lesaunier et al. [17] emphasized the need for further studies to enhance the reliability and clinical utility of AI, particularly in personalized treatment approaches such as individualized care planning.

A notable challenge pointed out by Zhang et al. [18] is the slow clinical adoption of AI technologies in IR, which may be attributed to regulatory barriers, ethical concerns, and the limited availability of sufficiently large and accurately labeled datasets for effective machine learning.

Emerging Trends and Future Directions: Reviews also highlight promising developments in AI integration, such as reinforcement learning and real-time image fusion. For example, Lee et al. [14] explored the potential of smartphones' built-in sensors for virtual and augmented reality applications in IR, which could significantly enhance needle-based procedures like biopsy and ablation. These technologies could lead to advancements in intra-procedural guidance, enabling more precise and targeted interventions with minimal patient risk.

Additionally, the potential for AI in personalized treatment strategies is becoming increasingly apparent. As AI technologies evolve, they are expected to play a significant role in predicting treatment outcomes and optimizing care plans. Geevarghese et al. [22] highlighted AI's ability to accelerate tumor detection and guide treatment strategies, furthering personalized medicine in IR.

Challenges in Clinical Implementation: Ethical considerations, regulatory barriers, and the need for reliable, clinically validated AI systems are common themes across the reviews. As noted by Warran et al. [20,21], there are concerns regarding the transparency of AI decision-making processes, and the need for appropriate validation of AI systems in clinical environments is critical. The slow adoption of AI in IR, as pointed out by Zhang et al. [18] and Mazaheri et al. [33], is another key challenge that needs to be addressed through further research, regulation, and technology deployment strategies.

Multidisciplinary Opportunities and Future Integration: The integration of AI, robotics, and multimodal imaging technologies holds the potential to revolutionize IR. This multidisciplinary approach will likely lead to enhanced treatment planning, real-time monitoring, and decision support systems that will complement the work of radiologists and improve patient outcomes. Posa et al. [28] emphasized the importance of engaging physicians in the development and application of AI, particularly in the field of interventional oncology, to ensure that AI technologies are effectively incorporated into clinical practice.

4.3. Navigating AI in Interventional Radiology: Unique Complexities and Future Directions in Comparison to Diagnostic Radiology

Radiology, including interventional radiology as a specialized branch, has seen significant advancements in its intersection with AI compared to other fields, such as digital pathology [38,39], where delays in digital standardization have also hindered AI development.

However, when we focus specifically on interventional radiology, a clear distinction emerges in how AI is being integrated compared to diagnostic radiology [40]. The role of AI differs significantly between these two branches.

In diagnostic radiology, AI primarily serves as a support tool for analyzing medical images. Advanced deep learning algorithms can detect abnormalities, such as tumors, fractures, or degenerative diseases, in images obtained from CT scans, MRI, or X-rays. This not only accelerates the diagnostic process but also reduces the risk of errors, improving both accuracy and efficiency for radiologists. Additionally, AI contributes to anatomical segmentation, image enhancement, and even disease progression prediction, further augmenting radiologists' capabilities.

In contrast, interventional radiology presents a more complex and active role for AI. Here, AI can assist physicians during minimally invasive, image-guided procedures by improving real-time imaging, aiding in the navigation of catheters and needles, and suggesting the safest and most efficient procedural approaches. In some cases, AI even has the potential to automate certain instrument movements, reducing the margin of error and improving patient outcomes.

Furthermore, interventional radiology integrates robotics, catheterization techniques, and percutaneous dosing methods, in addition to imaging. As a result, AI in this field requires stricter safety measures, particularly due to its real-time procedural implications. These considerations extend beyond clinical safety to include cybersecurity challenges, as AI-driven interventional technologies must ensure robust protection against potential cyber threats.

All of this contributes to a delay in standardization, as the complexity of real-time AI integration in interventional radiology requires more rigorous validation, regulatory approvals, and cybersecurity measures compared to diagnostic AI applications, as highlighted in several studies analyzed [11,12,17]. Studies have pointed out that the dynamic and high-risk nature of interventional radiology (IR) procedures introduces additional challenges in ensuring the reliability and safety of AI systems during live interventions [13,19]. The need for thorough validation in clinical environments, where AI-guided systems interact with multiple variables such as patient anatomy and real-time imaging, makes the integration process more complex than in diagnostic radiology, where AI typically works with static images and less variable conditions [17,25]. Moreover, ensuring that AI systems meet stringent regulatory and cybersecurity standards is essential to protect patient safety and prevent potential breaches in sensitive healthcare data, a concern raised in multiple studies [27,31].

On the comparative side of scientific production, a search on PubMed using the search key 3 in box 1 [41] reveals a substantial surge in interest in AI in diagnostic radiology starting in 2018 (with the exception of a few sporadic initiatives). A total of 126 studies were identified, 111 of which were published in the past five years, accounting for 88.1% of the total, with 34 of these being review articles. This dissemination activity is comparable to that in interventional radiology, which saw 114 studies, 109 of which were published in the last five years. Therefore, research trends show a significant rise in AI publications in both diagnostic and interventional radiology, especially in the past five years and during the pandemic. The high number of review articles reflects growing interest in AI applications, with both fields demonstrating similar dissemination patterns.

What sets them apart is the differing vision on the future development of interventional radiology [42]. On one hand, there is a strong belief within the scientific community that the future of interventional radiology should remain closely aligned with diagnostic radiology, fostering collaboration and integration [43]. On the other hand, some experts argue for a more independent trajectory for interventional radiology, emphasizing its distinct role and specialization [44]. This divide reflects ongoing debates on how best to position the field in relation to other radiological practices.

4.4. AI in Interventional Radiology: Pioneering Studies Tackling Challenges, Exploring Opportunities, and Driving Key Innovations

Recent studies [45–62] have also been considered (as well as valuable editorial mapping of the route) to explore how they align with the challenges identified in the overview of reviews. The integration of AI into IR offers promising advancements, but it also brings a set of significant challenges that need to be addressed thoughtfully. The potential for AI to enhance precision, optimize patient outcomes, and revolutionize treatment planning is undeniable. However, this integration also requires careful consideration of multiple factors to ensure its success and avoid unintended consequences.

One of the first challenges lies in data availability, particularly when it comes to deep learning (DL) models in fluoroscopy-guided interventions (FGIs). Obtaining large datasets of labeled clinical images for training AI models has traditionally been difficult. To address this, the use of synthetic data has been proposed, generated from advanced imaging techniques like CT angiography. While this approach promises to alleviate the data scarcity issue, it raises questions about the accuracy and reliability of synthetic images. AI models trained on synthetic data must be rigorously validated to ensure they can effectively support real-world clinical decision-making [46].

Another important challenge is the acceptance of AI in clinical practice, particularly in areas where it directly impacts patient communication and informed consent. Studies have shown that while AI tools, such as ChatGPT, can provide accurate and understandable information, they often fall short in offering comprehensive responses. This raises concerns among clinicians about relying on AI-generated outputs for critical tasks such as discussing risks and benefits with patients. Ensuring AI systems can deliver more than just surface-level answers and provide in-depth, context-sensitive information is crucial before they can be fully embraced in clinical settings [47].

Moreover, the integration of AI in IR demands a focus on continuous education and training. As AI technologies evolve rapidly, IR professionals must keep pace with these advancements, equipping themselves with the skills necessary to harness the full potential of these innovations. The challenge, however, is not just about acquiring technical knowledge but also about maintaining an ergonomic approach to minimize the risk of fatigue or strain from using new AI tools. Furthermore, there is a need to consider the economic implications of adopting AI. Innovations must not only improve patient care but also be economically viable to ensure widespread adoption and sustainability within healthcare systems [48]. AI's potential to disrupt the field of IR brings both opportunities and risks. While these innovations can greatly improve efficiency and patient outcomes, they also present a challenge in adapting to and managing technological disruptions. Interventional radiologists must not only accept these changes but actively engage in shaping how AI is developed, trained, and implemented. This requires leadership, collaboration, and an ongoing commitment to understanding the broader implications of AI on clinical practice and patient care [49]. The integration of AI into IR, while transformative, requires careful navigation to address challenges related to data, clinical practice, education, and economic feasibility. As AI continues to evolve, interventional radiologists must lead this change, balancing technological innovation with a focus on improved patient care and operational sustainability [50]. While technologies such as large language models (LLMs) and robotic systems show promise in enhancing procedural precision and patient understanding, there are still obstacles that need to be addressed for optimal clinical implementation.

One of the main challenges is the dependency on linguistic context, as demonstrated in a study comparing ERNIE Bot and ChatGPT in answering questions about liver cancer interventional radiology. Both technologies showed potential, but their performance varied significantly depending on the language. ERNIE Bot outperformed ChatGPT in the Chinese

context, while ChatGPT performed better in English [51]. This discrepancy suggests that LLMs must be adapted not only to the language but also to the specific cultural and healthcare contexts to be truly effective. Improving the ability of AI models to understand and respond accurately to medical questions in different languages and contexts remains a key challenge.

Furthermore, while robotic systems have demonstrated improvements in precision and efficiency during CT-guided procedures, there are still substantial hurdles to overcome. A robotic system designed for lesion detection and path planning showed reduced insertion time and improved accuracy compared to traditional freehand techniques [52]. However, integrating advanced robotic systems into clinical procedures requires overcoming technological challenges related to calibration, safety, and reliability of the devices. Additionally, the complexity of adapting robots to various patient types and clinical situations presents an additional obstacle to their widespread adoption.

Another challenge concerns the use of AI to simplify radiology reports and make information more accessible to patients. Although radiologists recognize the potential of AI in report simplification, most of them indicated that human oversight is necessary. In one study, the majority of radiologists expressed a preference for AI-generated simplifications to be reviewed by professionals [54]. This highlights a reluctance to fully rely on AI without adequate supervision, especially in a field where precision is crucial. The challenge is to find a balance between automation and human control that ensures optimal quality and safety.

Additionally, while deep learning models have shown high accuracy in identifying anatomical locations in digital subtraction angiography (DSA), further refinement is needed. Despite improving location classification in angiographic sequences, challenges remain related to reliability and the ability to handle complex or atypical cases [53]. The difficulty in managing non-standard or uniquely complex cases is one of the main barriers to the widespread adoption of AI in interventional radiology.

AI has also shown potential in automating collimation in interventional radiology, an area that is often underutilized due to its complexity. Collimation is essential for improving image quality and minimizing radiation exposure, but it requires manual adjustments that can be time-consuming and vary widely across procedures and operators. A tunable algorithm for automatic collimation, as demonstrated in a study, showed promising results in reducing radiation exposure while meeting the varying preferences of interventionalists [55]. Despite its potential, the challenge lies in ensuring that AI-driven solutions can be seamlessly integrated into fast-paced, high-stakes environments without disrupting the workflow or procedure quality.

AI's role in advancing diagnostic and therapeutic techniques, such as in the analysis of parapharyngeal space (PPS) lesions, has also shown promise. Advanced imaging, supported by AI algorithms, allows for precise lesion diagnosis and planning of interventions, such as CT-guided biopsies [56]. However, AI's ability to accurately differentiate between pathologies still requires further validation, especially in real-world clinical settings. While AI has contributed significantly to tumor segmentation, lymph node metastasis detection, and pathological classification, challenges remain in its application across all head and neck pathologies, particularly in rare or complex cases. In the specific field of Interventional Oncology (IO), there have been numerous advancements in recent years, showing immense potential for transforming cancer treatment and management. However, IO faces unique challenges as it bridges the gap between diagnostic radiology and clinical oncology. While promising, these advancements must overcome difficulties in integration with clinical workflows, as well as address the accuracy of AI and robotic systems in highly specialized cancer treatments [45].

The use of LLMs like GPT-4 and Claude-3-Opus to simplify IR reports has shown promising results in terms of readability and clinical relevance. However, errors that could undermine patient trust in the information provided have been observed [57]. Trust-breaking errors, such as those that may lead to confusion or misunderstanding, are a major concern when using AI in healthcare. To be adopted in real-world clinical settings, these technologies must be refined to minimize errors and ensure clear, accurate communication. Further recent studies highlight progress in areas such as immunotherapy, patient education, safety data analysis, exam preparation, and advanced imaging techniques. In cancer treatment, immunotherapy, particularly intratumoral approaches, represents a paradigm shift. By integrating machine learning algorithms and radiomics, AI allows for the analysis of complex imaging data to refine diagnoses, predict treatment responses, and tailor therapeutic strategies. These advancements have the potential to improve patient survival and quality of life while minimizing side effects, potentially reshaping clinical guidelines. However, further clinical studies are needed to confirm the efficacy of these promising approaches [58]. Patient education is another area where AI shows great promise. The use of GPT-4 to generate easily understandable educational material for interventional radiology procedures has yielded positive results. The AI-generated instructions were found to be more readable and accessible compared to traditional sources, improving patient comprehension and satisfaction. This approach can significantly enhance health literacy and patient-centered care, making it easier for patients to understand their treatment options and procedures. However, further work is needed to refine these materials for more complex procedures and to ensure consistency and accuracy across different types of interventional radiology treatments [59].

AI is also making strides in safety data analysis in IR. GPT-4 has been used to classify and summarize microwave ablation device safety event data accurately. The AI model demonstrated high accuracy in categorizing failures and generating meaningful summaries, helping clinicians interpret large volumes of safety data more effectively. This ability to streamline data processing could improve the efficiency and safety of IR procedures. Nevertheless, additional validation across a wider variety of devices and event types is necessary to improve the generalizability of these findings and address the limitations of the model when dealing with more diverse datasets [60].

Moreover, GPT-4 has been tested in medical exam preparation, particularly for the European Board of Interventional Radiology (EBIR) exam. The AI was able to generate exam questions at various difficulty levels, assisting medical students and professionals at different stages of their careers. The results showed that GPT-4 could help train medical students and interventional radiologists by generating relevant exam items, thereby aiding in exam preparation and professional development. One area of improvement, however, lies in refining the AI's ability to generate items with varying degrees of difficulty to better match the knowledge and experience levels of a diverse audience, ensuring that it meets the educational needs of both beginners and experts [61].

In advanced imaging, AI-enhanced techniques have been developed to generate synthetic contrast-enhanced images from non-contrast CT scans, particularly in the context of renal cancer treatments. This approach reduces the reliance on iodine-based contrast agents, which can be associated with adverse reactions and higher radiation doses. The AI models demonstrated competitive image quality and improved performance in downstream tasks, such as image segmentation, offering significant benefits for interventional procedures. While promising, these techniques would benefit from further refinement in handling various anatomical structures and improving the accuracy of image enhancement across diverse patient populations [62].

Table 3 highlights key advancements in AI technologies in interventional radiology. These innovations are improving precision, simplifying processes, and advancing diagnostic and therapeutic practices. However, challenges still remain for full clinical implementation.

Table 3. Key advancements in AI applications in interventional radiology.

Advancements	Description	Related Studies
AI Integration for Improved Precision	AI has demonstrated significant potential in enhancing precision, optimizing patient outcomes, and revolutionizing treatment planning in IR.	[46–50]
AI in Robotic Systems	Robotic systems have shown improvements in precision and efficiency, particularly in lesion detection and path planning during CT-guided procedures.	[52]
AI Simplification of Radiology Reports	AI simplifies complex radiology reports, improving accessibility and readability for both clinicians and patients, but requires oversight to ensure quality.	[54]
AI in Parapharyngeal Space (PPS) Lesion Diagnosis	AI-assisted imaging has shown potential for precise diagnosis of PPS lesions and improved intervention planning, such as for CT-guided biopsies.	[56]
Advancements in Interventional Oncology (IO)	IO innovations are transforming cancer treatment, bridging diagnostic radiology and oncology, though challenges remain in integration and accuracy.	[45]
AI in IR Report Simplification and Accuracy	Large language models like GPT-4 and Claude-3-Opus simplify IR reports for better readability, though errors need to be minimized to ensure trust and clarity.	[57]
Linguistic Context Dependency	AI's performance varies based on language and cultural context, impacting its application in liver cancer IR.	[51]
Deep Learning Model Refinement	While deep learning models aid anatomical identification, they face challenges in handling complex or atypical cases.	[53]
Automating Collimation	AI-driven collimation algorithms reduce radiation exposure and enhance image quality, but integrating them into fast-paced clinical workflows remains difficult.	[55]
Immunotherapy and Intratumoral Approaches	AI, through machine learning and radiomics, refines diagnosis and predicts treatment responses, optimizing personalized care in oncology.	[58]
Patient Education and Communication	GPT-4 aids in simplifying explanations of IR procedures, improving patient understanding and health literacy.	[59]
AI in Safety Data Analysis	GPT-4 categorizes and summarizes safety event data from IR devices, streamlining data analysis for improved patient safety.	[60]
Training and Exam Preparation	GPT-4 generates exam items for medical professionals, supporting EBIR exam preparation and training for various expertise levels.	[61]
Improved Imaging Techniques	AI-enhanced imaging generates synthetic contrast-enhanced images from non-contrast CT scans, improving image quality and reducing contrast agent use.	[62]

Table 4 outlines the main challenges in integrating AI into interventional radiology. These challenges range from data availability to the need for continuous education and training for professionals. Addressing these issues is crucial for the successful application of AI in clinical settings.

Table 4. Challenges in the integration of AI in interventional radiology.

Challenges	Description	Related Studies
Data Availability	Difficulty in obtaining large datasets of labeled clinical images for AI model training, especially for fluoroscopy-guided interventions (FGIs). Synthetic data could help, but it needs rigorous validation.	[46]
Acceptance in Clinical Practice	AI's ability to provide comprehensive, context-sensitive information for patient communication and informed consent remains a challenge. AI-generated outputs need more than surface-level responses.	[47]
Continuous Education and Training	IR professionals must keep up with rapidly evolving AI technologies, ensuring they can use these innovations effectively and ergonomically. Economic feasibility of AI adoption is also a key concern.	[48]
Linguistic Context Dependency	AI's performance varies based on language and cultural context. For instance, ERNIE Bot and ChatGPT had differing performances in liver cancer IR depending on the language (Chinese vs. English).	[51]
Robotic Systems Integration	While robotic systems enhance precision in CT-guided procedures, challenges related to calibration, safety, and adaptability to different patients remain.	[52]
Radiology Report Simplification	AI can simplify radiology reports, but human oversight is often preferred by radiologists to ensure accuracy and quality, reflecting hesitance to fully trust AI.	[54]
Deep Learning Model Refinement	While deep learning models are useful for anatomical identification, issues persist with handling complex or atypical cases, requiring further refinement.	[53]
Automating Collimation	AI-driven collimation algorithms can reduce radiation exposure and improve image quality, but their integration into fast-paced clinical workflows remains challenging.	[55]
AI in Parapharyngeal Space (PPS) Lesions	AI shows potential in lesion diagnosis and intervention planning but needs further validation to ensure accuracy, especially in complex or rare cases.	[56]
Trust and Accuracy of AI Systems	While LLMs like GPT-4 and Claude-3-Opus show promise in simplifying IR reports, trust-breaking errors need to be addressed for clinical implementation.	[57]
Advancements in Interventional Oncology (IO)	IO innovations are transforming cancer treatment, bridging diagnostic radiology and oncology, though challenges remain in integration and accuracy.	[45]
Navigating Technological Change in IR	The integration of AI in IR requires leadership, collaboration, and an ongoing commitment to understanding its broader implications for patient care and clinical practice.	[50]
Data Quality and Consistency in Patient Education	Despite positive results in using AI for patient education, challenges remain in ensuring that AI-generated content is accurate, consistent, and suitable for all patients.	[58]
AI Reliability in Data Classification	AI models like GPT-4 for safety event data classification are effective, but there are concerns about the model's reliability when handling complex or rare data.	[59]
AI Integration into Medical Education	AI's role in medical exam preparation is promising, yet further validation is required to ensure that AI-generated questions are clinically relevant and effective for training purposes.	[60]

Table 4. *Cont.*

Challenges	Description	Related Studies
Image Quality and Safety in Advanced Imaging	While AI can generate synthetic contrast-enhanced images, ensuring their quality and clinical validity without causing patient harm remains a significant challenge.	[61]
Safety and Efficacy of AI in High-Stakes Procedures	AI models for synthetic contrast imaging are beneficial, but there is a need to address concerns about safety, accuracy, and clinical efficacy in high-stakes procedures like renal cancer treatments.	[62]

It is particularly interesting to observe how the contributions of these cutting-edge research studies align, in some cases, directly with the opportunities and barriers outlined in Figure 4. This convergence highlights the relevance of recent advancements in artificial intelligence (AI) to the broader discussion on its integration into interventional radiology.

Tables 5 and 6 provide a structured comparison between the cutting-edge research (CER) and the narrative review of reviews (NRR), mapping their respective contributions to the identification of key opportunities and challenges. By analyzing these contributions side by side, it becomes possible to assess how different methodological approaches—one focused on direct evidence synthesis and the other on a broader synthesis of existing reviews—complement each other in providing a comprehensive perspective on AI implementation in interventional radiology.

This comparison underscores the need for a balanced approach that not only identifies the technical and clinical advantages of AI integration but also acknowledges the potential barriers related to regulatory frameworks, clinical acceptance, workflow adaptation, and long-term sustainability. By leveraging insights from both CER and NRR methodologies, we can obtain a more nuanced understanding of how AI can be effectively deployed while addressing the critical challenges associated with its adoption.

Table 5. Opportunities from NRR and contributions from CER.

Opportunities form NRR	Description of Contribution from CER	References
1. Improving Accuracy and Efficiency	AI plays a key role in enhancing diagnostic accuracy and operational efficiency within interventional radiology (IR). Deep learning models, for example, have been shown to improve the accuracy of anatomical location classification in imaging techniques like digital subtraction angiography (DSA). Additionally, the integration of AI into robotic systems for lesion detection and path planning helps ensure more precise targeting, reducing errors, and increasing procedural success rates. These advancements collectively improve patient outcomes and workflow efficiency.	[52,53]
2. Personalization of Treatment	AI is significantly advancing personalized treatment planning, particularly in interventional oncology (IO). AI's ability to improve tumor segmentation and detect lymph node metastasis allows for more tailored and accurate treatment strategies that are customized to individual patient profiles. For instance, AI is used to refine treatment planning for procedures like CT-guided biopsies, ensuring better-targeted interventions that align with the patient's specific condition. This approach is vital for enhancing patient care and improving outcomes, especially in complex or rare cases.	[56]

Table 5. *Cont.*

Opportunities form NRR	Description of Contribution from CER	References
3. Enhanced Intra-Procedural Guidance	AI is crucial for improving intra-procedural guidance in IR. For example, robotic systems used in procedures like CT-guided interventions have shown improvements in precision and accuracy, enabling real-time decision-making and guidance. AI algorithms that automatically control collimation during procedures also contribute by reducing radiation exposure while maintaining high image quality. This capability is crucial for ensuring both patient safety and procedural success, making AI an essential tool in modern IR workflows.	[52,55]
4. Automation and Robotics	The integration of AI-driven automation and robotic systems significantly enhances the accuracy, efficiency, and precision of interventional procedures. Robotic assistance not only improves procedural accuracy but also reduces clinician fatigue, particularly when performing repetitive or complex tasks. Moreover, AI's role in automating tasks like report generation or collimation streamlines workflows, allowing healthcare professionals to allocate more time for critical decision-making and patient care. These advancements support the optimization of clinical outcomes in IR.	[52]
5. Multimodal Integration	AI's capacity to integrate different modalities, such as imaging, robotics, and real-time data analysis, is revolutionizing IR. Recent innovations, such as generating synthetic contrast-enhanced images from non-contrast CT scans, reduce reliance on contrast agents while maintaining high diagnostic quality. By improving imaging accuracy and optimizing procedural planning, AI-driven multimodal integration enhances treatment strategies and reduces patient exposure to harmful substances, making it a game-changer in the field.	[62]

Table 6. Barriers from NRR and contributions from CER.

Barriers from NRR	Description of Contribution of CER	References
1. Ethical Considerations	Ethical challenges are a major barrier to AI adoption in IR. Concerns around data privacy, patient consent, and potential biases in AI models must be addressed to ensure that AI systems are ethically deployed. Additionally, the use of AI tools for tasks such as explaining procedures or risks to patients raises questions about whether these systems can provide contextually accurate and sensitive information. Clinicians may hesitate to fully rely on AI, especially when it comes to explaining complex medical situations to patients.	[47]
2. Regulatory Barriers and Intellectual Property	Regulatory hurdles pose significant challenges to the integration of AI in clinical practice. AI technologies require rigorous validation and approval processes to ensure their safety and effectiveness. However, the rapid pace of AI development often outstrips existing regulatory frameworks, causing delays in clinical adoption. Furthermore, intellectual property issues related to AI models and the sharing of new technologies complicate collaboration and innovation in healthcare. These obstacles must be addressed to facilitate the broader use of AI in interventional radiology.	[46]

Table 6. *Cont.*

Barriers from NRR	Description of Contribution of CER	References
3. Reliability and Clinical Utility	The reliability and clinical utility of AI systems remain significant barriers to their widespread adoption. While AI models have demonstrated promising results in controlled environments, their real-world performance may be inconsistent due to variations in data quality, model training, and the complexity of clinical cases. For example, AI models trained on synthetic data may not perform as effectively in real-world clinical scenarios, particularly in atypical or complex cases. This raises concerns among clinicians, who may be reluctant to trust AI outputs without human oversight, especially in critical applications like interventional procedures.	[46,54]
4. Slow Adoption	Slow adoption of AI technologies remains a key barrier in interventional radiology. While there is significant interest in AI's potential, many professionals are hesitant to integrate it into their practices. Concerns about the clinical utility of AI, the need for continuous training, and the disruption of established workflows all contribute to this resistance. Additionally, the financial and logistical challenges involved in adopting AI technologies, such as the cost of new equipment and system integration, hinder the widespread implementation of AI in clinical settings.	[48]

4.5. Emerging Directions in Interventional Radiology: AI Integration and the Path to Global Harmonization

Interventional radiology devices, from a medical device perspective, are complex systems with a high degree of interoperability. These devices can include, in addition to radiative imaging systems, components of robotics, catheterization, and others. Even without AI, they already have a highly intricate certification and integration process within the healthcare domain. A key example is cybersecurity, which has significant implications for patient safety, much more so than database management. In interventional radiology (IR), a cyberattack—whether altering the dose, compromising robotics, or other aspects—directly impacts patient safety and can cause harm.

The addition of AI to IR, which already brings its own challenges in integration, and even more so in the healthcare domain, introduces further questions and problems, on top of the well-known ones. Therefore, the role of national and international bodies, alongside national and international scientific societies, is crucial. National bodies must aim to establish as uniform and comprehensive an approach as possible, ideally unified, rather than dissimilar, as is currently the case for medical devices in general. Scientific societies must take concerted, impactful actions to develop broad and inclusive guidelines in this area. The works by Warren et al. [20,21] are crucial in guiding the integration of AI into IR by addressing key challenges and proposing structured solutions. They focus on the importance of clinician education in AI, the classification of AI models based on complexity, and the assessment of AI applications across different procedural stages (pre-, intra-, and post-procedural).

Additionally, these studies emphasize the need for rigorous risk evaluation, outlining a hierarchy of potential harms and providing a checklist for AI deployment in clinical IR environments. Their direction aligns with the broader goal of establishing clear regulatory frameworks and safety guidelines to ensure the responsible and effective implementation of AI in IR.

An important future direction is to focus on the two strategic points: medical devices (MD) and guidelines, which will involve scientific societies, certifying bodies, and national

or supranational legislative entities. This is also a strongly suggested future research direction, to observe what is happening internally. Achieving uniformity at the international level will be the next step. In fact, the integration of artificial intelligence (AI) into interventional radiology is gaining increasing attention, partly due to the growing use of advanced tools that combine AI with medical devices to support professionals during image-guided procedures. While much of the international guidelines focus primarily on diagnostic radiology, many of the principles developed by professional bodies and regulatory agencies can also be directly applied to interventional radiology, which, as a part of radiology, shares many challenges and opportunities in adopting AI technologies.

For instance, the Radiological Society of North America (RSNA) has developed a comprehensive strategy for AI use in radiology, which, although primarily focused on diagnostics, offers valuable insights for interventional radiology [63]. The RSNA emphasizes the importance of continuous clinical validation, balanced AI integration into workflows, and the monitoring of AI tool performance. One of the main recommendations is to use AI as support, not as a substitute for healthcare professionals—an essential principle for interventional radiology, where AI can optimize procedural precision but should always be used under expert supervision.

The American College of Radiology (ACR), in collaboration with the RSNA, has produced guidelines for implementing AI tools in radiology [64]. While the focus is on diagnostic applications, these guidelines offer relevant points for interventional radiology, particularly regarding the need for proper professional training to safely use AI in clinical procedures. The ACR stresses that AI adoption should be accompanied by continuous education for radiologists, ensuring they are prepared to adapt to technological changes while also managing risks, such as potential loss of practical skills in the face of automation.

The Society of Radiographers (SoR) [65] has also developed guidelines for AI use in diagnostics, which, although centered around clinical imaging, provide valuable suggestions on how AI should be integrated while maintaining human control. In interventional radiology, this principle of human oversight is crucial, as AI tools used for image guidance during procedures must always be monitored by qualified professionals to prevent potential errors during delicate interventions.

The Royal College of Radiologists (RCR) [66], with its recommendations for AI integration into radiological workflows, emphasizes the need for a gradual introduction of AI to avoid disruptions to existing workflows and ensure patient safety. This approach is particularly relevant in interventional radiology, where AI integration could influence the efficiency of real-time procedures, but it must be done in a way that does not compromise the quality of clinical decision-making.

Regulatory bodies like the U.S. Food and Drug Administration (FDA) have also played a critical role in shaping the landscape of AI in medical devices, including in interventional radiology [67–70]. The FDA has approved several AI-based software systems aimed at assisting interventional radiologists during image-guided procedures. These approvals underscore the need for regulatory oversight in AI tool deployment, ensuring that they meet high standards for safety and effectiveness before being used in clinical practice. The FDA's focus on Software as a Medical Device (SaMD) provides a clear regulatory pathway for AI tools used in IR, ensuring that these tools are developed with the same scrutiny as any other medical device.

The regulatory guidance from the FDA is complemented by ongoing developments in the European Union, where lawmakers are refining policies under the EU AI Act [71]. This legislation aims to regulate the development and use of AI across sectors, including healthcare. By adhering to these international guidelines, both diagnostic and interventional radiology practices can ensure that AI is used in a safe, effective, and ethically responsible

manner. An example of movement is the ESR initiative [72] producing a statement on how the EU AI Act applies to radiology and therefore also to IR. It outlines key policies and offers recommendations to policymakers and professionals for effectively implementing the legislation, addressing gaps and uncertainties. It covers areas such as AI literacy in radiology, classification of high-risk AI systems, data governance, transparency, human oversight, quality management, obligations of AI deployers, regulatory sandboxes, post-market monitoring, information sharing, and market surveillance. The statement highlights ESR's commitment to supporting the safe and effective integration of AI in radiology to benefit patients across Europe.

4.6. Limitations

This narrative review follows a structured methodology with well-defined inclusion and exclusion criteria, ensuring a focused and high-quality synthesis of available literature. While the exclusion of conference proceedings means that some emerging research or preliminary developments may not be captured, this approach prioritizes studies that have undergone rigorous peer review, enhancing the reliability of the findings. Additionally, by focusing on internationally published literature in English, the review ensures broad applicability and comparability across different healthcare settings. However, this may result in the omission of region-specific insights or localized best practices, highlighting an opportunity for future research to explore diverse clinical approaches and treatment protocols in various cultural and healthcare contexts.

5. Conclusions

The integration of artificial intelligence (AI) into interventional radiology (IR) presents both significant opportunities and challenges that will undoubtedly shape the future of this field. This technology has the potential to greatly enhance the precision, efficiency, and personalization of IR procedures. By automating complex tasks, such as catheter manipulation and needle placement, it improves procedural accuracy and reduces variability. Furthermore, its ability to integrate multiple imaging modalities supports personalized patient management, optimizing treatment plans and outcomes. Intra-procedural guidance also benefits from these advancements, enhancing precision through technologies like needle tracking and real-time image fusion.

As AI continues to evolve, so too does the potential for robotics and automation in IR. AI-guided robotic systems are set to play an increasingly autonomous role, transforming the function of interventional radiologists from active procedure performers to supervisors of intelligent systems. However, full autonomy in these systems has not yet been achieved, and there remain significant areas for further development and integration. Despite these promising advancements, challenges remain.

IR devices represent a highly complex and integrated system within healthcare. These devices, which include radiative imaging systems, robotics, catheterization, and other advanced technologies, require a detailed certification and integration process. This complexity is further amplified by the inclusion of AI, which introduces new challenges and considerations. As this technology becomes more deeply integrated into IR, it raises numerous questions that extend beyond traditional device integration concerns. The role of national and international bodies, along with scientific societies, is crucial in navigating this evolving landscape. These organizations must work together to establish uniform and comprehensive regulatory frameworks that ensure safety and efficacy across different healthcare systems.

Moving forward, a key direction for IR will focus on developing clear, standardized guidelines for both medical devices and AI applications. Collaboration between certifying

bodies, scientific societies, and legislative entities will be essential in ensuring global uniformity. Ongoing efforts, such as the EU AI Act, highlight the importance of addressing AI-specific challenges. In the context of radiology, initiatives like those from the European Society of Radiology (ESR) that align with the EU AI Act—focusing on transparency, data governance, human oversight, and post-market monitoring—serve as exemplary steps in the right direction.

The integration of these technologies makes interventional radiology more personalized, patient-centered, and safe. With concerted efforts, the integration of advanced technologies in IR can proceed with the necessary safeguards, ultimately benefiting patient outcomes and advancing the field responsibly.

Author Contributions: Conceptualization, D.G. and A.L.; methodology, D.G.; software, All authors; validation, All authors; formal analysis, D.G., A.L., G.L., N.F. and N.I.; investigation, D.G. and A.L.; resources, All authors; data curation, All authors; writing—original draft preparation, D.G. and A.L.; writing—review and editing, All authors; visualization, All authors; supervision, A.B., V.M., Y.W., D.G. and R.R.; project administration, A.L. and D.G.; funding acquisition, D.G. 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: Not applicable.

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

References

1. Available online: https://www.collinsdictionary.com/dictionary/english/interventional-radiology (accessed on 23 February 2025).
2. Available online: https://www.oxfordreference.com/display/10.1093/oi/authority.20110803100008707?p=emailAMeeGkkfeJyEM&d=/10.1093/oi/authority.20110803100008707 (accessed on 23 February 2025).
3. Chand, R.; Eltorai, A.E.; Healey, T.; Ahn, S. (Eds.) *Essential Interventional Radiology Review A Question and Answer Guide*; Springer: Berlin/Heidelberg, Germany, 2021.
4. Neri, E.; Erba, P.A. (Eds.) *Multimodality Imaging and Intervention in Oncology*; Sringer: Berlin/Heidelberg, Germany, 2023.
5. Kini, A.; Sharma, S.K. (Eds.) *Practical Manual of Interventional Cardiology*; Springer: Berlin/Heidelberg, Germany, 2021.
6. Patel, A.; Petrone, B.; Carter, K.R. Percutaneous Vertebroplasty and Kyphoplasty. In *StatPearls [Internet]*; StatPearls Publishing: Treasure Island, FL, USA, 2025. Available online: https://www.ncbi.nlm.nih.gov/books/NBK525963/ (accessed on 23 February 2025).
7. Pinto, A.; Giurazza, F.; Califano, T.; Rea, G.; Valente, T.; Niola, R.; Caranci, F. Interventional radiology in gynecology and obstetric practice: Safety issues. In *Seminars in Ultrasound, CT and MRI*; WB Saunders: Philadelphia, PA, USA, 2021; Volume 42, pp. 104–112. [CrossRef] [PubMed] [PubMed Central]
8. Montanarella, M.J.; Agarwal, A.; Moon, B. Peripherally Inserted Central Catheter (PICC) Line Placement. In *StatPearls [Internet]*; StatPearls Publishing: Treasure Island, FL, USA, 2025. Available online: https://www.ncbi.nlm.nih.gov/books/NBK573064/ (accessed on 23 February 2025).
9. Giurazza, F.; Niola, R. Interventional radiology in the 3rd millennium: Past, present and future perspectives. *Radiol. Med.* **2024**, *129*, 1429–1430. [CrossRef] [PubMed]
10. Brady, A.P.; Allen, B.; Chong, J.; Kotter, E.; Kottler, N.; Mongan, J.; Oakden-Rayner, L.; Pin-to Dos Santos, D.; Tang, A.; Wald, C.; et al. Developing, purchasing, implementing and monitoring AI tools in radiology: Practical considerations. A multi-society statement from the ACR, CAR, ESR, RANZCR & RSNA. *J. Med. Imaging Radiat. Oncol.* **2024**, *68*, 7–26. [CrossRef] [PubMed]
11. Cornelis, F.H.; Filippiadis, D.K.; Wiggermann, P.; Solomon, S.B.; Madoff, D.C.; Milot, L.; Bodard, S. Evaluation of navigation and robotic systems for percutaneous image-guided interventions: A novel metric for advanced imaging and artificial intelligence integration. *Diagn. Interv. Imaging*, 2025, *ahead of print*. [CrossRef] [PubMed]
12. Boeken, T.; Lim, H.D.; Cohen, E.I. The Role and Future of Artificial Intelligence in Robotic Image-Guided Interventions. *Tech. Vasc. Interv. Radiol.* **2024**, *27*, 101001. [CrossRef] [PubMed]

13. Vlastaris, K.; Alrez, A.; Friedland, S.; Randazzo, A.; Abboud, R.; Martin, C. The Transformative Impact of AI, Extended Reality, and Robotics in Interventional Radiology: Current Trends and Applications. *Tech. Vasc. Interv. Radiol.* **2024**, *27*, 101003. [CrossRef] [PubMed]
14. Lee, K.; Muniyandi, P.; Li, M.; Saccenti, L.; Christou, A.; Xu, S.; Wood, B.J. Smartphone Technology for Applications in Image-Guided Minimally Invasive Interventional Procedures. *Cardiovasc. Interv. Radiol.* **2025**, *48*, 142–156. [CrossRef] [PubMed] [PubMed Central]
15. Barat, M.; Crombé, A.; Boeken, T.; Dacher, J.N.; Si-Mohamed, S.; Dohan, A.; Chassagnon, G.; Lecler, A.; Greffier, J.; Nougaret, S.; et al. Imaging in France: 2024 Update. *Can. Assoc. Radiol. J.* 2024; *ahead of print*. [CrossRef] [PubMed]
16. Matsui, Y.; Ueda, D.; Fujita, S.; Fushimi, Y.; Tsuboyama, T.; Kamagata, K.; Ito, R.; Yanagawa, M.; Yamada, A.; Kawamura, M.; et al. Applications of artificial intelligence in interventional oncology: An up-to-date review of the literature. *Jpn. J. Radiol.* **2025**, *43*, 164–176. [CrossRef] [PubMed] [PubMed Central]
17. Lesaunier, A.; Khlaut, J.; Dancette, C.; Tselikas, L.; Bonnet, B.; Boeken, T. Artificial intelligence in interventional radiology: Current concepts and future trends. *Diagn. Interv. Imaging* **2025**, *106*, 5–10. [CrossRef] [PubMed]
18. Zhang, J.; Fang, J.; Xu, Y.; Si, G. How AI and Robotics Will Advance Interventional Radiology: Narrative Review and Future Perspectives. *Diagnostics* **2024**, *14*, 1393. [CrossRef] [PubMed] [PubMed Central]
19. Glielmo, P.; Fusco, S.; Gitto, S.; Zantonelli, G.; Albano, D.; Messina, C.; Sconfienza, L.M.; Mauri, G. Artificial intelligence in interventional radiology: State of the art. *Eur. Radiol. Exp.* **2024**, *8*, 62. [CrossRef] [PubMed] [PubMed Central]
20. Warren, B.E.; Bilbily, A.; Gichoya, J.W.; Chartier, L.B.; Fawzy, A.; Barragán, C.; Jaberi, A.; Mafeld, S. An Introductory Guide to Artificial Intelligence in Interventional Radiology: Part 2: Implementation Considerations and Harms. *Can. Assoc. Radiol. J.* **2024**, *75*, 568–574. [CrossRef] [PubMed]
21. Warren, B.E.; Bilbily, A.; Gichoya, J.W.; Conway, A.; Li, B.; Fawzy, A.; Barragán, C.; Jaberi, A.; Mafeld, S. An Introductory Guide to Artificial Intelligence in Interventional Radiology: Part 1 Foundational Knowledge. *Can. Assoc. Radiol. J.* **2024**, *75*, 558–567. [CrossRef] [PubMed]
22. Geevarghese, R.; Bodard, S.; Razakamanantsoa, L.; Marcelin, C.; Petre, E.N.; Dohan, A.; Kastler, A.; Frandon, J.; Barral, M.; Soyer, P.; et al. Interventional Oncology: 2024 Update. *Can. Assoc. Radiol. J.* **2024**, *75*, 658–670. [CrossRef] [PubMed]
23. Campbell, W.A., IV; Chick, J.F.B.; Shin, D.; Makary, M.S. Understanding ChatGPT for evidence-based utilization in interventional radiology. *Clin. Imaging* **2024**, *108*, 110098. [CrossRef] [PubMed]
24. O'Brien, A.J.; Vrazas, J.I.; Clements, W. Current applications of algorithmic artificial intelligence in interventional radiology: A review of the literature. *J. Med. Imaging Radiat. Oncol.* **2024**, *68*, 194–207. [CrossRef] [PubMed]
25. Gaddum, O.; Chapiro, J. An Interventional Radiologist's Primer of Critical Appraisal of Artificial Intelligence Research. *J. Vasc. Interv. Radiol.* **2024**, *35*, 7–14. [CrossRef] [PubMed]
26. Lanza, C.; Carriero, S.; Buijs, E.F.M.; Mortellaro, S.; Pizzi, C.; Sciacqua, L.V.; Biondetti, P.; Angileri, S.A.; Ianniello, A.A.; Ierardi, A.M.; et al. Robotics in Interventional Radiology: Review of Current and Future Applications. *Technol. Cancer Res. Treat.* **2023**, *22*, 15330338231152084. [CrossRef] [PubMed] [PubMed Central]
27. von Ende, E.; Ryan, S.; Crain, M.A.; Makary, M.S. Artificial Intelligence, Augmented Reality, and Virtual Reality Advances and Applications in Interventional Radiology. *Diagnostics* **2023**, *13*, 892. [CrossRef] [PubMed] [PubMed Central]
28. Posa, A.; Barbieri, P.; Mazza, G.; Tanzilli, A.; Natale, L.; Sala, E.; Iezzi, R. Technological Advancements in Interventional Oncology. *Diagnostics* **2023**, *13*, 228. [CrossRef] [PubMed] [PubMed Central]
29. Boeken, T.; Feydy, J.; Lecler, A.; Soyer, P.; Feydy, A.; Barat, M.; Duron, L. Artificial intelligence in diagnostic and interventional radiology: Where are we now? *Diagn. Interv. Imaging* **2023**, *104*, 1–5. [CrossRef] [PubMed]
30. Kallini, J.R.; Moriarty, J.M. Artificial Intelligence in Interventional Radiology. *Semin. Interv. Radiol.* **2022**, *39*, 341–347. [CrossRef] [PubMed] [PubMed Central]
31. Waller, J.; O'Connor, A.; Rafaat, E.; Amireh, A.; Dempsey, J.; Martin, C.; Umair, M. Applications and challenges of artificial intelligence in diagnostic and interventional radiology. *Pol. J. Radiol.* **2022**, *87*, e113–e117. [CrossRef] [PubMed] [PubMed Central]
32. Seah, J.; Boeken, T.; Sapoval, M.; Goh, G.S. Prime Time for Artificial Intelligence in Interventional Radiology. *Cardiovasc. Interv. Radiol.* **2022**, *45*, 283–289. [CrossRef] [PubMed] [PubMed Central]
33. Mazaheri, S.; Loya, M.F.; Newsome, J.; Lungren, M.; Gichoya, J.W. Challenges of Implementing Artificial Intelligence in Interventional Radiology. *Semin. Interv. Radiol.* **2021**, *38*, 554–559. [CrossRef] [PubMed] [PubMed Central]
34. Desai, S.B.; Pareek, A.; Lungren, M.P. Current and emerging artificial intelligence applications for pediatric interventional radiology. *Pediatr. Radiol.* **2022**, *52*, 2173–2177. [CrossRef] [PubMed]
35. D'Amore, B.; Smolinski-Zhao, S.; Daye, D.; Uppot, R.N. Role of Machine Learning and Artificial Intelligence in Interventional Oncology. *Curr. Oncol. Rep.* **2021**, *23*, 70. [CrossRef] [PubMed]
36. Gurgitano, M.; Angileri, S.A.; Rodà, G.M.; Liguori, A.; Pandolfi, M.; Ierardi, A.M.; Wood, B.J.; Carrafiello, G. Interventional Radiology ex-machina: Impact of Artificial Intelligence on practice. *Radiol. Med.* **2021**, *126*, 998–1006. [CrossRef] [PubMed] [PubMed Central]

37. Iezzi, R.; Goldberg, S.N.; Merlino, B.; Posa, A.; Valentini, V.; Manfredi, R. Artificial Intelligence in Interventional Radiology: A Literature Review and Future Perspectives. *J. Oncol.* **2019**, *2019*, 6153041. [CrossRef] [PubMed] [PubMed Central]
38. Giansanti, D. AI in Cytopathology: A Narrative Umbrella Review on Innovations, Challenges, and Future Directions. *J. Clin. Med.* **2024**, *13*, 6745. [CrossRef] [PubMed] [PubMed Central]
39. Pirrera, A.; Giansanti, D. Human-Machine Collaboration in Diagnostics: Exploring the Synergy in Clinical Imaging with Artificial Intelligence. *Diagnostics* **2023**, *13*, 2162. [CrossRef] [PubMed] [PubMed Central]
40. Auloge, P.; Garnon, J.; Robinson, J.M.; Dbouk, S.; Sibilia, J.; Braun, M.; Vanpee, D.; Koch, G.; Cazzato, R.L.; Gangi, A. Interventional radiology and artificial intelligence in radiology: Is it time to enhance the vision of our medical students? *Insights Into Imaging* **2020**, *11*, 127. [CrossRef]
41. Available online: https://pubmed.ncbi.nlm.nih.gov/?term=(diagnostic+radiology[Title/Abstract])+AND+((neural+network[Title/Abstract])+OR+(Artificial+Intelligence[Title/Abstract])+OR+(deep+learning[Title/Abstract])+OR+(ANN[Title/Abstract])+OR+(GAN[Title/Abstract])+)&sort=date&size=50 (accessed on 23 February 2025).
42. Trivedi, P.; Young, S. The Evolution of Interventional Radiology: How Far Will the Apple Land From the Tree? *J. Am. Coll. Radiol.* **2024**, *21*, 701–703. [CrossRef]
43. Majid, A.S.; White, S.B. The future of interventional radiology should be continued alignment with diagnostic radiology. *J. Am. Coll. Radiol.* **2024**, *21*, 704–705. [CrossRef] [PubMed]
44. Ryu, R.K. The future of interventional radiology is independent of diagnostic radiology. *J. Am. Coll. Radiol.* **2024**, *21*, 706–707. [PubMed]
45. Newbury, A.; Ferguson, C.; Valero, D.A.; Kutcher-Diaz, R.; McIntosh, L.; Karamanian, A.; Harman, A. Interventional oncology update. *Eur. J. Radiol. Open* **2022**, *9*, 100430. [CrossRef] [PubMed] [PubMed Central]
46. Fum, W.K.S.; Md Shah, M.N.; Raja Aman, R.R.A.; Abd Kadir, K.A.; Wen, D.W.; Leong, S.; Tan, L.K. Generation of fluoroscopy-alike radiographs as alternative datasets for deep learning in interventional radiology. *Phys. Eng. Sci. Med.* **2023**, *46*, 1535–1552. [CrossRef] [PubMed]
47. Hofmann, H.L.; Vairavamurthy, J. Large language model doctor: Assessing the ability of ChatGPT-4 to deliver interventional radiology procedural information to patients during the consent process. *CVIR Endovasc.* **2024**, *7*, 83. [CrossRef] [PubMed] [PubMed Central]
48. Cornelis, F.H.; Soyer, P.; Patlas, M.N. Interventional Radiology is Now at the Confluence of Expertise, Innovation, and Artificial Intelligence. *Can. Assoc. Radiol. J.* **2024**, *75*, 456–457. [CrossRef] [PubMed]
49. Najafi, A.; Cazzato, R.L.; Meyer, B.C.; Pereira, P.L.; Alberich, A.; López, A.; Ronot, M.; Fritz, J.; Maas, M.; Benson, S.; et al. CIRSE Position Paper on Artificial Intelligence in Interventional Radiology. *Cardiovasc. Interv. Radiol.* **2023**, *46*, 1303–1307. [CrossRef] [PubMed]
50. Bonnet, B.; Tselikas, L. Available online: https://www.sciencedirect.com/science/article/abs/pii/S2211568425000075?via=ihub (accessed on 23 February 2025).
51. Yuan, X.T.; Shao, C.Y.; Zhang, Z.Z.; Qian, D. Comparing the performance of ChatGPT and ERNIE Bot in answering questions regarding liver cancer interventional radiology in Chinese and English contexts: A comparative study. *Digit. Health* **2025**, *11*, 20552076251315511. [CrossRef] [PubMed] [PubMed Central]
52. Sato, T.; Kawai, T.; Shimohira, M.; Ohta, K.; Suzuki, K.; Nakayama, K.; Takikawa, J.; Kawaguchi, T.; Urano, M.; Ng, K.W.; et al. Robot-Assisted CT-guided Biopsy with an Artificial Intelligence-Based Needle-Path Generator: An Experimental Evaluation Using a Phantom Model. *J. Vasc. Interv. Radiol.* **2025**, S1051-0443(25)00040-5. [CrossRef] [PubMed]
53. Taghavi, R.M.; Shah, A.; Filkov, V.; Goldman, R.E. Deep Learning Models for Automatic Classification of Anatomic Location in Abdominopelvic Digital Subtraction Angiography. *J. Imaging Inform. Med.* 2025; *ahead of print*. [CrossRef] [PubMed]
54. Amin, K.S.; Davis, M.A.; Naderi, A.; Forman, H.P. Release of complex imaging reports to patients, do radiologists trust AI to help? *Curr. Probl. Diagn. Radiol.* **2025**, *54*, 147–150. [CrossRef] [PubMed]
55. Lee, B.C.; Rijhwani, D.; Lang, S.; van Oorde-Grainger, S.; Haak, A.; Bleise, C.; Lylyk, P.; Ruijters, D.; Sinha, A. Tunable and real-time automatic interventional x-ray collimation from semi-supervised deep feature extraction. *Med. Phys.* 2024, *ahead of print*. [CrossRef] [PubMed]
56. Vogl, T.J.; Burck, I.; Stöver, T.; Helal, R. Parapharyngeal Space: Diagnostic Imaging and Intervention. In *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der Bildgebenden Verfahren*; Georg Thieme Verlag KG: Stuttgart, Germany, 2024. [CrossRef] [PubMed]
57. Can, E.; Uller, W.; Vogt, K.; Doppler, M.C.; Busch, F.; Bayerl, N.; Ellmann, S.; Kader, A.; Elkilany, A.; Makowski, M.R.; et al. Large Language Models for Simplified Interventional Radiology Reports: A Comparative Analysis. *Acad. Radiol.* **2025**, *32*, 888–898. [CrossRef] [PubMed]
58. Sajan, A.; Lamane, A.; Baig, A.; Floch, K.L.; Dercle, L. The emerging role of AI in enhancing intratumoral immunotherapy care. *Oncotarget* **2024**, *15*, 635–637. [CrossRef] [PubMed] [PubMed Central]

59. Elhakim, T.; Brea, A.R.; Fidelis, W.; Paravastu, S.S.; Malavia, M.; Omer, M.; Mort, A.; Ramasamy, S.K.; Tripathi, S.; Dezube, M.; et al. Enhanced PROcedural Information READability for Patient-Centered Care in Interventional Radiology With Large Language Models (PRO-READ IR). *J. Am. Coll. Radiol.* **2025**, *22*, 84–97. [CrossRef] [PubMed]
60. Warren, B.E.; Alkhalifah, F.; Ahrari, A.; Min, A.; Fawzy, A.; Annamalai, G.; Jaberi, A.; Beecroft, R.; Kachura, J.R.; Mafeld, S.C. Feasibility of Artificial Intelligence Powered Adverse Event Analysis: Using a Large Language Model to Analyze Microwave Ablation Malfunction Data. *Can. Assoc. Radiol. J.* **2025**, *76*, 171–179. [CrossRef] [PubMed]
61. Ebel, S.; Ehrengut, C.; Denecke, T.; Gößmann, H.; Beeskow, A.B. GPT-4o's competency in answering the simulated written European Board of Interventional Radiology exam compared to a medical student and experts in Germany and its ability to generate exam items on interventional radiology: A descriptive study. *J. Educ. Eval. Health Prof.* **2024**, *21*, 21. [CrossRef] [PubMed]
62. Pinnock, M.A.; Hu, Y.; Bandula, S.; Barratt, D.C. Time conditioning for arbitrary contrast phase generation in interventional computed tomography. *Phys. Med. Biol.* **2024**, *69*, 115010.
63. Available online: https://pubs.rsna.org/page/ai/claim (accessed on 23 February 2025).
64. Available online: https://www.acr.org/News-and-Publications/Joint-Radiological-Society-Paper-Guides-Implementation-and-Use-of-AI-Tools-in-Radiology (accessed on 23 February 2025).
65. Available online: https://www.sor.org/learning-advice/professional-body-guidance-and-publications/documents-and-publications/policy-guidance-document-library/artificial-intelligence-guidance-for-clinical-imag (accessed on 23 February 2025).
66. Available online: https://www.rcr.ac.uk/our-services/all-our-publications/clinical-radiology-publications/integrating-artificial-intelligence-with-the-radiology-reporting-workflows-ris-and-pacs/ (accessed on 23 February 2025).
67. Available online: https://www.fda.gov/media/184856/download (accessed on 23 February 2025).
68. Available online: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device (accessed on 23 February 2025).
69. Available online: https://radiologybusiness.com/topics/healthcare-management/healthcare-policy/fda-clears-ai-software-aiding-interventional-radiologists-during-image-guided-procedures (accessed on 23 February 2025).
70. Available online: https://www.fda.gov/media/184830/download (accessed on 23 February 2025).
71. Available online: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf (accessed on 23 February 2025).
72. Kotter, E.; D'antonoli, T.A.; Cuocolo, R.; Hierath, M.; Huisman, M.; Klontzas, M.E.; Martí-Bonmatí, L.; May, M.S.; Neri, E.; Nikolaou, K.; et al. Guiding AI in radiology: ESR's recommendations for effective imple-mentation of the European AI Act. *Insights Imaging* **2025**, *16*, 33. [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

Deep Learning Models to Detect Anterior Cruciate Ligament Injury on MRI: A Comprehensive Review

Michele Mercurio [1,2,†], Federica Denami [1,*], Dimitra Melissaridou [3], Katia Corona [4], Simone Cerciello [5], Domenico Laganà [6,7], Giorgio Gasparini [1,2,†] on behalf of the IORS and Roberto Minici [7]

1 Department of Orthopaedic and Trauma Surgery, Magna Graecia University, "Renato Dulbecco" University Hospital, 88100 Catanzaro, Italy; michele.mercurio@unicz.it (M.M.); gasparini@unicz.it (G.G.)
2 Research Center on Musculoskeletal Health, MusculoSkeletal Health@UMG, Magna Graecia University, 88100 Catanzaro, Italy
3 1st Department of Orthopaedic Surgery, National and Kapodistrian University of Athens, Attikon Hospital, 12462 Athens, Greece; dimitramelissaridi@gmail.com
4 Department of Medicine and Health Sciences "Vincenzo Tiberio", University of Molise, 86100 Campobasso, Italy; katia.corona@unimol.it
5 School of Medicine, Saint Camillus University, 00131 Rome, Italy; simone.cerciello@unicamillus.org
6 Department of Experimental and Clinical Medicine, Magna Graecia University, 88100 Catanzaro, Italy; domenico.lagana@unicz.it
7 Radiology Unit, Department of Experimental and Clinical Medicine, Magna Graecia University, "Renato Dulbecco" University Hospital, 88100 Catanzaro, Italy; miniciroberto@gmail.com
* Correspondence: federica.denami@studenti.unicz.it; Tel.: +39-0961-3647002
† Members of the IORS are indicated in Acknowledgments.

Abstract: Magnetic resonance imaging (MRI) is routinely used to confirm the suspected diagnosis of anterior cruciate ligament (ACL) injury. Recently, many studies explored the role of artificial intelligence (AI) and deep learning (DL), a sub-category of AI, in the musculoskeletal field and medical imaging. The aim of this study was to review the current applications of DL models to detect ACL injury on MRI, thus providing an updated and critical synthesis of the existing literature and identifying emerging trends and challenges in the field. A total of 23 relevant articles were identified and included in the review. Articles originated from 10 countries, with China having the most contributions (n = 9), followed by the United State of America (n = 4). Throughout the article, we analyzed the concept of DL in ACL tears and provided examples of how these tools can impact clinical practice and patient care. DL models for MRI detection of ACL injury reported high values of accuracy, especially helpful for less experienced clinicians. Time efficiency was also demonstrated. Overall, the deep learning models have proven to be a valid resource, although still requiring technological developments for implementation in daily practice.

Keywords: artificial intelligence; deep learning; knee; ligament; sport; anterior cruciate ligament; magnetic resonance imaging

1. Introduction

The anterior cruciate ligament (ACL) is the most injured ligament in the knee. Sports injuries are the main source of tears with a higher incidence among male subjects [1]. The annual reported incidence in the United States is 1 in 3500 people [2,3]. The symptoms include swelling, pain, knee deformation, and difficulty walking [4]. An ACL tear can range in severity from partial to complete, and each grade affects the knee's stability and function differently [5]. A clinical exam, including Lachman and Pivot Shift tests, is essential for diagnosis [6]. Magnetic resonance imaging (MRI) is routinely used to confirm suspected

diagnosis and to assess for concomitant injuries [7,8]. MRI presents benefits such as high contrast and resolution, non-invasiveness, and multi-planar imaging [9,10]. However, the current diagnosis process requires time-consuming examination by radiologists, which is also error-prone at scale. Indeed, a radiologist is required to examine each slice of the MRI scan, looking for ACL ruptures and other secondary complications, including bone marrow edema and anterior tibial translation [11]. Recently, numerous studies have investigated the role of artificial intelligence (AI) and its subgroups in the field of musculoskeletal system. AI refers to a branch of computer systems that are capable of performing tasks that mimic human cognitive functions such as learning and problem solving by analyzing and comparing data [12,13]. One specific area of investigation is how AI can contribute to the accurate diagnosis and characterization of ACL tears through imaging examinations to minimize medical errors. In this context, deep learning (DL), a sub-category of AI, is frequently used in medical imaging (Figure 1).

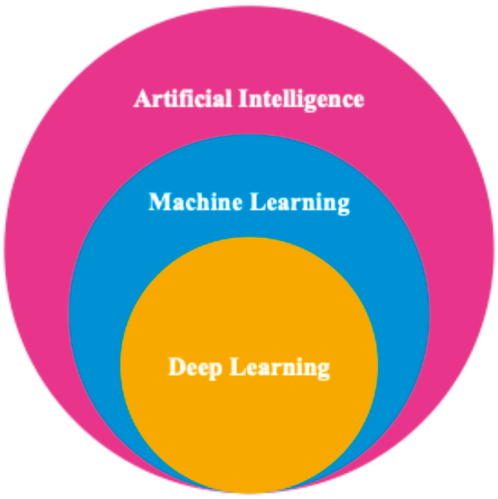


Figure 1. Machine learning and deep learning are subsets of artificial intelligence.

This tool consists of a series of inputs that pass through several interconnected layers of neurons that independently recognize different features and make predictions about large amounts of information [14,15]. Many DL target recognition systems are based on a convolutional neural network (CNN) because it can reduce the complexity of the overall network and training parameters and keep the data relatively constant in terms of scheduling, bias, and scaling (Figure 2).

Furthermore, the network structure is easy to train, optimize, and control [16]. To our knowledge, the use of DL models to detect ACL injury has been evaluated for a few years now, along with a few papers offering a comprehensive overview of this topic. Even though some reviews on DL models for detecting ACL injuries have been recently published [17], a new comprehensive review may be of significant value for several reasons. The field of AI, especially DL, is evolving rapidly, with the potential to transform musculoskeletal imaging [18]. A new review can include the latest studies capturing the most recent research and trends; evaluate newer models assessing the effectiveness of newer architectures like transformers or generative models; and consider larger and more diverse datasets, thus exploring how increased data availability and diversity impact model performance [19–21]. The aim of this study was to review the current applications of DL models to detect ACL injury on MRI, thus providing an updated and critical synthesis of the existing literature and identifying emerging trends and challenges in the field.

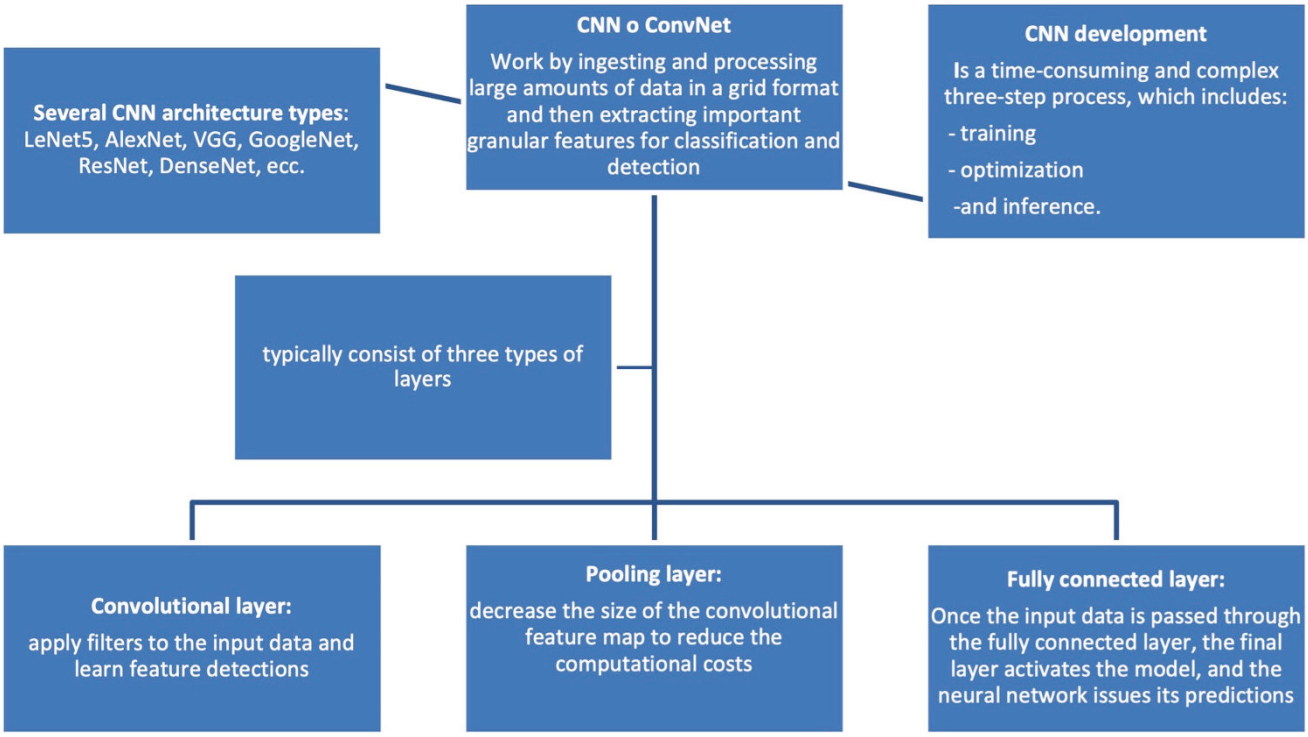


Figure 2. Convolutional Neural Network (CNN), also known as ConvNet. This is a specialized type of deep learning algorithm mainly designed for tasks that necessitate object recognition, including image classification, detection, and segmentation.

2. Materials and Methods

A review of the published literature was conducted and reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement [22]. The PubMed, MedLine, Scopus, and Cochrane Central databases were searched in December 2024. The search terms used to retrieve relevant articles were "deep learning", AND "artificial intelligence", AND "anterior cruciate ligament", AND "MRI", AND "diagnosis". Only articles published over the last five years were considered. Two authors (FD and RM) independently reviewed the titles and abstracts to identify articles for inclusion in the database and contacted a third lead author (MM) in cases of major discrepancies. The reference list of each included article and the gray literature available at our institution were checked for possible additional articles. Other reviews, editorials, letters to the editor, and expert opinion papers were also considered but not included. The included articles are listed in Table 1.

A methodological quality assessment was independently conducted by 3 authors (FD, RM, and MM) with the modified Newcastle–Ottawa Quality Assessment Scale (Table 2). Substantial interobserver agreement (Cohen kappa coefficients ranging between 0.59 and 0.74) was reported. Based on the total score, the quality was classified as "low" (0–3), "medium" (4–6), and "high" (7–9). Criterion number: 1, representativeness of exposed cohort; 2, selection of unexposed cohort; 3, ascertainment of exposure; 4, evidence that the outcome of interest was not present at baseline; 5, comparability of cohorts based on design or analysis; 6, assessment of outcome; 7, follow-up period long enough to capture the results; 8, adequacy of the follow-up of the cohorts. Each study was scored with a maximum of one or two points for each numbered item within the categories, based on the rules of the modified Newcastle–Ottawa scale.

Table 1. Included studies.

Authors and Year	Where	Journal	Method	Software	% Accuracy and % Specificity	Dataset
Awan et al. 2023 [4]	Pakistan	PeerJ Computer Science	MGACA	Python (version 9.3.12); Python Software Foundation, Wilmington, DE, USA	98%–NA	In-house dataset (15,265 images)
Awan et al. 2021 [23]	Pakistan	Diagnostics	ResNet-14 CNN	Python (version 3.6); Python Software Foundation, Wilmington, DE, USA	92–94%	In-house dataset (917 knees sagittal plane)
Chang et al. 2019 [7]	USA	Journal of Imaging Informatics in Medicine	CNN	Python (version 3.5; Python Software Foundation, Wilmington, DE, USA	96–100%	In-house dataset
Chen et al. 2022 [24]	Taiwan	JMIR AI.	CNN	Python (version 3.x); Python Software Foundation, Wilmington, Delaware, USA and PyTorch (version 1.1.x) Meta AI, Menlo Park, CA, USA	96–96%	In-house dataset (1000 cases)
Cheng et al. 2024 [16]	China	Journal of Orthopaedic Surgery and Research	SVM	Python (version 3.6); Python Software Foundation, Wilmington, DE, USA	93–93%	In-house dataset (526 patients)
Dung et al. 2023 [25]	Vietnam	Diagnostic and Interventional Imaging	DCLU-Net CNN	TensorFlow (Available online: https://www.tensorflow.org, Accessed on: 12 December 2024) Google, Mountain View, CA, USA PyTorch (Available online: https://pytorch.org, Accessed on: 12 December 2024): Meta AI, Menlo Park, CA, USA	90–89%	In-house dataset (expanded—247 patients)
Germann et al. 2020 [26]	Switzerland	Invest Radiol.	DCNN	TensorFlow (version 1.11) Google, Mountain View, CA, USA	AUC 93–93%	In-house dataset (5802 MRI)
Jeon et al. 2021 [27]	Japan	IEEE Journal of Biomedical and Health Informatics	CNN	PyTorch (Available online: https://pytorch.org, Accessed on: 12 December 2024): Meta AI, Menlo Park, CA, USA	AUC 98–95%	Chiba and Stanford datasets+ data augmentation
Joshi and Suganthi 2022 [28]	India	Diagnostics	CPDCNN	Python (version 3.x), Python Software Foundation Beaverton, OR, USA	96%–NA	Stanford dataset

Table 1. *Cont.*

Authors and Year	Where	Journal	Method	Software	% Accuracy and % Specificity	Dataset
Li et al. 2023 [29]	China	Frontiers in Bioengineering and Biotechnology	3D CNN	PyTorch (Available online: https://pytorch.org, Accessed on: 12 December 2024): Meta AI, Menlo Park, CA, USA	AUC 97% and 93% for each dataset–NA	In-house dataset + Stanford dataset
Li et al. 2021 [30]	China	Journal of Healthcare Engineering	pretrained VGG16	TensorFlow (Available online: https://www.tensorflow.org, Accessed on: 12 December 2024) Google, Mountain View, California, USA PyTorch (Available online: https://pytorch.org, Accessed on: 12 December 2024): Meta AI, Menlo Park, CA, USA	92–91%	In-house dataset (30 patients)
Liang et al. 2023 [9]	China	BMC Medical Imaging	Res-Net CNN	TensorFlow (Available online: https://www.tensorflow.org, Accessed on: 12 December 2024) Google, Mountain View, CA, USA PyTorch (Available online: https://pytorch.org, Accessed on: 12 December 2024): Meta AI, Menlo Park, CA, USA	81–65%	In-house dataset (468 images) + data augmentation
Liu et al. 2019 [31]	USA	Radiology: Artificial Intelligence	CNN	Python (version 2.7); Python Software Foundation, Wilmington, DE, USA	AUC 98–96%	In-house dataset (300 subjects)
Minamoto et al. 2022 [32]	Japan	BMC Musculoskeletal Disorders	CNN	Python (version 3.6.7); Python Software Foundation, Wilmington, DE, USA	88–86%	In-house dataset (100 epochs) + data augmentation

Table 1. *Cont.*

Authors and Year	Where	Journal	Method	Software	% Accuracy and % Specificity	Dataset
Namiri et al. 2020 [33]	USA	Radiology: Artificial Intelligence	2D e 3D CNN	TensorFlow (Available online: https://www.tensorflow.org, Accessed on: 12 December 2024) Google, Mountain View, CA, USA PyTorch (Available online: https://pytorch.org, Accessed on: 12 December 2024): Meta AI, Menlo Park, CA, USA	92% and 89–90% and 88%	In-house dataset (1243 knee MRI)
Richardson. 2021 [34]	USA	Current Problem in Diagnostic Radiology	CNN	Python (version 1.2.2); Python Software Foundation, Wilmington, DE, USA	99–99%	In-house dataset (2007 images)
Shin et al. 2022 [35]	Korea	Medicine	VGGNet model CNN	TensorFlow (Available online: https://www.tensorflow.org, Accessed on: 12 December 2024) Google, Mountain View, CA, USA PyTorch (Available online: https://pytorch.org, Accessed on: 12 December 2024): Meta AI, Menlo Park, CA, USA	94%–NA	In-house dataset (130 images)
Tran et al. 2022 [36]	France	European Radiology	CNN	TensorFlow (Available online: https://www.tensorflow.org, Accessed on: 12 December 2024) Google, Mountain View, CA, USA	AUC 94–91%	In-house dataset vs. two esternal dataset (Stanford dataset and KneeMRI)
Wang et al. 2024 [37]	China	Arthroscopy	CNN	TensorFlow (Available online: https://www.tensorflow.org, Accessed on: 12 December 2024) Google, Mountain View, CA, USA PyTorch (Available online: https://pytorch.org, Accessed on: 12 December 2024): Meta AI, Menlo Park, CA, USA	96–95%	Internal dataset (22,767 MRIs) vs. external validation dataset (4086 MRIs)

Table 1. *Cont.*

Authors and Year	Where	Journal	Method	Software	% Accuracy and % Specificity	Dataset
Wang et al. 2024 [38]	China	QIMS	CNN (YOLOv5m and ResNet-18)	PyTorch (version 1.11.0) Meta AI, Menlo Park, CA, USA	95–95%	OAI dataset [39] (1589 knees) vs. external (Stanford and kneeMRI dataset)
Wang et al. 2024 [40]	China	Tomography	SGNET model CNN	PyTorch (Available online: https://pytorch.org, Accessed on: 12 December 2024) Meta AI, Menlo Park, CA, USA	92–92%	Stanford dataset
Xue et al. 2024 [41]	China	Nature	U-Net CNN	Python (version 2.7); Python Software Foundation, Wilmington, DE, USA	99–97%	In-house dataset (862 partici-pants)
Zhang et al. 2020 [42]	China	Journal of Magnetic Resonance Imaging	3D DenseNet CNN	ITK-SNAP software (v. 3.6; Available online: http: //www.itksnap.org, Accessed on: 12 December 2024)	96–94%	In-house dataset (408 subjects) + data augmentation

Table 2. Quality assessment of included studies according to the modified Newcastle–Ottawa scale.

	Criteria								Total	Quality
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**		
Awan et al. (2021) [23]	1	1	1	1	1	1	1	1	8	High
Awan et al. (2023) [5]	1	1	1	1	1	1	1	0	7	High
Chang et al. (2019) [7]	1	1	1	1	1	1	0	1	7	High
Chen et al. (2022) [24]	1	1	1	1	1	1	1	1	8	High
Cheng et al. (2024) [16]	1	1	1	1	1	1	1	1	8	High
Dung et al. (2023) [25]	1	1	1	1	1	1	1	1	8	High
Germann et al. (2020) [26]	1	1	1	1	1	1	1	1	8	High
Jeon et al. (2021) [27]	1	1	1	1	1	1	1	1	8	High
Joshi, Suganthi (2022) [28]	1	1	1	1	1	1	1	1	8	High
Li et al. (2023) [29]	1	1	1	1	1	1	1	1	8	High
Li et al. (2021) [30]	1	1	1	1	1	1	1	1	8	High
Liang et al. (2023) [9]	1	1	1	1	1	1	1	1	8	High
Liu et al. (2019) [31]	1	1	1	1	1	1	1	1	8	High
Minamoto et al. (2022) [32]	1	1	1	1	1	1	0	1	7	High
Namiri et al. (2020) [33]	1	1	1	1	1	1	1	1	8	High
Richardson (2021) [34]	1	1	1	1	1	1	1	1	8	High
Shin et al. (2022) [35]	1	1	1	1	1	1	1	0	7	High
Tran et al. (2022) [36]	1	1	1	1	1	1	1	1	8	High
Wang et al. (2024) [38]	1	1	1	1	1	1	1	1	8	High
Wang et al. (2024) [40]	1	1	1	1	1	1	1	1	8	High
Wang et al. (2024) [37]	1	1	1	1	1	1	1	1	8	High
Xue et al. (2024) [41]	1	1	1	1	1	1	1	1	8	High
Zhang et al. (2020) [42]	1	1	1	1	1	1	1	1	8	High

3. Results and Discussion

Although ACL injury can be assessed clinically, MRI is the best non-invasive imaging technique utilized to confirm the diagnosis. Moreover, a precise diagnosis is pivotal to planning and performing successful ACL treatment [43,44]. In the Results Section, we consider the applications of DL models to detect ACL lesions on MRI.

A total of 42 relevant articles were identified through the initial search, resulting in 23 [5,7,9,16,23–38,40–42] studies that were included in the review (Figure 3).

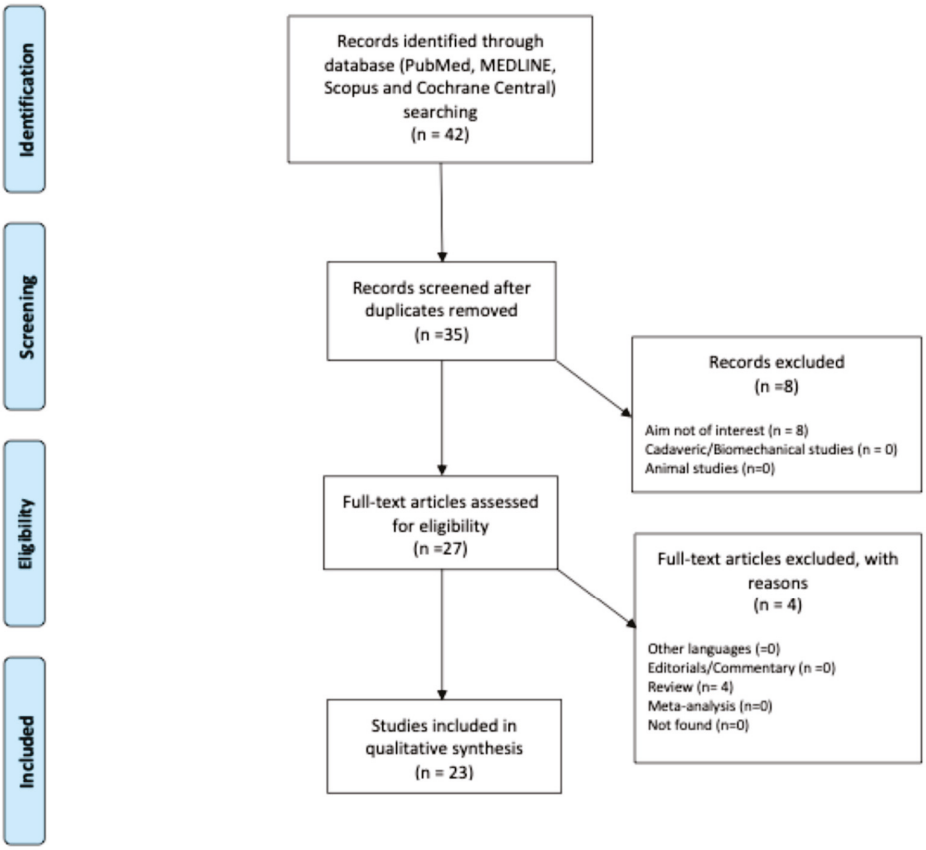


Figure 3. PRISMA flowchart.

The articles originated from 10 countries, with China having the most contributions (n = 9), followed by the United State of America (n =4), Japan (n = 2), Taiwan (n =2), and France, India, Vietnam, Korea, Taiwan, and Switzerland with 1 each.

All the studies only used CNN architectures in their methodology, not mentioning other methodologies like transformer-based and hybrid deep learning models. Specifically, only eight studies indicated the specific architecture type, as follows: MGACA, ResNet, DLCU-Net, CPDNN, VGGNet, SGNET, and DenseNet. All studies report the values of the accuracy and specificity of the models used, with values greater than 90%.

Another relevant element was the dataset used; the majority of the reports used an in-house dataset, with varying patient numbers, MRI images, MRI sequences, size images, and Tesla magnetic field (1.5 T vs. 3 T), while other studies used public dataset, such as the Chiba [27], Stanford (or MRNet or MRNetv-1.0) [45], and KneeMRI [46] datasets. For all studies, the examinations have been split into a training set, a validation set, and a hidden test set. The Chiba dataset include a total of 1177 MRI scans of knees collected from two institutions in Chiba, Japan, between 2014 and 2018. A 3.0 T MRI and a 1.5 T MRI l were used for Institute H and Institute K, respectively. The Stanford dataset, used in six papers, consists of 1370 knee MRI exams performed at Stanford University Medical Center. Examinations were performed using GE scanners, with a 3 T or a 1.5 T magnetic

field. The KneeMRI is a publicly available dataset collected by Stajduhar et al. that contains 917 sagittal PD-weighted examinations acquired with a Siemens Avanto 1.5 T scanner at Clinical Hospital Centre Rijeka, Croatia, from 2007 until 2014 [46].

Furthermore, four studies compare the accuracy and specificity of CNN's model between two MRI datasets, one in-house dataset and a public one, with comparable results.

Some authors, to mitigate constraints posed by limited dataset sizes, used domain adaptation methodologies and approaches such as data augmentation. Data augmentation [47] is a technique for artificially enlarging a training dataset by applying various transformations to the existing data. This technique is often used in machine learning and deep learning tasks, especially in the field of computer vision, to improve the generalization and robustness of the trained models. Common data augmentation techniques used are rotation, translation, scaling, flipping, shearing, zooming, brightness and contrast adjustment, and noise addition.

DL algorithms have been reported to improve the ability to diagnose cruciate ligament injuries using MRI. A first study applying AI to MRI [34] provided evidence that CNNs can be used to compare the diagnostic efficacy of different MRI sequences as follows: different fat-saturated (FS) and not-fat-saturated (NFS) images, including T1-weighted, T2-weighted, and STIR sequences, were considered. The results suggested that both FS and NFS sagittal images are suitable for the detection of ACL tears. The authors also emphasized that human variability is not a problem with a CNN, as the same result is expected every time for a given image set.

Shin et al. [35] created a CNN model for diagnosing ACL tears by using only one oblique sagittal image on which the largest ACL was observed, thus reporting a high diagnostic accuracy. Similarly, Minamoto et al. [32] evaluated the accuracy of a CNN system by a single sagittal image to compare the results with those of experienced human readers. One hundred sagittal MRI images from patients with and without ACL injuries were cropped and used for CNN training. The system demonstrated a sufficient capacity to define an ACL tear, with a higher specificity compared to human readers. Alternatively, other studies evaluated the role of the acquisition of multiple MRI slices. Chang et al. [7] used coronal MRI images of 260 subjects as input data as follows: 130 patients had ACL completely torn, whereas the other 130 had an intact ACL. The authors reported an accuracy of 96% of the CNN model to detect an ACL injury. Liu et al. [31] used whole T2-weighted fat-suppressed MRI images (coronal, sagittal, and axial planes) of 175 patients with complete ACL tear and 175 subjects with an intact ACL. The CNN model showed an area under the curve (AUC) of 0.98. Furthermore, several studies demonstrated that models constructed using multi-sequence MRI exhibited a significant performance advantage over their single-sequence counterparts [48–50].

Other investigators compared [26] the diagnostic performance of a DL model according to the magnetic field strength. The authors demonstrated that DL model sensitivity was comparable to that obtained by academic musculoskeletal radiologists (96% vs. 98%, respectively) at both 1.5 T and 3 T magnetic field strengths. However, the specificity of the DL model was significantly lower (93% vs. 100%, respectively). Furthermore, the subgroup analysis showed that the performance of the DL model decreased if MRI examinations were from outside institutions. An additional study, which developed a DL model based on a training dataset from 5 different centers, yielded an AUC of 0.99 and a sensitivity and specificity of 95.1% for both. The AI assistance significantly improved the accuracy of all clinicians, exceeding 96%. The most significant improvements in accuracy and time efficiency were reported for clinicians with moderately limited diagnostic expertise [37].

Furthermore, several studies reported application of 2D and 3D CNN models. The performance of a 3D CNN compared to a 2D CNN model was also investigated. The 2D

convolutional kernels can use context about the height and width of the slice to make predictions. However, since 2D CNNs use a single slice as input, they are inherently unable to use the context from adjacent slices. Three-dimensional CNNs address this problem, improved the performance, but comes at a computational cost due to the higher number of parameters used. Despite this, several studies showed no clear difference in accuracy between the two methods. Namiri et al. [33] presented a fully automated ACL segmentation and a classification framework that provided hierarchical severity staging of ACL injuries. Notably, a higher accuracy was observed with the 2D model. Both the 2D and 3D architectures displayed a relatively high degree of sensitivity and specificity for intact, fully torn, and reconstructed ACLs. Similarly, other authors showed that 3D DL-based diagnosis systems yielded comparable results to 2D models [29]. Jeon et al. [27] designed a 3D CNN model to automatically extract relevant features, focusing on key regions where ACL tears are likely to occur. The model demonstrated good performance in the diagnosis of ACL tears, with accuracy comparable to expert radiologists. Furthermore, the authors highlighted that the specific lightweight nature of the model allowed it to be used in settings with limited computational resources, such as mobile devices or clinics with less powerful hardware. Awan et al. [5] proposed a further method for the localization of a knee ACL injury region called multi-scale guided attention-based context aggregation (MGACA). The authors demonstrated the superiority of the proposed MGACA method in comparison to other localization models in terms of overfitting, underfitting, and generalization.

In the most recently published paper included in the current review [40], the authors proposed a DL model consisting of three main modules as follows: a Dual-Scale Data Augmentation module to enrich the training data on both the spatial and layer scales; a selective group attention module to capture relationships across the layer, channel, and space scales; and a fusion module to explore the inter-relationships among various perspectives to achieve the final classification. Results demonstrated the superior performance of the proposed SGNET model in ACL tear detection compared to the MRNet [45], DLD [36], ELNet [51], VIT [52], and Med3D [53] models, achieving an accuracy of 0.925, a sensitivity of 0.926, a specificity of 0.924, and an AUC of 0.975.

Traditional CNNs have demonstrated significant potential in recognizing patterns in images, such as ACL lesions. These networks can learn hierarchical features from raw images without requiring the manual extraction of specific characteristics like ligament continuity or the presence of bone marrow edema. The ability to perform end-to-end learning has simplified the diagnostic process and reduced inter-observer variability. A notable limitation of traditional CNNs is their difficulty in handling small datasets, which is common in medical imaging [42]. Overfitting is a common issue in such cases, where the model tends to memorize specific details from the training data without effectively generalizing to new examples. Additionally, CNNs are often criticized for their black-box nature, meaning that the decision-making process of the model is not easily interpretable [32]. This lack of transparency can be a significant barrier to clinical adoption, where understanding the rationale behind the model's diagnosis is crucial.

VGGNet, with its deep architecture consisting of 3×3 convolutional layers and 2×2 pooling operations, provides a simple yet effective approach for medical image analysis [35]. The most common versions are VGG16 and VGG19, which have 16 and 19 layers, respectively. Transfer learning using pre-trained models from large datasets like ImageNet allowed for the adaptation of VGGNet to detect ACL lesions, significantly improving performance on smaller datasets [30]. Among its advantages, it is easy to implement and adapt to different classification tasks. This works well with moderately sized datasets. Despite its advantages, VGGNet is computationally expensive due to the large number of parameters, making it challenging to implement in real-world clinical

settings. Additionally, the model is prone to overfitting on small datasets, which requires techniques like data augmentation or transfer learning from more generalized datasets to improve robustness. Futhermore, VGGNet's depth and number of parameters make it a network that requires a lot of memory and computational power. Compared to more modern networks, such as ResNet, it is less efficient [30].

ResNet introduced the concept of residual connections, allowing networks to be extremely deep without suffering from the issue of vanishing gradients. This is particularly beneficial for ACL lesion diagnosis, where complex tissue patterns require significant depth to be effectively captured [23]. The ability to build very deep networks without compromising performance has led to improved accuracy in detecting even partial lesions. Among its advantages, it allows us to train deep networks without vanishing gradient problems, and it has excellent performance in computer vision competitions. However, ResNet still requires significant computational resources, especially with deeper models like ResNet-50 or ResNet-101, and it may not be suitable for environments with limited hardware [9]. Furthermore, like other deep learning models, ResNet requires large annotated datasets to achieve optimal performance, a critical factor given the limited availability of clinical datasets for ACL lesions.

DenseNet employs dense connections between layers, which enhances gradient flow and promotes feature reuse. This is particularly useful for extracting fine-grained details, which are crucial for identifying small ACL lesions in MRI scans [42]. Furthermore, DenseNet is parameter-efficient, reducing the overall number of parameters compared to networks such as VGGNet. Among the advantages, it reduces the number of parameters needed without compromising the capacity of the network and improves generalization and feature retrieval. Despite its efficiency in terms of parameters, DenseNet is still computationally intensive, especially when used with deeper models. Memory consumption during both training and inference can be a challenge in environments with limited resources [42].

The choice of the most suitable CNN model for ACL lesion diagnosis depends on several factors, including the availability of computational resources, the size of the dataset, and the complexity of the lesions. Deep architectures such as ResNet and DenseNet offer excellent lesion detection capabilities, especially for complex or partial injuries, but they require significant computational power.

While a growing body of studies has investigated the applicability of DL models to the MRI detection of ACL injuries, these applications are still primarily limited to the experimental realm. Joshi et al. [28] showed limitations related to training and testing time, performance on imbalanced hyperparameter tuning, computational complexity, generalization capability, and dataset size. Indeed, one of the big limitations in training CNN models is data scarcity in medical imaging. Approaches such as data augmentation (technique for creating new samples from the existing data and introducing additional variations) and synthetic data generation (artificially generated data, such as simulation, generative models, or data generation algorithms, which can be used to supplement or replace real-world data in machine learning and other applications) to mitigate constraints posed by dataset sizes should be considered.

Limitations include class imbalance and variability difficulties associated with MRI scans across studies. Many ACL investigations employ different MRI machines, field strengths, and acquisition techniques, which could substantially affect model performance. Note that different authors use different datasets of varying sizes, which does not allow for standardization in comparing models.

In addition, almost all studies included in the current review analyzed datasets from multiple centers with varying MRI scan quality. Overall, the power of DL models cannot

be overlooked; in this context, Liang et al. [9] tested a CNN-based MRI system for ACL injury using arthroscopy as a standard reference and obtained an AUC of 0.89.

Regarding clinical applicability, reports indicate that DL models may not accurately identify partial ACL tears and may struggle to distinguish between complete and partial tears. Another limitation is that AI technologies may prove disproportionately expensive upon introduction [54]. Furthermore, the introduction of AI may raise concerns and skepticism among some clinicians and patients, as it represents a significant departure from traditional practices. The ongoing research will also inevitably spark debate on whether experienced clinicians or AI offers greater reliability [55]. Moreover, widespread AI implementation has brought to the forefront crucial ethical considerations, including legislative and privacy concerns.

Prospectives in the use of CNNs for the diagnosis of ACL injuries are very promising, with several directions for development that could revolutionize the diagnosis and treatment of these injuries. Neural networks could evolve by making themselves capable of detecting even the smallest or earliest ACL injuries, which might escape traditional visual diagnosis. This would lead to earlier diagnosis, reducing the risk of complications and combined injurie. They may also be able to distinguish between partial and complete injuries, and even suggest possible treatments based on severity. Another aspect could be the automatic detection of associated meniscal injuries. CNNs could evolve to detect both injuries simultaneously, reducing the need for multiple diagnostic investigations [3].

Currently, CNNs are mainly used to analyze MRI, but there are prospects for integrating computed tomography (CT) [56], ultrasound, and X-rays as well. Integrating these different techniques into a single AI system could improve the accuracy of diagnosis and reduce dependence on a single imaging modality.

Training CNNs requires large datasets, which are expensive and difficult to obtain. Self-supervised learning techniques could be an important step toward reducing the need for datasets, allowing models to learn from huge amounts of unlabeled data, such as medical images. Another possibility is the use of transfer learning, where models pretrained on other tasks (such as general image recognition) can be adapted to the specific problem of ACL diagnosis [54]. This could accelerate the adoption of CNNs in diagnosis and improve the ability of models to adapt to new contexts with less specific data.

CNNs could be integrated into clinical decision support platforms to assist physicians in the diagnostic process by automatically suggesting likely diagnoses based on MRI images. These systems could also suggest treatment plans, monitor lesion evolution, and personalize patient management.

CNNs could be used not only in the initial diagnosis but also in the continuous monitoring of the ACL injury during treatment and recovery [57,58]. For example, CNNs could analyze longitudinal (pre- and post-operative) MRI images to monitor the healing of the ligament and detect any complications or reinjuries. In addition, the use of CNNs could expand to the assessment of post-operative rehabilitation, allowing for the automatic assessment of the improvement or deterioration of the ligament during recovery. This would be useful for adapting the therapeutic plan in real time.

Neural networks could evolve into hybrid models, combining CNNs with other AI architectures, such as recurrent neural networks (RNNs) or transformers. These models could also be able to process the temporal sequences of images or clinical data (e.g., MRI images collected during different phases of treatment), allowing for a deeper and more dynamic understanding of the injury and its recovery. Generative Adversarial Networks (GANs) could be used to generate synthetic imaging data that improves model training by increasing the amount and variety of data without the need to collect new images.

Another emerging field could be the collaboration between CNNs and robotic surgery [59]. CNNs could analyze, in real time, the images obtained by robotic devices during ACL repair operations, helping surgeons to pinpoint the exact location of the tear and optimize the accuracy of the treatment. The use of augmented reality (AR) could also be integrated into CNNs to visualize the tears in 3D and help surgeons plan the surgery. CNNs could analyze the images in real time and overlay diagnostic information or treatment plans directly on the AR view of the area in question.

In the future, CNNs could evolve to consider not only medical images but also the genetic, physiological, or clinical data of the patients. This would allow for the personalization of ACL diagnosis and treatment, based on the individual characteristics of the patient (e.g., age, gender, physical activity, and medical history) [60]. CNN-based predictive models that estimate the likelihood of recurrence or development of post-operative complications may also emerge, with the aim of providing more targeted and tailored treatments for the patient.

The prospects of using CNNs for the diagnosis of ACL injuries are highly promising and destined to revolutionize medicine. Due to the integration with other emerging technologies and the continuous evolution of deep learning algorithms, the future of ACL diagnostics could be increasingly automated, precise, and accessible.

4. Conclusions

Deep learning models for the MRI detection of ACL injury reported high values of accuracy and specificity, which is especially helpful for less experienced clinicians. Time efficiency was also demonstrated. Overall, the deep learning models have proven to be a valid resource, though they still require technological developments for implementation in daily practice.

Author Contributions: Conceptualization, M.M. and F.D.; methodology, M.M., F.D. and D.M.; data curation, M.M. and F.D.; writing—original draft preparation, M.M. and F.D.; writing—review and editing, M.M., F.D., D.M., K.C., S.C., D.L. and R.M.; supervision, M.M., F.D., D.M., G.G. and R.M.; project administration, M.M. and F.D. 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 data presented in this study are available upon reasonable request from the corresponding author.

Acknowledgments: Michele Mercurio and Giorgio Gasparini are members of the Executive Board of the Italian Orthopaedic Research Society (IORS). Giorgio Gasparini is the current President of the IORS.

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

References

1. Gianotti, S.M.; Marshall, S.W.; Hume, P.A.; Bunt, L. Incidence of anterior cruciate ligament injury and other knee ligament injuries: A national population-based study. *J. Sci. Med. Sport* **2009**, *12*, 622–627. [CrossRef] [PubMed]
2. Evans, J.; Mabrouk, A.; Nielson, J. *l: Anterior Cruciate Ligament Knee Injury*; StatPearls: Treasure Island, FL, USA, 2024.
3. Siouras, A.; Moustakidis, S.; Giannakidis, A.; Chalatsis, G.; Liampas, I.; Vlychou, M.; Hantes, M.; Tasoulis, S.; Tsaopoulos, D. Knee injury detection using deep learning on MRI studies: A systematic review. *Diagnostics* **2022**, *12*, 537. [CrossRef] [PubMed]
4. Corona, K.; Cerciello, S.; Vasso, M.; Toro, G.; D'ambrosi, R.; Pola, E.; Ciolli, G.; Mercurio, M.; Panni, A.S. Age over 50 does not predict results in anterior cruciate ligament reconstruction. *Orthop. Rev.* **2022**, *14*, 37310. [CrossRef] [PubMed]

5. Awan, M.J.; Mohd Rahim, M.S.; Salim, N.; Nobanee, H.; Asif, A.A.; Attiq, M.O. MGACA-Net: A novel deep learning based multi-scale guided attention and context aggregation for localization of knee anterior cruciate ligament tears region in MRI images. *PeerJ Comput. Sci.* **2023**, *9*, e1483. [CrossRef]
6. Familiari, F.; Tollefson, L.V.; Izzo, A.; Mercurio, M.; LaPrade, R.F.; Di Vico, G. A High-Grade Lachman's Exam Predicts a Ramp Tear of the Medial Meniscus in Patients with Anterior Cruciate Ligament Tear: A Prospective Clinical and Radiological Evaluation. *J. Clin. Med.* **2024**, *13*, 683. [CrossRef]
7. Chang, P.D.; Wong, T.T.; Rasiej, M.J. Deep Learning for Detection of Complete Anterior Cruciate Ligament Tear. *J. Digit. Imaging* **2019**, *32*, 980–986. [CrossRef]
8. Mercurio, M.; Corona, K.; Galasso, O.; Cerciello, S.; Morris, B.J.; Guerra, G.; Gasparini, G. Soccer players show the highest seasonal groin pain prevalence and the longest time loss from sport among 500 athletes from major team sports. *Knee Surg. Sports Traumatol. Arthrosc.* **2022**, *30*, 2149–2157. [CrossRef]
9. Liang, C.; Li, X.; Qin, Y.; Li, M.; Ma, Y.; Wang, R.; Xu, X.; Yu, J.; Lv, S.; Luo, H. Effective automatic detection of anterior cruciate ligament injury using convolutional neural network with two attention mechanism modules. *BMC Med. Imaging* **2023**, *23*, 120. [CrossRef]
10. Minici, R.; Mercurio, M.; Iannò, B.; Galasso, O.; Gasparini, G.; Laganà, D. Advantages of the Use of Axial Traction Magnetic Resonance Imaging (MRI) of the Shoulder in Patients with Suspected Rota-Tor Cuff Tears: An Exploratory Pilot Study. *Healthc. Switz.* **2023**, *11*, 724. [CrossRef]
11. Ng, W.H.A.; Griffith, J.F.; Hung, E.H.Y.; Paunipagar, B.; Law, B.K.Y.; Yung, P.S.H. Imaging of the anterior cruciate ligament. *World J. Orthop.* **2011**, *2*, 75. [CrossRef]
12. Holmes, J.; Sacchi, L.; Bellazzi, R. Artificial intelligence in medicine. *Ann. R. Coll. Surg. Engl.* **2004**, *86*, 334–338.
13. Loftus, T.J.; Tighe, P.J.; Filiberto, A.C.; Efron, P.A.; Brakenridge, S.C.; Mohr, A.M.; Rashidi, P.; Upchurch, G.R.; Bihorac, A. Artificial intelligence and surgical decision-making. *JAMA Surg.* **2020**, *155*, 148–158. [CrossRef] [PubMed]
14. Deo, R.C. Machine learning in medicine. *Circulation* **2015**, *132*, 1920–1930. [CrossRef] [PubMed]
15. Santomartino, S.M.; Kung, J.; Yi, P.H. Systematic review of artificial intelligence development and evaluation for MRI diagnosis of knee ligament or meniscus tears. *Skelet. Radiol.* **2024**, *53*, 445–454. [CrossRef]
16. Cheng, Q.; Lin, H.; Zhao, J.; Lu, X.; Wang, Q. Application of machine learning-based multi-sequence MRI radiomics in diagnosing anterior cruciate ligament tears. *J. Orthop. Surg.* **2024**, *19*, 99. [CrossRef]
17. Kumar, Y.J.; Wee, S.Y.; Perhakaran, V.K. A Systematic Review on Deep Learning Model in Computer-aided Diagnosis for Anterior Cruciate Ligament Injury. *Curr. Med. Imaging* **2024**, *20*, e15734056295157. [CrossRef]
18. Guermazi, A.; Omoumi, P.; Tordjman, M.; Fritz, J.; Kijowski, R.; Regnard, N.-E.; Carrino, J.; Kahn, C.E.; Knoll, F.; Rueckert, D.; et al. How AI May Transform Musculoskeletal Imaging. *Radiology* **2024**, *310*, e230764. [CrossRef]
19. Althnian, A.; AlSaeed, D.; Al-Baity, H.; Samha, A.; Dris, A.B.; Alzakari, N.; Abou Elwafa, A.; Kurdi, H. Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. *Appl. Sci.* **2021**, *11*, 796. [CrossRef]
20. Mercurio, M.; Carlisi, G.; Ostojic, M.; Imbrogno, A.; Galasso, O.; Gasparini, G. The Protective Role of the FIFA 11+ Training Program on the Valgus Loading of the Knee in Academy Soccer Players Across a Season. *Healthcare* **2025**, *13*, 73. [CrossRef]
21. D'Ambrosi, R.; Corona, K.; Cerciello, S.; Guerra, G.; Mercurio, M.; Galasso, O.; Valli, F.; Abermann, E.; Fink, C. Combining an Anterolateral Complex Procedure With Anterior Cruciate Ligament Reconstruction Reduces Graft Reinjury Without Increasing the Rate of Complications: A Systematic Review and Meta-analysis of Randomized Controlled Trials. *Am. J. Sports Med.* **2025**, *12*, 3635465241285887. [CrossRef]
22. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Ann. Intern. Med.* **2009**, *151*, 264–269. [CrossRef] [PubMed]
23. Javed Awan, M.; Mohd Rahim, M.; Salim, N.; Mohammed, M.; Garcia-Zapirain, B.; Abdulkareem, K. Efficient Detection of Knee Anterior Cruciate Ligament from Magnetic Resonance Imaging Using Deep Learning Approach. *Diagnostics* **2021**, *11*, 105. [CrossRef] [PubMed]
24. Chen, K.-H.; Yang, C.-Y.; Wang, H.-Y.; Ma, H.-L.; Lee, O.K.-S. Artificial Intelligence–Assisted Diagnosis of Anterior Cruciate Ligament Tears From Magnetic Resonance Images: Algorithm Development and Validation Study. *JMIR AI* **2022**, *1*, e37508. [CrossRef] [PubMed]
25. Dung, N.T.; Thuan, N.H.; Van Dung, T.; Van Nho, L.; Tri, N.M.; Vy, V.P.T.; Hoang, L.N.; Phat, N.T.; Chuong, D.A.; Dang, L.H. End-to-end deep learning model for segmentation and severity staging of anterior cruciate ligament injuries from MRI. *Diagn. Interv. Imaging* **2023**, *104*, 133–141. [CrossRef]
26. Germann, C.; Marbach, G.; Civardi, F.; Fucentese, S.F.; Fritz, J.; Sutter, R.; Pfirrmann, C.W.A.; Fritz, B. Deep Convolutional Neural Network–Based Diagnosis of Anterior Cruciate Ligament Tears: Performance Comparison of Homogenous Versus Heterogeneous Knee MRI Cohorts With Different Pulse Sequence Protocols and 1.5-T and 3-T Magnetic Field Strengths. *Investig. Radiol.* **2020**, *55*, 499–506. [CrossRef]

27. Jeon, Y.S.; Yoshino, K.; Hagiwara, S.; Watanabe, A.; Quek, S.T.; Yoshioka, H.; Feng, M. Interpretable and lightweight 3-D deep learning model for automated ACL diagnosis. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 2388–2397. [CrossRef]
28. Joshi, K.; Suganthi, K. Anterior Cruciate Ligament Tear Detection Based on Deep Convolutional Neural Network. *Diagnostics* **2022**, *12*, 2314. [CrossRef]
29. Li, F.; Zhai, P.; Yang, C.; Feng, G.; Yang, J.; Yuan, Y. Automated diagnosis of anterior cruciate ligament via a weighted multi-view network. *Front. Bioeng. Biotechnol.* **2023**, *11*, 1268543. [CrossRef]
30. Li, Z.; Ren, S.; Zhou, R.; Jiang, X.; You, T.; Li, C.; Zhang, W. Deep Learning-Based Magnetic Resonance Imaging Image Features for Diagnosis of Anterior Cruciate Ligament Injury. *J. Healthc. Eng.* **2021**, *2021*, 4076175. [CrossRef]
31. Liu, F.; Guan, B.; Zhou, Z.; Samsonov, A.; Rosas, H.; Lian, K.; Sharma, R.; Kanarek, A.; Kim, J.; Guermazi, A.; et al. Fully Automated Diagnosis of Anterior Cruciate Ligament Tears on Knee MR Images by Using Deep Learning. *Radiol. Artif. Intell.* **2019**, *1*, 180091. [CrossRef]
32. Minamoto, Y.; Akagi, R.; Maki, S.; Shiko, Y.; Tozawa, R.; Kimura, S.; Yamaguchi, S.; Kawasaki, Y.; Ohtori, S.; Sasho, T. Automated detection of anterior cruciate ligament tears using a deep convolutional neural network. *BMC Musculoskelet. Disord.* **2022**, *23*, 577. [CrossRef] [PubMed]
33. Namiri, N.K.; Flament, I.; Astuto, B.; Shah, R.; Tibrewala, R.; Caliva, F.; Link, T.M.; Pedoia, V.; Majumdar, S. Deep Learning for Hierarchical Severity Staging of Anterior Cruciate Ligament Injuries from MRI. *Radiol. Artif. Intell.* **2020**, *2*, e190207. [CrossRef]
34. Richardson, M.L. MR Protocol Optimization With Deep Learning: A Proof of Concept. *Curr. Probl. Diagn. Radiol.* **2021**, *50*, 168–174. [CrossRef] [PubMed]
35. Shin, H.; Choi, G.S.; Chang, M.C. Development of convolutional neural network model for diagnosing tear of anterior cruciate ligament using only one knee magnetic resonance image. *Medicine* **2022**, *101*, e31510. [CrossRef] [PubMed]
36. Tran, A.; Lassalle, L.; Zille, P.; Guillin, R.; Pluot, E.; Adam, C.; Charachon, M.; Brat, H.; Wallaert, M.; d'Assignies, G.; et al. Deep learning to detect anterior cruciate ligament tear on knee MRI: Multi-continental external validation. *Eur. Radiol.* **2022**, *32*, 8394–8403. [CrossRef]
37. Wang, D.; Liu, S.; Ding, J.; Sun, A.; Jiang, D.; Jiang, J.; Zhao, J.; Chen, D.; Ji, G.; Li, N.; et al. A deep learning model enhances clinicians' diagnostic accuracy to more than 96% for anterior cruciate ligament ruptures on magnetic resonance imaging. *Arthrosc. J. Arthrosc. Relat. Surg.* **2024**, *40*, 1197–1205. [CrossRef]
38. Wang, M.; Yu, C.; Li, M.; Zhang, X.; Jiang, K.; Zhang, Z.; Zhang, X. One-stop detection of anterior cruciate ligament injuries on magnetic resonance imaging using deep learning with multicenter validation. *Quant. Imaging Med. Surg.* **2024**, *14*, 3405–3416. [CrossRef]
39. NIMH Data Archive—NDA. Available online: https://nda.nih.gov/oai%C2%BB (accessed on 24 February 2025).
40. Wang, X.; Wu, Y.; Li, J.; Li, Y.; Xu, S. Deep Learning-Assisted Automatic Diagnosis of Anterior Cruciate Ligament Tear in Knee Magnetic Resonance Images. *Tomography* **2024**, *10*, 1263–1276. [CrossRef]
41. Xue, Y.; Yang, S.; Sun, W.; Tan, H.; Lin, K.; Peng, L.; Wang, Z.; Zhang, J. Approaching expert-level accuracy for differentiating ACL tear types on MRI with deep learning. *Sci. Rep.* **2024**, *14*, 938. [CrossRef]
42. Zhang, L.; Li, M.; Zhou, Y.; Lu, G.; Zhou, Q. Deep Learning Approach for Anterior Cruciate Ligament Lesion Detection: Evaluation of Diagnostic Performance Using Arthroscopy as the Reference Standard. *J. Magn. Reson. Imaging* **2020**, *52*, 1745–1752. [CrossRef]
43. Mercurio, M.; Cerciello, S.; Corona, K.; Guerra, G.; Simonetta, R.; Familiari, F.; Galasso, O.; Gasparini, G. Factors Associated With a Successful Return to Performance After Anterior Cruciate Ligament Reconstruction: A Multiparametric Evaluation in Soccer Players. *Orthop. J. Sports Med.* **2024**, *12*, 23259671241275663. [CrossRef] [PubMed]
44. Mercurio, M.; Cofano, E.; Gasparini, G.; Galasso, O.; Familiari, F.; Sanzo, V.; Ciolli, G.; Corona, K.; Cerciello, S. Isolated ACL Reconstruction Versus Combined ACL and Anterolateral Ligament Reconstruction: Functional Outcomes, Return to Sport, and Survivorship: An Updated Meta-analysis of Comparative Studies. *Am. J. Sports Med.* **2025**, *53*, 971–980. [CrossRef]
45. Bien, N.; Rajpurkar, P.; Ball, R.L.; Irvin, J.; Park, A.; Jones, E.; Bereket, M.; Patel, B.N.; Yeom, K.W.; Shpanskaya, K.; et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med.* **2018**, *15*, e1002699. [CrossRef] [PubMed]
46. Štajduhar, I.; Mamula, M.; Miletić, D.; Ünal, G. Semi-automated detection of anterior cruciate ligament injury from MRI. *Comput. Methods Programs Biomed.* **2017**, *140*, 151–164. [CrossRef] [PubMed]
47. Keskin, D. Synthetic Data and Data Augmentation. Available online: https://medium.com/@dogankeskin01/synthetic-data-and-data-augmentation-c022029dd660 (accessed on 24 February 2025).
48. Wang, G.; He, L.; Yuan, C.; Huang, Y.; Liu, Z.; Liang, C. Pretreatment MR imaging radiomics signatures for response prediction to induction chemotherapy in patients with nasopharyngeal carcinoma. *Eur. J. Radiol.* **2018**, *98*, 100–106. [CrossRef]
49. Tsuchiya, M.; Masui, T.; Terauchi, K.; Yamada, T.; Katyayama, M.; Ichikawa, S.; Noda, Y.; Goshima, S. MRI-based radiomics analysis for differentiating phyllodes tumors of the breast from fibroadenomas. *Eur. Radiol.* **2022**, *32*, 4090–4100. [CrossRef]

50. Wu, M.; Xu, W.; Fei, Y.; Li, Y.; Yuan, J.; Qiu, L.; Zhang, Y.; Chen, G.; Cheng, Y.; Cao, Y.; et al. MRI-based clinical radiomics nomogram may predict the early response after concurrent chemoradiotherapy in locally advanced nasopharyngeal carcinoma. *Front. Oncol.* **2023**, *13*, 1192953. [CrossRef]
51. Tsai, C.-H.; Kiryati, N.; Konen, E.; Eshed, I.; Mayer, A. Knee injury detection using MRI with efficiently-layered network (ELNet). In Proceedings of the Third Conference on Medical Imaging with Deep Learning, Virtual, 6–8 July 2020; pp. 784–794.
52. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. Dinov2: Learning robust visual features without supervision. *arXiv* **2023**, arXiv:2304.07193.
53. Chen, S.; Ma, K.; Zheng, Y. Med3d: Transfer learning for 3d medical image analysis. *arXiv* **2019**, arXiv:1904.00625.
54. Familiari, F.; Galasso, O.; Massazza, F.; Mercurio, M.; Fox, H.; Srikumaran, U.; Gasparini, G. Artificial Intelligence in the Management of Rotator Cuff Tears. *Int. J. Environ. Res. Public. Health* **2022**, *19*, 16779. [CrossRef]
55. Cerciello, S.; Mercurio, M.; Corona, K.; Proietti, L.; Di Vico, G.; Giordano, M.C.; Morris, B.J. Posterior Cruciate Buckling Angle Variations Are Associated with Different Patterns of Medial Meniscus Tears in Anterior-Cruciate-Deficient Knees: Results of a Prospective Comparative Magnetic Imaging Resonance Study. *Healthcare* **2024**, *12*, 1553. [CrossRef] [PubMed]
56. Li, G.; Xie, Y.; Xu, B. Three-dimensional reconstruction with dual-source computed tomography for evaluating graft deformation and bone tunnel position following reconstruction of the anterior cruciate ligament. *Med. Eng. Phys.* **2022**, *110*, 103858. [CrossRef] [PubMed]
57. Han, M.; Singh, M.; Karimi, D.; Kim, J.-Y.; Flannery, S.W.; Ecklund, K.; Murray, M.M.; Fleming, B.C.; Gholipour, A.; Kiapour, A.M.; et al. LigaNET: A multi-modal deep learning approach to predict the risk of subsequent anterior cruciate ligament injury after surgery. *medRxiv* **2023**. [CrossRef]
58. Liuzzi, P.; Nesi, E.; Campagnini, S.; Mari, F.; Dimauro, I.; Carta, N.; Rocchi, J.; Bergamini, E.; Mannini, A.; Mariani, P. Predicting 60-day Recovery Outcomes After ACL Surgery Using Machine Learning. *Gait Posture* **2024**, *114*, S26. [CrossRef]
59. Park, S.G.; Park, J.; Choi, H.R.; Lee, J.H.; Cho, S.T.; Lee, Y.G.; Ahn, H.; Pak, S. Deep Learning Model for Real-time Semantic Segmentation During Intraoperative Robotic Prostatectomy. *Eur. Urol. Open Sci.* **2024**, *62*, 47–53. [CrossRef]
60. Baker, L.A.; Momen, M.; Chan, K.; Bollig, N.; Lopes, F.B.; Rosa, G.J.; Todhunter, R.J.; Binversie, E.E.; Sample, S.J.; Muir, P. Bayesian and machine learning models for genomic prediction of anterior cruciate ligament rupture in the canine model. *G3 Genes Genomes Genet.* **2020**, *10*, 2619–2628. [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

Clinical Applications of Artificial Intelligence (AI) in Human Cancer: Is It Time to Update the Diagnostic and Predictive Models in Managing Hepatocellular Carcinoma (HCC)?

Mario Romeo [1], Marcello Dallio [1,*], Carmine Napolitano [1], Claudio Basile [1], Fiammetta Di Nardo [1], Paolo Vaia [1], Patrizia Iodice [2] and Alessandro Federico [1]

[1] Hepatogastroenterology Division, Department of Precision Medicine, University of Campania Luigi Vanvitelli, 80138 Naples, Italy; mario.romeo@unicampania.it (M.R.); carmine.napolitano1@studenti.unicampania.it (C.N.); claudio.basile@studenti.unicampania.it (C.B.); fiammetta.dinardo@studenti.unicampania.it (F.D.N.); paolo.vaia@studenti.unicampania.it (P.V.); alessandro.federico@unicampania.it (A.F.)

[2] Oncology Division, Monaldi Hospital, 80131 Naples, Italy; patrizia.iodice@ospedalideicolli.it

* Correspondence: marcello.dallio@unicampania.it; Tel.: +39-0815664311

Abstract: In recent years, novel findings have progressively and promisingly supported the potential role of Artificial intelligence (AI) in transforming the management of various neoplasms, including hepatocellular carcinoma (HCC). HCC represents the most common primary liver cancer. Alarmingly, the HCC incidence is dramatically increasing worldwide due to the simultaneous "pandemic" spreading of metabolic dysfunction-associated steatotic liver disease (MASLD). MASLD currently constitutes the leading cause of chronic hepatic damage (steatosis and steatohepatitis), fibrosis, and liver cirrhosis, configuring a scenario where an HCC onset has been reported even in the early disease stage. On the other hand, HCC represents a serious plague, significantly burdening the outcomes of chronic hepatitis B (HBV) and hepatitis C (HCV) virus-infected patients. Despite the recent progress in the management of this cancer, the overall prognosis for advanced-stage HCC patients continues to be poor, suggesting the absolute need to develop personalized healthcare strategies further. In this "cold war", machine learning techniques and neural networks are emerging as weapons, able to identify the patterns and biomarkers that would have normally escaped human observation. Using advanced algorithms, AI can analyze large volumes of clinical data and medical images (including routinely obtained ultrasound data) with an elevated accuracy, facilitating early diagnosis, improving the performance of predictive models, and supporting the multidisciplinary (oncologist, gastroenterologist, surgeon, radiologist) team in opting for the best "tailored" individual treatment. Additionally, AI can significantly contribute to enhancing the effectiveness of metabolomics–radiomics-based models, promoting the identification of specific HCC-pathogenetic molecules as new targets for realizing novel therapeutic regimens. In the era of precision medicine, integrating AI into routine clinical practice appears as a promising frontier, opening new avenues for liver cancer research and treatment.

Keywords: artificial intelligence; hepatocellular carcinoma; precision medicine

1. Background

Despite the recent considerable therapeutic progress [1,2], primary hepatobiliary liver tumors (HBLTs) continue to play a significant role in the vast array of cancerous diseases, representing the sixth most common neoplasm for global incidence and ranking

third in terms of mortality [3]. Among the HBLTs, the hepatocellular carcinoma (HCC) constitutes the most prevalent histotype, accounting for approximately 80% of all hepatic malignancies [1,3,4].

A multifactorial picture, where various exogenous risk factors promote the onset of this neoplasm in individuals with a susceptible genetic background, depicts the faithful portrait of the complex pathogenesis of HCC [2,4–7].

Several etiological agents contribute to the chronic hepatitis fueling the progression to advanced fibrosis (AF) and liver cirrhosis stage, where the hazard of HCC dramatically increases [2]; a percentage ranging from 1% to 8% of cirrhotic patients develop HCC annually [2].

Interestingly, different leading etiological factors of the HCC are observed in various geographic areas. The hepatitis B virus (HBV) infection represents the major etiological factor in most parts of Asia (except Japan) [8,9], South America, and Africa [10]; The hepatitis C virus (HCV) infection is the predominant cause in Western Europe, North America, and Japan; and alcohol intake constitutes the leading factor in Central and Eastern Europe [11].

Even though viral-related (HBV and HCV) chronic liver diseases continue to represent a major global factor contributing to the HCC onset, in recent years, the growing prevalence of metabolic dysfunction-associated steatotic liver disease (MASLD), due to the spreading of unhealthy lifestyle habits [12–14], has promoted a noticeable shift in the etiological panorama of HCC in Western countries [15,16]. Additionally, the long-term effects of HBV vaccination in newborns, as well as the current availability of effective treatments for viral chronic liver diseases, have substantially contributed to this epidemic-etiological change, simultaneously with the increasing rise of MASLD [2,15]. MASLD currently represents the predominant hepatopathy, rapidly affirming itself as the most common cause of hepatic damage, liver cirrhosis, hepatic transplantation, and HCC [17,18].

Epidemiological projections for HCC in the MASLD context appear even more alarming, considering that obesity and diabetes, which represent crucial cardiometabolic factors configuring the "metabolic dysfunction", are well-recognized and independent risk factors for HCC [16,19–21], in addition to the reported occurrence of this neoplasm even in patients without AF or cirrhosis escaping the dedicated traditional HCC surveillance programs [22–24].

Altogether, this evidence stresses the urgent need for a tailored approach, based on the development of tools appropriately stratifying the HCC risk and aiming to implement the most suitable management strategy [25]. In this sense, the identification of HCC in the early stages appears crucial in opening different "resolutive" therapeutic scenarios, that would otherwise be precluded in the more advanced stages of the disease [26]. Nonetheless, the low survival rate, particularly in advanced HCC, combined with the high recurrence rate in these stages, has created a substantial obstacle in the battle against this cancer, reaching what seems to be a plateau in terms of management progress [27].

Therefore, the need to explore new effective approaches appears to be a crucial challenge that has substantially contributed to the adoption of Artificial Intelligence (AI) as a promising ally. In line with this, investigating the potential of AI-based applications for optimizing the identification, diagnostic, and therapeutic strategies for liver cancer represents an urgent need, especially considering the above-presented relative epidemiologic burden and the changes that recently arose in the etiology of this neoplasm.

This review introduces AI and its potential in oncology and aims to explore the current applications of AI in managing HCC across all phases (screening, diagnosis, treatment, and prognosis). In this review, the status of the AI-based approaches for diagnosis and treatment of HCC is presented, providing insights into whether this emerging tool can already be

considered a valuable "weapon", enlarging the available armamentarium against this neoplasm in routine clinical practice.

2. Artificial Intelligence: An Emerging Tool Revolutionizing the Management of Human Cancer

AI is defined as the ability of a machine to imitate and replicate cognitive processes typical of human intelligence, such as reasoning, learning, planning, and creativity. AI systems are designed to autonomously adapt their behavior by analyzing the outcomes of previous actions and making independent decisions based on these analyses [28].

In recent years, AI has become a topic of growing interest for the scientific community, with an exponential increase in terms of research, particularly in the oncology field [28–30]. AI mimics and implements processes inherent to human intelligence via technological approaches: deep learning (DL) and machine learning (ML).

Machine Learning (ML) focuses on developing algorithms and models that enable systems to learn from data, improving their performance over time without needing explicit programming [31,32]. The primary ML approaches are supervised learning, unsupervised learning, and reinforcement learning. ML, because of its focus on training algorithms to identify patterns and make predictions, holds promise for early disease diagnosis by analyzing large volumes of clinical data [33,34].

A potentially significant role of ML in cancer-related clinical practice has been revealed in radiological imaging. As the most relevant example, the ML application to magnetic resonance imaging (MRI) has been shown to enhance tumor detection in human breast cancer [35]. In particular, the use of both ML-supervised and unsupervised methods for tasks such as anatomical segmentation, lesion identification, texture analysis, and radiomics can contribute to the standardization of image interpretation, thereby improving both sensitivity and specificity, reducing false negatives, and potentially leading to decreased costs and shorter scan durations [35]. Moreover, a recent meta-analysis revealed a sensitivity and specificity comparable to that of a radiologists' "human reading" in detecting breast cancer in patients undergoing mammography for the "computer-aided detection" (CADe) and "computer-aided diagnosis" (CADx) ML algorithms [36]. However, the retrospective nature of the considered studies, as well as the relatively small sample sizes, represented serious limitations of this research, suggesting the need for large datasets to fully leverage the potential of ML [36]. Additionally, another significant limit of ML application is related to the inherent complexity of the decision-making process, which often complicates the understanding of how a particular response is reached, leading to the characterization of these models as a "black box" [28,29,31].

Deep Learning (DL), utilizes deep neural networks to learn from vast amounts of data [31,32]. A neural network is a computational system inspired by the human brain, designed to detect patterns within the data. These networks consist of nodes, known as "artificial neurons", which are organized into layers. These layers are categorized into three main types: (a) the first input layer, which receives raw data; (b) middle hidden layers, where complex calculations are performed to extract features and identify patterns within the data; (c) the final output layer, which provides the model's prediction.

A distinguishing characteristic of DL is the presence of multiple hidden layers in the neural networks, enabling the model to learn progressively more abstract and complex representations of the data [31,32].

In contrast to the traditional ML, DL is particularly adept at handling unstructured data (images and videos), thanks to the power of neural networks [37]. In this sense, DL can be mainly effective in processing complex data, such as medical images (both radiological and histopathological), molecular networks, and genetic sequences, with a

large potential range of applications in the oncological field, especially regarding genomics, transcriptomics, and radiomics [38–41]. Additionally, for the DL approaches utilizing transcriptome data, the utility in predicting primary tumor sites in cancer of unknown origin has been revealed [42,43]. Furthermore, "SCOPE", a neural network trained on the whole of the transcriptome "The Cancer Genome Atlas Program" ("TCGA") data, was shown to predict the origins of treatment-resistant metastatic cancers, including rare neoplasms, such as the metastatic adenoid cystic carcinoma [42]. Another relevant research recently presented a DL radio-clinical signature (DLCS) model combining imaging features with clinical risk factors in order to accurately predict the tumor response and to identify the risk of Overall Survival (OS) in locally advanced gastric cancer (LAGC), proposing a novel tool, potentially applicable to guide the personalized treatment approaches through computerized tumor-level characterization [44].

However, similarly to the above-presented ML-related application limits, DL models for image recognition require extensive and well-curated datasets, often consisting of thousands of cases, with the exigence of external validation being a crucial need [28,29,37]. In this sense, assembling such large datasets can be challenging, particularly when the data sources are represented by multiple healthcare centers. Moreover, the variability in the parameters derived from the images produced by scanners from different manufacturers poses another hurdle, as it can affect the consistency and reliability of the extracted features, ultimately complicating the training and generalization of the models across diverse clinical environments [32,37].

Despite these limitations, ML and DL-based approaches promisingly represent encouraging frontiers to be further explored and improved, potentially revolutionizing clinical strategies guiding the management of several human tumors, including HCC.

3. Artificial Intelligence in the Management of Hepatocellular Carcinoma

3.1. From Standard to Risk-Based HCC Surveillance: AI Contributes to Individual Stratification Risk

Among the several strategies aimed at improving the prognosis of HCC patients, major efforts have been made to characterize an adequate surveillance program, providing early identification of the disease for recognized high-risk individuals, thus ensuring a larger range of "curative"/"resolutive" therapeutic opportunities [26].

The term "HCC surveillance" is defined as a screening for HCC performed at regular intervals [45]. It is currently estimated that carrying out surveillance has a good cost/benefit ratio; therefore, it is recommended for certain categories of patients.

The standard, not-personalized (according to the tools estimating individual risk) surveillance program, recommended by the European Association for the Study of the Liver (EASL), includes a liver ultrasound (US) performed every six months, which should always be performed by an experienced medical operator [46]. Despite these recommendations, the HCC surveillance program can be optimized and may benefit from improvements, considering that less than half of the European patients who should receive surveillance do not receive it for reasons including lack of awareness, adherence, and access [47]. Additionally, for a considerable portion of patients, the currently available strategy is not effective in properly detecting early HCC, even considering the reported HCC occurrence in patients without AF or cirrhosis in MASLD [22,23], escaping the dedicated traditional HCC surveillance program's identification.

To optimize the current strategy, emerging evidence has proposed novel, alternative, risk-based surveillance models [48]. In this sense, risk-based surveillance, by tailoring the HCC surveillance strategy to the individual risk of developing HCC, promises to be a superior method compared to the standard approach [48].

The main challenge in this situation is selecting the relatively adequate diagnostic test and stratifying individual HCC risk, as well as exploring the potential contribution of AI.

Current guidelines suggest adopting MRI as a surveillance tool for HCC in selected patients, for whom the US is ineffective [46]. However, the specific indications, imaging sequences, and intervals for MRI surveillance remain unclear.

Altogether, the current knowledge status on this topic highlights the relevance of properly stratifying the patients for tailored HCC risk, rather than merely evaluating the general performance of a single diagnostic test, in a desirable scenario where the mutual binomial "optimal test—individual risk" should drive the strategy.

In this regard, different HCC risk stratification scores, including the Age–Male–ALBI–Platelets Score (aMAP) [49–51] and the Toronto HCC Risk Index (THRI), among others, have been proposed [52]. However, despite these encouraging findings, developing and validating regression models remains a complex challenge, especially considering the wide individual fluctuation of the HCC risk over time, and with the nonlinear changes being hard to estimate using the existing tools [53,54].

The rapid expansion of available electronic health record (EHR) longitudinal data has represented a useful aspect of exploiting AI, ensuring some progress both in viral and MASLD settings [25,55,56] (Figure 1).

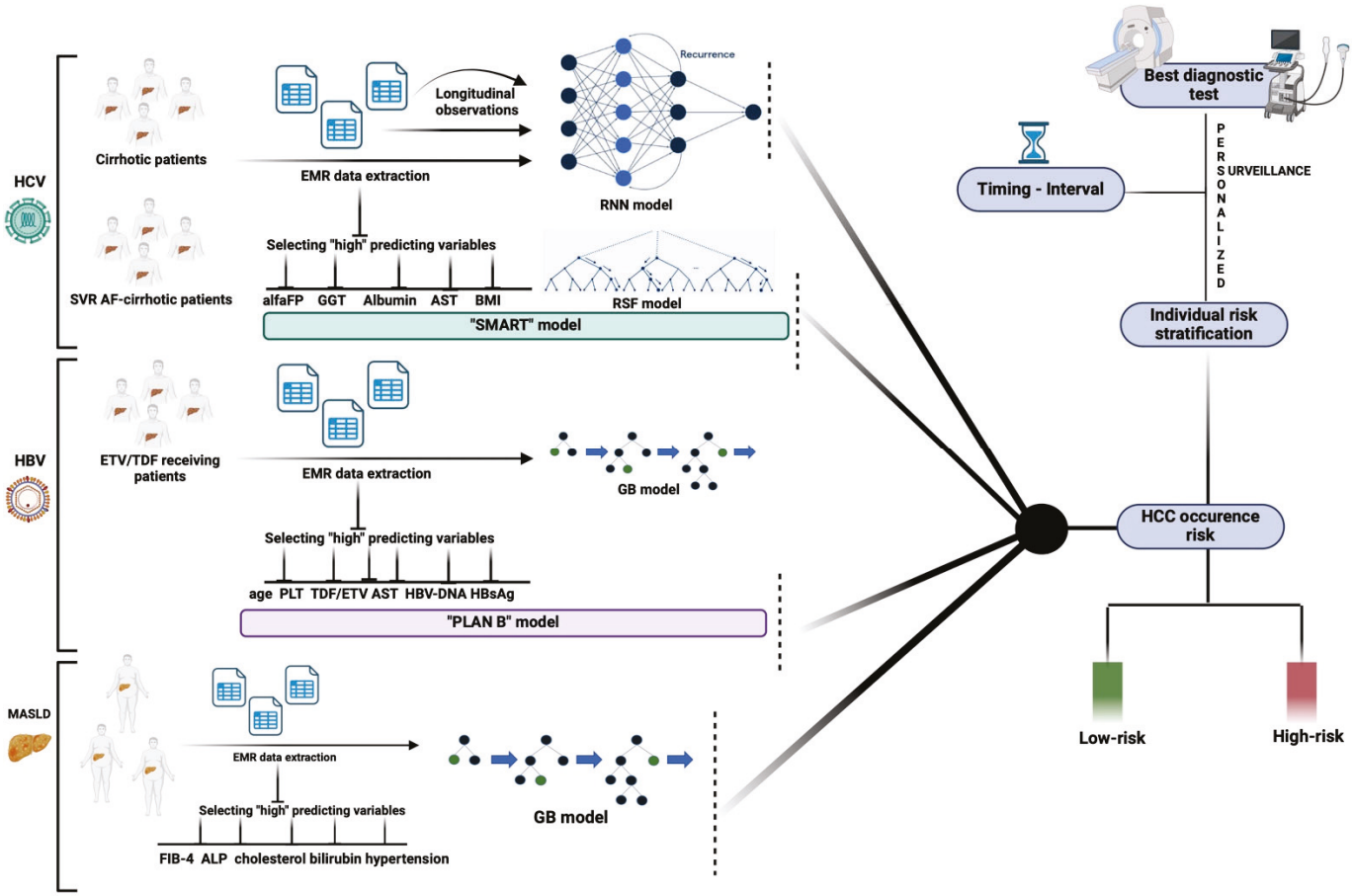


Figure 1. Potential AI contribution to developing tailored HCC surveillance models according to chronic liver disease etiologies. HBV: Hepatitis B virus; HCV: Hepatitis C virus; MASLD: Metabolic dysfunction-associated steatotic liver disease; EMR: Electronic medical record; alfa: alpha-fetoprotein; GGT: gamma-glutamyl transferase; BMI: body mass index; AST: aspartate aminotransferase; RNN: recurrent neural network; RSF: random survival forest; TDF: tenofovir; ETV: entecavir; PLT: platelet count; HBsAg: HBV-surface antigen; FIB-4: Fibrosis 4 score; GB: gradient-boosting.

Ioannou et al. introduced a novel approach for the identification of high-risk patients, which could be particularly useful when the limited resources do not allow for an adequate surveillance of all high-risk patients. The research included 48,151 patients with HCV–related cirrhosis in the national Veterans Health Administration, who had at least 3 years of follow-up after the diagnosis of cirrhosis. The DL recurrent neural network (RNN) models, using raw longitudinal data extracted directly from EHRs, outperformed the conventional linear regression models, suggesting that risk-based HCC outreach and surveillance strategies using RNN models could be used to identify patients with HCV-related cirrhosis with a high risk of developing HCC [56].

In this scenario, individualized HCC surveillance, using risk stratification scores in AF and cirrhotic HCV patients achieving sustained virological response (SVR), represents another relevant question. For this purpose, various risk stratification scores, including the "General Evaluation Score" (GES), have been progressively developed; and moreover, prospective research encouragingly revealed that the use of GES was associated with an improved early-stage detection and the reception of curative treatment [57,58]. Once again, AI appears to ensure a solid contribution even in this research challenge.

Minami et al., in a multicenter cohort study enrolling 1742 HCV-SVR patients, developed an ML-based random survival forest (RSF) model by using seven commonly measured parameters [age, platelet count, alpha-fetoprotein (AFP), gamma-glutamyl transferase (GGT), body mass index (BMI), albumin, and aspartate aminotransferase (AST) ("SMART" model)], showing a good predictive ability and potentially providing a personalized novel surveillance system [59].

For the chronic HBV infection, Kim HY, Lampertico P, et al. aimed to develop and validate an AI-assisted prediction model of HCC risk using a gradient-boosting machine (GBM) algorithm [60]. The research involved a derivation cohort of 6051 chronic HBV-infected patients who received HBV-antiviral therapy (entecavir or tenofovir) and two external validation cohorts. The emerging model, named PLAN-B, comprised ten baseline parameters: cirrhosis, age, platelet count, entecavir/tenofovir, sex, serum ALT and HBV-DNA, albumin and bilirubin levels, and HBeAg status, and it showed significant superiority over the previous models in both validation cohorts [60].

Finally, for the metabolic setting, very recently, Sarkar et al. proposed a predictive model of HCC for MASLD patients, who are becoming the predominant chronic liver patients. A characteristic of MASLD is the occurrence of HCC even in patients without AF, making diagnosis more difficult and less likely to receive a liver transplant. By evaluating a validation cohort of five different ML algorithms, including decision tree (DT), gradient boosted (GB), naïve bayes (NB), probabilistic neural network (PNN), and random forest (RF), the GB-ML model, using EMR data, predicted HCC development at a 92.06% accuracy, with an AUC of 0.97, an F1 score of 0.84, 98.34% specificity, and 74.41% sensitivity [55]. The strongest clinical predictors of HCC in this model were the validated noninvasive tool for liver fibrosis assessment "Fibrosis-4 score" (FIB-4), alkaline phosphatase (ALP), total cholesterol, bilirubin, and hypertension [55].

However, despite these efforts and the encouraging findings in different chronic liver disorders etiologies, the possible exclusive use of AI-based scores to predict the HCC risk remains undefined, mainly because of their dependency on the size and diversity of the training dataset, suggesting that stronger evidence is required to adopt this approach in clinical practice largely.

Therefore, the data supporting the use of other promising modalities (including abbreviated MRI, serum biomarkers, and models) is limited, and the last update of EASL clinical practice guidelines on HCC surveillance continues to recommend a US of the liver every 6 months for HCC screening [61].

3.2. AI Facilitates HCC Diagnosis: From Imaging to Histopathology and Transcriptomics

As in a domino, an accurate diagnosis is crucial for adequate grading, staging, and therapeutic approaches, crucially influencing treatment strategies and patient prognosis in HCC. In malignant liver nodules, changes in vascularisation occur and can be appreciated by contrast-enhanced imaging, representing the basis of the non-invasive diagnosis of HCC [45,61].

Relevantly, non-invasive diagnosis can be sufficient when the additional information provided by histopathology will not influence a treatment choice [61].

In some clinical scenarios, histopathology is required, considering the relatively crucial role in confirming the malignant nature of a liver lesion, as well as providing essential insights influencing the therapeutic decision-making (including cellular differentiation, vascular and lymphatic invasion, and metastatic potential) [61].

However, in a routine clinical practice, the approach to a HCC diagnosis recurrently appears nonlinear, making the identification of this neoplasm and the differential diagnosis with other primary or secondary HBLCs a real challenge, particularly in a non-cirrhotic context.

In this scenario, AI-based models may help clinicians to correctly acquire and interpret the obtained results (imaging or histopathologic), facilitating the diagnostic process.

3.2.1. AI Applications Supporting Classical Imaging and Radiomics in Diagnosing HCC

Classical imaging represents the non-invasive standard for HCC definition: in cirrhotic, HBV-infected, or previous HCC-affected patients, a liver nodule can be diagnosed when the major features of HCC are observed with dynamic contrast-enhanced CT, MRI, or contrast enhancement ultrasound (CEUS) following the LI-RADS v2018 recommendations [61]. Routinely, HCC appears chameleonic, and several of the typical major imaging features are faded and not easily identifiable, making the diagnosis of this neoplasm very hard, particularly in patients without a suggestive clinical history (HBV, cirrhosis, etc.). In this sense, radiomics, a technology based on the quantitative extraction of image characteristics from radiological imaging modalities, that permits the identification of crucial imaging biomarkers that cannot be detected visually [62], has made progress in testing the ability of AI frameworks to improve the diagnostic accuracy of US, CT, and MRI in the setting of HCC diagnosis.

Ultrasound. The US results represent the first-line imaging modality to identify liver lesions in the clinical workflow. In this context, AI has the potential to assist less experienced radiologists in improving their performance and lowering their dependence on sectional imaging in HCC diagnosis.

Emerging research has investigated the potential role of AI in improving the diagnostic accuracy of the US HCC diagnosis.

In a multicenter study, Yang et al. developed and externally validated a deep convolutional neural network (DCNN) using a large, multicenter database of US images from 13 hospital systems [63].

The resulting model ("DCNN-US") showed superior diagnostic sensitivity and specificity compared to that of 15-year skilled radiologists and an accuracy comparable to that of a contrast-enhanced CT. The developed model presented an area under the receiver, operating a characteristic curve (AUROC) of 0.92 for differentiating benign from malignant liver lesions. This performance was comparable to (a) the diagnostic accuracy of clinical radiologists (both achieving 76.0%) and (b) the accuracy of contrast-enhanced (CT), showing a diagnostic accuracy of 84.7%, only slightly inferior to MRI at 87.9% [63].

A very recent retrospective study from Thailand incorporated 26,288 US images from 5444 patients to train the YOLOv5 model, which utilizes DL techniques, for focal liver

lesions (FLLs) detection and classification of seven different types of FLLs, including HCC and regenerative nodules. The AI achieved an overall FLLs detection rate of 84.8% (95%CI: 83.3–86.4), with the HCC detection at 82.3% (95% CI: 77.1–87.5), preceded by the cholangiocarcinoma at 92.2% (95%CI: 88.0–96.4) and the focal fatty infiltration FFS at 89.7% (95%CI: 87.1–92.3). The specificities and NPVs for regenerative nodules were 100% and 99.9% (95%CI: 99.8–100.0), respectively [64].

CEUS has been reported to show superior diagnostic performance compared with classical US B-mode images in defining FFLs [65]. For this reason, combining AI with CEUS could be very useful to differentiate benign lesions from HCC or other malignant liver lesions. In 2018, Guo LH et al. proposed a two-stage multiple-view learning framework for CEUS that was a combination of deep canonical correlation analysis and multiple kernel learning (DCCA-MKL), only adopting three typical CEUS images selected from the three phases of the exam (arterial, portal venous and late) [66].

Their results show that this framework best discriminates between benign and malignant liver cancers. Moreover, it was also proved that the three-phase CEUS image-based CAD is feasible for liver cancers with the proposed model [66].

Altogether, the above-presented encouraging findings suggest that AI models can potentially assist physicians in the FLL detection and diagnosis during conventional (CE)US examinations, although further external validation is needed for large-scale clinical application. Table 1 summarizes the principal AI-based models supporting US imaging in diagnosing HCC with the relatively main applications.

Table 1. Principal AI-based models supporting US imaging in diagnosing HCC: main applications.

AI-Based Model	Description and Applications	Reference
DCNN-US	Deep convolutional neural network (DCNN) applicated to US imaging in differentiating benign from malignant liver lesions with an accuracy comparable to expert radiologists, CT-scan, and only slight inferior to MRI.	Yang et al. [63]
YOLOv5	DL techniques applicated to US imaging for detection and classification of seven different types of FLLs, including HCC and regenerative nodules.	Chaiteerakij et al. [64].
DCCA-MKL	A two-stage multiple-view learning framework for CEUS combining deep canonical correlation analysis and multiple kernel learning (DCCA-MKL), only adopting three typical CEUS images selected from the three phases of the exam (arterial, portal venous and late) to discriminates between benign and malignant FLLs.	Guo LH et al. [66]

HCC: hepatocellular carcinoma; DL: deep learning; FLLs: focal liver lesions; US: ultrasound; CEUS: contrast enhancement ultrasound; CT: computer tomography; MRI: magnetic resonance imaging.

CT and MRI. CT is routinely used for an accurate diagnosis of HCC, primarily when the results from CEUS are doubtful [61]. In an interesting and highly rated research, to enhance the decision-making of clinicians when diagnosing HCC in cirrhotic patients with indeterminate liver nodules, Mokrane et al. realized a radiomics signature based on 13,920 CT imaging features extracted from data from 178 cirrhotic patients. They used ML techniques for the training, calibration, and validation and reported great accuracy in distinguishing HCC from non-HCC lesions for this signature [67].

More recently, Ho Yu et al. aimed to improve the diagnostic process of HCC diagnosis, particularly targeting the considerable proportions of indeterminate observations recurrently observed in routine clinical practice [68]. The authors developed four DL

models for diagnosing HCC on CT via a training-validation-testing approach, identifying, the best-performing model: the Spatio-Temporal 3D Convolution Network (ST3DCN). Relevantly, this model showed better AUCs compared to standard-of-care radiological interpretation [68].

There is a scarcity of research on MRI in HCC due to the technically difficult application of AI in this type of imaging and the costs of manually designing features. A preliminary study by Hamm et al. consisted of the creation of a convolutional neural network (CNN)-based deep learning system (DLS), which demonstrated a 92% accuracy, a 92% sensitivity, and a 98% specificity in classifying liver lesions [69]. Another DL algorithm was created using CNN by *Zhen* et al. to classify liver lesions with MR (enhanced and unenhanced) images and clinical data [70]. CNNs were trained using data from 1210 patients with liver tumours, and models were validated in an external cohort of 201 patients. They also compared the models with the performance of three experienced radiologists. Combining unenhanced images with clinical data greatly improved the performance of the model in classifying HCC with an AUC of 0.985, and the sensitivity and specificity of these models were comparable to that of the three experienced radiologists [70].

Table 2 summarizes the principal AI-based models supporting CT and MRI imaging in diagnosing HCC with the relatively main applications.

Table 2. Principal AI-based models supporting CT and MRI imaging in diagnosing HCC.

AI-Based Model	Description and Applications	Performance and Accuracy	Reference
DeltaV-A_DWT1_LL_Variance-2D	Application of ML techniques on CT imaging features of cirrhotic patients for the training, calibrating, and validating a signature distinguishing HCC from non-HCC lesions with elevated accuracy by quantifying changes between arterial and portal venous phases	AUC: 0.740 (95%CI: 0.610–0.801)	Mokrane et al. [67]
ST3DCN	DL techniques applicated to CT imaging for excluding HCC, particularly targeting indeterminate observations	AUC: 0.919 (95%CI: 0.903–0.935) NPV: 0.966 (95% CI: 0.954–0.979)	Ho Yu et al. [68]
DCCA-MKL	DL algorithm created using CNN on MRI (enhanced and unenhanced) images and clinical data able to classify HCC with accuracy comparable to that of the experienced radiologists	AUC: 0.985; (95%CI: 0.960–1.000)	Zhen et al. [70]

HCC: hepatocellular carcinoma; DL: deep learning CT: computer tomography; MRI: magnetic resonance imaging; AUC: area under the receiver operating characteristic curve; CI: Confidence interval; NPV: Negative Predictive Value.

3.2.2. AI Applications to Histopathology in Correctly Diagnosing HCC

When the imaging results are inconclusive, histopathology could be crucial in achieving a correct diagnosis of HCC [71,72]. Various efforts are being made to utilize AI to enhance this aspect in order to achieve better disease management, increase the amount of information extracted from a biopsy sample, and potentially remove the bias introduced by manual scoring. The scientific community has focused on relative research on the potential to match or surpass pathologists in specific tasks, even in types of cancer other than HCC [73–75].

Limiting the HCC AI applications to human tissue, Lin et al. proposed a fusion between a multiphoton microscopy (MPM) and a DL algorithm for the classification of

a HCC differentiation based on a pre-trained VGG-16 framework, reporting an elevated (>90%) accuracy [76].

Subsequently, Chen et al. used histopathological hematoxylin- and eosin-stained slides from the Genomic Data Commons databases to train and externally validate a CNN ("inception V3") for an automatic classification of this cancer. The authors reported a performance level close to the ability of a 5-year-experienced pathologist, with a 96.0% accuracy for the benign and malignant classifications, and an 89.6% accuracy for good, moderate, and poor tumor differentiation. Additionally, in this study, the capability to predict somatic mutations was evaluated. The model was further trained to identify the ten most common and prognostic mutated genes in HCC, evidencing four of them, including *CTNNB1*, *FMN2*, *TP53*, and *ZFX4*, as predictable from histopathology images, with external the AUCs from 0.71 to 0.89 [77].

Similarly, in further research, aiming to develop and externally validate a CNN-based classification to distinguish HCC from adjacent normal tissues, Liao et al. highlighted the great accuracy and the superiority of a DCNN model compared with the "inception V3" in adopting HCC histopathological slides in order to realise the automatic diagnosis of this cancer, using the features obtained from the TCGA database, simultaneously predicting somatic mutations, thus revealing DL-based histopathology as a promising tool to free pathologists from dull routine practice [78].

More recently, a brilliant study by Kiani et al. aimed to evaluate the effect of DL-based assistance on the diagnostic performance of 11 pathologists, with varying levels of expertise. The model was developed to distinguish the HCC and the cholangiocarcinoma on hematoxylin- and eosin-stained whole-slide images (WSI), achieving accuracies of 0.885 on a validation set of 26 WSI and 0.842 on an independent test set of 80 WSI, respectively.

The model significantly impacted the diagnostic decisions of all pathologists. Furthermore, while the correct predictions contributed to improving the accuracy of the pathologists (Odds Ratio-OR = 4.289), the incorrect predictions reduced it (OR = 0.253), regardless of the experience of the pathologist or the difficulty of the case [79]. Despite the obvious and proven positive sides, the study showcased the potential unintended negative consequences of the use of an AI-based model in the clinical setting.

3.2.3. Besides the Imaging: Exploring the AI Applications to Transcriptomics-Based Approaches

Molecular biology represents yet another promising modern AI-applications field [80,81]. Its contributions to the development of blood tests and the proposal of biomarkers, that are able to identify (even as a screening tool) HCC, would revolutionize the management of this cancer, even in terms of socioeconomic healthcare strategies.

Moreover, Yu et al. identified a panel of fusion genes (created in DNA rearrangements involving the hepatocyte during neoplastic transformation) present in several types of human cancers, including HCC.

Among the eight fusion genes, MAN2A1-FER, TRMT11-GRIK2, and CCNH-C5orf30 were shown as the most frequent in HCC samples, and the relative fusion transcripts in the serum samples of liver cancer patients were identified as circulating, cell-free RNA. Furthermore, the distributions of these gene fusion RNA fragments largely matched those of the primary HCC samples [82].

In more recent research from the University of Pittsburgh, the authors identified the fusion transcripts present in the blood of 61 patients with HCC and 75 patients with liver diseases other than cancer [83]. A computer analysed the data and classified them using ML, generating the hypothesis that four of the fusion transcripts (MAN2A1-FER $\leq$ 40, CCNH-C5orf30 $\leq$ 38, SLC45A2-AMACR $\leq$ 41, and PTEN-NOLC1 $\leq$ 40) were associated with a high probability of cancer. The researchers then estimated that through the application

of this system, it would be possible to correctly predict the presence of the disease in about 83% of cases. A second system, based on the expression of two fusion transcripts (MAN2A1-FER $\leq$ 40, CCNH-C5orf30 $\leq$ 38) and serum AFP, showed a predictive capacity approaching 95% [83].

According to the authors, from a bench to bedside point of view, the developed tests may have multiple applications and translational repercussions.

First, these tools could be used as screening tests for the early diagnosis of HCC in patients who have risk factors for this type of tumour. Moreover, these tests could be used to determine whether the patients who discover liver nodules of an unknown nature should undergo a biopsy

Finally, they could be used to evaluate the effectiveness of treatment and the presence of residual disease, taking into consideration the fact that decreased (or zero) levels of fusion transcripts in patients treated for HCC have been observed. Further studies will be needed to validate these hypotheses and the actual reliability of the tests. Still, the simplicity of these methods makes them very interesting and authorises further investigation, even in terms of the AI applications in the prediction of HCC patients' outcomes.

3.3. AI Supports the Development of HCC Predictive Models Influencing Therapeutic Choices

The staging of HCC is essential for the guiding prognosis, treatment decisions, and research, which ultimately contribute to the enhancement of current clinical and epidemiological health practices [27,84].

Several staging systems for HCC have been developed across different regions [10,45,84]. The most widely used prognostic model in Western Europe is the Barcelona Clinic Liver Cancer (BCLC) staging system [26].

This model includes five stages (0, A, B, C, D), based on the extent of the primary lesion, performance status, vascular invasion, and extrahepatic spread [26].

According to this system, the prognosis is primarily influenced by the disease stage at the time of diagnosis and the liver dysfunction severity, impacted either by the tumor itself or by an underlying cirrhosis. For patients diagnosed at the early stages, surgical resection and LT are potentially curative options, with a reported 5-year survival rate of 60–80% [26].

Patients in stage A are considered suitable candidates for a radical surgical resection, while those with stage B or C disease are typically not eligible for resection, but may be candidates for liver transplantation, locoregional therapies such as transarterial chemoembolization (TACE), or systemic treatments [26]. For patients with stage D disease, management should be limited to palliative and supportive care [26].

The importance of implementing a personalized approach for patients with HCC at these stages, aimed at achieving long-term survival outcomes, has driven scientific research toward focusing on the development and training of AI models capable of predicting the likelihood of a treatment response at various stages of the disease.

These AI-driven tools hold significant potential for guiding therapeutic decisions, especially in advanced stages, and improving overall prognosis. Despite the recent, considerable progress [1,2], the HCC continues to represent a major therapeutic challenge, with a dramatically worsening prognosis in stages BCLC-B and C, where the locoregional and systemic treatments, respectively, constitute the predominant choices, impacted by a wide variability in terms of success rate [26,85,86]. Therefore, the potential AI support in predicting the response appears crucial to avoid the related adverse effects, as well as to select the patients who would effectively benefit from that approach.

In a multicenter clinical study, Peng et al. aimed to develop a DL-model for the preoperative prediction of the treatment response in BCLC-B patients receiving TACE [87]. The authors used a total of 789 CT images from three different hospitals and constructed

a predictive model leveraging transfer learning techniques with a residual convolutional neural network (ResNet50).

The model's performance was validated in two independent cohorts. The results demonstrated a high accuracy in predicting the response to TACE therapy, with areas under the curve (AUCs) of 0.97, 0.96, 0.95, and 0.96 for the complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD), respectively.

More recently, in a multicenter retrospective study, Zeng Q et al. developed an AI model able to estimate the atezolizumab-bevacizumab response signature (ABRS) expression directly from the histological slides and to evaluate whether the model predictions were associated with progression-free survival (PFS) [88]. This study indicates that an AI model applied to HCC digital slides can serve as a biomarker for PFS in patients treated with atezolizumab–bevacizumab, proposing a novel, potentially useful approach to the development of inexpensive and fast biomarkers for targeted therapies [88]. Furthermore, the previously mentioned diagnostic progress related to molecular biology and the combination of AI heatmaps with spatial transcriptomics provides insight into the molecular features associated with predictions, providing an understanding of the biological mechanisms that drive the responses to treatments.

In addition to the disease stage at the time of diagnosis and the related generally poor prognosis in advanced scenarios, the substantial risk of disease recurrence after treatment represents the other relevant variable, severely impacting the prognosis and influencing the HCC mortality rate. In this regard, the possibility of recurrence significantly influences the subsequent treatment decisions and ultimately affects the long-term survival outcomes of patients [89].

A recurrence of HCC is typically categorized as either "early" or "late", based on the time of diagnosis, with a standard cutoff point of two years [46,61]. A HCC recurrence within 2 years of treatment is defined as "early", and is generally caused by the occult intrahepatic spread of the primary neoplasm and related to the tumor burden [46,61]. A recurrence occurring after 2 years of treatment is defined as "late", and is related to de novo HCC, independent of the primary neoplasm. The early HCC recurrence has a significantly poorer prognosis and outcome than the late recurrence [46,61]. In this context, the application of AI to predict the risk of a recurrence in patients treated for HCC could play a crucial role in patient stratification. This would facilitate the implementation of targeted surveillance strategies and guide the decisions regarding the initiation of neoadjuvant or adjuvant therapies following treatment.

Despite the current clinical guidelines recommending surgical resection as the first-line treatment for patients with solitary HCC, the rate of postoperative recurrence remains notably high [26,90]. For this purpose, Ji GW et al. explored the potential of radiomics coupled with ML algorithms in the improvement of the predictive accuracy for the HCC recurrence. In this retrospective study, a total of 470 patients who underwent contrast-enhanced CT and a curative resection for solitary HCC were recruited from 3 independent departments [91]. Using the ML framework, the authors identified a three-feature radiomics signature showing a superior prognostic performance for the prediction of a HCC recurrence, with a C-index of 0.733–0.801, compared to the rival models and widely used staging systems [91]. In another study, Nam JY et al. presented a model for predicting a tumor recurrence after LT by adopting artificial intelligence ("MoRAL-AI"), including 563 patients who underwent LT for HCC at three large LT centers in Korea [92]. The algorithm was developed with a deep neural network (DNN). The largest weighted factors in MoRAL-AI were age, tumor size, age, and serum levels of the AFP and prothrombin time. The authors also proved the advantages of their model in the external validation cohort, with a significantly better discrimination function for predicting a HCC recurrence,

compared to other state-of-the-art predictive models, like the Milan criteria [92]. Similarly, Saillard et al. investigated the independent validation sets of two DL algorithms, based on digitized histological slides: the first DL-based algorithm ("SCHMOWDER") used an attention mechanism on the tumoral areas annotated by a pathologist, whereas the second ("CHOWDER") worked without human expertise [93]. Relevantly, the c-indices for the survival prediction of SCHMOWDER and CHOWDER reached 0.78 and 0.75, respectively. Both models outperformed a composite score, incorporating all of the baseline variables associated with survival, revealing the vascular spaces, the macro-trabecular architectural pattern, and the lack of immune infiltration as the most predictive values of poor survival in the tumor areas [93].

Similarly, Yamashita et al. developed and validated a DL-based system ("HCC-SurvNet") to provide risk scores for the disease recurrence after the primary resection, taken directly from hematoxylin- and eosin-stained digital slides of formalin-fixed, paraffin-embedded liver resections [94]. This model stratified the patients into low- and high-risk subgroups, both for OS and PFS (i.e., post-surgical recurrence), reaching concordance indices of 0.724 and 0.683 on the internal and external test cohorts, respectively, exceeding the performance of the standard Tumor-Node-Metastasis classification system [94,95].

More recently, Saito et al. aimed to predict the early recurrence of HCC after the resection, based on digital pathologic images of hematoxylin- and eosin-stained specimens and on ML applying a support vector machine (SVM) [96]. In this research, 158 HCC patients meeting the Milan criteria and receiving a surgical resection were included and further categorized into three groups: Group I, patients with HCC recurrence within 1 year after resection; Group II, patients with HCC recurrence between 1 and 2 years after resection; and Group III, patients with no HCC recurrence within 4 years after resection. Notably, the SVM-based prediction model distinguished the three groups with an elevated (89.9%) accuracy [96]. These findings revealed the relevance of models integrating digital pathology with ML, as a novel potential method for improving the postoperative follow-up, especially for patients at a high risk of early HCC recurrence.

Furthermore, a DL model predicting recurrence-free survival (RFS) on histological slides after LT and resection was developed on a total of 1,118 HCC patients from four independent cohorts, using a nucleus map set (n = 120) to train the U-net on capturing the nuclear architecture [97]. The training set and its nuclear architectural information extracted by the U-net were used to train the MobileNet V2-based classifier (MobileNetV2_HCC_class).

The authors reported the MobileNetV2_HCC_class was a strong predictor of RFS in the LT setting, maintaining a relatively higher discriminatory power than the other factors, even after resection, in the independent validation cohort. Interestingly, a pathological review revealed, among other things, the degree of cytological atypia and nuclear hyperchromasia as the most relevant predictor of histologic features in the areas presenting the highest recurrence [97]. Altogether, the above-presented findings suggest the potential utility of DL- and ML-based approaches on both the imaging data and the histological slides in refining the prognostic prediction of HCC patients and identifying the more adequate individual management after the diagnosis, as well as in the case of recurrence.

4. AI Applications in Routine Managing HCC: Are We Ready?

4.1. Main Limitations of AI Use in Routine Clinical Practice: Need for Robust Evidence

Despite the recent progress in AI applications for the routine clinical management of human cancer, as for other malignant diseases, even for HCC, a great number of challenges remain, ranging from standardizing and robustly evaluating AI algorithms in prospective studies by using large-scale datasets to establishing the consensus guidelines to ensure accurate and comprehensive reporting of data from the ML and DL research.

As previously shown, the performance of AI algorithms is severely influenced by the amount of data adopted for training. Unfortunately, most of the above-presented studies that were able to consider and enclose a large series of patients were retrospective and thus affected by the relative biases of non-perspective experimental design. In line with this, there is an urgent need to encourage the further deposition of large databases worldwide. In this sense, it is important not to limit this challenge to a restricted geographic area, as to date, the populations considered training the development and training of AI models have lacked sufficient ethnic and socioeconomic diversity. Finally, the heterogeneity of algorithms and software adopted in the currently available research evaluating the application of AI to the clinical management of HCC represents another relevant issue, suggesting that the external validation of all these experimental models is, at the same time, crucial and difficult.

4.2. Future Perspectives and Conclusions: How Can the Scientific Community Contribute?

This review of the recent literature highlights the need for further efforts to give AI a concrete supporting role in order for it to be implemented in the clinical practice of HCC management. Although many steps forward have been taken in the last decade, a correct and appropriate application of AI in the routine management of HCC, from diagnosis to treatment, still requires more work [98].

Furthermore, as previously stated, the status of AI applications in the various clinical practice phases appears extremely heterogeneous; while its use currently seems more promising in the areas of diagnosis and prediction of response to therapy, we are still far from its use in the risk prediction and the subsequent stratification of patients for "tailored" screenings.

Prospective studies that irrefutably demonstrate the improvement in the clinical management of patients are still needed, as well as the comparison of new AI models with those already present and used in clinical practice, without underestimating the potential benefits, even in terms of economic and research sources, of adopting transfer learning (TL) approaches [99]. TL represents a type of ML and is an emerging promising technique in which a model trained on one task is reused as the starting point for a new activity [99].

Furthermore, integrating AI-based algorithms into current clinical trials, to authorize the emerging research suggesting their validity as tools for accurately identifying early HCC and predicting therapeutic responses and recurrence, appears to be even more crucial [100,101]. Simultaneously, it is also essential to externally validate any model developed on a real-world population before the clinical implementation, to disavow the presence of any bias highlighted by the trial.

To achieve trials receiving these potential ameliorations, the sharing and deposition of large databases containing imaging, clinical variables, and biological samples, from various departments in different geographic areas, as well as the accessibility of all these data to the scientific community, together with the interpretation and subsequent model-realization performed by AI-dedicated personnel, appears to be crucial [102].

Furthermore, significant benefits can derive from the applications of collaborative-learning-based strategies, consisting of learning approaches involving groups of learners working together to solve a problem, complete a task, or create a product [103].

Moreover, despite the ongoing debate on AI "replacing" humans, the progress shows the increasing need for the pertinent use of these tools in a clinical setting, to insert input data coherently, and to interpret the output data correctly.

The eXplainable Artificial Intelligence (XAI) can significantly contribute to this challenge, representing an encouraging impulse for further progress and a novel frontier in this research field, particularly in reference to interpretations and applications of multiparamet-

ric tools [104–106]. XAI represents a rapidly emerging branch of ML, aiming to unravel the predictions of complex models, especially those required in healthcare when the diagnoses, recommendations, and treatment choices might rely on the decisions made by AI-based systems [107]. In pioneering research, Lacalamita et al. aimed to develop a supervised learning framework based on a hierarchical community detection and AI models to distinguish the patients and controls, using publicly available microarray data [104]. The authors identified 20 gene communities that discriminated between healthy and cancerous samples with an elevated accuracy, validating the performance even on an independent dataset. Subsequently, after selecting two communities based on relevant biological functions, an XAI-based approach assessed the contribution of single genes to the relative classification task [104].

These findings proposed an effective tool contributing to the recognition of gene communities, helping reveal crucial mechanisms sustaining the HCC onset, and identifying biomarkers potentially useful in a clinical practice, both in terms of prognosis prediction and early diagnosis.

Moreover, another research investigated an innovative XAI framework to identify the key genetic biomarkers for a HCC prognosis [105]. In this study, the authors adopted robust AI models, validated against extensive gene expression datasets, and, by subsequently applying an XAI approach, demonstrated, at the same time, the predictive accuracy and the clinical relevance of the evidenced biomarkers (TOP3B, SSBP3, and COX7A2L) through explainable metrics [105]. Altogether, these results suggest that the application of XAI in biomarker discovery represents relevant progress in HCC research, offering a more complete understanding of the disease pathogenesis and concrete support to the currently available traditional diagnostic strategies.

Furthermore, a more recent study aimed to assess the efficacy of combining the automated ML (AutoML) with XAI in the identification of metabolomic biomarkers that can differentiate between HCC and liver cirrhosis in HCV-infected patients [106]. Particularly, the ML-TPOT tool was adopted to optimize the preparation of features and data, and the XAI-TreeSHAP approach was used to interpret the model by assessing the individual contribution of each metabolite to the categorization process. The authors reported a superior performance of the TPOT tool in distinguishing between HCC and cirrhosis compared to the other evaluated AutoML approaches, as well as the relative capability in identifying key metabolites, including L-valine, glycine, and DL-isoleucine, subsequently confirmed by the TreeSHAP analysis. TreeSHAP provided a comprehensive explanation of the contribution of these metabolites to the model's predictions, thus increasing the interpretability of these findings [106].

To summarize, this evidence highlights that the combination of AI-based models with XAI has the potential to enhance the identification of biomarkers and to develop precise, easily interpretable, AI-driven solutions for diagnosing HCC.

In addition to the interpretation of AI-based-models-provided results, another relevant issue influencing the application of these tools in routine clinical practice is warranting the safety of the HCC patients and the relative data, constituting simultaneously a medical and an ethical question [108]. To objectivize the applications of trustworthy AI principles in research, the Assessment List for Trustworthy Artificial Intelligence (ALTAI) was recently proposed as a checklist (including seven items: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination, and fairness; environmental and societal well-being; accountability) to help ensure that users can benefit from AI without being exposed to unnecessary risks by indicating a set of concrete steps for self-assessment [109].

Given this context, it is crucial to implement and diffuse trustworthy AI worldwide, by representing an approach to AI development that prioritizes safety and transparency for the people who interact with it; while ensuring its complete integration with XAI and traditional AI emerges as paramount when conceiving, designing, and validating all novel models aiming to improve the current management strategies of HCC [108,109].

Therefore, rather than imagining a context in which the clinician takes a backseat, the conscientious activity of new, highly trained figures appears to be essential to take full advantage of all the novel AI-dependent tools that will be developed in the future, without forgetting the main and noble aim of improving the HCC patient care.

In conclusion, now is not the time to update the diagnostic and predictive models' armamentarium in HCC management. Although some important scientific battles have been won, the "war" against this human cancer is ongoing. In this scenario, the preconditions for a new era of precision medicine, where the "war efforts" of the scientific community, supported by the advantages offered by the application of AI to routine clinical practice, will lead to winning strategies, appear increasingly clear and give rise to hope for a significant improvement in the prognosis and quality of life of HCC-affected patients.

Author Contributions: M.R.: guarantor of the article, conceptualization, methodology, formal analysis, investigation, and writing the original draft; M.D.: conceptualization, methodology, formal analysis, investigation, and writing of the original draft; C.N., C.B., F.D.N., P.V. and P.I.: investigation, resources, data curation, and writing of the original draft; A.F.: conceptualization, data curation, supervision, and project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

Abbreviations

The following abbreviations are used in this manuscript:

HBLTs	hepatobiliary liver tumors
HCC	hepatocellular carcinoma
AF	advanced fibrosis
HBV	Hepatitis B virus
HCV	Hepatitis C virus
MASLD	Metabolic Dysfunction-Associated Steatotic Liver Disease
AI	artificial intelligence
ML	Machine learning
DL	Deep learning
MRI	magnetic resonance imaging
CAD	computer-aided detection
CADx	computer-aided diagnosis
TCGA	The Cancer Genome Atlas
DLCS	deep learning radio-clinical signature
LAGC	locally advanced gastric cancer
OS	overall survival
CP	Child-Pugh score
US	Ultrasound
EASL	European Association for the Study of the Liver
CT	computer tomography
NC-AMRI	non-contrast abbreviated magnetic resonance imaging

CEUS	contrast enhancement ultrasound
SVR	sustained virological response
RFS	random survival forest
AFP	alpha-fetoprotein
GGT	gamma-glutamyl transferase
BMI	body mass index
ALP	alkaline phosphatase
FFL	focal liver lesion
BCLC	Barcelona Clinic Liver Cancer
TACE	transarterial chemoembolization
TL	transfer learning
XAI	eXplainable Artificial Intelligence
AutoML	automated machine learning

References

1. Brown, Z.J.; Tsilimigras, D.I.; Ruff, S.M.; Mohseni, A.; Kamel, I.R.; Cloyd, J.M.; Pawlik, T.M. Management of Hepatocellular Carcinoma: A Review. *JAMA Surg.* **2023**, *158*, 410–420. [CrossRef] [PubMed]
2. Kulik, L.; El-Serag, H.B. Epidemiology and Management of Hepatocellular Carcinoma. *Gastroenterology* **2019**, *156*, 477–491.e1. [CrossRef] [PubMed]
3. Balogh, J.; Victor, D.; Asham, E.H.; Burroughs, S.G.; Boktour, M.; Saharia, A.; Li, X.; Ghobrial, R.M.; Monsour, H.P. Hepatocellular Carcinoma: A Review. *J. Hepatocell. Carcinoma* **2016**, *3*, 41–53. [CrossRef] [PubMed]
4. Toh, M.R.; Wong, E.Y.T.; Wong, S.H.; Ng, A.W.T.; Loo, L.-H.; Chow, P.K.-H.; Ngeow, J. Global Epidemiology and Genetics of Hepatocellular Carcinoma. *Gastroenterology* **2023**, *164*, 766–782. [CrossRef]
5. Dallio, M.; Masarone, M.; Romeo, M.; Tuccillo, C.; Morisco, F.; Persico, M.; Loguercio, C.; Federico, A. PNPLA3, TM6SF2, and MBOAT7 Influence on Nutraceutical Therapy Response for Non-Alcoholic Fatty Liver Disease: A Randomized Controlled Trial. *Front. Med.* **2021**, *8*, 734847. [CrossRef]
6. Romeo, M.; Dallio, M.; Scognamiglio, F.; Ventriglia, L.; Cipullo, M.; Coppola, A.; Tammaro, C.; Scafuro, G.; Iodice, P.; Federico, A. Role of Non-Coding RNAs in Hepatocellular Carcinoma Progression: From Classic to Novel Clinicopathogenetic Implications. *Cancers* **2023**, *15*, 5178. [CrossRef]
7. Dallio, M.; Ventriglia, L.; Romeo, M.; Scognamiglio, F.; Diano, N.; Moggio, M.; Cipullo, M.; Coppola, A.; Ziogas, A.; Netea, M.G.; et al. Environmental Bisphenol A Exposure Triggers Trained Immunity-Related Pathways in Monocytes. *Front. Immunol.* **2023**, *14*, 1270391. [CrossRef]
8. Giri, S.; Singh, A. Epidemiology of Hepatocellular Carcinoma in India—An Updated Review for 2024. *J. Clin. Exp. Hepatol.* **2024**, *14*, 101447. [CrossRef]
9. Enomoto, H.; Akuta, N.; Hikita, H.; Suda, G.; Inoue, J.; Tamaki, N.; Ito, K.; Akahane, T.; Kawaoka, T.; Morishita, A.; et al. Etiological Changes of Liver Cirrhosis and Hepatocellular Carcinoma-Complicated Liver Cirrhosis in Japan: Updated Nationwide Survey from 2018 to 2021. *Hepatol. Res.* **2024**, *54*, 763–772. [CrossRef]
10. Rashed, W.M.; Kandeil, M.A.M.; Mahmoud, M.O.; Ezzat, S. Hepatocellular Carcinoma (HCC) in Egypt: A Comprehensive Overview. *J. Egypt. Natl. Cancer Inst.* **2020**, *32*, 5. [CrossRef]
11. Llovet, J.M.; Kelley, R.K.; Villanueva, A.; Singal, A.G.; Pikarsky, E.; Roayaie, S.; Lencioni, R.; Koike, K.; Zucman-Rossi, J.; Finn, R.S. Hepatocellular Carcinoma. *Nat. Rev. Dis. Primers* **2021**, *7*, 6. [CrossRef] [PubMed]
12. Chan, W.-K.; Chuah, K.-H.; Rajaram, R.B.; Lim, L.-L.; Ratnasingam, J.; Vethakkan, S.R. Metabolic Dysfunction-Associated Steatotic Liver Disease (MASLD): A State-of-the-Art Review. *J. Obes. Metab. Syndr.* **2023**, *32*, 197–213. [CrossRef] [PubMed]
13. Dallio, M.; Romeo, M.; Gravina, A.G.; Masarone, M.; Larussa, T.; Abenavoli, L.; Persico, M.; Loguercio, C.; Federico, A. Nutrigenomics and Nutrigenetics in Metabolic-(Dysfunction) Associated Fatty Liver Disease: Novel Insights and Future Perspectives. *Nutrients* **2021**, *13*, 1679. [CrossRef] [PubMed]
14. Dallio, M.; Sangineto, M.; Romeo, M.; Cipullo, M.; Coppola, A.; Mammone, S.; Di Gioia, G.; Masarone, M.; Persico, M.; Serviddio, G.; et al. The Influence of Acute Lifestyle Changes on NAFLD Evolution in a Multicentre Cohort: A Matter of Body Composition. *Nutr. Diabetes* **2024**, *14*, 33. [CrossRef] [PubMed]
15. McGlynn, K.A.; Petrick, J.L.; El-Serag, H.B. Epidemiology of Hepatocellular Carcinoma. *Hepatology* **2021**, *73* (Suppl. S1), 4–13. [CrossRef]
16. Koshy, A. Evolving Global Etiology of Hepatocellular Carcinoma (HCC): Insights and Trends for 2024. *J. Clin. Exp. Hepatol.* **2025**, *15*, 102406. [CrossRef]

17. Rinella, M.E.; Lazarus, J.V.; Ratziu, V.; Francque, S.M.; Sanyal, A.J.; Kanwal, F.; Romero, D.; Abdelmalek, M.F.; Anstee, Q.M.; Arab, J.P.; et al. A Multisociety Delphi Consensus Statement on New Fatty Liver Disease Nomenclature. *Hepatology* **2023**, *78*, 1966–1986. [CrossRef]
18. Rinella, M.E.; Sookoian, S. From NAFLD to MASLD: Updated Naming and Diagnosis Criteria for Fatty Liver Disease. *J. Lipid Res.* **2024**, *65*, 100485. [CrossRef]
19. Kim, H.; Lee, D.S.; An, T.H.; Park, H.-J.; Kim, W.K.; Bae, K.-H.; Oh, K.-J. Metabolic Spectrum of Liver Failure in Type 2 Diabetes and Obesity: From NAFLD to NASH to HCC. *Int. J. Mol. Sci.* **2021**, *22*, 4495. [CrossRef]
20. Meroni, M.; Longo, M.; Dongiovanni, P. Cardiometabolic Risk Factors in MASLD Patients with HCC: The Other Side of the Coin. *Front. Endocrinol.* **2024**, *15*, 1411706. [CrossRef]
21. Palma, R.; Pronio, A.; Romeo, M.; Scognamiglio, F.; Ventriglia, L.; Ormando, V.M.; Lamazza, A.; Pontone, S.; Federico, A.; Dallio, M. The Role of Insulin Resistance in Fueling NAFLD Pathogenesis: From Molecular Mechanisms to Clinical Implications. *J. Clin. Med.* **2022**, *11*, 3649. [CrossRef]
22. Provera, A.; Vecchio, C.; Sheferaw, A.N.; Stoppa, I.; Pantham, D.; Dianzani, U.; Sutti, S. From MASLD to HCC: What's in the Middle? *Heliyon* **2024**, *10*, e35338. [CrossRef] [PubMed]
23. Phoolchund, A.G.S.; Khakoo, S.I. MASLD and the Development of HCC: Pathogenesis and Therapeutic Challenges. *Cancers* **2024**, *16*, 259. [CrossRef] [PubMed]
24. Dallio, M.; Sangineto, M.; Romeo, M.; Villani, R.; Romano, A.D.; Loguercio, C.; Serviddio, G.; Federico, A. Immunity as Cornerstone of Non-Alcoholic Fatty Liver Disease: The Contribution of Oxidative Stress in the Disease Progression. *Int. J. Mol. Sci.* **2021**, *22*, 436. [CrossRef]
25. Ioannou, G.N. Epidemiology and Risk-Stratification of NAFLD-Associated HCC. *J. Hepatol.* **2021**, *75*, 1476–1484. [CrossRef]
26. Reig, M.; Forner, A.; Rimola, J.; Ferrer-Fàbrega, J.; Burrel, M.; Garcia-Criado, Á.; Kelley, R.K.; Galle, P.R.; Mazzaferro, V.; Salem, R.; et al. BCLC Strategy for Prognosis Prediction and Treatment Recommendation: The 2022 Update. *J. Hepatol.* **2022**, *76*, 681–693. [CrossRef]
27. Sidali, S.; Trépo, E.; Sutter, O.; Nault, J.-C. New Concepts in the Treatment of Hepatocellular Carcinoma. *United Eur. Gastroenterol. J.* **2022**, *10*, 765–774. [CrossRef]
28. Hamet, P.; Tremblay, J. Artificial Intelligence in Medicine. *Metabolism* **2017**, *69S*, S36–S40. [CrossRef]
29. Mintz, Y.; Brodie, R. Introduction to Artificial Intelligence in Medicine. *Minim. Invasive Ther. Allied Technol.* **2019**, *28*, 73–81. [CrossRef]
30. Patel, S.K.; George, B.; Rai, V. Artificial Intelligence to Decode Cancer Mechanism: Beyond Patient Stratification for Precision Oncology. *Front. Pharmacol.* **2020**, *11*, 1177. [CrossRef]
31. Domingos, P. A Few Useful Things to Know About Machine Learning. *Commun. ACM* **2012**, *55*, 78–87. [CrossRef]
32. Bini, S.A. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? *J. Arthroplast.* **2018**, *33*, 2358–2361. [CrossRef] [PubMed]
33. 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]
34. Kharchenko, P.V.; Silberstein, L.; Scadden, D.T. Bayesian Approach to Single-Cell Differential Expression Analysis. *Nat. Methods* **2014**, *11*, 740–742. [CrossRef]
35. Reig, B.; Heacock, L.; Geras, K.J.; Moy, L. Machine Learning in Breast MRI. *J. Magn. Reson. Imaging* **2020**, *52*, 998–1018. [CrossRef]
36. Hickman, S.E.; Woitek, R.; Le, E.P.V.; Im, Y.R.; Mouritsen Luxhøj, C.; Aviles-Rivero, A.I.; Baxter, G.C.; MacKay, J.W.; Gilbert, F.J. Machine Learning for Workflow Applications in Screening Mammography: Systematic Review and Meta-Analysis. *Radiology* **2022**, *302*, 88–104. [CrossRef]
37. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
38. Tran, K.A.; Kondrashova, O.; Bradley, A.; Williams, E.D.; Pearson, J.V.; Waddell, N. Deep Learning in Cancer Diagnosis, Prognosis and Treatment Selection. *Genome Med.* **2021**, *13*, 152. [CrossRef]
39. Oh, D.-Y.; He, A.R.; Bouattour, M.; Okusaka, T.; Qin, S.; Chen, L.-T.; Kitano, M.; Lee, C.-K.; Kim, J.W.; Chen, M.-H.; et al. Durvalumab or Placebo plus Gemcitabine and Cisplatin in Participants with Advanced Biliary Tract Cancer (TOPAZ-1): Updated Overall Survival from a Randomised Phase 3 Study. *Lancet Gastroenterol. Hepatol.* **2024**, *9*, 694–704. [CrossRef]
40. Liu, S.; Yang, Z.; Li, G.; Li, C.; Luo, Y.; Gong, Q.; Wu, X.; Li, T.; Zhang, Z.; Xing, B.; et al. Multi-Omics Analysis of Primary Cell Culture Models Reveals Genetic and Epigenetic Basis of Intratumoral Phenotypic Diversity. *Genom. Proteom. Bioinform.* **2019**, *17*, 576–589. [CrossRef]
41. Zeng, W.Z.D.; Glicksberg, B.S.; Li, Y.; Chen, B. Selecting Precise Reference Normal Tissue Samples for Cancer Research Using a Deep Learning Approach. *BMC Med. Genom.* **2019**, *12*, 21. [CrossRef]
42. Grewal, J.K.; Tessier-Cloutier, B.; Jones, M.; Gakkhar, S.; Ma, Y.; Moore, R.; Mungall, A.J.; Zhao, Y.; Taylor, M.D.; Gelmon, K.; et al. Application of a Neural Network Whole Transcriptome-Based Pan-Cancer Method for Diagnosis of Primary and Metastatic Cancers. *JAMA Netw. Open* **2019**, *2*, e192597. [CrossRef] [PubMed]

43. Zhao, Y.; Pan, Z.; Namburi, S.; Pattison, A.; Posner, A.; Balachander, S.; Paisie, C.A.; Reddi, H.V.; Rueter, J.; Gill, A.J.; et al. CUP-AI-Dx: A Tool for Inferring Cancer Tissue of Origin and Molecular Subtype Using RNA Gene-Expression Data and Artificial Intelligence. *EBioMedicine* **2020**, *61*, 103030. [CrossRef]
44. Hu, C.; Chen, W.; Li, F.; Zhang, Y.; Yu, P.; Yang, L.; Huang, L.; Sun, J.; Chen, S.; Shi, C.; et al. Deep Learning Radio-Clinical Signatures for Predicting Neoadjuvant Chemotherapy Response and Prognosis from Pretreatment CT Images of Locally Advanced Gastric Cancer Patients. *Int. J. Surg.* **2023**, *109*, 1980–1992. [CrossRef]
45. Ayuso, C.; Rimola, J.; Vilana, R.; Burrel, M.; Darnell, A.; García-Criado, Á.; Bianchi, L.; Belmonte, E.; Caparroz, C.; Barrufet, M.; et al. Diagnosis and Staging of Hepatocellular Carcinoma (HCC): Current Guidelines. *Eur. J. Radiol.* **2018**, *101*, 72–81. [CrossRef]
46. European Association for the Study of the Liver. EASL Clinical Practice Guidelines: Management of Hepatocellular Carcinoma. *J. Hepatol.* **2018**, *69*, 182–236. [CrossRef]
47. Wolf, E.; Rich, N.E.; Marrero, J.A.; Parikh, N.D.; Singal, A.G. Use of Hepatocellular Carcinoma Surveillance in Patients with Cirrhosis: A Systematic Review and Meta-Analysis. *Hepatology* **2021**, *73*, 713–725. [CrossRef]
48. Singal, A.G.; Sanduzzi-Zamparelli, M.; Nahon, P.; Ronot, M.; Hoshida, Y.; Rich, N.; Reig, M.; Vilgrain, V.; Marrero, J.; Llovet, J.M.; et al. International Liver Cancer Association (ILCA) White Paper on Hepatocellular Carcinoma Risk Stratification and Surveillance. *J. Hepatol.* **2023**, *79*, 226–239. [CrossRef]
49. Yang, Y.; Zhou, Y.; Zhang, X.; Xin, Y.; Chen, Y.; Fan, Q.; Li, X.; Wei, X.; Li, Q.; Zhou, X.; et al. Using the aMAP Risk Score to Predict Late Recurrence Following Radiofrequency Ablation for Hepatocellular Carcinoma in Chinese Population: A Multicenter Study. *J. Hepatocell Carcinoma* **2021**, *8*, 837–850. [CrossRef]
50. Xie, D.; Shi, J.; Zhou, J.; Fan, J.; Gao, Q. Clinical Practice Guidelines and Real-Life Practice in Hepatocellular Carcinoma: A Chinese Perspective. *Clin. Mol. Hepatol.* **2023**, *29*, 206–216. [CrossRef]
51. Fan, R.; Papatheodoridis, G.; Sun, J.; Innes, H.; Toyoda, H.; Xie, Q.; Mo, S.; Sypsa, V.; Guha, I.N.; Kumada, T.; et al. aMAP Risk Score Predicts Hepatocellular Carcinoma Development in Patients with Chronic Hepatitis. *J. Hepatol.* **2020**, *73*, 1368–1378. [CrossRef] [PubMed]
52. Sharma, S.A.; Kowgier, M.; Hansen, B.E.; Brouwer, W.P.; Maan, R.; Wong, D.; Shah, H.; Khalili, K.; Yim, C.; Heathcote, E.J.; et al. Toronto HCC Risk Index: A Validated Scoring System to Predict 10-Year Risk of HCC in Patients with Cirrhosis. *J. Hepatol.* **2017**, *68*, 92–99. [CrossRef] [PubMed]
53. Sherman, M. HCC Risk Scores: Useful or Not? *Semin. Liver Dis.* **2017**, *37*, 287–295. [CrossRef]
54. Fujiwara, N.; Friedman, S.L.; Goossens, N.; Hoshida, Y. Risk Factors and Prevention of Hepatocellular Carcinoma in the Era of Precision Medicine. *J. Hepatol.* **2018**, *68*, 526–549. [CrossRef]
55. Sarkar, S.; Alurwar, A.; Ly, C.; Piao, C.; Donde, R.; Wang, C.J.; Meyers, F.J. A Machine Learning Model to Predict Risk for Hepatocellular Carcinoma in Patients with Metabolic Dysfunction-Associated Steatotic Liver Disease. *Gastro Hep Adv.* **2024**, *3*, 498–505. [CrossRef]
56. Ioannou, G.N.; Tang, W.; Beste, L.A.; Tincopa, M.A.; Su, G.L.; Van, T.; Tapper, E.B.; Singal, A.G.; Zhu, J.; Waljee, A.K. Assessment of a Deep Learning Model to Predict Hepatocellular Carcinoma in Patients with Hepatitis C Cirrhosis. *JAMA Netw. Open* **2020**, *3*, e2015626. [CrossRef]
57. Shiha, G.; Waked, I.; Soliman, R.; Elbasiony, M.; Gomaa, A.; Mikhail, N.N.H.; Eslam, M. GES: A Validated Simple Score to Predict the Risk of HCC in Patients with HCV-GT4-Associated Advanced Liver Fibrosis after Oral Antivirals. *Liver Int.* **2020**, *40*, 2828–2833. [CrossRef]
58. Shiha, G.; Hassan, A.; Mousa, N.; El-Domiaty, N.; Mikhail, N.; Gameaa, R.; Kobtan, A.; El Bassat, H.; Sharaf-Eldin, M.; Waked, I.; et al. Individualized HCC Surveillance Using Risk Stratification Scores in Advanced Fibrosis and Cirrhotic HCV Patients Who Achieved SVR: Prospective Study. *Aliment. Pharmacol. Ther.* **2025**, *61*, 99–108. [CrossRef]
59. Minami, T.; Sato, M.; Toyoda, H.; Yasuda, S.; Yamada, T.; Nakatsuka, T.; Enooku, K.; Nakagawa, H.; Fujinaga, H.; Izumiya, M.; et al. Machine Learning for Individualized Prediction of Hepatocellular Carcinoma Development after the Eradication of Hepatitis C Virus with Antivirals. *J. Hepatol.* **2023**, *79*, 1006–1014. [CrossRef]
60. Kim, H.Y.; Lampertico, P.; Nam, J.Y.; Lee, H.-C.; Kim, S.U.; Sinn, D.H.; Seo, Y.S.; Lee, H.A.; Park, S.Y.; Lim, Y.-S.; et al. An Artificial Intelligence Model to Predict Hepatocellular Carcinoma Risk in Korean and Caucasian Patients with Chronic Hepatitis B. *J. Hepatol.* **2022**, *76*, 311–318. [CrossRef]
61. Sangro, B.; Argemi, J.; Ronot, M.; Paradis, V.; Meyer, T.; Mazzaferro, V.; Jepsen, P.; Golfieri, R.; Galle, P.; Dawson, L.; et al. EASL Clinical Practice Guidelines on the Management of Hepatocellular Carcinoma. *J. Hepatol.* **2024**, *82*, 315–374. [CrossRef] [PubMed]
62. Hu, W.; Yang, H.; Xu, H.; Mao, Y. Radiomics Based on Artificial Intelligence in Liver Diseases: Where We Are? *Gastroenterol. Rep.* **2020**, *8*, 90–97. [CrossRef] [PubMed]
63. Yang, Q.; Wei, J.; Hao, X.; Kong, D.; Yu, X.; Jiang, T.; Xi, J.; Cai, W.; Luo, Y.; Jing, X.; et al. Improving B-Mode Ultrasound Diagnostic Performance for Focal Liver Lesions Using Deep Learning: A Multicentre Study. *EBioMedicine* **2020**, *56*, 102777. [CrossRef] [PubMed]

64. Chaiteerakij, R.; Ariyaskul, D.; Kulkraisri, K.; Apiparakoon, T.; Sukcharoen, S.; Chaichuen, O.; Pensuwan, P.; Tiyarattanachai, T.; Rerknimitr, R.; Marukatat, S. Artificial Intelligence for Ultrasonographic Detection and Diagnosis of Hepatocellular Carcinoma and Cholangiocarcinoma. *Sci. Rep.* **2024**, *14*, 20617. [CrossRef]
65. Kudo, M.; Ueshima, K.; Osaki, Y.; Hirooka, M.; Imai, Y.; Aso, K.; Numata, K.; Kitano, M.; Kumada, T.; Izumi, N.; et al. B-Mode Ultrasonography versus Contrast-Enhanced Ultrasonography for Surveillance of Hepatocellular Carcinoma: A Prospective Multicenter Randomized Controlled Trial. *Liver Cancer* **2019**, *8*, 271–280. [CrossRef]
66. Guo, L.-H.; Wang, D.; Qian, Y.-Y.; Zheng, X.; Zhao, C.-K.; Li, X.-L.; Bo, X.-W.; Yue, W.-W.; Zhang, Q.; Shi, J.; et al. A Two-Stage Multi-View Learning Framework Based Computer-Aided Diagnosis of Liver Tumors with Contrast Enhanced Ultrasound Images. *Clin. Hemorheol. Microcirc.* **2018**, *69*, 343–354. [CrossRef]
67. Mokrane, F.-Z.; Lu, L.; Vavasseur, A.; Otal, P.; Peron, J.-M.; Luk, L.; Yang, H.; Ammari, S.; Saenger, Y.; Rousseau, H.; et al. Radiomics Machine-Learning Signature for Diagnosis of Hepatocellular Carcinoma in Cirrhotic Patients with Indeterminate Liver Nodules. *Eur. Radiol.* **2020**, *30*, 558–570. [CrossRef]
68. Yu, P.L.H.; Chiu, K.W.-H.; Lu, J.; Lui, G.C.S.; Zhou, J.; Cheng, H.-M.; Mao, X.; Wu, J.; Shen, X.-P.; Kwok, K.M.; et al. Application of a Deep Learning Algorithm for the Diagnosis of HCC. *JHEP Rep.* **2025**, *7*, 101219. [CrossRef]
69. Hamm, C.A.; Wang, C.J.; Savic, L.J.; Ferrante, M.; Schobert, I.; Schlachter, T.; Lin, M.; Duncan, J.S.; Weinreb, J.C.; Chapiro, J.; et al. Deep Learning for Liver Tumor Diagnosis Part I: Development of a Convolutional Neural Network Classifier for Multi-Phasic MRI. *Eur. Radiol.* **2019**, *29*, 3338–3347. [CrossRef]
70. Zhen, S.-H.; Cheng, M.; Tao, Y.-B.; Wang, Y.-F.; Juengpanich, S.; Jiang, Z.-Y.; Jiang, Y.-K.; Yan, Y.-Y.; Lu, W.; Lue, J.-M.; et al. Deep Learning for Accurate Diagnosis of Liver Tumor Based on Magnetic Resonance Imaging and Clinical Data. *Front. Oncol.* **2020**, *10*, 680. [CrossRef]
71. Fu, Y.; Jung, A.W.; Torne, R.V.; Gonzalez, S.; Vöhringer, H.; Shmatko, A.; Yates, L.R.; Jimenez-Linan, M.; Moore, L.; Gerstung, M. Pan-Cancer Computational Histopathology Reveals Mutations, Tumor Composition and Prognosis. *Nat. Cancer* **2020**, *1*, 800–810. [CrossRef] [PubMed]
72. Kather, J.N.; Pearson, A.T.; Halama, N.; Jäger, D.; Krause, J.; Loosen, S.H.; Marx, A.; Boor, P.; Tacke, F.; Neumann, U.P.; et al. Deep Learning Can Predict Microsatellite Instability Directly from Histology in Gastrointestinal Cancer. *Nat. Med.* **2019**, *25*, 1054–1056. [CrossRef] [PubMed]
73. Campanella, G.; Hanna, M.G.; Geneslaw, L.; Miraflor, A.; Werneck Krauss Silva, V.; Busam, K.J.; Brogi, E.; Reuter, V.E.; Klimstra, D.S.; Fuchs, T.J. Clinical-Grade Computational Pathology Using Weakly Supervised Deep Learning on Whole Slide Images. *Nat. Med.* **2019**, *25*, 1301–1309. [CrossRef]
74. Ehteshami Bejnordi, B.; Veta, M.; Johannes van Diest, P.; van Ginneken, B.; Karssemeijer, N.; Litjens, G.; van der Laak, J.A.W.M.; The CAMELYON16 Consortium; Hermsen, M.; Manson, Q.F.; et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women with Breast Cancer. *JAMA* **2017**, *318*, 2199–2210. [CrossRef]
75. Lu, M.Y.; Chen, T.Y.; Williamson, D.F.K.; Zhao, M.; Shady, M.; Lipkova, J.; Mahmood, F. AI-Based Pathology Predicts Origins for Cancers of Unknown Primary. *Nature* **2021**, *594*, 106–110. [CrossRef]
76. Lin, H.; Wei, C.; Wang, G.; Chen, H.; Lin, L.; Ni, M.; Chen, J.; Zhuo, S. Automated Classification of Hepatocellular Carcinoma Differentiation Using Multiphoton Microscopy and Deep Learning. *J. Biophotonics* **2019**, *12*, e201800435. [CrossRef]
77. Chen, M.; Zhang, B.; Topatana, W.; Cao, J.; Zhu, H.; Juengpanich, S.; Mao, Q.; Yu, H.; Cai, X. Classification and Mutation Prediction Based on Histopathology H&E Images in Liver Cancer Using Deep Learning. *NPJ Precis. Oncol.* **2020**, *4*, 14. [CrossRef]
78. Liao, H.; Long, Y.; Han, R.; Wang, W.; Xu, L.; Liao, M.; Zhang, Z.; Wu, Z.; Shang, X.; Li, X.; et al. Deep Learning-Based Classification and Mutation Prediction from Histopathological Images of Hepatocellular Carcinoma. *Clin. Transl. Med.* **2020**, *10*, e102. [CrossRef]
79. Kiani, A.; Uyumazturk, B.; Rajpurkar, P.; Wang, A.; Gao, R.; Jones, E.; Yu, Y.; Langlotz, C.P.; Ball, R.L.; Montine, T.J.; et al. Impact of a Deep Learning Assistant on the Histopathologic Classification of Liver Cancer. *NPJ Digit. Med.* **2020**, *3*, 23. [CrossRef]
80. Dey, S.S.; Kester, L.; Spanjaard, B.; Bienko, M.; van Oudenaarden, A. Integrated Genome and Transcriptome Sequencing of the Same Cell. *Nat. Biotechnol.* **2015**, *33*, 285–289. [CrossRef]
81. Macaulay, I.C.; Haerty, W.; Kumar, P.; Li, Y.I.; Hu, T.X.; Teng, M.J.; Goolam, M.; Saurat, N.; Coupland, P.; Shirley, L.M.; et al. G&T-Seq: Parallel Sequencing of Single-Cell Genomes and Transcriptomes. *Nat. Methods* **2015**, *12*, 519–522. [CrossRef] [PubMed]
82. Yu, Y.-P.; Tsung, A.; Liu, S.; Nalesnick, M.; Geller, D.; Michalopoulos, G.; Luo, J.-H. Detection of Fusion Transcripts in the Serum Samples of Patients with Hepatocellular Carcinoma. *Oncotarget* **2019**, *10*, 3352–3360. [CrossRef] [PubMed]
83. Yu, Y.-P.; Liu, S.; Geller, D.; Luo, J.-H. Serum Fusion Transcripts to Assess the Risk of Hepatocellular Carcinoma and the Impact of Cancer Treatment through Machine Learning. *Am. J. Pathol.* **2024**, *194*, 1262–1271. [CrossRef] [PubMed]
84. Tellapuri, S.; Sutphin, P.D.; Beg, M.S.; Singal, A.G.; Kalva, S.P. Staging Systems of Hepatocellular Carcinoma: A Review. *Indian J. Gastroenterol.* **2018**, *37*, 481–491. [CrossRef] [PubMed]

85. Haber, P.K.; Castet, F.; Torres-Martin, M.; Andreu-Oller, C.; Puigvehí, M.; Miho, M.; Radu, P.; Dufour, J.-F.; Verslype, C.; Zimpel, C.; et al. Molecular Markers of Response to Anti-PD1 Therapy in Advanced Hepatocellular Carcinoma. *Gastroenterology* **2023**, *164*, 72–88.e18. [CrossRef]
86. Sangro, B.; Melero, I.; Wadhawan, S.; Finn, R.S.; Abou-Alfa, G.K.; Cheng, A.-L.; Yau, T.; Furuse, J.; Park, J.-W.; Boyd, Z.; et al. Association of Inflammatory Biomarkers with Clinical Outcomes in Nivolumab-Treated Patients with Advanced Hepatocellular Carcinoma. *J. Hepatol.* **2020**, *73*, 1460–1469. [CrossRef]
87. Peng, J.; Kang, S.; Ning, Z.; Deng, H.; Shen, J.; Xu, Y.; Zhang, J.; Zhao, W.; Li, X.; Gong, W.; et al. Residual Convolutional Neural Network for Predicting Response of Transarterial Chemoembolization in Hepatocellular Carcinoma from CT Imaging. *Eur. Radiol.* **2020**, *30*, 413–424. [CrossRef]
88. Zeng, Q.; Klein, C.; Caruso, S.; Maille, P.; Allende, D.S.; Mínguez, B.; Iavarone, M.; Ningarhari, M.; Casadei-Gardini, A.; Pedica, F.; et al. Artificial Intelligence-Based Pathology as a Biomarker of Sensitivity to Atezolizumab-Bevacizumab in Patients with Hepatocellular Carcinoma: A Multicentre Retrospective Study. *Lancet Oncol.* **2023**, *24*, 1411–1422. [CrossRef]
89. Nevola, R.; Ruocco, R.; Criscuolo, L.; Villani, A.; Alfano, M.; Beccia, D.; Imbriani, S.; Claar, E.; Cozzolino, D.; Sasso, F.C.; et al. Predictors of Early and Late Hepatocellular Carcinoma Recurrence. *World J. Gastroenterol.* **2023**, *29*, 1243–1260. [CrossRef]
90. Akahoshi, K.; Shindoh, J.; Tanabe, M.; Watanabe, S.; Takamizawa, H.; Eguchi, S.; Endo, I.; Kubo, S.; Taketomi, A.; Nagano, H.; et al. Questionnaire Survey of Japanese Board-Certified Expert Hepatobiliary and Pancreatic Surgeons and Instructors on the Surgical Indications for Hepatocellular Carcinoma. *J. Hepato-Biliary-Pancreat. Sci.* **2024**, *31*, 143–151. [CrossRef]
91. Ji, G.-W.; Zhu, F.-P.; Xu, Q.; Wang, K.; Wu, M.-Y.; Tang, W.-W.; Li, X.-C.; Wang, X.-H. Machine-Learning Analysis of Contrast-Enhanced CT Radiomics Predicts Recurrence of Hepatocellular Carcinoma after Resection: A Multi-Institutional Study. *EBioMedicine* **2019**, *50*, 156–165. [CrossRef] [PubMed]
92. Nam, J.Y.; Lee, J.-H.; Bae, J.; Chang, Y.; Cho, Y.; Sinn, D.H.; Kim, B.H.; Kim, S.H.; Yi, N.-J.; Lee, K.-W.; et al. Novel Model to Predict HCC Recurrence after Liver Transplantation Obtained Using Deep Learning: A Multicenter Study. *Cancers* **2020**, *12*, 2791. [CrossRef] [PubMed]
93. Saillard, C.; Schmauch, B.; Laifa, O.; Moarii, M.; Toldo, S.; Zaslavskiy, M.; Pronier, E.; Laurent, A.; Amaddeo, G.; Regnault, H.; et al. Predicting Survival After Hepatocellular Carcinoma Resection Using Deep Learning on Histological Slides. *Hepatology* **2020**, *72*, 2000–2013. [CrossRef]
94. Yamashita, R.; Long, J.; Saleem, A.; Rubin, D.L.; Shen, J. Deep Learning Predicts Postsurgical Recurrence of Hepatocellular Carcinoma from Digital Histopathologic Images. *Sci. Rep.* **2021**, *11*, 2047. [CrossRef]
95. Faria, S.C.; Szklaruk, J.; Kaseb, A.O.; Hassabo, H.M.; Elsayes, K.M. TNM/Okuda/Barcelona/UNOS/CLIP International Multidisciplinary Classification of Hepatocellular Carcinoma: Concepts, Perspectives, and Radiologic Implications. *Abdom. Imaging* **2014**, *39*, 1070–1087. [CrossRef]
96. Saito, A.; Toyoda, H.; Kobayashi, M.; Koiwa, Y.; Fujii, H.; Fujita, K.; Maeda, A.; Kaneoka, Y.; Hazama, S.; Nagano, H.; et al. Prediction of Early Recurrence of Hepatocellular Carcinoma after Resection Using Digital Pathology Images Assessed by Machine Learning. *Mod. Pathol.* **2021**, *34*, 417–425. [CrossRef]
97. Liu, Z.; Liu, Y.; Zhang, W.; Hong, Y.; Meng, J.; Wang, J.; Zheng, S.; Xu, X. Deep Learning for Prediction of Hepatocellular Carcinoma Recurrence after Resection or Liver Transplantation: A Discovery and Validation Study. *Hepatol. Int.* **2022**, *16*, 577–589. [CrossRef]
98. Calderaro, J.; Seraphin, T.P.; Luedde, T.; Simon, T.G. Artificial Intelligence for the Prevention and Clinical Management of Hepatocellular Carcinoma. *J. Hepatol.* **2022**, *76*, 1348–1361. [CrossRef]
99. Yu, Z.; Shen, D.; Jin, Z.; Huang, J.; Cai, D.; Hua, X.-S. Progressive Transfer Learning. *IEEE Trans. Image Process.* **2022**, *31*, 1340–1348. [CrossRef]
100. Cruz Rivera, S.; Liu, X.; Chan, A.-W.; Denniston, A.K.; Calvert, M.J.; SPIRIT-AI and CONSORT-AI Working Group; SPIRIT-AI and CONSORT-AI Steering Group. SPIRIT-AI and CONSORT-AI Consensus Group Guidelines for Clinical Trial Protocols for Interventions Involving Artificial Intelligence: The SPIRIT-AI Extension. *Nat. Med.* **2020**, *26*, 1351–1363. [CrossRef]
101. Liu, X.; Cruz Rivera, S.; Moher, D.; Calvert, M.J.; Denniston, A.K. SPIRIT-AI and CONSORT-AI Working Group Reporting Guidelines for Clinical Trial Reports for Interventions Involving Artificial Intelligence: The CONSORT-AI Extension. *Nat. Med.* **2020**, *26*, 1364–1374. [CrossRef] [PubMed]
102. Bauchner, H.; Golub, R.M.; Fontanarosa, P.B. Data Sharing: An Ethical and Scientific Imperative. *JAMA* **2016**, *315*, 1237–1239. [CrossRef] [PubMed]
103. Laal, M.; Ghodsi, S.M. Benefits of Collaborative Learning. *Procedia-Soc. Behav. Sci.* **2012**, *31*, 486–490. [CrossRef]
104. Lacalamita, A.; Serino, G.; Pantaleo, E.; Monaco, A.; Amoroso, N.; Bellantuono, L.; Piccinno, E.; Scalavino, V.; Dituri, F.; Tangaro, S.; et al. Artificial Intelligence and Complex Network Approaches Reveal Potential Gene Biomarkers for Hepatocellular Carcinoma. *Int. J. Mol. Sci.* **2023**, *24*, 15286. [CrossRef] [PubMed]
105. Gutierrez-Chakraborty, E.; Chakraborty, D.; Das, D.; Bai, Y. Discovering Novel Prognostic Biomarkers of Hepatocellular Carcinoma Using eXplainable Artificial Intelligence. *bioRxiv* **2023**. bioRxiv:2023.11.27.568859. [CrossRef]

106. Yagin, F.H.; El Shawi, R.; Algarni, A.; Colak, C.; Al-Hashem, F.; Ardigò, L.P. Metabolomics Biomarker Discovery to Optimize Hepatocellular Carcinoma Diagnosis: Methodology Integrating AutoML and Explainable Artificial Intelligence. *Diagnostics* **2024**, *14*, 2049. [CrossRef]
107. Kalyakulina, A.; Yusipov, I.; Moskalev, A.; Franceschi, C.; Ivanchenko, M. eXplainable Artificial Intelligence (XAI) in Aging Clock Models. *Ageing Res. Rev.* **2024**, *93*, 102144. [CrossRef]
108. Díaz-Rodríguez, N.; Del Ser, J.; Coeckelbergh, M.; López De Prado, M.; Herrera-Viedma, E.; Herrera, F. Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation. *Inf. Fusion* **2023**, *99*, 101896. [CrossRef]
109. Radclyffe, C.; Ribeiro, M.; Wortham, R.H. The Assessment List for Trustworthy Artificial Intelligence: A Review and Recommendations. *Front. Artif. Intell.* **2023**, *6*, 1020592. [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

The Integration of Deep Learning in Radiotherapy: Exploring Challenges, Opportunities, and Future Directions through an Umbrella Review

Andrea Lastrucci [1], Yannick Wandael [1], Renzo Ricci [1], Giovanni Maccioni [2] and Daniele Giansanti [2,*]

[1] Department of Allied Health Professions, Azienda Ospedaliero-Universitaria Careggi, 50134 Florence, Italy; andrea.lastrucci@unifi.it (A.L.); wandaely@aou-careggi.toscana.it (Y.W.); riccire@aou-careggi.toscana.it (R.R.)
[2] Centre TISP, Istituto Superiore di Sanità, 00161 Rome, Italy; gvnnmaccioni@gmail.com
* Correspondence: daniele.giansanti@iss.it

Abstract: This study investigates, through a narrative review, the transformative impact of deep learning (DL) in the field of radiotherapy, particularly in light of the accelerated developments prompted by the COVID-19 pandemic. The proposed approach was based on an umbrella review following a standard narrative checklist and a qualification process. The selection process identified 19 systematic review studies. Through an analysis of current research, the study highlights the revolutionary potential of DL algorithms in optimizing treatment planning, image analysis, and patient outcome prediction in radiotherapy. It underscores the necessity of further exploration into specific research areas to unlock the full capabilities of DL technology. Moreover, the study emphasizes the intricate interplay between digital radiology and radiotherapy, revealing how advancements in one field can significantly influence the other. This interdependence is crucial for addressing complex challenges and advancing the integration of cutting-edge technologies into clinical practice. Collaborative efforts among researchers, clinicians, and regulatory bodies are deemed essential to effectively navigate the evolving landscape of DL in radiotherapy. By fostering interdisciplinary collaborations and conducting thorough investigations, stakeholders can fully leverage the transformative power of DL to enhance patient care and refine therapeutic strategies. Ultimately, this promises to usher in a new era of personalized and optimized radiotherapy treatment for improved patient outcomes.

Keywords: deep learning; radiotherapy; digital radiology; artificial intelligence

1. Introduction

1.1. Background

Radiotherapy is a clinical sector based on the therapeutic use of ionizing radiation. It represents one of the most important methods of anti-tumor therapy in the various oncological sectors [1,2]. Among the sectors in which it is most used, we find the sector of gynecological tumors, of the rectum and anus, of the lungs, and of the head and neck area. However, the new equipment allows this technological method to also be used in many other sectors, such as, to name a few: sarcomas, primary and secondary tumors of the brain and spinal cord, breast cancer, pancreatic tumors, primary and secondary liver tumors, and other categories. This therapy can be used both exclusively and in combination with other therapeutic methods.

The technological approach is based on a range of solutions that offer different opportunities, and the most recent results have made it possible to develop different radiotherapy modalities with the objective of increasing focus on ensuring better control of the disease and minimizing the medium- and long-term adverse effects. All this requires an increasingly continuous investment in high technological specialization and new experimentation.

Among the dominant methods, we find, for example [3–7]:

1. *Traditional radiotherapy* The patient undergoes radiotherapy every day for five days a week, and for a number of weeks that varies based on the disease and parameters, which depend on the tumor to be treated [3].
2. *Conformal radiotherapy* It differs from the previous one due to the use of a linear accelerator equipped with a multi-leaf collimator, which allows for better focusing on the neoplasm to be treated while minimizing damage to the surrounding healthy tissues [4].
3. *Intensity-modulated beam radiotherapy (IMRT)* It has the use of a multi-leaf collimator in common with conformal radiotherapy. In this case, however, the collimator blades move over the area to be irradiated during irradiation with a pre-established sequence monitored by the computer, while the machine delivers the radiation beam. This allows for greater precision than conformal radiotherapy [5].
4. *Stereotactic radiotherapy* It is based on greater attention and accuracy to the immobilization of the patient [6]. It is indicated for very particular clinical cases and a very high specialization of the centers where it is used. In this case, particular accelerators are used, which use a particular device called a gamma knife (gamma ray scalpel).
5. *Hadrontherapy* This technology uses subatomic particles, such as protons and ions, capable of irradiating the disease with extreme precision and with different biological effectiveness [6]. In this case, as in the previous one, the therapeutic indications are diverted toward very selected cases [7].

These different methods are now the subject of numerous 360-degree comparative studies by the scientific community [8,9], with a particular focus on innovative dose minimization approaches [10].

1.2. Introduction and Potential Applications of Deep Learning in Radiation Therapy

Regardless of which of the five technological methods will be used, radiotherapy is preceded by complex planning. For example, important parameters come into play, such as [11]:

- The PTV (planning target volume), which represents the target volume associated with the body region to be irradiated.
- The GTV (gross tumor volume), which is associated with the actual tumor mass and is related to the PTV.
- The CTV (clinical target volume), which includes the neighboring regions with healthy tissue. This parameter is also related to PTV.

PTV has a central role in radiotherapy, and for some time we have been trying to support its selection with algorithms and/or image processing software [11]. The development of deep learning (DL) has brought new life to all this.

Deep learning, through the use of neural networks with multiple interconnected layers, allows the processing of an enormous amount of data and learning from them [12,13].

Software solutions for contouring based on DL and automated segmentation of anatomical structures are becoming increasingly widespread and seem to offer clear advantages in clinical application [14]. The use of DL-based auto-contouring tools can save a significant amount of time in the procedure, minimize manual corrections, represent a third-party solution for standardizing procedures, and minimize collateral exposure of healthy tissue. In recent years, we have witnessed important developments in this area also driven by the boom in interest that has spread toward the use of artificial intelligence in the health domain sector with the explosion of the COVID-19 pandemic [15,16].

1.3. The Rationale and the Purpose of the Study

Some important questions are resonating around the introduction of AI and, in particular, DL in this area, such as:

- *What is the state of development and diffusion of DL in radiotherapy?*
- *What are the opportunities and the obstacles encountered and the challenges to overcome them?*

Systematic reviews are an important basis for making a map point since they are an index of the consolidation of emerging themes and an indirect index of the themes that require greater concentration of scholars.

The objective of this study is, therefore, to carry out, through a narrative review, an umbrella review [17,18], which allows an analysis of the systematic reviews produced to date in this area to answer these questions.

Performing an umbrella review based on a narrative review of systematic reviews is vital for consolidating evidence and identifying both emerging themes and patterns. It enables a nuanced understanding of the research landscape, an exploration of heterogeneity, and the identification of gaps [17,18].

1.4. Organization of the Study

The study is structured into five sections, including the introduction in Section 1.

Section 2 outlines the umbrella review design, illustrating the methodology employed to identify each systematic review included in the study.

Following this is Section 3, presenting the results, which are divided into subsections.

Section 3.1 conducts an analysis of trends in scientific production in this field over time.

Section 3.2 provides a detailed analysis, categorizing each study from the overview, supported by tables (Table 1 and Table 2).

Additionally, Section 3.2 includes an excerpt from each analyzed study.

Section 4 is devoted to the discussion, organized into seven subsections. Section 4.1 presents preliminary insights, offering initial observations and understandings.

Section 4.2 critically explores emerging opportunities, examining potential paths for further investigation or development.

In Section 4.3, specific areas warranting deeper examination are identified and discussed.

Section 4.4 elaborates on key areas where a collective need for more in-depth exploration is recognized. This subsection is further subdivided into detailed paragraphs, providing targeted analyses derived from the literature.

Section 4.5 introduces and offers commentary on a comprehensive synoptic diagram, conclusively summarizing the umbrella review project. Following this, the takeaway message (Section 4.6) and limitations (Section 4.7) of the study are presented.

Finally, Section 5 is dedicated to the conclusions.

2. Methods

This umbrella review, based on a narrative review, used the ANDJ standardized checklist designed for narrative reviews [19]. Such a narrative checklist is a methodological tool that provides detailed and structured guidance during the review process. It aids in standardizing the review process by establishing key criteria for use during the analysis, making the process of constructing the study transparent.

The PubMed and Scopus databases were used in the overview. A qualification methodology was used to choose the studies based on an assessment of qualified parameters [20]. Based on [20], we evaluated each contribution based on six key parameters:

1. Clarity of study rationale in the introduction,
2. Appropriateness of the work's design,
3. Clarity in describing methods,
4. Clear presentation of results,
5. Justification and alignment of conclusions with results,
6. Adequate disclosure of conflicts of interest by authors,
7. The scoring system involves assigning graded scores (1 = min; 5 = max) to each one of the first five parameters based on the quality of each criterion.

For the last parameter, a binary assessment (Yes/No) was conducted regarding the disclosure of conflicts.

To preselect studies:

- Each of the first five parameters must obtain a minimum score of 3,

- The last parameter must be marked "Yes" for conflict disclosure.

Only peer-reviewed studies were considered.

We used the keys "radiotherapy" and "deep learning" combined in AND logic and with searches both in the title/abstract and the full text.

We identified 18 studies [21–38].

3. Results

Below is an analysis of the trends of the studies in this field, reported in Section 3.1, a detailed analysis, with a categorization of the overview of each study supported by tables (Table 1 and Table 2), in Section 3.2, and a summary in Section 3.3.

3.1. The Trends in the Studies on Deep Learning in Radiotherapy

A search was conducted using specific criteria (see Box 1) on the PubMed database, resulting in a total of 1003 studies on the use of DL in radiotherapy from 2013 to the present. Figure 1 shows the increasing number of articles found in PubMed related to the application of DL in radiotherapy, based on the selected search parameters. Figure 2 presents the distribution of article types, with reviews (n = 87) being more prevalent than systematic reviews (n = 19), in relation to DL in radiotherapy.

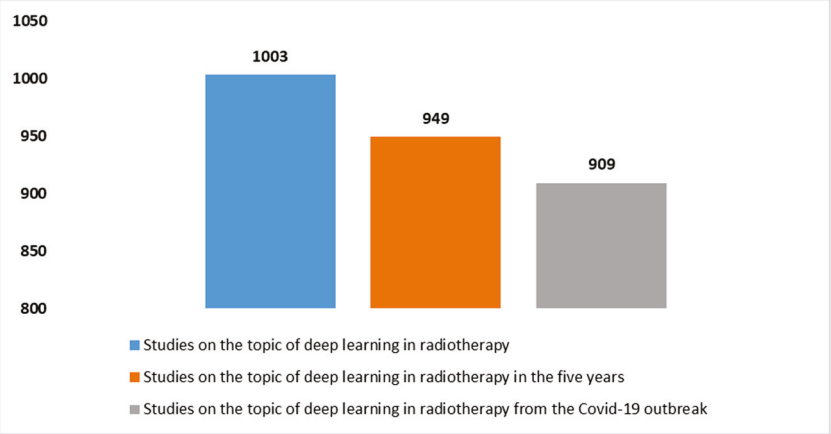


Figure 1. Studies focusing on deep learning and radiotherapy.

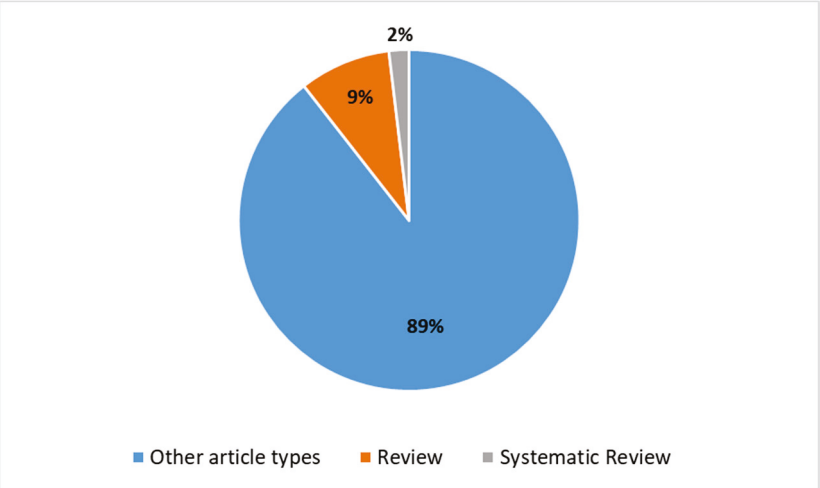


Figure 2. Article types focusing on deep learning and radiotherapy.

Box 1. The proposed composite keys.

(radiotherapy[Title/Abstract]) AND (deep learning[Title/Abstract])
(radiotherapy[Title/Abstract])

Research in this field has significantly accelerated during two distinct periods, as depicted in Figure 1. The first surge occurred in the past five years (from 2019 to now), with 94.6% of all indexed articles on this topic being published on PubMed. This period highlighted the growing interest and collaborative efforts to enhance knowledge and understanding of DL in radiotherapy.

The onset of the COVID-19 pandemic marked an even faster growth phase, with nearly all published articles (90.6%) since 2020 focusing on this topic.

The COVID-19 pandemic has caused a surge in the use of DL in the medical field, particularly in radiotherapy. This increase in research activity highlights the ability of the academic and research communities to respond quickly and effectively to the challenges of the pandemic. It also showcases their ability to drive innovation during times of crisis.

The rise in publications on DL in radiotherapy on PubMed is a result of advancements in technology and a growing recognition of DL's potential to improve the radiotherapy workflow. By streamlining processes, DL is at the forefront of transforming radiotherapy practices, making them more personalized, accurate, and efficient. The emphasis placed by researchers on this subject has been markedly intense from 2020 to the present, as demonstrated in Figure 3. This figure indeed reveals that interest among researchers has significantly surged since the COVID-19 pandemic, with approximately 90% of articles concerning the application of DL in radiotherapy being published since the onset of the COVID-19 pandemic (n = 897).

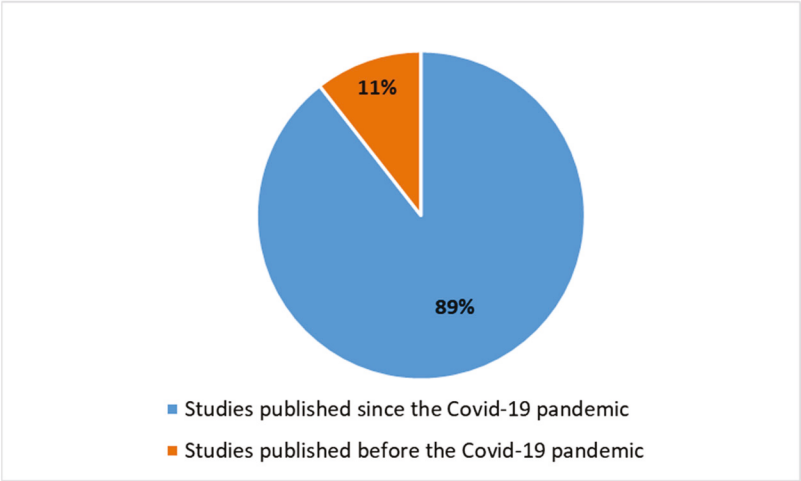


Figure 3. Number of articles published about the use of DL in RT pre- and post-COVID-19 pandemic.

In conclusion, the growing body of literature on DL in radiotherapy, especially in light of the COVID-19 pandemic, emphasizes the important role of DL in shaping the radiotherapy workflow. This trend not only showcases the field's ability to adapt and innovate in response to emerging challenges, but also signifies the increasing acknowledgment of DL as a crucial element in continuously improving and refining radiotherapy techniques.

3.2. Outcome from the Analysis

3.2.1. General Findings from the Analysis

The collective message that emerged from the analysis of the systematic reviews underscores a profound development in the medical field, particularly within the domain of radiotherapy, propelled by the advent of cutting-edge technologies, such as artificial intelligence (AI), deep learning (DL), and convolutional neural networks (CNNs) [38]. These

systematic reviews meticulously delineate the multifaceted impact of these technologies on various aspects of healthcare, encompassing clinical practice, the education of healthcare professionals, and the broader healthcare landscape [25].

A salient theme elucidated throughout these reviews is the immense potential of DL to catalyze a paradigm shift in radiotherapy workflows. DL, especially in its manifestation within the field of AI, emerges as a pivotal player, offering substantial enhancements across diverse stages of radiotherapy—ranging from pre-treatment planning to real-time treatment delivery and post-treatment evaluation [38]. Through a comprehensive analysis of the systematic reviews, it becomes evident that DL's scope within radiotherapy spans a myriad of topics. Notably, DL methodologies empower significant advancements in predicting treatment outcomes [38], streamlining and automating contour delineation [21,33,35,37], and optimizing treatment planning procedures [21,24,32,35,37].

Moreover, the integration of DL into radiomic analyses represents a promising avenue for clinicians, enabling the extraction of intricate features from medical imaging data. This integration is thoroughly evaluated in several systematic reviews [22,26,27,36,38], which underscore its potential to enhance diagnostic precision and prognostic assessments. Furthermore, these reviews illuminate that the development of radiomics not only aims to augment the accuracy of diagnostic and prognostic evaluations but also endeavors to tailor radiotherapy interventions to the unique characteristics of individual tumors, potentially leading to superior clinical outcomes and mitigated side effects.

The imperative of standardizing DL integration looms large in the specialist literature, alongside considerations regarding user-friendliness and technical nuances [22,26,34]. These systematic reviews underscore the critical necessity of establishing clear guidelines for the seamless integration of DL into existing systems, thereby addressing challenges pertaining to user interface, user experience, and the technical intricacies associated with the deployment of AI technologies in healthcare.

Each systematic review, including those conducted by Boldrini et al. [38] and Huang et al. [25], accentuates the exigency for further research to comprehensively grasp the clinical implications and intricacies surrounding the integration of DL into routine clinical practice. These studies advocate for intensified efforts to discern how DL can be effectively assimilated into standard daily protocols. Additionally, several systematic reviews underscore the paramount importance of publicly available, high-quality databases and rigorous validation processes to bolster the development of DL across diverse radiotherapy applications [23,26,28,29,32–34]. Notably, the review by Booth et al. [32] draws attention to the prevailing challenges of small sample sizes and a high risk of bias in current research endeavors, thereby emphasizing the imperatives for enhanced study quality and refined research methodologies.

In essence, these systematic reviews collectively paint a compelling narrative of a transformative shift in the medical landscape, precipitated by the seamless integration of AI and DL into clinical practice within radiotherapy workflows. A pivotal determinant for augmenting the efficacy of DL in this context lies in the availability of expansive, shared datasets. Such repositories of high-quality data serve as indispensable assets in optimizing DL's performance and fostering the development of more precise, reliable applications in the realm of radiotherapy [29,33,34].

Table 1 reports the key findings from the analysis of the overview of systematic reviews.

Table 1. Key findings from the analysis of the overview of systematic reviews.

Review Study	Key Findings from the Analysis
Almeida et al. (2020) [21]	This study showcases DL's potential to revolutionize prostate cancer treatment planning by automating the contouring process, leveraging CT and MRI for improved efficiency and consistency. This paper also underscores the hurdles of limited patient datasets and the need for further development before full clinical adoption is feasible.
Kothari et al. (2020) [22]	Focusing on the prognostic value of radiomics in NSCLC treated with radiotherapy, this systematic review notes heterogeneity in study methodologies and recommends standardizing radiomics features and employing robust methods and DL to improve future model performance.
Chlap et al. (2021) [23]	This study explores the critical role of data augmentation in enhancing DL models for radiology and radiotherapy, categorizing techniques for CT and MRI images to offset the need for large datasets. The study emphasizes data augmentation's value in bolstering algorithm performance and validating clinical applicability amidst dataset limitations.
Spadea et al. (2021) [24]	This study assesses DL-based synthetic CT generation across three clinical applications, including its use in MR-based treatment planning, IGRT, and PET correction. The study examines the contributions, challenges, and future potential of DL-based sCT methods, evaluating their readiness for clinical implementation.
Huang et al. (2021) [25]	This study explores the pivotal role of computer technology and data expansion in advancing AI, offering radiation oncologists efficient tools to enhance radiotherapy and prompting a crucial need for understanding DL principles for effective clinical application. This paper also discusses AI's potential growth in radiation oncology.
Walls et al. (2021) [26]	This study investigates radiomics' role in lung cancer radiotherapy decisions, noting the absence of validated biomarkers for personalized treatment. Despite linking radiomic indicators to clinical outcomes, challenges, such as data standardization and validation, hinder practical application, emphasizing the need for improved research methodologies.
Kim et al. (2021) [27]	This meta-analysis evaluates the AI-assisted MRI's diagnostic ability for distinguishing true progression from non-progression in brain metastasis post-radiotherapy, finding sensitivity and specificity rates of 77% and 74%, respectively. Despite these findings, the current diagnostic reliability of AI-assisted MRI is insufficient for clinical application.
Avanzo et al. (2021) [28]	This paper investigates AI applications in Italian imaging research from 2015 to 2020, revealing MRI as the predominant modality, notably in neurological diseases and cancer diagnosis. The study highlights a surge in AI research, particularly in classification and segmentation tasks, emphasizing the necessity for collaborative frameworks and shared databases.
Yang et al. (2022) [29]	This study analyzes the DL models for cervical cancer CT image segmentation, showing high accuracy in segmenting clinical target volumes and organs-at-risk. Despite the efficient performance, the study emphasizes the necessity for public, high-quality databases, and extensive validation for future radiotherapy applications.
Rusanov et al. (2022) [30]	This study investigates DL's role in improving CBCT image quality for online ART, focusing on updating patient anatomy to optimize treatment parameters despite traditional CBCT limitations. This review evaluates DL strategies for CBCT correction and synthetic CT generation, concluding with recommendations for clinicians and DL practitioners.
Booth et al. (2022) [31]	This study assesses the accuracy of DL models in monitoring biomarkers for glioblastoma treatment response, revealing promising diagnostic performance in differentiating tumor progression from mimics using MRI features. Despite moderate sensitivity and specificity, the studies suffer from small sample sizes and a high bias risk, indicating a need for improved study quality to refine research methodologies.
Hasan et al. (2022) [32]	This study examines CNNs' application in ENT radiology, revealing their high accuracy in tasks such as structure identification, pathology detection, and tumor segmentation for radiotherapy across various subspecialties. The study highlights the potential of CNN's application to revolutionize clinical practice by automating and improving diagnostic and treatment planning processes.
Liu et al. (2023) [33]	This study assesses DL algorithms' effectiveness in contouring organs-at-risk in HNC radiation planning, demonstrating DSC and indicating DL's potential to automate contouring and enhance precision in radiotherapy plans. The study emphasizes the importance of quality datasets to optimize DL's performance, potentially reducing oncologists' workload.

Table 1. *Cont.*

Review Study	Key Findings from the Analysis
Tan et al. (2023) [34]	This study explores the efficacy of DL models in predicting radiotherapy-induced toxicity across multiple cancer types, emphasizing advanced techniques, such as ensemble learning and transfer learning, and underscores the necessity for larger datasets and standardized methodologies to improve research outcomes.
Franzese et al. (2023) [35]	This study analyzes DL's role in HNC radiotherapy, emphasizing organ-at-risk segmentation's prominence and advocating for assessing AI's clinical impact and confidence levels for predictions. The study concludes by highlighting AI's potential to automate HNC radiotherapy workflows.
Eidex et al. (2023) [36]	This systematic review highlights DL's role on MRI-guided radiation therapy in enhancing tumor segmentation, deriving X-ray attenuation from MRI, and improving tumor characterization and motion tracking, with recent trends focusing on multi-modal, visual transformer, and diffusion models.
Chen et al. (2024) [37]	This study investigates unpaired image-to-image translation in medical imaging, showcasing its applications in segmentation and clinical tasks but noting limitations, such as limited external validation and scarce pre-trained models, hindering immediate clinical application.
Boldrini et al. (2024) [38]	This systematic review explores the impact of AI, DL, and radiomics on IGRT, revealing their potential in diagnosis, treatment optimization, and outcome prediction, though further research is needed to establish their clinical impact and integration into standard protocols.

AI: artificial intelligence; ART: adaptive radiation therapy; CBCT: cone-beam computed tomography; CNN: convolutional neural network; CT: computed tomography; DSC: dice similarity coefficient; DL: deep learning; HNC: head and neck cancer; IGRT: image-guided radiation therapy; MRI: magnetic resonance imaging; NSCLC: non-small-cell lung cancer; PET: positron emission tomography.

3.2.2. Emerging Categorization from the Analysis

Delving into greater detail, we identified the following areas of interest and/or emphasis, with a focus on the categorization, as reported in Table 2.

Table 2. Emerging categorization from the systematic reviews analyzed.

Area of Interest	Focus on the Categorization
Automating contouring process	Several systematic reviews (Almeida et al. [21], Huang et al. [25], Avanzo et al. [28], Yang et al. [29], Liu et al. [33], Franzese et al. [34], and Chen et al. [37]) explore the increasing role of DL on image segmentation and automating the contouring process, permitting a reduction in the workload for physicians and enabling more precise radiotherapy plans.
Use of radiomics	Kothari et al. [22], Walls et al. [26], Kim et al. [27], Eidex et al. [36], and Boldrini et al. [38] discuss the increasing role of radiomics in guiding clinical decisions before, during, and after radiotherapy, highlighting the need of further research for its integration into clinical practice.
Synthetic CT	Spadea et al. [24] and Rusanov et al. [30] explore the innovative application of DL to generate and enhance synthetic CT (sCT) images for improved radiation therapy planning and execution. This approach addresses traditional imaging limitations by providing high-quality, accurate sCT images for a variety of clinical applications in radiotherapy.
Application of DL for ART	Sapdea et al. [24] and Rusanov et al. [30] investigate the use of DL to improve the quality of cone-beam CT images for guiding healthcare professionals in online ART.
Data augmentation techniques for DL models	Chlap et al. [23] and Tan et al. [34] examine data augmentation techniques for the development and improvement of DL models with various applications in radiotherapy and emphasize their necessity due to the reliance on large datasets for training.

Table 2. *Cont.*

Area of Interest	Focus on the Categorization
Use of DL for prediction of side effects and clinical outcome	Several systematic reviews (Huang et al. [25], Walls et al. [26], Booth et al. [31], Tan et al. [34], Franzese et al. [35], and Boldrini et al. [38]) discuss the application and use of DL predictions of the outcome of radiotherapy and use DL to predict the toxicity and outcomes after radiotherapy.
Improvements in treatment planning process	Several systematic reviews (Almeida et al. [21], Hasan et al. [32], Liu et al. [33], Franzese et al. [35], Chen et al. [37], and Boldrini et al. [38]) focus on the growing role of DL techniques in radiotherapy planning optimization, highlighting the significant innovation due to introduction of DL in daily practice workflow.
Image fusion	Huang et al. [25] discuss the utilization of DL techniques for enhancing medical image registration across various imaging modalities (multi-time and/or multimode registration).

ART: adaptive radiation therapy; CT: computed tomography; DL: deep learning.

These emerging themes/patterns collectively indicate a trend toward the integration of DL to enhance and improve radiotherapy workflow. The systematic reviews also highlight the need for further research, standardization, and overcoming challenges for broader adoption of this technology.

3.3. In-Depth Analysis of the Detected Studies: A Comprehensive Overview

To complement our overview, after having identified the themes and focus elements of the studies, here, we report a more far-reaching summary of each individual systematic review.

Almeida et al. [21] highlighted the potential of DL in improving treatment planning for prostate cancer, focusing on automating the contouring process to enhance speed and consistency while maintaining quality. They examined various network architectures based on computed tomography (CT) or magnetic resonance imaging (MRI), noting the field's rapid growth but also the limitations due to small patient datasets. Despite promising results, there remains a significant gap before these technologies can be fully integrated into clinical practice.

The systematic review and meta-analysis proposed by Kothari et al. [22] assessed the prognostic value of radiomics-based models in non-small-cell lung cancer (NSCLC) treated with radical radiotherapy, finding modest capabilities with a Harrell's Concordance Index (C-index) of 0.57. They highlighted significant heterogeneity in feature selection and model development across studies and suggested that future research focus on standardized radiomics features, robust methodology, and DL to enhance model performance.

Chlap et al. [23] examined data augmentation techniques for DL models in radiology and radiotherapy, highlighting their necessity due to the reliance on large datasets for training. They categorized augmentation methods applied to CT and MRI medical images into basic, deformable, and DL-based techniques, aiming to enhance model performance and provide insights into their clinical validity. The study underscores the importance of data augmentation in improving DL algorithms where large datasets are scarce.

Spadea et al. [24] evaluated DL-based synthetic CT (sCT) generation methods, grouped into three main clinical applications: (i) replacing CT in MR-based treatment planning, (ii) aiding cone-beam CT in image-guided adaptive radiotherapy, and (iii) creating attenuation maps for positron emission tomography (PET) correction. They covered research from January 2014 to December 2020, detailing each category's contributions, challenges, and achievements, and assessed the clinical readiness and potential future trends of DL-based sCT methods.

Huang et al. [25] highlighted that the widespread use of computers and data explosion has significantly fueled artificial intelligence (AI) development, with DL algorithms, such as convolutional neural networks (CNNs), offering radiation oncologists promising tools to streamline radiotherapy and reduce their workload. This facilitates more time for advanced

decision-making. As DL evolves closer to clinical application, it is crucial for oncologists to understand its principles for effective use. This paper delved into AI's development, basic concepts, and its potential in radiation oncology, suggesting further growth prospects for DL in this field.

Walls et al. [26] explored the utility of radiomics in guiding clinical decisions for lung cancer radical radiotherapy, highlighting the lack of validated biomarkers for personalized treatment in the face of the disease's significant mortality. Their systematic review included and analyzed 44 studies that established a connection between radiomic indicators, notably texture features and kurtosis, and clinical outcomes, including disease management, patient survival, and treatment-related toxicity. Despite these insights, obstacles, such as the need for standardized data, improved reporting, and external validation through prospective studies, pose barriers to the practical application of radiomics, emphasizing the critical need for enhanced evaluative frameworks and research methodologies in the field.

Kim et al. [27] conducted a systematic review and meta-analysis to tackle the diagnostic challenge of distinguishing true progression from non-progression in brain metastasis after stereotactic radiotherapy or surgery using MRI. This analysis, encompassing 7 studies with a total of 485 patients, found that AI-assisted MRI, through radiomics, achieved sensitivity and specificity rates of 77% and 74%, respectively. While the overall quality of the studies was deemed favorable, the current diagnostic capabilities of AI-assisted MRI fall short of the reliability needed for clinical use, underscoring the necessity for further research utilizing more sophisticated methodologies and expanded datasets.

Avanzo et al. [28] explored the landscape of AI applications in imaging research in Italy, covering the period of 2015–2020. They highlighted MRI as the most prevalent imaging modality, with a significant focus on neurological diseases and cancer diagnosis. The review noted a dramatic increase in AI research, particularly in classification and segmentation tasks, with a mix of machine learning and DL approaches. The findings underscore the burgeoning interest and need for collaborative frameworks, shared databases, and research guidelines in the AI imaging domain within Italy.

Yang et al. [29] proposed a systematic review and meta-analysis on DL models for cervical cancer CT image segmentation, reviewing 1893 articles and including 14 in the analysis. The study included 14 articles and revealed high accuracy in segmenting clinical target volumes and organs-at-risk, with dice similarity coefficient (DSC) scores ranging from 0.83 to 0.92. Despite good performance and efficiency (segmentation times between 15 s and 2 min), the findings highlighted the need for public, high-quality databases and further large-scale validation for future radiotherapy applications.

Rusanov et al. [30] explored the application of DL in enhancing cone-beam CT (CBCT) image quality for online adaptive radiation therapy (ART), addressing how updated patient anatomy can optimize treatment parameters despite CBCT's traditional quality limitations. This comprehensive review, covering January 2018 to April 2022, evaluated DL strategies for CBCT correction and synthetic CT generation, emphasizing study designs, DL techniques, image quality, and dosimetric accuracy. They concluded with recommendations for clinicians and DL practitioners, identifying literature gaps and advocating for the integration of state-of-the-art DL methods in radiation oncology.

Booth et al. [31] evaluated the accuracy and quality of machine learning (ML) models, specifically DL, in monitoring biomarkers to assess treatment response in glioblastoma, based on articles from September 2018 to January 2021. Despite the promising diagnostic performance of ML models in differentiating between tumor progression and mimics in glioblastoma using MRI features, the studies suffered from small sample sizes, high bias risk, and applicability concerns, highlighting the need for improved study quality and design. The meta-analysis of ten studies showed moderate sensitivity, specificity, and balanced accuracy, indicating potential yet calling for refinement in research methodologies.

Hasan et al. [32] explored the application of CNNs in ENT radiology, a field where their use is less common compared to other radiology disciplines, possibly due to unfamiliarity within the otolaryngology community. Through a systematic review of thirty articles up

to October 2020, they demonstrated CNNs' high accuracy in tasks such as identifying structures, detecting pathology, and segmenting tumors for radiotherapy in various ENT subspecialties. This study underscores the potential of CNNs to significantly impact clinical practice in ENT radiology by automating and enhancing diagnostic and treatment planning processes.

The systematic review and meta-analysis proposed by Liu et al. [33] investigated the effectiveness of DL algorithms for contouring organs-at-risk in head and neck cancer radiation treatment planning, assessing 22 studies from a pool of 149. The analysis revealed high dice similarity coefficient (DSC) scores across various organs-at-risk, indicating DL algorithms' accuracy in automating the contouring process, thus potentially reducing the workload for oncologists and enabling more precise radiotherapy plans. The study highlighted the necessity for high-quality datasets and further algorithmic improvements to optimize DL's performance in clinical settings.

Tan et al. [34] evaluated DL models' effectiveness in predicting radiotherapy-induced toxicity, analyzing fourteen studies across various cancer types (prostate (n = 2), head and neck cancer (HNC; n = 4), liver (n = 2), lung (n = 4), cervical (n = 1), and esophagus (n = 1)). The review highlighted the utilization of advanced techniques, such as ensemble learning, data augmentation, and transfer learning, in model development. The authors concluded that while DL models show promise in toxicity prediction, future research needs larger, more diverse datasets and standardized methodologies to enhance research consistency and outcomes.

Franzese et al. [35] investigated the role of DL in the radiotherapy workflow for HNC, analyzing 62 selected articles from 2016 to 2022 that covered the entire RT workflow: contouring, planning, and delivery. This systematic review highlighted the significant focus on organ-at-risk segmentation and the need for studies to assess the clinical impact of AI and provide confidence levels for AI predictions. The conclusion underscored AI's potential in automating the complex radiotherapy workflow for HNC and called for interdisciplinary research to align AI development with clinical requirements.

Eidex et al. [36] analyzed 197 studies up to 31 December 2022, on MRI-guided radiation therapy and DL, categorizing them into image segmentation, synthesis, radiomics, and real-time MRI to fully support ART. This systematic review discussed the clinical significance and challenges of DL in enhancing tumor segmentation, deriving X-ray attenuation from MRI, and improving tumor characterization and motion tracking, emphasizing the rise of multi-modal, visual transformer, and diffusion models in recent trends.

Chen et al. [37] investigated the use of unpaired image-to-image (I2I) translation in medical imaging. Out of 461 studies, 55 were included, showcasing I2I's role in segmentation, unpaired domain adaptation, denoising, and clinical applications, such as automatic contouring for MRI, CT, and radiotherapy planning. Despite its potential, the limited external validation and the scarcity of publicly available pre-trained models restrict the immediate practical application of these methods in clinical practice.

Boldrini et al. [38] investigated the influence of AI, DL, and radiomics on image-guided radiation therapy (IGRT) in radiation treatments, highlighting their promising roles in diagnosis, treatment optimization, and outcome prediction. Through a systematic search of electronic databases, 84 papers were analyzed, revealing significant contributions of AI and radiomics to IGRT, across 23 and 61 papers, respectively. Despite the reliance on retrospective data, AI and radiomics were shown to significantly enhance IGRT across all radiotherapy workflow phases, indicating a need for further research to solidify their clinical impact and integration into standard treatment protocols.

4. Discussion

The section is structured into multiple paragraphs. Section 4.1 presents early insights, providing initial observations and understandings. Section 4.2 delves into the opportunities that have arisen, exploring potential avenues for further investigation or development. In Section 4.3, specific areas warranting deeper examination are identified and discussed.

Section 4.4 elaborates on key areas where a collective need for more in-depth exploration is perceived. This section is subdivided into detailed subsections, offering targeted analyses derived from the literature. Section 4.5 introduces and provides commentary on a comprehensive synoptic diagram, conclusively summarizing the umbrella review project. Following this, the takeaway message (Section 4.6) and limitations (Section 4.7) of the study are presented.

4.1. Early Insights and Discussion

Radiation therapy has a history dating back to the early last century. The initial studies in this field date back to 1903, and since then, 226,830 studies have been published on PubMed (search key in Box 1, position 2). Over the years, with the advancement of technological innovations, various techniques, such as traditional radiotherapy, conformal radiotherapy, IMRT, stereotactic radiotherapy, and hadrontherapy, have been gradually established, and researchers are still exploring new methods [3–7].

Only a small fraction, approximately 0.44%, of these studies (equivalent to 1003, not perceptible in a potential graph) have addressed the introduction of DL, and this trend has only emerged since 2013. The significant push for research in the application of DL in radiation therapy was fueled by the COVID-19 pandemic (Figure 1). From the onset of the pandemic until now, 909 studies have been produced, representing over 90% of the total. Considering the last five years of scientific production, marked almost entirely by the COVID-19 pandemic, the ratio of total scientific production in radiation therapy, considering and not considering DL, has increased from 0.44% to 1.6%, quadrupling the interest and demonstrating a growing focus of researchers on these issues. This trend is now visually perceptible (Figure 4).

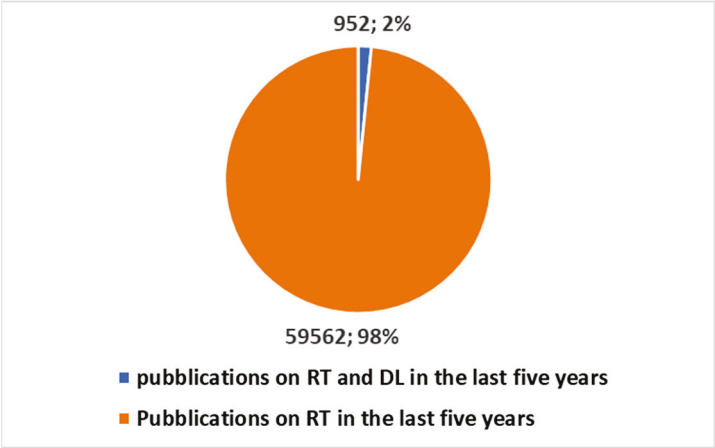


Figure 4. Articles on RT and DL versus articles on RT in the last five years.

The trend of reviews and systematic reviews, as shown in Figure 5, emphasizes developments in research and the consolidation of specific themes over the past five years through Cartesian graphs. Indeed, these two types of studies serve as indicators of both researchers' interest and the solidification around topics of interest, reflecting the establishment of medical knowledge. The trends of both types of studies follow a similar pattern, with systematic reviews naturally exhibiting a slight delay compared to the initial studies, as expected. This synchronization in trends implies that as primary studies emerge, systematic reviews follow, providing a comprehensive and reflective overview of the evolving research landscape. The alignment of these trends signifies the ongoing interest of researchers and the firm establishment of knowledge within the medical field.

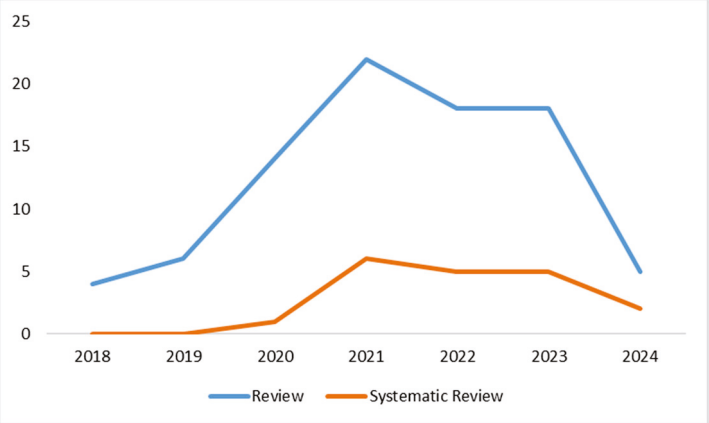


Figure 5. Temporal trends of reviews and systematic reviews published on the PubMed database focusing on deep learning and radiotherapy.

The accumulation of research trends, evolving themes, and the growing body of knowledge in a particular field serves as a compelling motivation and justification for our umbrella review study. An umbrella review [17,18] involves a comprehensive examination and synthesis of findings from multiple systematic reviews on a specific subject. In this context, our decision to conduct an umbrella review was specifically centered on systematic reviews. By delving into these studies, we aimed to offer a comprehensive overview of the existing body of research, focusing on the highest editorial scientific tools—the systematic reviews—recognized for their thorough analyses, which systematically gather and evaluate evidence, contributing to a comprehensive understanding of the subject matter. The overview revealed opportunities and specific areas requiring in-depth investigations, and perceived global areas needing further research and attention.

4.2. Opportunities Explored: Critical Reflections on DL in Radiotherapy

The intersection of deep learning (DL) and radiotherapy presents a dynamic yet critically transformative force, unveiling a plethora of opportunities across multifaceted dimensions of cancer care. Recent investigations [21–38] shed light on the expansive vistas that DL illuminates.

DL stands poised to revolutionize the landscape of radiotherapy, offering a myriad of transformative prospects across diverse applications. From automating intricate procedures to refining prognostic models and enhancing imaging methodologies, DL emerges as a disruptive force for innovation. Noteworthy advancements include its application in prostate cancer treatment planning, where DL streamlines the contouring process, promising efficiency and consistency [21]. This not only accelerates processes but also maintains high-quality standards, suggesting a paradigm shift in precision medicine for prostate cancer patients. Similarly, in non-small-cell lung cancer (NSCLC), DL-driven radiomics models enhance prognostic accuracy in radical radiotherapy [22], demonstrating the potential to optimize treatment strategies and outcomes, thus emphasizing DL's pivotal role in personalized cancer care.

Augmentation techniques for DL models in radiology and radiotherapy address the perennial challenge of limited large datasets, potentially improving performance [23]. DL-based synthetic CT generation methods pave the way for MR-based treatment planning, adaptive radiotherapy, and PET correction [24,25], offering precision and efficiency while optimizing treatment parameters based on updated patient anatomy.

Radiomics assumes a central role in guiding clinical decisions for lung cancer radical radiotherapy, leveraging advanced imaging features for personalized interventions [26]. The integration of radiomic indicators signifies a shift toward more targeted and effective lung cancer treatments. Moreover, in brain metastasis diagnosis, AI-assisted MRI shows

promise for enhancing diagnostic capabilities [27], albeit requiring further refinement for clinical deployment.

Italy has witnessed a surge in AI applications in imaging research, particularly in classification and segmentation tasks [28], highlighting the necessity of collaborative frameworks and shared databases to harness the full potential of AI in advancing Italian imaging research.

DL models exhibit promising accuracy in segmenting clinical target volumes for cervical cancer CT images, offering precision in treatment planning [29]. Improving cone-beam CT (CBCT) image quality for online adaptive radiation therapy emerges as a critical opportunity [30], necessitating the integration of state-of-the-art DL methods for optimized treatment.

Monitoring biomarkers through DL in glioblastoma treatment response assessment holds promise for personalized therapeutic strategies [31], potentially reshaping treatment evaluation paradigms.

In ENT radiology, convolutional neural networks (CNNs) emerge as high-accuracy tools for automating diagnostic and treatment planning processes [32]. Similarly, DL algorithms show potential in automating the contouring process for head and neck cancer radiation treatment planning [33], thereby refining precision and reducing oncologists' workload, leading to enhanced treatment efficiency.

DL models also demonstrate promise in predicting radiotherapy-induced toxicity across various cancer types [34], offering avenues for refining toxicity prediction models and tailoring personalized treatment plans. Moreover, DL's potential in automating complex radiotherapy workflows for head and neck cancer (HNC) presents transformative opportunities [35], with applications in organ-at-risk segmentation and workflow automation for enhanced HNC care.

DL in MRI-guided radiation therapy holds potential for improving tumor segmentation, deriving X-ray attenuation from MRI, and enhancing tumor characterization and motion tracking [36], promising precision and innovation.

Unpaired image-to-image translation in medical imaging emerges as a potential game-changer [37], with opportunities for exploring novel applications and methodologies in clinical practice, leading to advancements in segmentation, adaptation, and denoising.

The integration of AI, DL, and radiomics significantly influences image-guided radiation therapy (IGRT) across all workflow phases [38], offering prospects for further applications in diagnosis, treatment optimization, and outcome prediction for enhanced IGRT.

In summary, the transformative opportunities brought forth by DL in radiotherapy are extensive and profound. Collaborative endeavors, standardized methodologies, and a dedication to surmounting existing challenges are imperative to fully capitalize on these opportunities. As technology progresses, the potential to improve patient outcomes and streamline oncological care becomes increasingly tangible, awaiting exploration and implementation. In essence, the panorama of opportunities ushered in by DL in radiotherapy is vast and transformative, necessitating critical reflection and concerted efforts to realize its full potential [21–38]. Table 3 provides a synthesis of these emerging opportunities.

Table 3. Thematic areas within the emerging opportunities.

Opportunity	Application Area	Key Insights
Automating Contouring in Prostate Cancer Treatment Planning	Prostate Cancer Treatment Planning	DL automates contouring for speed and consistency, promising a paradigm shift in precision medicine [21].
Enhanced Prognostic Precision in NSCLC	NSCLC Radical Radiotherapy	DL-driven radiomics models refine prognostic precision, optimizing treatment strategies for NSCLC patients [22].

Table 3. *Cont.*

Opportunity	Application Area	Key Insights
Augmentation Techniques for Improved DL Performance	Radiology and Radiotherapy Augmentation	Basic, deformable, and DL-based augmentation methods enhance DL performance, addressing limited large datasets [23].
DL-Based Synthetic CT Generation Opportunities	MR-Based Treatment Planning and Adaptive Radiotherapy	DL-based synthetic CT generation optimizes treatment parameters based on updated patient anatomy, offering precision and efficiency [24].
Radiomics Guiding Clinical Decisions in Lung Cancer Radiotherapy	Lung Cancer Radical Radiotherapy	Radiomics guides personalized interventions, incorporating advanced imaging features for improved clinical decisions [26].
AI-Assisted MRI for Brain Metastasis Diagnosis	Brain Metastasis Diagnosis	AI-assisted MRI enhances diagnostic capabilities, showing potential for influencing timely and accurate treatment decisions [27].
AI Applications in Italian Imaging Research Landscape	Imaging Research in Italy	Italy experienced a surge in AI applications, particularly in classification and segmentation tasks, emphasizing the need for collaborative frameworks [28].
DL Precision in Cervical Cancer CT Image Segmentation	Cervical Cancer CT Image Segmentation	DL models exhibit high accuracy in segmenting clinical target volumes, enhancing precision in treatment planning [29].
Enhancing CBCT Image Quality for Online Adaptive Radiation Therapy	Online Adaptive Radiation Therapy	Opportunities in optimizing treatment parameters based on updated patient anatomy, addressing literature gaps in CBCT image quality enhancement [30].
Monitoring Biomarkers with DL in Glioblastoma Treatment Response	Glioblastoma Treatment Response Assessment	DL offers a promising avenue for monitoring biomarkers, potentially leading to personalized therapeutic strategies [31].
CNNs in ENT Radiology	ENT Radiology	CNNs automate diagnostic and treatment planning processes, showcasing high accuracy within the otolaryngology community [32].
DL Automation in Head and Neck Cancer Radiation Treatment Planning	Head and Neck Cancer Radiation Treatment Planning	DL algorithms promise to automate contouring, refining precision and reducing the workload for oncologists [33].
DL Models Predicting Radiotherapy-Induced Toxicity	Radiotherapy-Induced Toxicity Prediction	DL models exhibit promise in predicting toxicity, with opportunities for refining accuracy, specificity, and applicability [34].
DL in Complex Radiotherapy Workflow for HNC	Complex Radiotherapy Workflow for Head and Neck Cancer (HNC)	DL's potential in workflow automation presents opportunities for enhanced HNC care, particularly in organ-at-risk segmentation [35].
DL in MRI-Guided Radiation Therapy	MRI-Guided Radiation Therapy	Opportunities in enhancing tumor segmentation, deriving X-ray attenuation from MRI, and improving tumor characterization and motion tracking [36].
Unpaired Image-to-Image Translation in Medical Imaging	Medical Imaging	Potential game-changer in segmentation, unpaired domain adaptation, denoising, and automatic contouring, with opportunities for exploration [37].
Integration of AI, DL, and Radiomics in IGRT	Image-Guided Radiation Therapy (IGRT)	Significant influence on IGRT across all workflow phases, with opportunities for further exploration in diagnosis, treatment optimization, and outcome prediction [38].

4.3. Limitations Explored: Critical Reflections on DL in Radiotherapy

While the overview presented promising opportunities, it is imperative to confront the specific limitations existing within the field and underscore the areas requiring further scrutiny and refinement. The impact of AI and radiomics on image-guided radiation therapy (IGRT) appears substantial, promising advancements in treatment planning and precision radiotherapy [38]. However, the reliance on retrospective data in the literature casts doubt on the validity of the findings, necessitating rigorous external validation studies to establish their credibility. Confirmation of the potential of these tools and their

alignment with clinical outcomes and gold-standard treatment strategies is paramount for their practical deployment. Automated contouring technology driven by DL algorithms emerges as a potent tool for contouring head and neck organs-at-risk (OARs), purportedly reducing the workload of radiation oncologists and facilitating precision radiotherapy [33]. However, the identified need for constructing high-quality datasets highlights the imperative of optimizing datasets to bolster DL algorithm performance. Ongoing research endeavors should prioritize algorithm refinement and innovative approaches to ensure consistent accuracy.

Toxicity prediction employing DL techniques demonstrates consistent performance, suggesting their potential utility in radiotherapy [34]. However, the call for future research with large and diverse datasets and standardized study methodologies underscores the necessity of overcoming challenges to enhance the reliability and consistency of research outputs in this domain. The automation of the radiotherapy workflow for head and neck cancer (HNC) treatment holds transformative potential [35]. Interdisciplinary collaboration, involving both clinicians and computer scientists, is advocated for future studies, ensuring the effective alignment of AI technologies with clinical needs, and fostering a comprehensive understanding of their clinical impact.

In ENT radiology, the deployment of AI methodologies shows promise in various applications, including nodule and tumor identification, anatomical variation identification, and tumor segmentation [32]. This evolving field anticipates further evolution and integration into everyday practice. However, to fulfill its potential, continuous refinement of technologies and methodologies is indispensable. The translation of medical images through image-to-image (I2I) models presents valuable applications for medical physicists [37]. Nevertheless, the dearth of external validation studies and pre-trained models hampers immediate practical application. Initiatives aimed at overcoming these limitations, such as increased external validation and broader availability of pre-trained models, are imperative for advancing these techniques in practice.

Deep learning models demonstrate high accuracy in the automatic segmentation of cervical cancer CT images, offering promise for future radiotherapy applications. Nonetheless, the need for public high-quality databases and large-scale research verification underscores the importance of collaborative efforts to enhance the reliability and applicability of these models.

Future research endeavors should prioritize the development of improved tools for the evaluation of radiomics studies, addressing issues such as standardization of input scan data, quality of reporting, and external validation in randomized clinical trials. Machine learning models utilizing MRI features for distinguishing progression from mimics demonstrate good diagnostic performance, but the need for improvement in study quality and design is evident [31]. Future studies should focus on refining methodologies to strengthen the diagnostic capabilities of AI in MRI-based applications. The evaluation of DL-based synthetic CT (sCT) generation reveals its potential benefits and clinical readiness [24]. Future studies should delve deeper into the clinical applicability of these methods, ensuring they meet the required standards for safe and effective use. As artificial intelligence models trained using augmented data move toward clinical implementation, continuous refinement and validation are essential. The review anticipates the evolving landscape of AI techniques and aims to instill confidence in the validity of the models produced. The diagnostic performance of AI-assisted MRI remains inadequate for reliable use in clinical practice [27]. The identified need for future studies with improved methodologies and larger training sets emphasizes the ongoing efforts required to enhance the reliability and effectiveness of AI-assisted MRI. The surge of interest in AI applied to imaging in Italy highlights the need for collaborative frameworks, shared databases, and research guidelines [28]. These initiatives are crucial to fully harness the potential of AI applications in medical imaging. Radiomics-based models for lung cancer demonstrate modest prognostic capabilities [22]. Future research should focus on standardizing radiomics features, robust feature selection, and incorporating DL techniques to improve the predictive accuracy of

imaging-based models. The comprehensive review outlining the development and application of AI in radiation oncology emphasizes the need for ongoing research and clarifies the potential for further DL development in this field. The recognition of this potential paves the way for continued exploration and advancements. Although models have achieved satisfactory results, there is acknowledgment of the need for improvement before safe and effective clinical application. This recognition underscores the commitment to continuous refinement and enhancement in the journey toward practical implementation. Overall, the discourse on DL in radiotherapy reflects a dynamic landscape with transformative potential. While opportunities are abundant, the identified limitations underscore the importance of ongoing research, interdisciplinary collaboration, and standardization efforts. Addressing these challenges will be instrumental in unlocking the full potential of DL in this field, ultimately enhancing patient outcomes and advancing the field of radiotherapy.

Table 4 reports the specific emerging suggestions for a broader investigation.

Table 4. Thematic areas and key suggestions for a broader investigation for each reference.

Thematic Area	Key Suggestions for Broader Investigation	References
Integration of AI and Radiomics in IGRT	Conduct further studies to confirm the impact of AI and radiomics on IGRT, emphasizing the need for evidence beyond retrospective data.	[38]
Advanced Contouring Technologies	Focus on constructing high-quality datasets for automated contouring technology using DL algorithms in head and neck OARs. Enhance DL performance through algorithm optimization and innovation.	[33]
Toxicity Prediction in Radiotherapy	Utilize large and diverse datasets for toxicity prediction, emphasizing the standardization of study methodologies to improve the consistency of research outcomes.	[34]
Automation of HNC Treatment Workflow	Align the development of AI technologies in HNC treatment with clinical needs by conducting interdisciplinary studies involving clinicians and computer scientists.	[35]
Application of CNN Methodology in ENT Radiology	Explore the potential uses of CNN methodology in ENT radiology, including nodule and tumor identification, anatomical variation identification, and tumor segmentation. Encourage continued evolution of technologies in everyday practice.	[32]
Image-to-Image Models and External Validation	Address the scarcity of external validation studies for image-to-image (I2I) models, emphasizing the need for publicly available pre-trained models to enhance the immediate applicability of proposed methods.	[37]
Automatic Segmentation in Cervical Cancer	Despite good accuracy in automatic segmentation of CT images for cervical cancer, future investigations should focus on obtaining public high-quality databases and conducting large-scale research verification.	[29]
Recommendations for Clinical Practice	Provide recommendations for clinicians and DL practitioners based on literature trends and the current state-of-the-art DL methods in radiation oncology.	[30]
Diagnostic Performance Using MRI Features	Recognize the good diagnostic performance of ML models using MRI features but suggest improvements in study quality and design for enhanced reliability.	[31]
DL-Based Synthetic CT Generation	Evaluate the clinical readiness of DL-based synthetic CT generation methods and suggest further initiatives for their potential implementation.	[24]
Techniques Using Augmented Data	Explore techniques using augmented data in clinical settings and build confidence in the validity of models produced.	[23]

Table 4. *Cont.*

Thematic Area	Key Suggestions for Broader Investigation	References
Future Studies for AI-Assisted MRI	Acknowledge the need for future studies with improved methodologies and larger training sets to enhance the diagnostic performance of AI-assisted MRI in clinical practice.	[27]
AI Applied to Imaging in Italy	Recognize the unprecedented interest in AI applied to imaging in Italy and suggest initiatives for building common frameworks, databases, collaborations, and guidelines for research on AI.	[28]
Standardized Radiomics Features and Deep Learning	Advocate for future research focusing on standardized radiomics features, robust feature selection, and deep learning techniques to improve prognostic capabilities in lung cancer models.	[22]
Development of AI in Radiation Oncology	Explore the development and basic concepts of AI in radiation oncology based on different task categories of DL algorithms. Clarify the potential for further DL development in the field.	[25]
Continuous Improvement of Models	Acknowledge the satisfactory results achieved by models but highlight the need for continuous improvement before safe and effective clinical practice.	[21–38]

4.4. Navigating Critical Focus Areas: Global Explorations and Expansion of Literature Analysis in Deep Learning Applications in Radiotherapy

From the emerging studies, critical areas that demand further research in this field came to light. These aspects primarily highlight the strong interconnection between radiotherapy and digital radiology, suggesting the need for deeper investigations considering this interdependence.

4.4.1. Navigating Critical Focus Areas in Deep Learning Applications in Radiotherapy

The extensive analysis conducted across a multitude of studies underscores with utmost gravity the critical imperatives that lie at the heart of advancing DL within the realm of radiotherapy [21–38]. These pivotal focal points, which include the paramount importance of standardization, the establishment of robust frameworks, unfettered access to diverse and comprehensive datasets, rigorous external validation protocols, and the adherence to methodological standardization, stand as the bedrock upon which the reliability, efficacy, and broader applicability of DL models in this intricate domain hinge.

First and foremost, the resounding call for "standardization" reverberates throughout the corpus of literature, as elucidated in [36]. The glaring absence of standardized protocols governing data collection methodologies and analysis procedures poses a formidable barrier, significantly impeding the validity and practical utility of DL technologies across radiotherapy applications [36]. This unequivocal realization underscores the urgent necessity for a cohesive, uniform approach that ensures coherence and comparability across diverse research endeavors and clinical implementations alike.

Furthermore, the fervent advocacy for collaborative "frameworks and guidelines" echoes resoundingly, particularly in the context of burgeoning interest witnessed in nations such as Italy [28,35]. The burgeoning landscape of AI applications within Italy's dynamic imaging research sphere underscores the indispensable need for collaborative frameworks, shared repositories of data, and meticulously crafted research guidelines aimed at harnessing the transformative potential of AI within the realm of medical imaging [28]. Similarly, concerted initiatives are urgently called on for the establishment of common frameworks, centralized repositories of data, and collaborative networks within Italy, all aimed at fostering the rapid advancement of AI research [35]. These compelling insights underscore the pivotal role played by collaborative endeavors and meticulously structured frameworks in driving forward the frontiers of DL application within the intricate domain of medical imaging and radiotherapy.

Moreover, the profound significance of unfettered access to "diverse and standardized datasets" emerges as a recurrent, overarching theme, particularly in the context of confronting the formidable challenges posed by limited data availability [23,28]. The acute shortage of expansive, heterogeneous datasets poses a formidable impediment to the widespread adoption of DL methodologies across various applications, including but not limited to prostate cancer treatment planning and radiology, thereby impeding the refinement and robustness of DL models [23]. This pressing need for diverse datasets is further underscored within Italy's dynamic imaging research landscape, where collaborative frameworks are identified as indispensable vehicles for harnessing the burgeoning interest witnessed in AI applications [28]. These profound insights serve to underscore the indispensable role played by comprehensive datasets in augmenting the efficacy, generalizability, and real-world applicability of DL models across the expansive spectrum of radiotherapy applications.

Equally crucial is the unequivocal recognition of "external validation" as an indispensable linchpin for ensuring the reliability, validity, and immediate applicability of DL models [27]. The dearth of comprehensive external validation studies, particularly in the context of image-to-image models, is acknowledged as a glaring deficiency, underscoring the imperative need for further validation endeavors aimed at bolstering the immediate applicability and real-world efficacy of proposed methodologies [27]. External validation thus emerges as a pivotal step in cementing the credibility, robustness, and generalizability of DL models, thereby instilling much-needed confidence in their real-world performance across diverse radiotherapy applications.

4.4.2. Harmonizing Challenges between Digital Radiology and Radiotherapy

The above considerations align with the broader trends seen in the integration of artificial intelligence (AI) into digital radiology [39–41], a phenomenon that can be said to encompass and permeate various sectors, including radiation therapy [41–46].

The integration of artificial intelligence (AI) into digital radiology, encompassing radiotherapy as an integral component, has evolved into a prevalent and transformative theme. Recent studies, exemplified by Giansanti and Di Basilio [39], accentuate the challenges, acceptance, and consensus associated with the incorporation of AI in both digital radiology and radiotherapy. Their work not only highlights the technological advancement but also emphasizes a paradigm shift that significantly influences medical imaging practices, presenting crucial challenges and areas that warrant in-depth studies and research.

Moreover, the comprehensive exploration of regulatory aspects pertaining to AI in digital radiology, which is equally applicable to radiotherapy, is evident in Giansanti's comprehensive review [40]. This review underscores substantial challenges and bottlenecks in the regulatory landscape, emphasizing the need for a thorough understanding of these aspects in both domains.

In the realm of radiation oncology, the impact of digital radiology has been acknowledged since the early 1990s [41]. Aznar et al. [42] provided a contemporary perspective on radiation oncology in the digital era, emphasizing its transformative nature and the escalating reliance on digital technologies within radiation oncology practices. The pivotal role of Digital Imaging and Communications in Medicine—Radiation Therapy (DICOM-RT) in radiology informatics was highlighted by Law and Liu [43]. Their discussion underscores the significance of DICOM-RT in radiation therapy, emphasizing the importance of standardized protocols for managing and sharing medical imaging data. The longstanding integration of digital radiology into radiotherapy, anchored in [44], has been a fundamental aspect of treatment planning and delivery. Van den Berge et al. [44] elaborated on the significance and challenges associated with digital radiology in radiotherapy, tracing its evolution over the years. This interconnection is sustained with the introduction of DL, where its advent significantly influences both digital radiology and radiation therapy [45]. Sahiner et al. [45] reinforced the interplay between digital radiology and radiotherapy, emphasizing the potential for DL to revolutionize image analysis and treatment planning.

The utilization of digital radiology in the planning and delivery of radiation therapy remains a subject of continued interest [46]. Kalet and Austin-Seymour [46] delved into the intricate role of digital radiology in radiation therapy, emphasizing the critical importance of accurate imaging data for effective treatment planning and delivery.

4.4.3. Insights into Key Areas: A Comprehensive Exploration

The need for standardization, frameworks and guidelines, access to diverse and standardized datasets, and the importance of external validation have been specifically mentioned in some studies but are a vaguely perceived necessity in all the analyzed studies. All of this is in line with the common sentiment in digital radiology, upon which, as previously seen, DL applied to RT builds its foundation.

Focus on Ethics

The issues related to standardization and the need for guidelines and shared frameworks are strongly highlighted, for example, with emphasis on ethics in [40,47–50].

Mudgal et al. [47] conducted an in-depth study focusing exclusively on ethical considerations surrounding AI deployment. Their review covered ethical issues related to harmless and justifiable AI integration across training, incorporation into the health sector, and regulatory frameworks. Emphasizing the need for centralized and non-dispersive data management, the authors advocated for thorough evaluation, refinement, purification, and representation of all demographics. Transparency and security were highlighted as pivotal, and the authors stressed the importance of following authorization processes set by competent authorities.

In a related vein, Harvey et al. [48] explored the transformative potential of AI in digital radiology, addressing nascent phases and untested aspects within clinical spaces. The study outlined challenges, including the evolving legal–regulatory environment, impacting AI introduction. Issues such as FDA approval pathways, government oversight, privacy concerns, ethical dilemmas, and practical considerations in radiologist practice were discussed. Notably, the study underscored the need for careful considerations in nuclear medicine, encompassing reliability, safety, non-maleficence, beneficence, justice, fairness, data privacy, security, confidentiality, bias minimization, clear communication, autonomy, and clarification ability.

Jaremko et al. [49] contributed a study for the Canadian Association of Radiologists, focusing on both regulatory and ethical facets. Their work presented a comprehensive framework outlining legal and ethical issues, with specific reference to the Canadian health domain.

Currie et al. [50] centered their review specifically on nuclear medicine. They highlighted the substantial opportunities associated with "ethical AI," emphasizing its potential to enhance productivity, workflow, research capabilities, and clinical applications.

Focus on Regulatory Frameworks

The issues related to standardization and the need for guidelines and shared frameworks are strongly emphasized, along with the necessity for standardizing datasets, for instance, with a focus on regulatory frameworks in [39,48,49,51–55].

Two studies focused on regulatory frameworks in both the European Union and the US. Pesapane et al. [53] analyzed medical device (MD) regulations in both regions, noting the European emphasis on cybersecurity, data protection, and MDs' integration, while the US FDA prioritized data processing, consent, and user/consumer consent. Muehlematter [52] expressed concerns about the approval processes for AI-based MDs in both regions, highlighting an increase in approved MDs but suggesting a lack of a well-defined pathway. Transparency and a dedicated database for MDs were recommended.

Harvey et al. [48] examined the US context, addressing the FDA's innovative approach to streamline AI approval by adopting a total product lifecycle method. Jaremko et al. [49]

provided a study on Canada, discussing regulatory and ethical recommendations for patient data, algorithms, and healthcare practices.

The concept of AI in digital radiology being treated as medical devices (MDs) was emphasized in [52,53]. Arora et al. [54] highlighted AI's recognition as "Software as a Medical Device (SaMD)" and its interconnection with the Internet of Things, genetic data, and patient records. Allen et al. [55] warned about biases in commercial algorithms affecting gender, ethnicity, and social factors in healthcare. They proposed a solution involving collaboration among institutions to develop robust shared datasets and algorithms with standards to mitigate biases.

Focus on Bottlenecks

On these topics, including the need for evaluation of the retrieval of validated external datasets, we found the identification of bottlenecks in [39,40,56–59]. Three studies on regulatory aspects [56–58] provided complementary and comprehensive insights, emphasizing the critical need for external datasets. Alexander et al. [57] examined the impact of workload on decision correctness in radiology, underscoring the importance of regulating workloads with a scientifically sound approach. They cautioned that inadequate regulation could pose more risks than having no regulation at all, especially when considering the influence of AI on decision speediness.

Mezrich [56] delved into legal liability associated with errors linked to AI use. Highlighting critical issues, the review focused on enforcing AI-based product liability laws in the US, uncovering ambiguities that could significantly impact the integration of AI into the health domain and erode trust among stakeholders.

Ebrahimian et al. [58] concentrated on FDA-regulated AI algorithms, reviewing 127 regulated software to classify available information. They observed a growing number of FDA-regulated medical devices (MDs) from 2008 to 2021. Crucially, their review emphasized a lack of sufficient public data on validation/testing datasets for various algorithms, rendering applications in healthcare unjustifiable due to potential generalization and biases. The consensus across these reviews underscores the urgent need for transparent and comprehensive external datasets to ensure the safe and effective integration of AI in healthcare.

Focus on Consensus and Acceptance

An alternative viewpoint on these matters is discussed in [39,59–72], emphasizing the significance of consensus-building and acceptance initiatives. These initiatives are not solely targeted at professionals involved in the field but extend to patients, caregivers, and patient associations. The recognition of the importance of involving a broader spectrum of stakeholders underscores a holistic approach to address and incorporate diverse perspectives, fostering a more inclusive and well-rounded understanding of the challenges and advancements in the discussed topics.

Three studies [59–61] focused on patient questionnaires, exploring various aspects of artificial intelligence (AI) integration. Lennartz et al. [59] investigated opinions on AI integration throughout the medical workflow, while Zhang et al. [60] conducted interviews on AI in diagnostics, revealing positive considerations with some cybersecurity concerns. Ongega et al. [61] examined perceived perspectives on AI, emphasizing patients' importance, social factors' impact, and the utility of questionnaires as sensors.

Hendrix et al. [62] concentrated on primary care providers, highlighting the importance of sensitivity and other parameters in AI-supported clinical reports. AI was deemed suitable for triage roles without radiologist validation.

Other studies [63–68] explored the opinions of other stakeholders, including radiography students. Abuzaid et al. [63] found enthusiasm for AI integration in training programs but concerns about job security. Abuzaid et al. [64] in another study focused on magnetic resonance imaging applications, with participants acknowledging AI's potential to improve workflow. Giansanti et al. [65] gathered opinions on post-pandemic AI use, emphasizing

the need for a structured questionnaire for scientific societies. Abuzaid et al. [66] in a third article investigated opinions on AI integration into radiology workflow, revealing low awareness and the importance of tailored training.

Alelyani et al. [67] explored attitudes toward AI among radiologists, radiographers, technologists, and students, noting varying levels of awareness and the importance of specific AI training in medical schools. The European Society of Radiology [68] extended the investigation internationally, reporting favorable positions on AI and detailed expectations for 5–10 years.

Caparros Galan et al. [69] focused on student opinions, finding a belief in AI's potential to reform the radiology workflow without jeopardizing radiologists' work. Di Basilio et al.'s study [70] involved professionals with diverse backgrounds and highlighted the importance of survey administration procedures.

Two surveys sponsored by different scientific societies were reported in [71,72]. Diaz et al. [71] surveyed international medical physicists, exploring training aspects and opinions on AI introduction. Coppola et al. [72] conducted a nationwide survey on Italian radiologists, examining their interactions with AI, ethical concerns, job loss risks, policy needs, and opinions.

Further Areas of Focus with the Point of View of the Challenges

Further domains exploring these themes as challenges encompass algorithms [73], examining the pivotal role of the responsible professional [74], delving into tools, datasets, and workflows [75], understanding the dynamics of teamwork [76], and exploring educational aspects [77]. Each of these areas merits a dedicated review [39] for a comprehensive understanding of the challenges posed by these critical components in the integration of artificial intelligence in healthcare. The challenges in algorithm development, also with a focus on the need for standardization, were explored by Fazal et al. [73]. The study highlighted macro-areas of challenges, including the need to decrease false-positive rates for computer-aided detection and to enhance the understanding of AI reasoning. Moreover, the study emphasized, from a standardization viewpoint, the importance of well-defined responsibilities in case of algorithm-induced errors. Gampala et al. [74] delved into challenges related to the professional role and responsibilities in digital radiology in the context of AI integration. The study discussed how AI could simplify various activities in the working chain, emphasizing the need for clinical teams to become familiar with guiding principles for data management and curation and to standardize it.

Ahmad [75] focused on challenges in engineering and machinery aspects, such as hardware, software, and workflow impact. The study reported on new activities, such as selecting AI products and vendors, piloting AI algorithms, and creating proprietary AI algorithms that need standardization.

Martin-Noguerol et al. [76] highlighted challenges in teamwork concerning the collaboration between engineers, system developers, and radiologists. Communication between radiologists and data scientists was deemed crucial for successful collaborative work and needs to be regulated.

Pesapane et al. faced the importance of education [77] and underscored the significance of continuous, specialized training, considering the evolving nature of AI and the need for well-defined and standardized programs.

4.5. Comprehensive Synoptic Overview

This extensive investigation, conducted via an umbrella review using a narrative review methodology, has allowed for the illumination of intricate themes regarding the application and integration of DL techniques within RT. This comprehensive exploration (Figure 6A) extends beyond mere thematic revelations (Figure 6B,C), encapsulating trends (Figure 6D), nuanced emerging opportunities (Figure 6E), needs for broader investigation areas (Figure 6F), and key focus areas for global explorations in DL for RT suggested from the overview (Figure 6G). The last passage suggested a further elaboration of the overview

with complementary information. The complementation with a non-systematic review of recent contributions has made it possible to perform an evaluation and comparison by linking the technological challenges faced in the reality of digital radiology and AI, strictly interconnected with the investigated field of the integration of DL in RT (Figure 6H), and a comprehensive exploration of the key areas identified has been provided (Figure 6J). This exploration has facilitated a series of focused discussions on the detected key focus areas for global exploration: ethical considerations (Figure 6I), regulatory frameworks (Figure 6K), identification of bottlenecks (Figure 6L), the cultivation of consensus and acceptance within the field (Figure 6M), and the focus with the point of view of the challenges (Figure 6N).

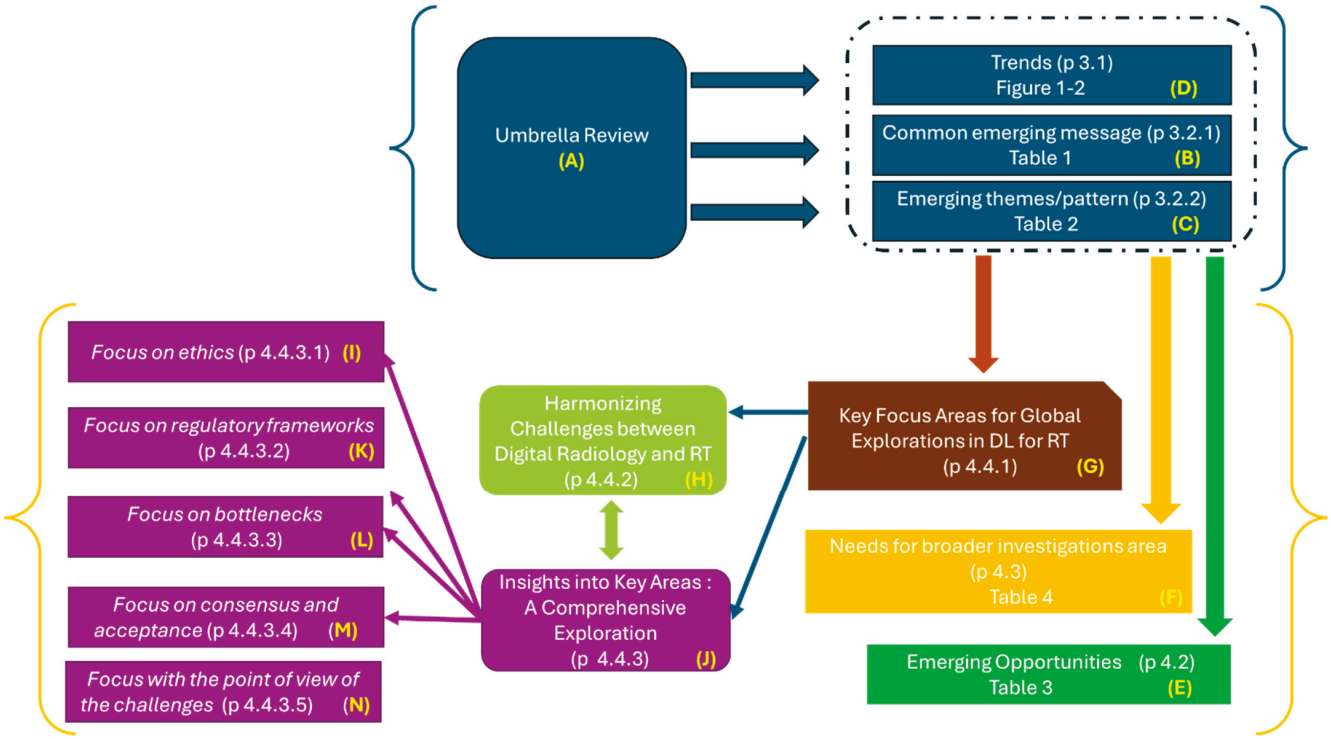


Figure 6. Synoptic diagram representing the flow and the key points of the study.

As emerged from the synoptic diagram, the study of recent literature has allowed for a detailed exploration of the key areas identified by the systematic literature reviews initially analyzed in this umbrella review.

To summarize, the comprehensive overview, combined with the additional analytical discussion, successfully addressed the critical issues arising from the original hypothesis. However, it is important to emphasize that although a narrative review does not inherently aim for such outcomes, the in-depth nature of this investigation greatly enhances the understanding and development of the topic under discussion.

4.6. Takeway Message

Significant initiatives have involved the integration of DL with radiotherapy. These initiatives experienced a notable acceleration during the COVID-19 pandemic. This period highlighted significant development opportunities alongside specific niches needing improvements and further scientific exploration. Furthermore, there is a widespread call for in-depth examination regarding standardization, guidelines, access to dataset databases, and the corresponding external validations.

4.7. Limitations

This study utilized the PubMed and Scopus databases, focusing on systematic review studies through an umbrella review narrative. Further targeted research on emerging

themes is encouraged. Exploring national databases can also enhance knowledge in this domain by identifying local initiatives, with particular emphasis on standardization and regulatory compliance.

5. Conclusions

In conclusion, the synthesis of the findings shed light on the profound potential of DL in revolutionizing radiotherapy. The acceleration driven in this field (as in others) by the COVID-19 pandemic has led to significant advancements that indeed have the potential to offer great opportunities in the healthcare domain. However, attention must be paid to specific areas that require further research.

The overview not only highlighted key areas poised for further exploration but also underscored the nuanced interplay between digital radiology and radiation therapy. This interdependence aids in addressing cross-cutting areas that require further developments in the realms of both research and regulation.

As technological advancements continue to progress, concerted collaborative efforts and thorough investigations are essential to fully harness the transformative power of DL. Ultimately, such endeavors hold the promise of significantly improving patient care and refining treatment strategies for optimal outcomes.

Funding: This research received no external fundings. The APC funded by Giovanni Maccioni.

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

References

1. Available online: https://www.cancer.gov/about-cancer/treatment/types/radiation-therapy (accessed on 20 April 2024).
2. Available online: https://www.nhs.uk/conditions/radiotherapy/ (accessed on 20 April 2024).
3. Available online: https://www.epa.gov/radtown/radiation-therapy (accessed on 20 April 2024).
4. Fraass, B.A. The development of conformal radiation therapy. *Med. Phys.* **1995**, *22 Pt 2*, 1911–1921. [CrossRef] [PubMed]
5. Taylor, A.; Powell, M.E. Intensity-modulated radiotherapy—What is it? *Cancer Imaging* **2004**, *4*, 68–73. [CrossRef] [PubMed] [PubMed Central]
6. Guckenberger, M.; Baus, W.W.; Blanck, O.; Combs, S.E.; Debus, J.; Engenhart-Cabillic, R.; Gauer, T.; Grosu, A.L.; Schmitt, D.; Tanadini-Lang, S.; et al. Definition and quality requirements for stereotactic radiotherapy: Consensus statement from the DEGRO/DGMP Working Group Stereotactic Radiotherapy and Radiosurgery. *Strahlenther Onkol.* **2020**, *196*, 417–420. [CrossRef] [PubMed] [PubMed Central]
7. Amaldi, U. Hadrontherapy in the world and the programmes of the TERA Foundation. *Tumori* **1998**, *84*, 188–199. [CrossRef] [PubMed]
8. Geng, H.; Giaddui, T.; Cheng, C.; Zhong, H.; Ryu, S.; Liao, Z.; Yin, F.F.; Gillin, M.; Mohan, R.; Xiao, Y. A comparison of two methodologies for radiotherapy treatment plan optimization and QA for clinical trials. *J. Appl. Clin. Med. Phys.* **2021**, *22*, 329–337. [CrossRef] [PubMed] [PubMed Central]
9. Arslan, A.; Sengul, B. Comparison of radiotherapy techniques with flattening filter and flattening filter-free in lung radiotherapy according to the treatment volume size. *Sci. Rep.* **2020**, *10*, 8983. [CrossRef] [PubMed] [PubMed Central]
10. Chaikh, A.; Giraud, J.Y.; Perrin, E.; Bresciani, J.P.; Balosso, J. The choice of statistical methods for comparisons of dosimetric data in radiotherapy. *Radiat. Oncol.* **2014**, *9*, 205. [CrossRef] [PubMed] [PubMed Central]
11. Unkelbach, J.; Bortfeld, T.; Cardenas, C.E.; Gregoire, V.; Hager, W.; Heijmen, B.; Jeraj, R.; Korreman, S.S.; Ludwig, R.; Pouymayou, B.; et al. The role of computational methods for automating and improving clinical target volume definition. *Radiother. Oncol.* **2020**, *153*, 15–25. [CrossRef] [PubMed]
12. Lin, H.; Xiao, H.; Dong, L.; Teo, K.B.; Zou, W.; Cai, J.; Li, T. Deep learning for automatic target volume segmentation in radiation therapy: A review. *Quant. Imaging Med. Surg.* **2021**, *11*, 4847–4858. [CrossRef] [PubMed] [PubMed Central]
13. Isaksson, L.J.; Summers, P.; Mastroleo, F.; Marvaso, G.; Corrao, G.; Vincini, M.G.; Zaffaroni, M.; Ceci, F.; Petralia, G.; Orecchia, R.; et al. Automatic Segmentation with Deep Learning in Radiotherapy. *Cancers* **2023**, *15*, 4389. [CrossRef] [PubMed] [PubMed Central]
14. Niraula, D.; Cui, S.; Pakela, J.; Wei, L.; Luo, Y.; Ten Haken, R.K.; El Naqa, I. Current status and future developments in predicting outcomes in radiation oncology. *Br. J. Radiol.* **2022**, *95*, 20220239. [CrossRef] [PubMed] [PubMed Central]
15. Wahid, K.A.; Glerean, E.; Sahlsten, J.; Jaskari, J.; Kaski, K.; Naser, M.A.; He, R.; Mohamed, A.S.R.; Fuller, C.D. Artificial Intelligence for Radiation Oncology Applications Using Public Datasets. *Semin. Radiat. Oncol.* **2022**, *32*, 400–414. [CrossRef] [PubMed] [PubMed Central]
16. Santoro, M.; Strolin, S.; Paolani, G.; Della Gala, G.; Bartoloni, A.; Giacometti, C.; Ammendolia, I.; Morganti, A.G.; Strigari, L. Recent Applications of Artificial Intelligence in Radiotherapy: Where We Are and Beyond. *Appl. Sci.* **2022**, *12*, 3223. [CrossRef]

17. Choi, G.J.; Kang, H. The umbrella review: A useful strategy in the rain of evidence. *Korean J. Pain* **2022**, *35*, 127–128. [CrossRef]
18. Which Review Is That? A Guide to Review Types. Available online: https://unimelb.libguides.com/whichreview/umbrellareview (accessed on 25 February 2024).
19. ANDJ Checklist. Available online: https://www.elsevier.com/__data/promis_misc/ANDJ%20Narrative%20Review%20Checklist.pdf (accessed on 3 June 2023).
20. Giansanti, D. An Umbrella Review of the Fusion of fMRI and AI in Autism. *Diagnostics* **2023**, *13*, 3552. [CrossRef] [PubMed]
21. Almeida, G.; Tavares, J.M.R.S. Deep Learning in Radiation Oncology Treatment Planning for Prostate Cancer: A Systematic Review. *J. Med. Syst.* **2020**, *44*, 179. [CrossRef] [PubMed]
22. Kothari, G.; Korte, J.; Lehrer, E.J.; Zaorsky, N.G.; Lazarakis, S.; Kron, T.; Hardcastle, N.; Siva, S. A systematic review and meta-analysis of the prognostic value of radiomics based models in non-small cell lung cancer treated with curative radiotherapy. *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.* **2021**, *155*, 188–203. [CrossRef] [PubMed]
23. Chlap, P.; Min, H.; Vandenberg, N.; Dowling, J.; Holloway, L.; Haworth, A. A review of medical image data augmentation techniques for deep learning applications. *J. Med. Imaging Radiat. Oncol.* **2021**, *65*, 545–563. [CrossRef] [PubMed]
24. Spadea, M.F.; Maspero, M.; Zaffino, P.; Seco, J. Deep learning based synthetic-CT generation in radiotherapy and PET: A review. *Med. Phys.* **2021**, *48*, 6537–6566. [CrossRef] [PubMed]
25. Huang, D.; Bai, H.; Wang, L.; Hou, Y.; Li, L.; Xia, Y.; Yan, Z.; Chen, W.; Chang, L.; Li, W. The Application and Development of Deep Learning in Radiotherapy: A Systematic Review. *Technol. Cancer Res. Treat.* **2021**, *20*, 15330338211016386. [CrossRef] [PubMed]
26. Walls, G.M.; Osman, S.O.S.; Brown, K.H.; Butterworth, K.T.; Hanna, G.G.; Hounsell, A.R.; McGarry, C.K.; Leijenaar, R.T.H.; Lambin, P.; Cole, A.J.; et al. Radiomics for Predicting Lung Cancer Outcomes Following Radiotherapy: A Systematic Review. *Clin. Oncol.* **2022**, *34*, e107–e122. [CrossRef] [PubMed]
27. Kim, H.Y.; Cho, S.J.; Sunwoo, L.; Baik, S.H.; Bae, Y.J.; Choi, B.S.; Jung, C.; Kim, J.H. Classification of true progression after radiotherapy of brain metastasis on MRI using artificial intelligence: A systematic review and meta-analysis. *Neuro-Oncol. Adv.* **2021**, *3*, vdab080. [CrossRef] [PubMed]
28. Avanzo, M.; Porzio, M.; Lorenzon, L.; Milan, L.; Sghedoni, R.; Russo, G.; Massafra, R.; Fanizzi, A.; Barucci, A.; Ardu, V.; et al. Artificial intelligence applications in medical imaging: A review of the medical physics research in Italy. *Phys. Medica* **2021**, *83*, 221–241. [CrossRef] [PubMed]
29. 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]
30. Rusanov, B.; Hassan, G.M.; Reynolds, M.; Sabet, M.; Kendrick, J.; Rowshanfarzad, P.; Ebert, M. Deep learning methods for enhancing cone-beam CT image quality toward adaptive radiation therapy: A systematic review. *Med. Phys.* **2022**, *49*, 6019–6054. [CrossRef]
31. Booth, T.C.; Grzeda, M.; Chelliah, A.; Roman, A.; Al Busaidi, A.; Dragos, C.; Shuaib, H.; Luis, A.; Mirchandani, A.; Alparslan, B.; et al. Imaging Biomarkers of Glioblastoma Treatment Response: A Systematic Review and Meta-Analysis of Recent Machine Learning Studies. *Front. Oncol.* **2022**, *12*, 799662. [CrossRef] [PubMed]
32. Hasan, Z.; Key, S.; Habib, A.R.; Wong, E.; Aweidah, L.; Kumar, A.; Sacks, R.; Singh, N. Convolutional Neural Networks in ENT Radiology: Systematic Review of the Literature. *Ann. Otol. Rhinol. Laryngol.* **2023**, *132*, 417–430. [CrossRef] [PubMed]
33. Liu, P.; Sun, Y.; Zhao, X.; Yan, Y. Deep learning algorithm performance in contouring head and neck organs at risk: A systematic review and single-arm meta-analysis. *Biomed. Eng. Online* **2023**, *22*, 104. [CrossRef]
34. Tan, D.; Mohd Nasir, N.F.; Abdul Manan, H.; Yahya, N. Prediction of toxicity outcomes following radiotherapy using deep learning-based models: A systematic review. *Cancer Radiother. J. Soc. Fr. Radiother. Oncol.* **2023**, *27*, 398–406. [CrossRef] [PubMed]
35. Franzese, C.; Dei, D.; Lambri, N.; Teriaca, M.A.; Badalamenti, M.; Crespi, L.; Tomatis, S.; Loiacono, D.; Mancosu, P.; Scorsetti, M. Enhancing Radiotherapy Workflow for Head and Neck Cancer with Artificial Intelligence: A Systematic Review. *J. Pers. Med.* **2023**, *13*, 946. [CrossRef] [PubMed]
36. Eidex, Z.; Ding, Y.; Wang, J.; Abouei, E.; Qiu RL, J.; Liu, T.; Wang, T.; Yang, X. Deep learning in MRI-guided radiation therapy: A systematic review. *J. Appl. Clin. Med. Phys.* **2024**, *25*, e14155. [CrossRef] [PubMed]
37. Chen, J.; Chen, S.; Wee, L.; Dekker, A.; Bermejo, I. Deep learning based unpaired image-to-image translation applications for medical physics: A systematic review. *Phys. Med. Biol.* **2023**, *68*, 05TR01. [CrossRef] [PubMed]
38. Boldrini, L.; D'Aviero, A.; De Felice, F.; Desideri, I.; Grassi, R.; Greco, C.; Iorio, G.C.; Nardone, V.; Piras, A.; Salvestrini, V. Artificial intelligence applied to image-guided radiation therapy (IGRT): A systematic review by the Young Group of the Italian Association of Radiotherapy and Clinical Oncology (yAIRO). *Radiol. Medica* **2024**, *129*, 133–151. [CrossRef] [PubMed]
39. Giansanti, D.; Di Basilio, F. The Artificial Intelligence in Digital Radiology: Part 1: The Challenges, Acceptance and Consensus. *Healthcare* **2022**, *10*, 509. [CrossRef] [PubMed] [PubMed Central]
40. Giansanti, D. The Regulation of Artificial Intelligence in Digital Radiology in the Scientific Literature: A Narrative Review of Reviews. *Healthcare* **2022**, *10*, 1824. [CrossRef]
41. Stafford, P.M.; Martin, E.E.; Chu, J.C.; Davidson, B.A.; Hanks, G.E. Digital imaging in the radiation oncology environment: A personal computer local area network solution. *J. Digit. Imaging* **1991**, *4*, 177–184. [CrossRef] [PubMed]

42. Aznar, M.C.; Bacchus, C.; Coppes, R.P.; Deutsch, E.; Georg, D.; Haustermans, K.; Hoskin, P.; Krause, M.; Lartigau, E.F.; Löck, S.; et al. Radiation oncology in the new virtual and digital era. *Radiother. Oncol.* **2021**, *154*, A1–A4. [CrossRef] [PubMed] [PubMed Central]
43. Law, M.Y.; Liu, B. Informatics in radiology: DICOM-RT and its utilization in radiation therapy. *Radiographics* **2009**, *29*, 655–667. [CrossRef] [PubMed]
44. Van den Berge, D.L.; De Ridder, M.; Storme, G.A. Imaging in radiotherapy. *Eur. J. Radiol.* **2000**, *36*, 41–48. [CrossRef] [PubMed]
45. Sahiner, B.; Pezeshk, A.; Hadjiiski, L.M.; Wang, X.; Drukker, K.; Cha, K.H.; Summers, R.M.; Giger, M.L. Deep learning in medical imaging and radiation therapy. *Med. Phys.* **2019**, *46*, e1–e36. [CrossRef] [PubMed] [PubMed Central]
46. Kalet, I.J.; Austin-Seymour, M.M. The use of medical images in planning and delivery of radiation therapy. *J. Am. Med. Inform. Assoc.* **1997**, *4*, 327–339. [CrossRef] [PubMed] [PubMed Central]
47. Mudgal, K.S.; Das, N. The ethical adoption of artificial intelligence in radiology. *BJR Open* **2020**, *2*, 20190020. [CrossRef] [PubMed]
48. Harvey, H.B.; Gowda, V. Regulatory Issues and Challenges to Artificial Intelligence Adoption. *Radiol. Clin. N. Am.* **2021**, *59*, 1075–1083. [CrossRef]
49. Jaremko, J.L.; Azar, M.; Bromwich, R.; Lum, A.; Alicia Cheong, L.H.; Gibert, M.; Laviolette, F.; Gray, B.; Reinhold, C.; Cicero, M.; et al. Canadian Association of Radiologists White Paper on Ethical and Legal Issues Related to Artificial Intelligence in Radiology. *Can. Assoc. Radiol. J.* **2019**, *70*, 107–118. [PubMed]
50. Currie, G.; Hawk, K.E. Ethical and Legal Challenges of Artificial Intelligence in Nuclear Medicine. *Semin. Nucl. Med.* **2021**, *51*, 120–125. [CrossRef] [PubMed]
51. Lee, E.E.; Torous, J.; De Choudhury, M.; Depp, C.A.; Graham, S.A.; Kim, H.C.; Paulus, M.P.; Krystal, J.H.; Jeste, D.V. Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **2021**, *6*, 856–864. [CrossRef] [PubMed]
52. Muehlematter, U.J.; Daniore, P.; Vokinger, K.N. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): A comparative analysis. *Lancet Digit. Health* **2021**, *3*, e195–e203. [CrossRef] [PubMed]
53. Pesapane, F.; Volonté, C.; Codari, M.; Sardanelli, F. Artificial intelligence as a medical device in radiology: Ethical and regulatory issues in Europe and the United States. *Insights Imaging* **2018**, *9*, 745–753. [CrossRef] [PubMed]
54. Arora, A. Conceptualising Artificial Intelligence as a Digital Healthcare Innovation: An Introductory Review. *Med. Devices* **2020**, *13*, 223–230. [CrossRef] [PubMed]
55. Allen, B.; Dreyer, K. The Role of the ACR Data Science Institute in Advancing Health Equity in Radiology. *J. Am. Coll. Radiol.* **2019**, *16*, 644–648. [CrossRef]
56. Mezrich, J.L. Is Artificial Intelligence (AI) a Pipe Dream? Why Legal Issues Present Significant Hurdles to AI Autonomy. *AJR Am. J. Roentgenol.* **2022**, *219*, 152–156. [CrossRef] [PubMed]
57. Alexander, R.; Waite, S.; Bruno, M.A.; Krupinski, E.A.; Berlin, L.; Macknik, S.; Martinez-Conde, S. Mandating Limits on Workload, Duty, and Speed in Radiology. *Radiology* **2022**, *304*, 274–282. [CrossRef]
58. Ebrahimian, S.; Kalra, M.K.; Agarwal, S.; Bizzo, B.C.; Elkholy, M.; Wald, C.; Allen, B.; Dreyer, K.J. FDA-regulated AI Algorithms:Trends, Strengths, and Gaps of Validation Studies. *Acad. Radiol.* **2022**, *29*, 559–566. [CrossRef] [PubMed]
59. Lennartz, S.; Dratsch, T.; Zopfs, D.; Persigehl, T.; Maintz, D.; Hokamp, N.G.; dos Santos, D.P. Use and Control of Artificial Intelligence in Patients Across the Medical Workflow: Single-Center Questionnaire Study of Patient Perspectives. *J. Med. Int. Res.* **2021**, *23*, e24221. [CrossRef] [PubMed]
60. Zhang, Z.; Citardi, D.; Wang, D.; Genc, Y.; Shan, J.; Fan, X. Patients' perceptions of using artificial intelligence (AI)-based technology to comprehend radiology imaging data. *Health Inform. J.* **2021**, *27*, 14604582211011215. [CrossRef] [PubMed]
61. Ongena, Y.P.; Haan, M.; Yakar, D.; Kwee, T.C. Patients' views on the implementation of artificial intelligence in radiology: Development and validation of a standardized questionnaire. *Eur. Radiol.* **2020**, *30*, 1033–1040. [CrossRef] [PubMed]
62. Hendrix, N.; Hauber, B.; Lee, C.I.; Bansal, A.; Veenstra, D.L. Artificial intelligence in breast cancer screening: Primary care provider preferences. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 1117–1124. [CrossRef] [PubMed]
63. Abuzaid, M.M.; Elshami, W.; McConnell, J.; Tekin, H.O. An extensive survey of radiographers from the Middle East and India on artificial intelligence integration in radiology practice. *Health Technol.* **2021**, *11*, 1045–1050. [CrossRef] [PubMed]
64. Abuzaid, M.; Tekin, H.; Reza, M.; Elhag, I.; Elshami, W. Assessment of MRI technologists in acceptance and willingness to integrate artificial intelligence into practice. *Radiography* **2021**, *27*, S83–S87. [CrossRef] [PubMed]
65. Giansanti, D.; Rossi, I.; Monoscalco, L. Lessons from the COVID-19 Pandemic on the Use of Artificial Intelligence in Digital Radiology: The Submission of a Survey to Investigate the Opinion of Insiders. *Healthcare* **2021**, *9*, 331. [CrossRef] [PubMed]
66. Abuzaid, M.M.; Elshami, W.; Tekin, H.; Issa, B. Assessment of theWillingness of Radiologists and Radiographers to Accept the Integration of Artificial Intelligence into Radiology Practice. *Acad. Radiol.* **2020**, *29*, 87–94. [CrossRef]
67. Alelyani, M.; Alamri, S.; Alqahtani, M.S.; Musa, A.; Almater, H.; Alqahtani, N.; Alshahrani, F.; Alelyani, S. Radiology Community Attitude in Saudi Arabia about the Applications of Artificial Intelligence in Radiology. *Healthcare* **2021**, *9*, 834. [CrossRef] [PubMed]
68. European Society of Radiology (ESR). Impact of artificial intelligence on radiology: A EuroAIM survey among members of the European Society of Radiology. *Insights Imaging* **2019**, *10*, 105. [CrossRef] [PubMed]
69. Galán, G.C.; Portero, F.S. Percepciones de estudiantes de Medicina sobre el impacto de la inteligencia artificial en radiología. *Radiología*, 2021; *in press*.

70. Di Basilio, F.; Esposisto, G.; Monoscalco, L.; Giansanti, D. The Artificial Intelligence in Digital Radiology: Part 2: Towards an Investigation of acceptance and consensus on the Insiders. *Healthcare* **2022**, *10*, 153. [CrossRef] [PubMed]
71. Diaz, O.; Guidi, G.; Ivashchenko, O.; Colgan, N.; Zanca, F. Artificial intelligence in themedical physics community: An international survey. *Phys. Med.* **2021**, *81*, 141–146. [CrossRef] [PubMed]
72. Coppola, F.; Faggioni, L.; Regge, D.; Giovagnoni, A.; Golfieri, R.; Bibbolino, C.; Miele, V.; Neri, E.; Grassi, R. Artificial intelligence: Radiologists' expectations and opinions gleaned from a nationwide online survey. *Radiol. Med.* **2021**, *126*, 63–71. [CrossRef] [PubMed]
73. Fazal, M.I.; Patel, M.E.; Tye, J.; Gupta, Y. The past, present and future role of artificial intelligence in imaging. *Eur. J. Radiol.* **2018**, *105*, 246–250. [CrossRef]
74. Gampala, S.; Vankeshwaram, V.; Gadula, S.S.P. Is Artificial Intelligence the New Friend for Radiologists? A Review Article. *Cureus* **2020**, *12*, e11137. [CrossRef] [PubMed]
75. Ahmad, R. Reviewing the relationship between machines and radiology: The application of artificial intelligence. *Acta Radiol. Open* **2021**, *10*, 2058460121990296. [CrossRef] [PubMed]
76. Martín-Noguerol, T.; Paulano-Godino, F.; López-Ortega, R.; Górriz, J.M.; Riascos, R.F.; Luna, A. Artificial intelligence in radiology: Relevance of collaborative work between radiologists and engineers for building a multidisciplinary team. *Clin. Radiol.* **2021**, *76*, 317–324. [CrossRef] [PubMed]
77. Pesapane, F. How scientific mobility can help current and future radiology research: A radiology trainee's perspective. *Insights Imaging* **2019**, *10*, 85. [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

Systematic Review

A Systematic Review: The Role of Artificial Intelligence in Lung Cancer Screening in Detecting Lung Nodules on Chest X-Rays

Puteri Norliza Megat Ramli [1,2], Azimatun Noor Aizuddin [2,*], Norfazilah Ahmad [2], Zuhanis Abdul Hamid [1] and Khairil Idham Ismail [3]

[1] Institut Kanser Negara, Ministry of Health, Putrajaya 62250, Malaysia; p131318@siswa.ukm.edu.my (P.N.M.R.); drzuhanis@nci.gov.my (Z.A.H.)
[2] Department of Public Health Medicine, Faculty of Medicine, Universiti Kebangsaan Malaysia, Cheras 56000, Wilayah Persekutuan Kuala Lumpur, Malaysia; norfazilah@hctm.ukm.edu.my
[3] Malaysian Health Technology Assessment Section (MaHTAS), Medical Development Division, Ministry of Health, Putrajaya 62590, Malaysia; khairilidham@moh.gov.my
* Correspondence: azimatunnoor@hctm.ukm.edu.my

Abstract: Background: Lung cancer remains one of the leading causes of cancer-related deaths worldwide. Artificial intelligence (AI) holds significant potential roles in enhancing the detection of lung nodules through chest X-ray (CXR), enabling earlier diagnosis and improved outcomes. **Methods**: Papers were identified through a comprehensive search of the Web of Science (WOS), Scopus, and Ovid Medline databases for publications dated between 2020 and 2024. Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, 34 studies that met the inclusion criteria were selected for quality assessment and data extraction. **Results**: AI demonstrated sensitivity rates of 56.4–95.7% and specificities of 71.9–97.5%, with the area under the receiver operating characteristic (AUROC) values between 0.89 and 0.99, compared to radiologists' mean area under the curve (AUC) of 0.81. AI performed better with larger nodules (>2 cm) and solid nodules, showing higher AUC values for calcified (0.71) compared to non-calcified nodules (0.55). Performance was lower in hilar areas (30%) and lower lung fields (43.8%). A combined AI-radiologist approach improved overall detection rates, particularly benefiting less experienced readers; however, AI showed limitations in detecting ground-glass opacities (GGOs). **Conclusions**: AI shows promise as a supplementary tool for radiologists in lung nodule detection. However, the variability in AI results across studies highlights the need for standardized assessment methods and diverse datasets for model training. Future studies should focus on developing more precise and applicable algorithms while evaluating the effectiveness and cost-efficiency of AI in lung cancer screening interventions.

Keywords: artificial intelligence; chest X-ray; lung nodule; lung cancer; radiologist

1. Introduction

1.1. Background

Lung cancer remains a significant global public health issue, with substantial variation in incidence and mortality rates across different regions. While the American Cancer Society has reported a decrease in the cancer burden in the United States, China is experiencing a shift in its cancer burden epidemiology, increasingly resembling that of the United States [1]. It is estimated that approximately one-quarter of all cancers could be prevented if risk factors such as tobacco smoking and obesity were effectively controlled [2]. However, early

diagnosis of the disease is crucial, as about 93% of patients are diagnosed with stage III and IV lung cancer, with an average survival time of just 18 weeks after diagnosis [3].

The National Lung Screening Trial (NLST) demonstrated that lung cancer screening has the potential to enable early detection of the disease. The study reported a 20% reduction in lung cancer mortality among participants who underwent low-dose computed tomography (LDCT) compared to those who underwent chest radiography screening [4]. However, the implementation patterns and benefits of LDCT screening vary significantly across different groups. Factors such as race, ethnicity, rurality, and socioeconomic status have emerged as major determinants of access and use [5]. Notably, the rural population undergoes fewer LDCT scans compared to the urban population [6]. In the ASEAN region, the scarcity of imaging equipment compared to global standards exacerbates these disparities [7]. Geographic accessibility is another constraint, with 5% of the target population lacking a screening facility within 40 miles, and no facilities available in rural areas at any distance [8].

Thus, new methods for interpreting chest X-rays (CXRs) using AI have been developed to address the disparities in lung cancer screening. Computer-assisted detection (CAD) developed through AI has enhanced lung cancer nodule detection on chest X-rays (CXRs) for general physicians and radiologists, leading to increased accuracy, specificity, and positive predictive value (PPV) [9]. The authors of [10] demonstrated that applying a deep learning-based automatic detection algorithm (DLAD) to CXRs improved the average area under the alternative free-response receiver operating characteristic curve for all observers from 0.67 to 0.76. The DLAD also helped detect more lung cancers (53% vs. 40%, $p < 0.001$) [10]. The performance of AI models in diagnosing CXRs has been very impressive, with median area under the curve (AUC) values ranging from 0.979 [11] to 0.983 [12] for detecting CXRs with major thoracic diseases.

While AI shows great promise in lung cancer screening, several significant challenges impact its implementation. Technical limitations include AI's difficulty in detecting small nodules and nodules with varying densities [13]. Additionally, AI systems often underperform when applied to populations that differ from the datasets used for training [14]. Poor image quality [15] and complex anatomical structures [16] further compromise performance, particularly in detecting nodules that overlap with other structures. Moreover, AI accuracy decreases with increasing case complexity [17], and challenges arise in comparing current images with past radiographs [18]. These implementation challenges highlight the need for enhanced AI systems capable of delivering consistent performance across diverse clinical settings and patient populations.

However, it is also important to realize that AI in medical imaging was never meant to replace radiologists but rather to serve as a tool that can help them. Research revealed that AI sensitivity for detecting early-stage lung cancer can increase to 83.3% when combined with radiologists [19]. The application of AI in CXRs has the potential to help in the early identification of lung cancer and lower the mortality rate. Furthermore, AI-assisted CXR analysis could help to address some of the disparities in lung cancer screening. It could provide a less costly, potentially more available first-screening test that might be available to regions that have limited access to LDCTs or few radiologists.

1.2. Objective of the Review

The primary objective of this systematic literature review (SLR) is to comprehensively assess the performance of AI algorithms compared to radiologists in detecting lung nodules in CXRs. This review aims to:

1. Evaluate and compare the performance of AI algorithms against radiologists in detecting lung nodules in CXRs.

2. Analyze performance variations of AI algorithms and radiologists based on nodule sizes, morphology, and locations of the nodules.
3. Identify and examine factors contributing to discrepancies between AI and radiologist interpretations in lung nodule detection.
4. Investigate patterns of missed lung nodules when comparing AI algorithms with radiologists in chest radiograph interpretation.

2. Materials and Methods

2.1. Reference Source (PRISMA)

This study adopts the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) as the primary reference source (Figure 1). PRISMA, developed by the authors of [20], was selected due to its advantages in several aspects. Firstly, PRISMA provides researchers with clear guidelines for producing and writing SLR articles. The PRISMA checklist will guide the documentation of search strategies, study selection criteria, and data extraction processes, ensuring transparency and completeness in reporting. In addition, PRISMA supports quality control of the selected articles [20]. Referring to PRISMA has enabled researchers to develop this SLR based on four main methodological processes: namely (1) formation of research questions, (2) systematic search strategy (identification, screening, and eligibility), (3) article quality assessment, and (4) data extraction and analysis. By following PRISMA guidelines, this review aims to comprehensively present the methodological report on the current state of the study, allowing readers to evaluate the findings and their reliability for practical application.

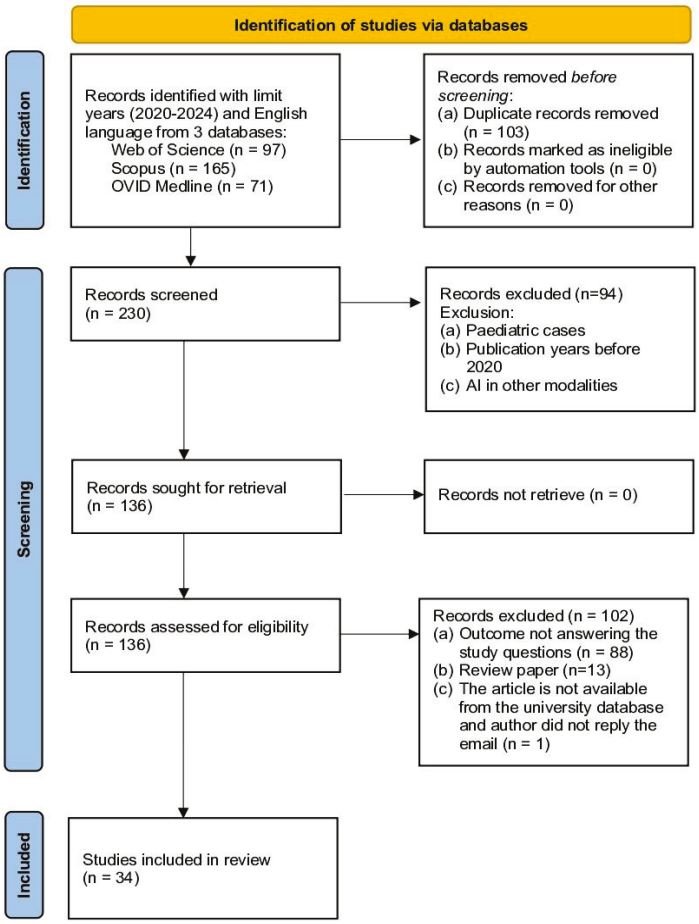


Figure 1. Study selection process following PRISMA guidelines.

2.2. Formation of Research Questions

Referring to the mnemonic PICO (Problem, Intervention, Comparison, and Outcome), the research question for this SLR was formulated. PICO is particularly suitable for medical investigations as it offers a clear and systematic means of drawing questions, which, in turn, will increase the focus and relevance of the investigation [21]. PICO improves the efficiency of literature searches, the solidity of study designs, and the dependability of evidence syntheses. In this SLR, P refers to lung cancer, I refers to the use of AI in chest radiographs, C refers to a radiologist, and O refers to lung nodule detection. This reference to the mnemonic PICO has helped the researcher to produce these research questions and the motivation behind these questions (Table 1).

Table 1. Research questions and motivations for evaluating AI performance in CXRs versus radiologists in lung nodule detection.

Research Questions	Motivation
How does the overall performance of AI algorithms compare to that of radiologists in detecting lung nodules in CXRs?	To assess the current state of AI technology in medical imaging and determine if AI can match or surpass human performance in this critical diagnostic task.
How does the performance of AI systems and radiologists vary across different nodule sizes, morphology types, and locations of lung nodules?	To identify the strengths and limitations of AI systems and radiologists across different nodule characteristics in diverse clinical scenarios.
What factors contribute to discrepancies between AI and radiologist interpretations?	To pinpoint areas where AI and human expertise differ, which can lead to insights for improving both AI algorithms and radiologist training.
What are the key patterns of missed lung nodule findings when comparing AI algorithms to radiologists in CXR interpretation across varying levels of radiologist experience and different types of nodules?	To identify specific areas where either AI or radiologists tend to make errors, considering factors such as nodule characteristics and radiologist experience.

Notes: Abbreviation: AI, Artificial Intelligence; CXRs, chest X-rays; Research questions were formulated using the PICO framework; Motivation was aligned with current clinical needs in radiology practice.

2.3. Systematic Search Strategy

The next process involves a systematic search strategy. This process helps researchers conduct a comprehensive material search, and it has three main steps: identification, screening, and eligibility. The search utilized three electronic databases, namely Web of Science (WOS), Scopus, and Ovid Medline.

2.3.1. Identification

The first author conducted the initial identification process to determine suitable keywords for this SLR. Based on the set of research questions formed, some keywords have been selected, such as “lung cancer”, “pulmonary cancer”, “lung nodule”, “pulmonary nodule”, “artificial intelligence”, “chest x-ray”, “chest radiograph”, and “radiologist”. Next, to maximize the number of articles that can be obtained, researchers have searched for synonyms of related words and variations to the keywords. These have been searched in two main sources: referring to the keywords of previous studies and keywords suggested by databases such as Ovid Medline.

This process has produced some additional keywords such as "lung neoplasm", "pulmonary neoplasm", "deep learning", "machine intelligence", "computer vision system", "computational intelligence", and "thoracic radiography". In addition, researchers have also produced search strings for this SLR based on functions such as Boolean Operator, Phrase Searching, Truncation, Wild Card, and Field Code. The final search was performed on 23 July 2024. The search brought a total of 392 documents identified from the three databases (WOS Core Collection = 116, Scopus = 168, Ovid Medline = 108). The number was reduced to 333 articles using language (English) and the publication year (2020–2024) filters (WOS 97, Scopus 165, Ovid Medline 71). A total of 230 articles were selected after the duplication was removed and taken to the next process, which was screening.

2.3.2. Screening

Screening is a process of filtering articles comprising several standards that the researcher uses to choose articles of interest. Two reviewers were independently involved in the screening process. Among the 230 articles that were identified and screened, 94 articles were removed based on the exclusion criteria, including pediatric cases, papers published before 2020, and articles discussing AI in other medical imaging techniques. From this screening process, there was a reduction in the articles to 136 for further determination of their relevance to the study.

2.3.3. Eligibility

The eligibility process for the selection of articles is another important step of the systematic review in which the researchers take their time to assess the suitability of the selected articles about the objectives of the study. The process includes the evaluation of the title and abstract of each article and, if required, the main text. The first step is to check and ensure that the chosen articles are relevant to the research objectives and questions. In this review, the eligibility screening resulted in the removal of 102 articles for various reasons: they did not achieve the aims of the study, review articles rather than primary research, and the full text of the articles was not available in the university's database. After the elimination process based on these criteria, 34 articles were considered to meet the inclusion criteria in the next step of the review, which was the quality assessment of the articles that were selected.

2.4. Evaluation of Article Quality

Next, the 34 articles selected earlier were evaluated for their quality. This process was carried out by the lead author of this SLR and assisted by two co-authors. The evaluation method by the authors of [22] was used to determine the selected articles. This method is based on six main criteria, which are as follows:

QA1. Is the purpose of the study clearly stated?

QA2. Is the interest and the usefulness of the work presented?

QA3. Is the study methodology established?

QA4. Are the concepts of the approach clearly defined?

QA5. Is the work compared and measured with other similar work?

QA6. Are the limitations of the work mentioned?

A scoring system was used to determine whether the selected articles were of good quality. Every time the researcher gave a yes answer to the evaluated criteria, 1.0 marks were given. If the researcher chooses an incomplete answer, 0.5 marks are given. If the researcher chooses the answer no, zero marks will be given. The condition for articles to be selected in this SLR is that the minimum score recorded should reach at least 3.0 and above [22]. Based on this process, the researchers found that all 34 articles that were evaluated passed the minimum score of 3.0 and were selected to be reviewed.

2.5. Data Extraction and Synthesis

Two reviewers independently reviewed the selected articles and extracted relevant data systematically. Most of the data collection was conducted on the results and discussion sections of the selected articles, with other sections examined where the need for context or clarity exists. The authors have utilized Petal.ai to assist with initial data extraction and manually validate all extracted data points to ensure accuracy. Data synthesis was conducted using a narrative descriptive approach since the studies included in the review are heterogeneous. The findings were organized into four main themes aligned with our research objectives. Data from all independent research were summed up in a common table for further comparison and analysis. A meta-analysis was not conducted because the study was rather diverse in terms of study designs, populations, and outcome measures. The authors used Claude.ai to improve clarity and readability under the authors' supervision and final approval.

3. Results

3.1. Overview of the Study

The number of research publications in AI-assisted lung nodule detection on chest radiographs, especially from 2020 to 2024, in which the number of articles has grown from 6 in 2020 to 13 in 2023 (Table 2). This SLR synthesizes the cross-national trends (Table 3) and findings from work conducted throughout this period (Table 4). These studies were conducted with participants from various geographical areas, such as Asia, including South Korea [13,17,23–33], Japan [16,18,34], Taiwan [35,36], and India [14,37,38]; Egypt [39], the United States [15,37,40,41], Israel [42], and Australia [43]; and European countries, such as Germany [44], France [45], The Netherlands [46,47], the United Kingdom [48,49], and Switzerland [50].

Table 2. Publication trends in AI-assisted lung nodule detection research.

Year	Number of Research	References
2020	6	[13,15,30,32,33,38]
2021	6	[35,44,49,50]
2022	5	[14,25,31,36,37,40]
2023	13	[16,17,23,24,26–29,34,39,43,45,48]
2024	4	[18,42,46,47]

Notes: This table shows an overview of the distribution of 34 included studies published over 2020–2024. References from the same year indicate different studies published in the same calendar year.

3.2. Study Design and Population

From the analysis of 34 selected articles, the majority employed retrospective cross-sectional study designs ($n = 26$) [15–18,24,26,28–39,41–46,48,49]. The remaining studies utilized various other methodologies, including retrospective cohort studies ($n = 3$) [25,40,47] and retrospective case-control studies ($n = 2$) [13,27].

Out of all of these study designs, only one used a randomized control trial approach [23], one conducted an experimental study using mannequin-like models or anthropomorphic phantoms [50], and one used a prospective cross-sectional study design [14]. The distribution of study designs shows that authors employed cross-sectional studies as the primary source of AI performance assessment for lung detection in chest radiographs.

Variations in sample size and the representation of study subjects were observed in the 34 studies under analysis. The largest study was conducted on 759,611 chest radiographs of 538,390 patients in five regional centers from India [38], and other select studies were conducted on 100–300 participants [24,27,35]. Some of the large sample studies sourced

participants from health screenings, with 10,476 participants from a tertiary hospital [23] and 6452 participants from a healthcare screening center [25]. Control subjects were generally selected using case-control studies, like in one study with 50 subjects of lung cancer patients and 50 controls [27] and another one with 168 patients in case groups and 50 controls [13]. Certain investigations targeted higher-risk people from NLST [15] or patients with lung cancer confirmed pathologically [18].

The studies used different criteria for defining the populations they included and excluded from their research. The majority of studies involved patients who were 18 years of age and older [23,26,35,39], and a few of them were designed exclusively for patients of a certain age, namely 55–74 years, in the case of lung cancer [15]. Some of the studies confirmed nodules by CT scans or pathology [16,28,30,33], while others excluded cases according to their specific characteristics, such as extensive infiltrative lesions [31], poor image quality [39], or multiple nodules [33]. Some investigations included data collected from more than one institution or source, such as more than thirty-five centers in India [14], seven medical centers in The Netherlands [47], national institutions, and public datasets [34,36].

Table 3. Characteristics of selected articles.

Author and Year	Country	Study Design	Study Population	Study Setting
Nam et al., 2020 [13]	Republic of Korea	Retrospective case-control study	168 patients (mean age 71.9 ± 9.5 years) with 187 initially undetected lung cancer nodules, plus 50 normal controls; CT or X-ray validated; from March 2017 until December 2018	A single center
Sim et al., 2020 [30]	Republic of Korea	Retrospective cross-sectional study	800 radiographs (200 per center: 150 cancer cases, 50 normal); from October 2015 until September 2017	Four tertiary hospitals
Lee et al., 2020 [32]	Republic of Korea	Retrospective cross-sectional study	Two cohorts: validation (10,289 X-rays from 10,206 individuals) and screening (100,576 X-rays from 50,098 individuals)	Seoul National University Hospital
Koo et al., 2020 [33]	Republic of Korea	Retrospective cross-sectional study	378 patients (61% male; mean age 61.4 years) with ≤three pathologically proven nodules; excluding nodules < 5 mm or calcified; from January 2016 until December 2018	Tertiary hospital
Majkowska et al., 2020 [38]	India	Retrospective cross-sectional study	Two datasets: DS1 (759,611 X-rays from 538,390 patients) with natural prevalence; ChestX-ray14 (112,120 X-rays from 30,805 patients) enriched for abnormalities	Five Indian centers (DS1) and NIH database (ChestX-ray14)
Yoo et al., 2020 [15]	United States	Retrospective cross-sectional study	5485 high-risk smokers (age 55–74 years; ≥30 pack-years; 55.2% male) receiving three annual screenings; from August 2002 until April 2004	NLST multicenter trial (21 sites)

Table 3. *Cont.*

Author and Year	Country	Study Design	Study Population	Study Setting
Teng et al., 2021 [35]	Taiwan	Retrospective cross-sectional study	100 subjects with CXR (47 with CT-validated nodules or masses, 53 normal); mean age 55.07 ± 13.80 years	A single center
Yoo et al., 2021 [41]	United States	Retrospective cross-sectional study	519 screening X-rays from 294 patients (98 cancer cases, 196 controls); from August 2002 until April 2004	NLST trial dataset
Homayounieh et al., 2021 [44]	Germany	Retrospective cross-sectional study	Adult patients of both genders with PA chest radiographs; including cases with nodules, non-nodular abnormalities, and normal findings	Two sites (A and B)
Tam et al., 2021 [49]	United Kingdom	Retrospective cross-sectional study	400 cases (200 tumors, 200 controls); mean ages 72.6 and 61.8 years; seven years retrospective review	NHS setting
Peters et al., 2021 [50]	Switzerland	Experiment study	61 chest phantoms with 140 solid lung nodules were placed in artificial lung parenchyma (53 phantoms with lung nodules and eight without nodules).	A phantom study using PACS-workstation
Nam et al., 2022 [25]	Republic of Korea	Retrospective cohort study	6452 health checkup participants (52% male; mean age 57.6 ± 13.0 years); a subset of 3073 with CT validation	Seoul National University Healthcare Screening Center
Bae et al., 2022 [31]	Republic of Korea	Retrospective cross-sectional study	111 patients with PA chest radiographs including 49 patients with 83 pulmonary nodules; excluding complex pathologies and CT-only visible nodules; from April 2016 until December 2019	A single center
Chiu et al., 2022 [36]	Taiwan	Retrospective cross-sectional study	Adults (≥20 years) with single lung nodule confirmed by biopsy/resection; plus, cases from four external datasets; from 2011 until 2014	Taipei Veterans General Hospital
Kaviani et al., 2022 [37]	United States	Retrospective cross-sectional study	2407 adults (≥21 years; 1248 males; mean age 39 years); de-identified CXR data; from 2015 until 2021	Eight sites (three in India, five in the US); including quaternary and community hospitals
Govindarajan et al., 2022 [14]	India	Prospective cross-sectional study	65,604 CXRs (median age 42 years); PA/AP views; excluding incomplete/artifact cases	35 centers across Six Indian states

Table 3. *Cont.*

Author and Year	Country	Study Design	Study Population	Study Setting
Ahn et al., 2022 [40]	United States	Retrospective cohort study	Adult patients with specific radiograph findings (pneumonia, nodules, pneumothorax, pleural effusion)	Deaconess Medical Centre and Massachusetts General Hospital
Nam et al., 2023 [23]	Republic of Korea	Randomized control trial	10,476 adults (>18 years) undergoing chest radiography; median age 59 years; 5121 men; from June 2020 until December 2021	Tertiary hospital health screening center
You et al., 2023 [24]	Republic of Korea	Retrospective cross-sectional study	300 patients: 100 normal cases (50% male, mean age 46.5 years) and 200 with pulmonary nodules (57% male, mean age 60.0 years)	A single tertiary hospital
Hwang et al., 2023 [26]	Republic of Korea	Retrospective cross-sectional study	73 first-visit outpatients (median age 70 years) with AI-detected incidental nodules; excluding known thoracic cases	Hospital outpatient clinic
Huh et al., 2023 [27]	Republic of Korea	Retrospective case-control study	100 cases (50 pathologically confirmed lung cancers, 50 normal); December 2015 until February 2021	Seoul National University Hospital (tertiary)
Kim et al., 2023 [28]	Republic of Korea	Retrospective cross-sectional study	Training set (n = 998; 54% male; mean age 54.2 years) plus two validation sets (n = 246 and n = 205); including normal and nodule cases; from November 2015 until July 2019	Severance Hospital, Pusan National University Hospital and Dongsan Medical Centre
Kwak et al., 2023 [29]	Republic of Korea	Retrospective cross-sectional study	Patients with incidentally detected, pathologically proven resectable lung cancer; referred from various departments; from March 2020 until February 2022	A single center
Lee et al., 2023 [17]	Republic of Korea	Retrospective cross-sectional study	120 cases (60 biopsy-proven cancers from 5647 registry, 60 controls); expert-annotated with CT validation; from December 2015 until February 2021	Seoul National University Hospital, Korea
Ueno et al., 2023 [16]	Japan	Retrospective cross-sectional study	388 patients with suspected lung cancer; required both CT and CXR (PA view) within one-month intervals; from June 2020 until May 2022	Single institution,

Table 3. *Cont.*

Author and Year	Country	Study Design	Study Population	Study Setting
Higuchi et al., 2023 [34]	Japan	Retrospective cross-sectional study	5800 CXRs total: 800 from Fukushima (50% normal, 50% nodules) and 5000 from the NIH dataset	Fukushima health centre and NIH database
Farouk et al., 2023 [39]	Egypt	Retrospective cross-sectional study	150 adults with varied lung pathologies; requiring both X-ray and CT; excluding poor-quality images and complex cases; from May 2021 until July 2022	Ain Sham University
Tang et al., 2023 [43]	Australia	Retrospective cross-sectional study	Volunteer radiologists evaluating three AI interfaces vs. no AI in paired-reader design	Royal Melbourne Hospital
Bennani et al., 2023 [45]	France	Retrospective cross-sectional study	500 patients (52% female; mean age 54 $\pm$ 19 years) with paired CXR and CT within 72 h; From emergency, inpatient, and day hospital settings; from January 2010 until December 2020	Cochin Hospital, Paris, France
Maiter et al., 2023 [48]	United Kingdom	Retrospective cross-sectional study	5722 radiographs from 5592 adults (53.8% female; median age 59 years); from July 2020 until February 2021	Sheffield Teaching Hospitals (tertiary)
Hamanaka et al., 2024 [18]	Japan	Retrospective cross-sectional study	Patients who underwent malignant lung tumor resection; from November 2021 until July 2023	Shin-Yurigaoka General Hospital
Kirshenboim et al., 2024 [42]	Israel	Retrospective cross-sectional study	683 chest radiographs from an initial pool of 50,286; excluding interpreted cases and duplicates	A single center
Topff et al., 2024 [46]	The Netherland	Retrospective cross-sectional study	25,104 radiographs from 21,039 adults ($\geq$18 years; mean age 61.1 $\pm$ 16.2 years); Institution 1 (April 2021–February 2022) and Institution 2 (January–December 2018)	Two institutions
van Leeuwen et al., 2024 [47]	The Netherland	Retrospective cohort study	Two cohorts: 95 hand radiographs (age range 0–18 years; from January 2017–January 2022) and 140 chest radiographs (January 2012–May 2022); CT-validated reference standard	Seven Dutch hospitals

3.3. Performance in Lung Nodule Detection of AI Algorithms and Radiologists

Numerous studies demonstrate that AI algorithms are more sensitive than radiologists, with similar or slightly lower specificities in lung nodule detection [13,23,40]. The authors of [13] reported that AI has a significantly higher sensitivity (70% vs. 47%) and specificity (94% vs. 78%), as well as a significantly higher area under the receiver operating character-

istic (AUROC) (0.899 vs. 0.634–0.663) compared to the radiologist. Similarly, another study reported superior AI sensitivity of 56.4% compared to 23.2% of radiologists, while both had relatively high specificity of 99.0% and 98.2%, respectively [23]. Additional analysis of the AI also indicated impressive AUC, where one study showed an AUC of 0.99 [32] while the authors of [37] also demonstrated a high AUC of 0.935 for missed findings. These improvements were particularly perceived to be significant in the identification of smaller nodules and those located in the "danger zone" [18,24].

However, a more detailed comparison of other studies shows variation in AI performance. This variability was described by [26] that depending on the specific model and data characteristics, AI sensitivities can vary from 44.1% to 95.7% and specificity from 71.9% to 97.5%. Variability in performance has also been observed for the various types of nodules, including a study that reported an AUC of 0.71 on calcified nodules but only 0.55 on non-calcified clinically important nodules [37]. This variability is further demonstrated across different studies. The authors of [28] also showed that the sensitivity of the AI model was 91.5%, in contrast to the mean sensitivity of radiologists at 77.5%, whereas the authors of [36] suggested that the sensitivity of AI was 79% with 3.04 FPs per image compared to the pooled radiologists' sensitivity of 76%.

In radiograph classification, the sensitivity of AI was 53.4%, and specificity was 84.3%, while the performance of radiologists resulted in 38.8% sensitivity and 94.1% specificity [25]. Overall, CXR classification AUC was achieved at 0.890 by AI, 0.816 by radiologists, and a combined 0.866 [39]. However, one study reported lower values of AI sensitivity of 35% compared to the pooled radiologists' sensitivity of 50% [50]. The use of AI systems alone could be insufficient and will be prone to over, or misdiagnosis in some of the studies [42]. Nevertheless, the most effective use of the AI system is in combination with radiologist interpretation [30,49].

The integration of AI with radiologists demonstrates an enhanced diagnostic performance. Several studies that described AI as a second reader or assistant tool show that the performance advantage was higher for junior readers [33,41,44]. The authors of [33] and [49] identified an increase in radiologist sensitivity from 78 to 91% and from 88.6 to 93.9%, respectively. Another study reported that the sensitivity ranges from 38.8 to 45.1% [25], while the authors of [44] noted that the sensitivity mean of junior radiologists increases from 40 to 52% with the help of AI. The previous studies highlighted that by implementing the AI model, the sensitivity of radiologists has risen from 77.5% to 92.1% [28]. When applying per nodule analysis, the AUC of AI was 0.875, compared to the radiologist's AUC of 0.772 and a combined AUC of 0.834 [39].

3.4. AI and Radiologist Performance Across Nodule Size, Morphological Type, and Location

A study found AI improved the identification of actionable lung nodules, specifically around the missed small-sized nodules [23]. Specifically, AI was 45% accurate in identifying nodules under 1 cm in size, while human readers were only 25% accurate [13]. AI algorithms demonstrated higher accuracy with nodules of higher visibility scores, especially when nodules exceeded 2 cm [24]. However, it was found that DLAD systems improved when detecting nodules larger than 15 mm [16]. The clinical data showed that, on average, resected lung cancers were 2.6 cm in size [29].

The ability to detect lung nodules also depends on the morphological characteristics of the nodules. Several studies included nodules of various types, such as solid, part-solid, and ground-glass opacity (GGO) [16,29,33]. DLAD systems reported significantly higher sensitivity for all the solid nodules compared with part-solid and pure GGNs [16]. For calcified nodules, AI systems had higher AUC values of 0.71 than non-calcified clinically important nodules (AUC 0.55) [37]. However, both AI and radiologists have difficulty detecting GGO-predominant nodules due to their subtle radiographic features [33]. Al-

though part-solid nodules have diagnostic challenges, a study shows that AI can detect these lesions, including invasive adenocarcinomas measuring 0.75 cm [29].

The detection performance also differentiated significantly according to the location of nodules within the chest radiograph. Studies revealed AI has higher sensitivity and specificity in detecting apical lung nodules than para-mediastinal or retro-diaphragmatic locations, which are considered "danger zones" (DZ) [24]. The authors of [13] identified that the AI system could detect lung nodules better in the apical and hilar regions than the radiologist, but both have difficulty in retro-cardiac and retro-diaphragmatic areas, with only 21% to 22% of these lesions being detected. The detection rate was overall lower in the lower lung fields at 43.8% and hilar regions at 30% [18]. In particular, smooth, well-marginated nodules without marked overlap of the internal structure were found to have significantly higher detection rates [33]. It was also found that central nodules were particularly challenging for AI-CAD systems, where sensitivity was found to be as low as 9% as compared to 64% for peripheral nodules [50].

The presence of overlapping anatomical structures provided the greatest impact on the accuracy of detection, both in the AI systems and for human readers. It was further found that 86.4% of the cases were nodules missed due to being masked by other lesions, and the majority of them were hidden by bony structures such as ribs and costochondral junctions [33]. Both AI systems and radiologists consistently struggled with nodules obscured by various anatomical structures such as bone, heart, diaphragm, and mediastinum [27]. Particularly, difficulties were observed in areas contacting the left clavicle [41], behind the heart, and hilum [49]. Anatomical overlap was most prominent in the frontal view, where nodules could be masked by anatomical normal structures [32]. To mitigate these issues, a deep learning-based synthesized bone-suppressed (DLBS) model was developed to improve nodule detection in regions where ribs and clavicles overlap without the need for special equipment or extra-radiation exposure [28]. AI also assisted less experienced individuals in detecting nodules that overlapped with anatomical structures, while readers with experience were able to detect these nodules with and without AI [41].

3.5. Factors Contributing to AI-Radiologist Discrepancies

Lung nodule detection presents significant challenges in both AI systems and radiologists due to nodule characteristics, anatomical structure, and technical reasons. AI systems exceed radiologists in the overall performance of detecting small nodules, but they tend to produce more false positives in their interpretations [32]. However, both AI and radiologists share common challenges. They equally struggle to identify ground-glass opacity nodules based on their subtlety and fail to examine anatomically difficult regions such as para-mediastinal, retro-cardiac, and retro-diaphragmatic areas [13,33]. Further complicating these challenges is the interpretation of chest radiographs with low contrast resolution and overlapping anatomical structures, obscuring subtle findings for both AI systems and human readers [31].

With the human factor in radiological interpretation, the performance is affected by multiple variables. The authors of [13] identified that environmental conditions such as reading time, fatigue levels, and years of experience impact radiologist performance. The impact of AI depends on both radiologist experience and case complexity [44]. Studies observe particularly high variability in performance across these experience levels and how AI assistance impacts them. For instance, AI can increase sensitivity in less experienced readers and specificity in more experienced radiologists [41].

Integration of AI into clinical practice is promising to reduce missed diagnoses but also has its challenges. A study reported that a radiologist's accuracy can be reduced by up to 18%, and sensitivity can be reduced by 30% because of distracting findings, including

COPD and small pleural effusions [49]. Different AI user interfaces (UIs) also had a major impact on how radiologists performed and interpreted AI outputs [43]. A study also described that occasional correct AI detections were rejected by readers, even where AI assistance improves outcomes [45].

The transition to practical AI in real-world clinical implementation reveals several practical limitations of current AI systems. The authors of [50] described the key challenges, including a lack of well-curated datasets, generalization, and lack of adequate publication standards for reproducibility. Another study demonstrates the lack of generalizability in training and test datasets for the AI, and the demographic differences may affect performance in real-world patient cohorts [48]. Hamanaka and Oda [18] noted the system's main limitation, which is it cannot compare past radiographs with new ones, a key feature used in clinical practice. This research of [47] attributed the discrepancy between the results of AI and those of radiologists to the lack of clinical information and the difference between the imaging reading platform of AI and clinical practice.

3.6. Pattern of Missed Lung Nodule Findings: AI Versus Radiologist

Several studies have shown the ability of the designed AI algorithm to identify missed nodules. The authors of [30] confirmed the possibility of AI-improved detection of 15% of missed nodules in standalone radiologist sessions. Similarly, it was highlighted the fact that AI revealed 54% (69/131) of undiagnosed lung nodules that were missed in routine clinical practice [37]. Radiologist interpretation integrated with AI was most effective, where AI was able to identify eight out of 15 cases that all the radiologists missed. Their combination was found to have reduced the missed tumors by 60% to 65.4% [49]. Another study discovered that AI performed well in identifying right mid- and left-lower lung nodules, which are often missed by radiologists [44].

These findings were followed by other patterns in nodule detection by AI and radiologists. Comparatively, the authors of [36] showed different detection patterns. AI would have missed 68 nodules, which were detected by radiologists, while radiologists failed to detect 91 nodules detected by AI. The authors also highlighted that none of the AI or radiologists was able to identify nodules in 89 patients during early assessment [36]. The authors of [50] further described that while the adopted AI-CAD systems achieved detection of the lesions that the radiologists missed, the overall sensitivity was poorer than that of radiologists. Meanwhile, a study observed comparable disparities in obscured regions where AI and radiologists exhibited similar miss rates, with lung nodules missed by AI in 38.7% and by radiologists in 38% [39].

Nevertheless, specific challenges were reported in AI missed cases. Several cases of AI failures with GGO-predominant nodules were reported [33]. For instance, in the case of a 75-year-old man, AI failed to detect the tumors, as well as two radiologists. Further, AI occasionally missed subsolid nodules with spiculated edges in the left upper lobe [33]. The authors of [48] found that AI failed to detect findings in real-world patient cohorts more than radiologists for lung nodule cases. Lung cancer detection specifically showed varying miss rates. The AI missed 12/48 lung cancer cases, whereas radiologists failed to identify 7/48 cases, and four cases were missed by both AI and radiologists [15].

Another study stated that AI failed to identify all 18 cases of adenocarcinoma in situ (AIS) and three additional cases because of the flat appearance of tumors in radiographic interpretation [18]. Concurrently, AI-reported missed findings in 21.1% (5289/25,104) of cases, of which 0.9% (47/5289) were clinically significant; among the missed findings were 71.4% (25/35) of lung nodules [46]. It was found that AI had a detection rate of around 1% for unexpected nodules in initial CXRs, 70% of which were true positive nodules, and 20.5% needed further action [26].

Table 4. AI versus radiologist performance in lung nodule detection across study parameters and outcomes.

Author and Year	Performance			Nodule Characteristics (Morphology Type, Size, and Location)		Discrepancies Factors Between AI and Radiologists in Lung Nodule Detection	The Pattern of Missed Lung Nodule Finding
	AI	Radiologist	Combined	AI	Radiologist		
Nam et al., 2020 [13]	• Sensitivity: 70% • Specificity: 94% • False positive rate: 21.1% (46 of 218) • AUROC: 0.899 Software: Lunit INSIGHT CXR version 1.0.1.1 (Lunit Inc., Seoul, Republic of Korea)	• Sensitivity: 47% • Specificity: 78% • False positive rate: 19.0% (166 of 872) • AUROC: 0.634–0.663	AI-aided radiologists improve AUROC to 0.685–0.724	• Detection rate of nodule conspicuity category 3: 93% • Consistent performance across different size nodules (AI detects 45% of <1.0 cm nodules compared to radiologists with a 25% detection rate) • Better detection at apical and hilar region compared to radiologist • Struggle with retro-cardiac and retro-diaphragmatic area	• Detection rate of nodule conspicuity category 3: 70% • Improved detection with larger nodules • Struggle with apical, hilar, retro-cardiac, and retro-diaphragmatic area	• AI challenges: • Small nodules (Category 0) • Density of nodule Radiologist challenges: • Certain location • An environmental condition during reading • Level of fatigue • Years of experience	N/A
Sim et al., 2020 [30]	Sensitivity: 67.3% FPPI: 0.2 Software: Samsung Auto Lung Nodule Detection-ALND, version 1.00 (Suwon, Republic of Korea)	Sensitivity: 65.1% FPPI: 0.2	Radiologists' sensitivity with DCNN: 70.3% FPPI: 0.18	N/A	• The lower detection rate in smaller nodules (1–2 cm) compares to larger nodules (2–3 cm) with/without DCNN. • The greater nodule detection rate in the peripheral lung compared to in the lung projecting over the mediastinum or diaphragm with or without DCNN	The DCNN software improved the sensitivity of radiologists, irrespective of radiologist experience, nodule characteristics, or the vendor of the radiography acquisition system	AI helped find 15% of missed nodules from standalone sessions

Table 4. *Cont.*

Author and Year	Performance			Nodule Characteristics (Morphology Type, Size, and Location)		Discrepancies Factors Between AI and Radiologists in Lung Nodule Detection	The Pattern of Missed Lung Nodule Finding
	AI	Radiologist	Combined	AI	Radiologist		
Lee et al., 2020 [32]	Validation test performance: • Sensitivity: 90% • Specificity: 97% • NPV: 100% • PPV: 2.7% • AUC: 0.99 • False-positive rate: 3.1% Screening cohort performance: 1. Cancer-Positive Radiographs: • Sensitivity: 40% (39/98) Specificity: 97% • NPV: 100% • PPV: 1.3% • AUC: 0.78 2. Visible Lung Cancers: • Sensitivity: 83% (39/47) • Specificity: 97% • NPV: 100% • PPV: 1.3% • AUC: 0.97 3. Visible Lung Cancers: • Sensitivity: 100% (28/28) • Specificity: 97% • NPV: 100% • PPV: 0.9% • AUC: 0.99 (95% CI: 0.99–0.99) Software: Lunit INSIGHT CXR version 4.7.2 (Lunit Inc., Seoul, Republic of Korea)	• Sensitivity: 60% • Specificity: 100% • NPV: 100% • PPV:19% • False-positive rate: 0.3%	N/A	N/A	N/A	AI performance improved with increased visibility of lung cancer in radiographs	N/A

Table 4. *Cont.*

Author and Year	Performance			Nodule Characteristics (Morphology Type, Size, and Location)		Discrepancies Factors Between AI and Radiologists in Lung Nodule Detection	The Pattern of Missed Lung Nodule Finding
	AI	Radiologist	Combined	AI	Radiologist		
Koo et al., 2020 [33]	• Sensitivity: 67.3% • Specificity: 86.2% • False positives per image: 0.2 • AUROC for abnormality detection: 0.87–0.96 Software: Lunit INSIGHT CXR, version 1.00 (Lunit Inc., Seoul, Republic of Korea)	• Average radiologist sensitivity without AI: 88.6% • Expert radiologists' sensitivity: 40% to 87% • False positives per image: 0.063 • AUROC for Group 1 Radiologist: 0.93 • AUROC for Group 2 Radiologist: 0.96	• Average radiologist sensitivity with AI: 93.9% (5.2% increase) • False positives per image with AI: 0.032 • The number of false positives per image fell significantly when radiologists were assisted by the DCNN	DCNN is better at the detection of nodules larger than 5 mm, solid, round-shaped, well-marginated, and laterally located nodules. DCNN struggles with masked nodules due to overlapping structure and GGO-dominant nodules	The radiologist missed a masked nodule due to overlapping with bony lesions and a GGO-dominant nodule in the upper left lobe	Radiologist interpretation errors due to: • Poor spatial resolution • Lesional masking by anatomical structures • Inadequate radiological skills • Overwork DCNN limitations: • Unreliable differentiation between true and false positives • Performance affected by nodule size and nature	• AI missed a GGO-dominant nodule in a 75-year-old man (also missed by two radiologists) • AI missed subsolid nodule due to GGO predominance • One radiologist failed to re-detect a nodule when AI data were available • AI missed a subsolid nodule with a spiculated margin in the left upper lobe • Missed spiculated nodule, diagnosed as adenocarcinoma
Majkowska et al., 2020 [38]	Sensitivity • CXR14 dataset: 82.4% • DS1 dataset: 44.1% Specificity • CXR14 dataset: 84.9% • DS1 dataset: 97.5% PPV • CXR14 dataset: 49.2% • DS1 dataset: 77.7% AUC • CXR14 dataset: 0.91 • DS1 dataset: 0.72 CNN architecture: Xception	Sensitivity • CXR14 dataset: 79.7% • DS1 dataset: 40.1% Specificity • CXR14 dataset: 82.3% • DS1 dataset: 96.7% PPV • CXR14 dataset: 44.3% • DS1 dataset: 72.4% AUC: N/A	N/A	N/A	N/A	• Lack of ground truth labels leads to inconsistent performance across different datasets • Risk of mislabeling or missing critical findings of difficult cases due to single-reader approach limitations and majority-vote approach	N/A

Table 4. *Cont.*

Author and Year	Performance			Nodule Characteristics (Morphology Type, Size, and Location)		Discrepancies Factors Between AI and Radiologists in Lung Nodule Detection	The Pattern of Missed Lung Nodule Finding
	AI	Radiologist	Combined	AI	Radiologist		
Yoo et al., 2020 [15]	All radiograph • Sensitivity: 86.2% • Specificity: 85% Digital radiograph • Sensitivity: 96.0% • Specificity: 93.2% Malignant nodule detection: • Sensitivity: 100% • Specificity: 90.9% • NPV: 100% • PPV: 8.2% Software: Lunit INSIGHT CXR, version not stated (Lunit Inc., Seoul, Republic of Korea)	All radiograph • Sensitivity: 87.7% • Specificity: 86.7% Digital radiograph • Sensitivity: 88.0% • Specificity: 82.8% Malignant nodule detection: • Sensitivity: 94.1% • Specificity: 91% • NPV: 99.9% • PPV: 7.8%	N/A	N/A	N/A	• The inferior quality of chest radiographs led to the underperformance of AI • AI struggled with non-calcified nodule detection • AI was trained on data obtained after 2010, while NLST data were from 2002 to 2004 • Ground truth labels created without CT correlation • NLST radiologist performance based on pooled data (may not reflect real-world variation)	Missed cases in nodule detection • AI 9/65 cases • Radiologist 8/65 cases • Both: 1 case Missed cases in cancer detection • AI 12/48 cases • Radiologist 7/48 cases • Both 4 cases Missed cases in malignant nodules • AI 2/34 cases • Radiologist 2/34 cases • Both 0 cases
Teng et al., 2021 [35]	• Sensitivity of mass probability algorithm: 81% • Specificity of mass probability algorithm: 90.57% • AI mass probability algorithm AUC: 0.916 (95% CI 0.891–0.937) • Sensitivity of heat map algorithm: 38.30% • Specificity of heat map algorithm: 98.11% • Heat map algorithm AUC: 0.778 (95% CI 0.743–0.811) Software: QUIBIM Chest X-ray Classifier (Valencia, Spain)	• Sensitivity of trained radiographers: 77.30% • Specificity of trained radiographers: 78.30%	N/A	N/A	N/A	N/A	N/A

Table 4. *Cont.*

Author and Year	Performance			Nodule Characteristics (Morphology Type, Size, and Location)		Discrepancies Factors Between AI and Radiologists in Lung Nodule Detection	The Pattern of Missed Lung Nodule Finding
	AI	Radiologist	Combined	AI	Radiologist		
Yoo et al., 2021 [41]	N/A	Radiology residents • Sensitivity: 61% • Specificity: 88% • FPPI: 0.15 Radiologist • Sensitivity: 76% • Specificity: 79% • FPPI: 0.24	Radiology residents with AI • Sensitivity: 72% • Specificity: 88% • FPPI: 0.12 Radiologist with AI • Sensitivity: 76% • Specificity: 86% • FPPI: 0.17 Software: Lunit INSIGHT CXR version 2.4.11.0 (Lunit Inc., Seoul, Republic of Korea)	• AI identified subtle lesions missed on prior radiographs which is useful for low visibility and overlapping lesions • Has lower performance on conventional radiographs	• Improved with AI to detect subtle lesions, especially in less-experienced reader	AI benefits for different experience levels: • Improved sensitivity of less experienced reader • Improve the specificity of more experienced readers	• Notably, 7 out of 98 cancer-positive patients had missed lung cancer on prior CXRs • AI detected more missed lung cancers than resident radiologists (71% vs. 39%, p = 0.021) • No significant improvement for experienced radiologists (51% vs. 57%, p = 0.63) • AI still missed some lung nodules on CXRs
Homayounieh et al., 2021 [44]	N/A	• Junior radiologist mean sensitivity: 40% • Senior radiologist mean sensitivity: 51% • Junior radiologist mean specificity: 95% • Senior radiologist mean specificity: 90% • Junior radiologist means partial AUC: 0.73 • Senior radiologist means partial AUC: 0.71	• Junior radiologist mean sensitivity: 52% • Senior radiologist mean sensitivity: 60% • Junior radiologist mean specificity: 96% • Senior radiologist mean specificity: 94% • Junior radiologist means partial AUC: 0.78 • Senior radiologist means partial AUC: 0.76 Software: AI-Rad Companion Chest X-ray Algorithm (Siemens Healthineers AG, Erlangen, Germany)	AI improved detection across all difficulty nodule levels (easy, moderate, and challenging)	N/A	• The impact of AI assistance is not constant depending on the radiologist's experience and the complexity of the cases • Challenging nodules likely difficult for both AI and humans	• AI detected right mid- and left-lower lung nodules missed by radiologists

Table 4. *Cont.*

Author and Year	Performance			Nodule Characteristics (Morphology Type, Size, and Location)		Discrepancies Factors Between AI and Radiologists in Lung Nodule Detection	The Pattern of Missed Lung Nodule Finding
	AI	Radiologist	Combined	AI	Radiologist		
Tam et al., 2021 [49]	• Sensitivity: 80% • Specificity: 93% • Accuracy: 87% Software: Red Dot (Behold.ai, London, UK)	• Mean sensitivity: 78% • Mean specificity: 96% • Mean accuracy: 87%	• Mean sensitivity: 91% • Mean specificity: 89% • Mean accuracy: 90%	Overall sensitivity increases with tumor size 1–2 cm	N/A	• Radiologist has reduced accuracy by 18% and sensitivity by 30% due to distracting findings (e.g., COPD, small pleural effusion) • Another factor for missed cases includes small tumor size, poor film quality, and hidden tumor	• AI detected 8 out of 15 cases missed by all radiologists • AI detected 70.2% of tumors missed by at least one radiologist • Combined AI and radiologists reduced missed tumors by 60–65.4%
Peters et al., 2021 [50]	• Sensitivity: 35% • Specificity: 84% Software: InferRead®DR (Infervision Technology Ltd., Beijing, China)	• Pooled sensitivity: 50% • Pooled specificity: 86%	N/A	• AI-CAD better performing on peripheral nodules (sensitivity of peripheral vs central: 0.64 vs. 0.094) (p = 0.025)	• Radiologist sensitivity by nodule size: - 10–12 mm: 0.80 - 8 mm: 0.66 (p = 0.04) - 5 mm: 0.14 (p < 0.001) • Radiologist sensitivity by nodule location (peripheral vs central): 0.66 vs. 0.48 (p = 0.004)	• AI-CAD systems use chest phantoms instead of human patients, which leads to difficulties with ancillary findings • Factors contributing to discrepancies include the need for good datasets, limitations in generalization, and insufficient publication standards for reproducibility	AI-CAD identified lesions missed by radiologists although had reduced overall sensitivity compared to radiologists
Nam et al., 2022 [25]	• Sensitivity: 53.4% • Specificity 84.3% • AUROC: 0.748 • Detection rate: 37.7% • False positive rate 0.36/radiograph Software: Lunit INSIGHT CXR version 2.0, Lunit Inc, Seoul, Republic of Korea	• Sensitivity: 38.8% • Specificity: 94.1% • Pooled AUROC: 0.671 • Detection rate 32% • False positive rate 0.09/radiograph	• Radiologists' sensitivity to AI: 45.1% • Radiologists' specificity with AI: 92.2% • The detection rate with AI is 38.9% • Only a minimal increase of 0.24% detection rate when radiologists aided with AI	The median nodule detection size was 1.2 cm	N/A	N/A	N/A

Table 4. *Cont.*

Author and Year	Performance			Nodule Characteristics (Morphology Type, Size, and Location)		Discrepancies Factors Between AI and Radiologists in Lung Nodule Detection	The Pattern of Missed Lung Nodule Finding
	AI	Radiologist	Combined	AI	Radiologist		
Bae et al., 2022 [31]	N/A	Sensitivity use of CXR alone: • Reader 1 (R1): 60.2% (50/83) • Reader 2 (R2): 53% (44/83)	Sensitivity withBSt-DE: • Reader 1: 89.2% (74/83) • Reader 2: 83.1% (69/83) Sensitivity with BSp-DL: • Reader 1: 78.3% (65/83) • Reader 2: 71.1% (59/83)	BSt-DE has good performance with small nodules, bone overlapping, and peripheral location	N/A	• Challenges in 2D image interpretation due to low contrast resolution and overlapping anatomical structures • AI training data limitations • Radiologist experience and interpretation variability	
Chiu et al., 2022 [36]	• AI model sensitivity: 79% with false positive per image 3.04 • The AI model detected nodules 19 (0–199) days early Deep Learning architecture: You Only Look Once version 4 (YOLOv4)	• Expert screening sensitivity is approximately 76% • The radiologists detected nodules only eight (0–263) days early	N/A	• AI has low sensitivity (50%) with smaller nodules < 1 cm. • Improve detection with larger nodules	N/A	• The AI model has potential overestimation due to the influence of the training data (data size and quality) • The study used written reports by the radiologists, which might underestimate the doctor's performance	• AI-detected lung nodules early in 91 cases missed by radiologists • Radiologists detected 68 cases of nodules missed by AI • 89 patients' nodules were not detected early by either AI or radiologists
Kaviani et al., 2022 [37]	• AI AUC for missed findings up to 0.935 • AI AUC for calcified nodules: 0.71 • AI AUC for non-calcified, clinically important nodules: 0.55 • AI identified >99% of nodules with 0.2 false positives per image Software: qXR (Qure.Ai Technology, Mumbai, India)	N/A	N/A	AI has higher performance on detection of calcified nodules—calcified nodules (AUC 0.71) vs. non-calcified nodules (AUC 0.55)	N/A	Radiologist challenges: • Often subtle or difficult to detect • Some detected by AI, not in original reports AI challenges: • Lower AUCs for some missed findings due to the increased complexity of subtle findings	AI detected 69/131 (53%) of missed lung nodules

Table 4. *Cont.*

Author and Year	Performance			Nodule Characteristics (Morphology Type, Size, and Location)		Discrepancies Factors Between AI and Radiologists in Lung Nodule Detection	The Pattern of Missed Lung Nodule Finding
	AI	Radiologist	Combined	AI	Radiologist		
Govindarajan et al., 2022 [14]	AI performance in categorizing normal and abnormal CXR (use radiologist as reference standard) • Sensitivity: 87.9% • Specificity: 82.9% • AUROC: 87.1% • NPV: 98.9% AI performance in categorizing CXR into sub abnormality (in this case lung nodule) • Sensitivity: 71.9% • Specificity: 95.5% • AUROC: 91.5% • NPV: 99.8% Software: qXR (Qure.Ai Technology, Mumbai, India)	N/A	Turnaround time of radiologists is reduced with AI assistance specifically: • Without AI (pre-qXR): 83.028 min average TAT • With AI (post-qXR): 50.287 min average TAT • Overall reduction: 32.741 min (40.63% decrease)	N/A	N/A	• AI has poor generalization due to target domain divergence leading to AI-radiologist discrepancies	N/A
Ahn et al., 2022 [40]	• Sensitivity: 81.6% • Specificity: 73.1% • AUROC: 0.858 Software: Lunit INSIGHT CXR, version 3.1.2.0 (Lunit Inc., Seoul, Republic of Korea)	• Sensitivity: 56.7% • Specificity: 88.5% • AUROC: 0.724	Reading time improved with AI assistance with an overall reduction was 3.9 s ($p < 0.001$)	N/A	N/A	Factors that influence reader performance with AI: • Size of finding • Extent of findings • Number of findings Impact of extensive AI output: • May slow down readers • Could contribute to interpretation discrepancies	N/A
Nam et al., 2023 [23]	• Sensitivity for actionable lung nodules: 56.4% • Specificity for actionable lung nodules: 99.0% • False-referral rate: 45.9% • Positive-report rate: 2.3% Software: Lunit INSIGHT CXR version 2.0.2.0 (Lunit Inc., Seoul, Republic of Korea)	• Sensitivity for actionable lung nodules: 23.2% • Specificity for actionable lung nodules: 98.2% • False-referral rate: 56% • Positive-report rate: 1.9%	AI-based CAD improved actionable lung nodule detection (odds ratio: 2.4)	AI improved the detection rate of • Solid nodules that are larger than 8 mm • Subsolid nodules where the solid component is larger than 6 mm • Small-sized nodules that might be overlooked during normal review • Malignant lung nodules with a higher detection rate compared to those without AI	N/A	N/A	N/A

Table 4. *Cont.*

Author and Year	Performance			Nodule Characteristics (Morphology Type, Size, and Location)		Discrepancies Factors Between AI and Radiologists in Lung Nodule Detection	The Pattern of Missed Lung Nodule Finding
	AI	Radiologist	Combined	AI	Radiologist		
You et al., 2023 [24]	• Sensitivity of danger zone (DZ): 64.2% • Sensitivity of non-danger zone (NDZ): 83.2% Software: Med-Chest X-ray System, version 1.0.1, (VUNO, Seoul, Republic of Korea)	N/A	Non-expert radiologists: improved DZ nodule detection with DLD	DLD has better performance for: • Apical lung nodules vs. para-mediastinal/retro-diaphragmatic • Higher nodule visibility scores • Nodules > 2 cm DLD has poorer performance in DZ nodules vs. NDZ (despite DZ nodules being larger)	N/A	Factors affecting DLD performance: • Trained based on human-provided chest radiographs • Potential human bias in training data Data quantity vs. performance: • Large training dataset (15,809 CXRs) • Still showed poorer DZ nodule detection	N/A
Hwang et al., 2023 [26]	• Sensitivity: 44.1–95.7% • Specificity: 71.9–97.5% Software: Lunit INSIGHT CXR version 3, Lunit Inc., Seoul, Republic of Korea)	N/A	N/A	False positive due to: • Bone overlaps/bony lesions • Vascular structure • Lymph nodes • Atelectasis • Fluid collection (pulmonary edema/pulmonary effusion)	N/A	Imaging challenges: • Different pathologies may present similar findings • Difficulty in differentiation	• The AI-detected unexpected nodules in ~1% of initial CXRs with 70% of true positive nodules and 20.5% needing management
Huh et al., 2023 [27]	N/A	Expert performance (seven experts) • Sensitivity: 67.3% • Specificity: 100%	• Radiologist with DLAD sensitivity: 72.8% • Radiologist with DLAD specificity: 94.2–97.7 Software: Lunit INSIGHT CXR version 4.7.2 (Lunit Inc., Seoul, Republic of Korea)	Limitations in the detection of those non-visible malignant nodules or obscured by anatomical structure (bone, heart, diaphragm, and mediastinum)	Limitations in the detection of those non-visible malignant nodules or obscured by anatomical structure (bone, heart, diaphragm, and mediastinum)	• The human interpretation introduces intra/inter-observer variability, leading to AI evaluation biases • The characteristic of expert-determined vs gold standard affects the accuracy of AI performance evaluation • Characteristics of experts and radiologists such asDetection of only clearly visible lung nodules can lead to interpretation discrepancies	N/A

Table 4. *Cont.*

Author and Year	Performance			Nodule Characteristics (Morphology Type, Size, and Location)		Discrepancies Factors Between AI and Radiologists in Lung Nodule Detection	The Pattern of Missed Lung Nodule Finding
	AI	Radiologist	Combined	AI	Radiologist		
Kim et al., 2023 [28]	• DLBS model sensitivity: 91.5% (109/119) • Original model sensitivity: 79.8% (95/119)	Radiologists' mean sensitivity: 77.5%	• Radiologists with DLBS model sensitivity: 92.1% • Radiologist performance with DLBS assistance can reduce false-positive markings per image	DLBS model improved detection across nodule types by suppressing bone structures (rib, clavicle, and other anatomic structures)	N/A	N/A	N/A
Kwak et al., 2023 [29]	• AI detection rate: 100% (13/13) • Stage I detection: 7 patients (53.8%)—requiring only surgery • Median abnormality score: 78% (range: 26–95%) Software: Lunit INSIGHT CXR, version 2 and 3 (Lunit Inc., Seoul, Republic of Korea)	N/A	N/A	Nodule Characteristics detected by AI: • Median size of resected lung cancer: 2.6 cm • Notable detection of small invasive adenocarcinoma (0.75 cm) • Detect both solid and part-solid nodules	N/A	N/A	N/A
Lee et al., 2023 [17]	High Diagnostic Accuracy (full dataset) AI Performance: • Per-radiograph sensitivity: 80% • Per-lesion sensitivity: 73% • Specificity: 97% • AUC: 0.88 • AUFROC: 0.85 Low Diagnostic Accuracy (10% dataset) AI Performance: • Per-radiograph sensitivity: 58% • Per-lesion sensitivity: 47% • Specificity: 95% • AUC: 0.77 • AUFROC: 0.71 CNN architecture: ResNet34	• Per-lesion sensitivity: 53% • Specificity: 88% • AUC: 0.77 • AUFROC: 0.71	Readers with high diagnostic accuracy AI: • Per-lesion sensitivity: 82% • Specificity: 94% • AUC: 0.82 • AUFROC: 0.79 Readers with low diagnostic accuracy AI (no significant improvement): • Per-lesion sensitivity: 50% • Specificity: 93% • AUC: 0.75 • AUFRC: 0.72	N/A	N/A	• The performance of both AIs reduced significantly as diagnostic difficulty increased • However, high diagnostic accuracy AI maintained better performance across all difficulty	N/A

Table 4. *Cont.*

Author and Year	Performance			Nodule Characteristics (Morphology Type, Size, and Location)		Discrepancies Factors Between AI and Radiologists in Lung Nodule Detection	The Pattern of Missed Lung Nodule Finding
	AI	Radiologist	Combined	AI	Radiologist		
Ueno et al., 2023 [16]	• Sensitivity of AI: 68.9% • Sensitivity of AI: 68.9% • False positive rate of AI: 0.168 Software: EIRL Chest X-ray Lung Nodule (LPIXEL Inc., Tokyo, Japan)	N/A	N/A	• DLAD has a lower detection rate for small-sized nodules and part solid or pure ground-glass nodules (GGNs) • DLAD has a higher detection rate for solid nodules and nodules larger than 15 mm • Reduce detection for medial side of right lower lung, lateral side of right upper lung, and left lower lung area overlapping with heart • Additional factors affecting detection include nodules with a visibility score of two (not well visible)	N/A	Challenges for radiologists: • Manual detection difficulties • Varying nodule visibility • Presence of overlapping structures DLAD detection difficulties: • Small-sized nodules • Subsolid nodules • Nodules with overlapping structures	N/A
Higuchi et al., 2023 [34]	Sensitivity • Type A dataset: 75% • Type B dataset: 72% Specificity • Type A dataset: 60% • Type B dataset: 74% AUC • Type A dataset: 0.74 • Type B dataset: 0.79 CNN architecture: CheXNet model	• Sensitivity: 47.1% • Specificity: 96.4% • AUC: 0.72	N/A	N/A	N/A	A massive teaching dataset is needed to increase the accuracy of the AI	N/A
Farouk et al., 2023 [39]	• Detection rate: 61.3% • AUC per nodule analysis: 0.875 • AUC per CXR classification: 0.890 Software: Lunit INSIGHT CXR version 2.4.11.0 (Lunit Inc., Seoul, Republic of Korea)	• Detection rate: 61.3% • AUC per nodule analysis: 0.772 • AUC per CXR classification: 0.816	• Detection rate: 62% • AUC per nodule analysis: 0.834 • AUC per CXR classification: 0.866	• Detection rate for right middle lung zone: 21.3% • Detection rate for left lower lung zone: 11.2% • High sensitivity of solid nodule: 91.4% (superior to visual and combined visual/AI) • Poor performance on GGNs compares to combined visual or AI	• Detection rate for right middle lung zone: 22.4% (22.1% if combined with AI) • Detection rate for left lower lung zone: 12.9% (11.6% if combined with AI)	AI and radiologists cannot detect nodules at blind or obscured areas on CXR	• AI missed lung nodules in 38.7% of cases • Radiologists missed lung nodules in 38% of cases

Table 4. *Cont.*

Author and Year	Performance			Nodule Characteristics (Morphology Type, Size, and Location)		Discrepancies Factors Between AI and Radiologists in Lung Nodule Detection	The Pattern of Missed Lung Nodule Finding
	AI	Radiologist	Combined	AI	Radiologist		
Tang et al., 2023 [43]	N/A	Sensitivity of radiologists without AI: 64% Specificity of radiologists without AI: 88% AUC of radiologist without AI: 0.82	Sensitivity of radiologist with • UI-A (text-only): 67% • UI-B (text + AI confidence score): 60% • UI-C (text + AI confidence score + image overlay): 63% Specificity of radiologist with • UI-A (text-only): 85% • UI-B (text + AI confidence score): 81% • UI-C (text + AI confidence score + image overlay): 90% AUC of radiologist with • UI-A (text-only): 0.87 (significant improvement, $p < 0.001$) • UI-B (text + AI confidence score): 0.77 (slight decrease, not significant) • UI-C (text + AI confidence score + image overlay): 0.80 (slight decrease, not significant) Software: Annalise Enterprise CXR (Annalise.ai, Sydney, Australia)	N/A	N/A	• Different UIs affect radiologist performance • Radiologist performance varied with different AI output UIs	N/A

Table 4. *Cont.*

Author and Year	Performance			Nodule Characteristics (Morphology Type, Size, and Location)		Discrepancies Factors Between AI and Radiologists in Lung Nodule Detection	The Pattern of Missed Lung Nodule Finding
	AI	Radiologist	Combined	AI	Radiologist		
Bennani et al., 2023 [45]	N/A	Sensitivity without AI • Thoracic radiologist 51.8% • General Radiologist 38.1% • Radiologist residents 36.1% • All readers 42.0% Specificity without AI • Thoracic radiologists 89.2% • General Radiologist 90.8% • Radiologist residents 88.3% • All readers 89.4%	Sensitivity with AI • Thoracic radiologists 55.9% • General Radiologist 55.7% • Radiologist residents 49.7% • All readers 53.8% Specificity with AI • Thoracic radiologists 90.6% • General Radiologist 91.8% • Radiologist residents 92.1% All readers 91.5% Software: ChestView version 1.2.0 (Gleamer, Paris, France)	AI increase detection of small nodules that are usually overlooked	N/A	Reader-AI interaction: • Some correct AI detections were rejected by readers • Possible mistrust of AI suggestions • Reading time: 25 s decrease with AI assistance	• AI missed the left lower lobe nodule (confirmed by CT) • Only one thoracic radiologist detected the missed nodule • AI reported a false-negative for the left lower lobe
Maiter et al., 2023 [48]	Compared with suspicious nodules: • Sensitivity 54.5% • Specificity: 83.2% • PPV: 5.5% • NPV: 99% • Accuracy: 82.7% • FPPI: 0.18 Compared with cancer diagnosis: • Sensitivity 60.9% • Specificity: 83.3% • PPV: 5.6% • NPV: 99.2% • Accuracy: 82.9% • FPPI: 0.18 Software: ALND version 1.0 (Samsung Electronics, Suwon, Republic of Korea)	• Clinical reports sensitivity: 80.0% • Clinical reports specificity: 98% • Clinical reports PPV: 35.7% • Clinical reports NPV: 99.7% • Clinical reports accuracy: 97.5% • FPPI: 0.02	N/A	N/A	N/A	AI challenges: • Underperformance in real-world patient cohort • Lack of generalizability in training/testing datasets • Low PPV may lead to over-investigation and limit clinical practice translation • Demographic differences may affect performance	• AI missed more findings than radiologists in lung nodule detection

Table 4. *Cont.*

Author and Year	Performance			Nodule Characteristics (Morphology Type, Size, and Location)		Discrepancies Factors Between AI and Radiologists in Lung Nodule Detection	The Pattern of Missed Lung Nodule Finding
	AI	Radiologist	Combined	AI	Radiologist		
Hamanaka et al., 2024 [18]	AI detection rate: • 51.7% of all patients (90/174) • 57.7% excluding adenocarcinoma in situ cases Software: EIRL X-ray Lung Nodule version 1.12 (LPIXEL Inc., Tokyo, Japan)	N/A	N/A	• AI sensitivity by lung area: - Lung apices: 60% - Upper lung field: 63.6% - Middle lung field: 78.6% - Lower lung field: 43.8% - Hilar area: 30% • AI sensitivity by tumor diameter: - ≤10 mm: 0 - 11–15 mm: 0.38 - 16–20 mm: 0.52 - ≥21 mm: >83% • Lower lung fields were less frequently detected	N/A	• AI challenges include overlapping anatomical structures, small nodules (<0.7 cm), oversight, and specific tumors such as adenocarcinoma in situ (AIS). • Larger total histopathological tumor size improves AI detection • Physician challenges include detecting smaller tumors (<1 cm) and GGNs • AI is unable to make comparative diagnoses with past radiographs	• AI missed all 18 cases of AIS • AI missed 3 cases due to flat-appearing tumors on radiographs
Kirshenboim et al., 2024 [42]	• Sensitivity: 92% • Specificity: 95% Software: Lunit INSIGHT CXR version 2.0.2.0 (Lunit Inc., Seoul, Republic of Korea)	N/A	• AI improves radiologist performance, especially for less skilled interpreters • Experienced chest radiologists have similar performance with/without AI	N/A	N/A	AI software limitations: • The false interpretation of nodules is due to subtle findings such as nipple shadows, skeletal abnormality, foreign objects, or other non-pulmonary findings • May be insufficient without a formal radiology report • Potential for over-diagnosis or misdiagnosis	• Clinician commented on 386 CXRs (56.5%). • True nodule detection by clinician were 113 CXRs (16.5%) • 31 CXRs (4.5%)—incorrectly identified false nodules as real nodules • 242 CXRs (35%)—clinician no mention of AI-detected nodules • Of these, 68 patients (10%) needed follow-up workup

Table 4. *Cont.*

Author and Year	Performance			Nodule Characteristics (Morphology Type, Size, and Location)		Discrepancies Factors Between AI and Radiologists in Lung Nodule Detection	The Pattern of Missed Lung Nodule Finding
	AI	Radiologist	Combined	AI	Radiologist		
Topff et al., 2024 [46]	N/A	N/A	N/A	• The AI software (ChestEye Quality, Oxipit, Vilnius, Lithuania) identified nodular opacities, the most frequent unreported findings, and solitary lung lesions	N/A	• AI software flagged reports with less descriptive language • AI limitations include no access to the patient's prior imaging, which could hinder the ability to detect missed findings in follow-up examinations	• AI detected discrepancies between imaging and report in 21.1% of cases (5289/25,104) • Clinically relevant missed findings were 0.9% of cases (47/5289) with lung nodules 71.4% (25/35) • Institution-confirmed clinically relevant missed findings were 74.5% (35/47)
van Leeuwen et al., 2024 [47]	• Sensitivity: 64% to 89% • Specificity: 80% to 99% • AUC ranged: 0.86 to 0.93 *** Software used	• Sensitivity: 89% • Specificity: 80% • Mean AUC: 0.81	N/A	• Decrease performance with lower nodule conspicuity • Nodule size had a limited correlation with performance for most algorithms	Decrease performance with lower nodule conspicuity	The difference in image reading platforms compare to clinical practice and absence of clinical information may contribute to discrepancies between AI and radiologists	N/A

Notes: Abbreviation: N/A, Not available; AI, Artificial Intelligence; AUC, Area under the curve; AUROC, Area under the receiver operating characteristic; BSp-DL: Bone suppression deep learning, BSt-DE, Bone subtraction dual energy; CAD, Computer-assisted detection; CI, confidence interval; COPD, Chronic obstructive pulmonary disease; CNN, Convolutional Neural Network; CT, Computed tomography; CXR, Chest X-ray; DCNN, Deep convolutional neural network; DLAD, Deep learning-based automatic detection; DLD, Deep learning detection; DZ, Danger zone; EMR, Electronic medical record; FPPI, False positive per image; FRCR, Fellow of Royal College of Radiologist; GGO, Ground-glass opacity; MDT, Multidisciplinary team; NDZ, Non-danger zone; NPV, Negative predictive value; NLST, National Lung Screening Trial; PACS, Picture archiving and communication system; PET-CT, Positron emission tomography-computed tomography; PPV, Positive predictive value; UI, User interface. *** Software used in this study are as follows: Annalise Enterprise CXR v3.1, Annalise.ai, Sydney, Australia; InferRead DR Chest v1.0.0.1, Infervision, Beijing, China; Lunit INSIGHT CXR v3.1.4.4, Lunit Inc., Seoul, Republic of Korea; Milvue Suite-Smart Urgences v1.24, Milvue, Paris, France; ChestEye v2.6, Oxipit, Vilnius, Lithuania; AI-Rad Companion Chest X-ray v9, Siemens Healthineers, Erlangen, Germany; Med-Chest X-ray v1.1.x, VUNO, Seoul, Republic of Korea.

4. Discussion

The integration of AI in chest radiographs for lung cancer screening plays an important role due to the high global burden of lung cancer and the disparities of traditional screening approaches using LDCT. The evidence presented shows that AI algorithms generally obtain higher sensitivity (ranging from 56.4% to 95.7%) than radiologists (ranging from 23.2% to 76%) while maintaining comparable specificity [13,23,25,26,29,36]. This performance characteristic is highly important in screening settings, where early detection is critical. However, the 44.1% to 95.7% variability in the sensitivity of the AI model [26] is concerning and asks pertinent questions about standardization and reliability. The amount of variability in the AI performance is strongly related to the specific model, implementation context, and characteristics of the data used for training. Therefore, this highlights the need for standardized protocols to assess the AI performance and training datasets that accurately reflect the population to be screened.

The differential performance of AI systems across nodule characteristics has big implications for screening programs. AI performs better than radiologists for larger nodules (i.e., >2 cm) and solid nodules [13,16,24,37] but has considerable variation in performance for different nodule morphology and location. Current AI implementations may miss subtle early-stage cancers that are critical targets in screening programs based on the lower detection rates for GGO nodules [16,33] and nodules in anatomically challenging locations (such as the "danger zone") [24]. Since these subtle findings are mostly early-stage lung cancer with a better prognosis if diagnosed early, this limitation is particularly concerning. Future AI development should, therefore, address these performance gaps to maximize the efficacy of such screening programs in capturing the full range of actionable findings.

The screening context offers both opportunities and challenges in the complementary relationship of AI and radiologists. Using AI as a second reader increased sensitivity by 5.3% to 15% [25,28,33,44,49] and may benefit the screening programs, especially for less experienced readers. Yet, the variable performance across different clinical contexts [26] and levels of experience [13,41,44] implies that thoroughly evaluating the integration of AI into current screening workflows is necessary. While there is potential for AI to improve missed diagnoses, it must be balanced with a risk of over-investigation, particularly given the reported PPV of 5.1% [48]. Future research and policy efforts should seek to develop evidence-based guidelines for AI-assisted screening that are consistent across diverse practice settings.

Current AI systems exhibit technical and practical limitations that may limit screening implementation. The AI system is unable to compare with prior radiographs [18], which is one of the critical features of screening programs. It also draws attention to the poor generalizability of this model to other populations of patients and the effect of demographic variables on AI performance [48]. Our findings demonstrate that successful implementation will require an understanding of both technical and human factors, such as variability in the dependence of performance on the different AI user interfaces [43]. On the other hand, radiologists tend to reject the correct AI detections occasionally [45]. Thus, robust validation studies, user-centered design, and performance monitoring are needed to realize the full potential of AI in lung cancer screening.

The role in lung cancer screening programs looks promising concerning AI but with careful optimization. The high negative predictive value (98.9%) for normal vs abnormal chest radiography [14] may make this a useful triaging tool in high-volume screening settings. Despite this, it is clear that the risk of missed diagnosis, in particular for nodule types and locations [18,33], dictates that AI should supplement, but not replace, the role of the radiologist. Future development in this area should be targeted toward better detection of the subtle findings, better generalizability between populations, and seamless integration

into clinical workflow while maintaining the balance between sensitivity and specificity, which is needed for a successful lung cancer screening program. The transformative potential of AI in the early detection and management of lung cancer will depend on ongoing research, multi-stakeholder collaboration, and evidence-based policy initiatives.

There are several key limitations to this systematic review. The included studies were predominantly retrospective cross-sectional and thus may introduce selection bias and do not include the impact of AI on real-time clinical decision-making. Longitudinal follow-up studies are missing to rate patient outcomes in early cancer detection and mortality reduction. Most of the previous studies adopt the binary classification approach (detection or non-detection) on clinical diagnosis, which may oversimplify this complexity. In addition, there are only limited studies on the effect of AI on radiologist training, a lack of standardized reporting on performance-influencing factors, and few studies on practical clinical implementation difficulties, all of which limit the generalizability and clinical translation of findings.

5. Conclusions

This review shows the great promise of AI algorithms to improve lung nodule detection on chest radiographs. AI systems have better sensitivity than radiologists, especially for smaller or subtle nodules, and can be an excellent instrument in lung cancer screening programs. The impact of how those tools can be deployed in parallel with radiologists and produce a combined effect is what makes AI for radiology important in enhancing diagnostic accuracy and efficiency. Despite this, many obstacles in AI implementation remain, including inconsistency in AI performance across studies and characteristics of nodules studied, variability in evaluation methods used, and a balance between high sensitivity and an acceptable false-positive rate.

More effort should be directed toward future studies constructing screening-oriented AI algorithms that yield more consistent and generalizable results. More long-term comparative investigations of the effectiveness of AI-assisted screening for the patients' outcomes and cost-effectiveness analyses are needed. AI deployment in clinical settings must enhance the value that both human expertise and AI can bring to clinical practices as technology advances. Data privacy and algorithmic bias, among other ethical considerations, should also be effectively addressed. The proposed strategies aim to enhance early detection rates and patients' well-being regarding lung cancer diagnosis by using AI appropriately in radiology practice.

Author Contributions: Conceptualization, P.N.M.R. and A.N.A.; methodology, P.N.M.R., Z.A.H. and K.I.I.; validation, A.N.A., N.A. and Z.A.H.; formal analysis, P.N.M.R.; investigation, P.N.M.R.; resources, P.N.M.R.; data curation, P.N.M.R.; writing—original draft preparation, P.N.M.R.; writing—review, and editing, A.N.A., N.A., Z.A.H. and K.I.I.; visualization, P.N.M.R.; supervision, P.N.M.R., A.N.A. and N.A.; project administration, P.N.M.R. and A.N.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: Not applicable.

Data Availability Statement: The data in this systematic review are primary data obtained from scholarly articles that are cited in this review paper and published within the studies. The extraction and analysis protocols are reported in detail in the Materials and Methods section of the study. Specific additional data or analysis protocols may be requested from the corresponding author.

Acknowledgments: During the preparation of this work, the Petal.ai tool was used to assist with initial data extraction, with all data points manually verified against original source articles by the authors. The authors also used Claude.AI as writing assistance to improve clarity and language presentation. After using this AI tool, authors reviewed and edited the content as needed and took full responsibility for the content of the published article.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AUROC	Area under the receiver operating characteristic
AUC	Area under the curve
CAD	Computer-aided detection
CXR	Chest X-ray
DLAD	Deep learning-based automatic detection
GGO	Ground-glass opacity
LDCT	Low-dose computed tomography
NPV	Negative predictive value
PPV	Positive predictive value

References

1. Xia, C.; Dong, X.; Li, H.; Cao, M.; Sun, D.; He, S.; Yang, F.; Yan, X.; Zhang, S.; Li, N. Cancer statistics in China and United States, 2022: Profiles, trends, and determinants. *Chin. Med. J.* **2022**, *135*, 584–590. [CrossRef] [PubMed]
2. Teh, H.; Woon, Y. Burden of cancers attributable to modifiable risk factors in Malaysia. *BMC Public Health* **2021**, *21*, 410. [CrossRef]
3. Rajadurai, P.; How, S.H.; Liam, C.K.; Sachithanandan, A.; Soon, S.Y.; Tho, L.M. Lung cancer in Malaysia. *J. Thorac. Oncol.* **2020**, *15*, 317–323. [CrossRef]
4. National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.* **2011**, *365*, 395–409. [CrossRef]
5. Dwyer, L.L.; Vadagam, P.; Vanderpoel, J.; Cohen, C.; Lewing, B.; Tkacz, J. Disparities in lung cancer: A targeted literature review examining lung cancer screening, diagnosis, treatment, and survival outcomes in the United States. *J. Racial Ethn. Health Disparities* **2024**, *11*, 1489–1500. [CrossRef] [PubMed]
6. Thames, M.; Liu, J.; Colditz, G.; James, A. Urban–Rural Disparities in Access to Low-Dose Computed Tomography Lung Cancer Screening in Missouri and Illinois. *PCD Staff* **2020**, *17*, E140.
7. Alberto, N.R.I.; Alberto, I.R.I.; Puyat, C.V.M.; Antonio, M.A.R.; Ho, F.D.V.; Dee, E.C.; Mahal, B.A.; Eala, M.A.B. Disparities in access to cancer diagnostics in ASEAN member countries. *Lancet Reg. Health–West. Pac.* **2023**, *32*, 100667. [CrossRef]
8. Sahar, L.; Douangchai Wills, V.L.; Liu, K.K.; Fedewa, S.A.; Rosenthal, L.; Kazerooni, E.A.; Dyer, D.S.; Smith, R.A. Geographic access to lung cancer screening among eligible adults living in rural and urban environments in the United States. *Cancer* **2022**, *128*, 1584–1594. [CrossRef] [PubMed]
9. Ueda, D.; Yamamoto, A.; Shimazaki, A.; Walston, S.L.; Matsumoto, T.; Izumi, N.; Tsukioka, T.; Komatsu, H.; Inoue, H.; Kabata, D. Artificial intelligence-supported lung cancer detection by multi-institutional readers with multi-vendor chest radiographs: A retrospective clinical validation study. *BMC Cancer* **2021**, *21*, 1120. [CrossRef] [PubMed]
10. Jang, S.; Song, H.; Shin, Y.J.; Kim, J.; Kim, J.; Lee, K.W.; Lee, S.S.; Lee, W.; Lee, S.; Lee, K.H. Deep learning–based automatic detection algorithm for reducing overlooked lung cancers on chest radiographs. *Radiology* **2020**, *296*, 652–661. [CrossRef] [PubMed]
11. Chassagnon, G.; Vakalopoulou, M.; Paragios, N.; Revel, M.-P. Artificial intelligence applications for thoracic imaging. *Eur. J. Radiol.* **2020**, *123*, 108774. [CrossRef] [PubMed]
12. Hwang, E.J.; Park, S.; Jin, K.-N.; Im Kim, J.; Choi, S.Y.; Lee, J.H.; Goo, J.M.; Aum, J.; Yim, J.-J.; Cohen, J.G. Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw. Open* **2019**, *2*, e191095. [CrossRef]
13. Nam, J.G.; Hwang, E.J.; Kim, D.S.; Yoo, S.-J.; Choi, H.; Goo, J.M.; Park, C.M. Undetected lung cancer at posteroanterior chest radiography: Potential role of a deep learning–based detection algorithm. *Radiol. Cardiothorac. Imaging* **2020**, *2*, e190222. [CrossRef]

14. Govindarajan, A.; Govindarajan, A.; Tanamala, S.; Chattoraj, S.; Reddy, B.; Agrawal, R.; Iyer, D.; Srivastava, A.; Kumar, P.; Putha, P. Role of an automated deep learning algorithm for reliable screening of abnormality in chest radiographs: A prospective multicenter quality improvement study. *Diagnostics* **2022**, *12*, 2724. [CrossRef] [PubMed]
15. Yoo, H.; Kim, K.H.; Singh, R.; Digumarthy, S.R.; Kalra, M.K. Validation of a deep learning algorithm for the detection of malignant pulmonary nodules in chest radiographs. *JAMA Netw. Open* **2020**, *3*, e2017135. [CrossRef] [PubMed]
16. Ueno, M.; Yoshida, K.; Takamatsu, A.; Kobayashi, T.; Aoki, T.; Gabata, T. Deep learning-based automatic detection for pulmonary nodules on chest radiographs: The relationship with background lung condition, nodule characteristics, and location. *Eur. J. Radiol.* **2023**, *166*, 111002. [CrossRef]
17. Lee, J.H.; Hong, H.; Nam, G.; Hwang, E.J.; Park, C.M. Effect of human-AI interaction on detection of malignant lung nodules on chest radiographs. *Radiology* **2023**, *307*, e222976. [CrossRef] [PubMed]
18. Hamanaka, R.; Oda, M. Can Artificial Intelligence Replace Humans for Detecting Lung Tumors on Radiographs? An Examination of Resected Malignant Lung Tumors. *J. Pers. Med.* **2024**, *14*, 164. [CrossRef]
19. Goncalves, S.; Fong, P.-C.; Blokhina, M. Artificial intelligence for early diagnosis of lung cancer through incidental nodule detection in low-and middle-income countries-acceleration during the COVID-19 pandemic but here to stay. *Am. J. Cancer Res.* **2022**, *12*, 1. [PubMed]
20. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef]
21. Richardson, W.S.; Wilson, M.C.; Nishikawa, J.; Hayward, R.S. The well-built clinical question: A key to evidence-based decisions. *ACP J. Club* **1995**, *123*, A12–A13. [CrossRef] [PubMed]
22. Abouzahra, A.; Sabraoui, A.; Afdel, K. Model composition in Model Driven Engineering: A systematic literature review. *Inf. Softw. Technol.* **2020**, *125*, 106316. [CrossRef]
23. Nam, J.G.; Hwang, E.J.; Kim, J.; Park, N.; Lee, E.H.; Kim, H.J.; Nam, M.; Lee, J.H.; Park, C.M.; Goo, J.M. AI Improves Nodule Detection on Chest Radiographs in a Health Screening Population: A Randomized Controlled Trial. *Radiology* **2023**, *307*, e221894. [CrossRef]
24. You, S.; Park, J.H.; Park, B.; Shin, H.-B.; Ha, T.; Yun, J.S.; Park, K.J.; Jung, Y.; Kim, Y.N.; Kim, M. The diagnostic performance and clinical value of deep learning-based nodule detection system concerning influence of location of pulmonary nodule. *Insights Into Imaging* **2023**, *14*, 149. [CrossRef] [PubMed]
25. Nam, J.G.; Kim, H.J.; Lee, E.H.; Hong, W.; Park, J.; Hwang, E.J.; Park, C.M.; Goo, J.M. Value of a deep learning-based algorithm for detecting Lung-RADS category 4 nodules on chest radiographs in a health checkup population: Estimation of the sample size for a randomized controlled trial. *Eur. Radiol.* **2022**, *32*, 213–222. [CrossRef]
26. Hwang, S.H.; Shin, H.J.; Kim, E.-K.; Lee, E.H.; Lee, M. Clinical outcomes and actual consequence of lung nodules incidentally detected on chest radiographs by artificial intelligence. *Sci. Rep.* **2023**, *13*, 19732. [CrossRef] [PubMed]
27. Huh, J.E.; Lee, J.H.; Hwang, E.J.; Park, C.M. Effects of expert-determined reference standards in evaluating the diagnostic performance of a deep learning model: A malignant lung nodule detection task on chest radiographs. *Korean J. Radiol.* **2023**, *24*, 155. [CrossRef]
28. Kim, H.; Lee, K.H.; Han, K.; Lee, J.W.; Kim, J.Y.; Im, D.J.; Hong, Y.J.; Choi, B.W.; Hur, J. Development and validation of a deep learning–based synthetic bone-suppressed model for pulmonary nodule detection in chest radiographs. *JAMA Netw. Open* **2023**, *6*, e2253820. [CrossRef] [PubMed]
29. Kwak, S.H.; Kim, E.K.; Kim, M.H.; Lee, E.H.; Shin, H.J. Incidentally found resectable lung cancer with the usage of artificial intelligence on chest radiographs. *PLoS ONE* **2023**, *18*, e0281690. [CrossRef]
30. Sim, Y.; Chung, M.J.; Kotter, E.; Yune, S.; Kim, M.; Do, S.; Han, K.; Kim, H.; Yang, S.; Lee, D.-J. Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs. *Radiology* **2020**, *294*, 199–209. [CrossRef] [PubMed]
31. Bae, K.; Oh, D.Y.; Yun, I.D.; Jeon, K.N. Bone suppression on chest radiographs for pulmonary nodule detection: Comparison between a generative adversarial network and dual-energy subtraction. *Korean J. Radiol.* **2022**, *23*, 139. [CrossRef] [PubMed]
32. Lee, J.H.; Sun, H.Y.; Park, S.; Kim, H.; Hwang, E.J.; Goo, J.M.; Park, C.M. Performance of a deep learning algorithm compared with radiologic interpretation for lung cancer detection on chest radiographs in a health screening population. *Radiology* **2020**, *297*, 687–696. [CrossRef] [PubMed]
33. Koo, Y.H.; Shin, K.E.; Park, J.S.; Lee, J.W.; Byun, S.; Lee, H. Extravalidation and reproducibility results of a commercial deep learning-based automatic detection algorithm for pulmonary nodules on chest radiographs at tertiary hospital. *J. Med. Imaging Radiat. Oncol.* **2020**, *65*, 15–22. [CrossRef]
34. Higuchi, M.; Nagata, T.; Iwabuchi, K.; Sano, A.; Maekawa, H.; Idaka, T.; Yamasaki, M.; Seko, C.; Sato, A.; Suzuki, J. Development of a novel artificial intelligence algorithm to detect pulmonary nodules on chest radiography. *Fukushima J. Med. Sci.* **2023**, *69*, 177–183. [CrossRef] [PubMed]

35. Teng, P.-H.; Liang, C.-H.; Lin, Y.; Alberich-Bayarri, A.; González, R.L.; Li, P.-W.; Weng, Y.-H.; Chen, Y.-T.; Lin, C.-H.; Chou, K.-J. Performance and educational training of radiographers in lung nodule or mass detection: Retrospective comparison with different deep learning algorithms. *Medicine* **2021**, *100*, e26270. [CrossRef]
36. Chiu, H.-Y.; Peng, R.H.-T.; Lin, Y.-C.; Wang, T.-W.; Yang, Y.-X.; Chen, Y.-Y.; Wu, M.-H.; Shiao, T.-H.; Chao, H.-S.; Chen, Y.-M. Artificial intelligence for early detection of chest nodules in X-ray images. *Biomedicines* **2022**, *10*, 2839. [CrossRef] [PubMed]
37. Kaviani, P.; Digumarthy, S.R.; Bizzo, B.C.; Reddy, B.; Tadepalli, M.; Putha, P.; Jagirdar, A.; Ebrahimian, S.; Kalra, M.K.; Dreyer, K.J. Performance of a Chest Radiography AI Algorithm for Detection of Missed or Mislabeled Findings: A Multicenter Study. *Diagnostics* **2022**, *12*, 2086. [CrossRef]
38. Majkowska, A.; Mittal, S.; Steiner, D.F.; Reicher, J.J.; McKinney, S.M.; Duggan, G.E.; Eswaran, K.; Cameron Chen, P.-H.; Liu, Y.; Kalidindi, S.R. Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology* **2020**, *294*, 421–431. [CrossRef] [PubMed]
39. Farouk, S.; Osman, A.M.; Awadallah, S.M.; Abdelrahman, A.S. The added value of using artificial intelligence in adult chest X-rays for nodules and masses detection in daily radiology practice. *Egypt. J. Radiol. Nucl. Med.* **2023**, *54*, 142. [CrossRef]
40. Ahn, J.S.; Ebrahimian, S.; McDermott, S.; Lee, S.; Naccarato, L.; Di Capua, J.F.; Wu, M.Y.; Zhang, E.W.; Muse, V.; Miller, B. Association of artificial intelligence–aided chest radiograph interpretation with reader performance and efficiency. *JAMA Netw. Open* **2022**, *5*, e2229289. [CrossRef]
41. Yoo, H.; Lee, S.H.; Arru, C.D.; Doda Khera, R.; Singh, R.; Siebert, S.; Kim, D.; Lee, Y.; Park, J.H.; Eom, H.J.; et al. AI-based improvement in lung cancer detection on chest radiographs: Results of a multi-reader study in NLST dataset. *Eur. Radiol.* **2021**, *31*, 9664–9674. [CrossRef] [PubMed]
42. Kirshenboim, Z.; Gilat, E.K.; Carl, L.; Bekker, E.; Tau, N.; Klug, M.; Konen, E.; Marom, E.M. Retrospectively assessing evaluation and management of artificial-intelligence detected nodules on uninterpreted chest radiographs in the era of radiologists shortage. *Eur. J. Radiol.* **2024**, *170*, 111241. [CrossRef] [PubMed]
43. Tang, J.S.; Lai, J.K.; Bui, J.; Wang, W.; Simkin, P.; Gai, D.; Chan, J.; Pascoe, D.M.; Heinze, S.B.; Gaillard, F. Impact of different artificial intelligence user interfaces on lung nodule and mass detection on chest radiographs. *Radiol. Artif. Intell.* **2023**, *5*, e220079. [CrossRef]
44. Homayounieh, F.; Digumarthy, S.; Ebrahimian, S.; Rueckel, J.; Hoppe, B.F.; Sabel, B.O.; Conjeti, S.; Ridder, K.; Sistermanns, M.; Wang, L. An artificial intelligence–based chest X-ray model on human nodule detection accuracy from a multicenter study. *JAMA Netw. Open* **2021**, *4*, e2141096. [CrossRef]
45. Bennani, S.; Regnard, N.-E.; Ventre, J.; Lassalle, L.; Nguyen, T.; Ducarouge, A.; Dargent, L.; Guillo, E.; Gouhier, E.; Zaimi, S.-H. Using AI to improve Radiologist performance in detection of abnormalities on chest Radiographs. *Radiology* **2023**, *309*, e230860. [CrossRef]
46. Topff, L.; Steltenpool, S.; Ranschaert, E.R.; Ramanauskas, N.; Menezes, R.; Visser, J.J.; Beets-Tan, R.G.; Hartkamp, N.S. Artificial intelligence-assisted double reading of chest radiographs to detect clinically relevant missed findings: A two-centre evaluation. *Eur. Radiol.* **2024**, *34*, 5876–5885. [CrossRef] [PubMed]
47. van Leeuwen, K.G.; Schalekamp, S.; Rutten, M.J.; Huisman, M.; Schaefer-Prokop, C.M.; de Rooij, M.; van Ginneken, B.; Maresch, B.; Geurts, B.H.; van Dijke, C.F. Comparison of commercial AI software performance for radiograph lung nodule detection and bone age prediction. *Radiology* **2024**, *310*, e230981. [CrossRef]
48. Maiter, A.; Hocking, K.; Matthews, S.; Taylor, J.; Sharkey, M.; Metherall, P.; Alabed, S.; Dwivedi, K.; Shahin, Y.; Anderson, E. Evaluating the performance of artificial intelligence software for lung nodule detection on chest radiographs in a retrospective real-world UK population. *BMJ Open* **2023**, *13*, e077348. [CrossRef]
49. Tam, M.; Dyer, T.; Dissez, G.; Morgan, T.N.; Hughes, M.; Illes, J.; Rasalingham, R.; Rasalingham, S. Augmenting lung cancer diagnosis on chest radiographs: Positioning artificial intelligence to improve radiologist performance. *Clin. Radiol.* **2021**, *76*, 607–614. [CrossRef]
50. Peters, A.A.; Decasper, A.; Munz, J.; Klaus, J.; Loebelenz, L.I.; Hoffner, M.K.M.; Hourscht, C.; Heverhagen, J.T.; Christe, A.; Ebner, L. Performance of an AI based CAD system in solid lung nodule detection on chest phantom radiographs compared to radiology residents and fellow radiologists. *J. Thorac. Dis.* **2021**, *13*, 2728. [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.

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 978-3-7258-8434-6