Published in Journals: Cancers, Diagnostics, Journal of Clinical Medicine and Tomography

Topic Reprint

AI in Medical Imaging and Image Processing

Volume II

Edited by
Karolina Nurzynska, Michał Strzelecki, Adam Piórkowski and Rafał Obuchowicz

mdpi.com/topics

MDPI

AI in Medical Imaging and Image Processing—Volume II

AI in Medical Imaging and Image Processing—Volume II

Topic Editors

Karolina Nurzynska
Michał Strzelecki
Adam Piórkowski
Rafał Obuchowicz

MDPI

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

Topic Editors

Karolina Nurzynska
Department of Algorithmics and Software
Silesian University of Technology
Gliwice
Poland

Michał Strzelecki
Institute of Electronics
Lodz University of Technology
Lodz
Poland

Adam Piórkowski
Department of Biocybernetics and Biomedical Engineering
AGH University of Krakow
Krakow
Poland

Rafał Obuchowicz
Department of Diagnostic Imaging
Jagiellonian University Medical College
Krakow
Poland

Editorial Office
MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Topic, published open access by the journals *Cancers* (ISSN 2072-6694), *Diagnostics* (ISSN 2075-4418), *Journal of Clinical Medicine* (ISSN 2077-0383) and *Tomography* (ISSN 2379-139X), freely accessible at: https://www.mdpi.com/topics/P281D89UL6.

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.

Volume II
ISBN 978-3-7258-4515-6 (Hbk)
ISBN 978-3-7258-4516-3 (PDF)
https://doi.org/10.3390/books978-3-7258-4516-3

Set
ISBN 978-3-7258-4517-0 (Hbk)
ISBN 978-3-7258-4518-7 (PDF)

Cover image courtesy of Karolina Nurzynska

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

Contents

About the Editors . . . . . **ix**

Ismail Taskent, Bunyamin Ece and Mehmet Ali Narsat
Are Preoperative CT Findings Useful in Predicting the Duration of Laparoscopic Appendectomy in Pediatric Patients? A Single Center Study
Reprinted from: *J. Clin. Med.* **2024**, *13*, 5504, https://doi.org/10.3390/jcm13185504 . . . . . **1**

Guillaume Herpe, Helena Nelken, Tanguy Vendeuvre, Jeremy Guenezan, Clement Giraud, Olivier Mimoz, et al.
Effectiveness of an Artificial Intelligence Software for Limb Radiographic Fracture Recognition in an Emergency Department
Reprinted from: *J. Clin. Med.* **2024**, *13*, 5575, https://doi.org/10.3390/jcm13185575 . . . . . **15**

Lei Zhang, Rong Zhang, Zhongjie Zhu, Pei Li, Yongqiang Bai and Ming Wang
Lightweight MRI Brain Tumor Segmentation Enhanced by Hierarchical Feature Fusion
Reprinted from: *Tomography* **2024**, *10*, 1577–1590, https://doi.org/10.3390/tomography10100116 **25**

Betül Tiryaki Baştuğ, Gürkan Güneri, Mehmet Süleyman Yıldırım, Kadir Çorbacı and Emre Dandıl
Fully Automated Detection of the Appendix Using U-Net Deep Learning Architecture in CT Scans
Reprinted from: *J. Clin. Med.* **2024**, *13*, 5893, https://doi.org/10.3390/jcm13195893 . . . . . **39**

Matthias Welsner, Henning Navel, Rene Hosch, Peter Rathsmann, Florian Stehling, Annie Mathew, et al.
Opportunistic Screening for Low Bone Mineral Density in Adults with Cystic Fibrosis Using Low-Dose Computed Tomography of the Chest with Artificial Intelligence
Reprinted from: *J. Clin. Med.* **2024**, *13*, 5961, https://doi.org/10.3390/jcm13195961 . . . . . **55**

Mikhail Fufin, Vladimir Makarov, Vadim I. Alfimov, Vladislav V. Ananev and Anna Ananeva
Pulmonary Fissure Segmentation in CT Images Using Image Filtering and Machine Learning
Reprinted from: *Tomography* **2024**, *10*, 1645–1664, https://doi.org/10.3390/tomography10100121 **68**

Emre Dandıl, Betül Tiryaki Baştuğ, Mehmet Süleyman Yıldırım, Kadir Çorbacı and Gürkan Güneri
MaskAppendix: Backbone-Enriched Mask R-CNN Based on Grad-CAM for Automatic Appendix Segmentation
Reprinted from: *Diagnostics* **2024**, *14*, 2346, https://doi.org/10.3390/diagnostics14212346 . . . . **88**

Irina Cristina Barca, Vasile Potop and Stefan Sorin Arama
Monitoring Progression in Hypertensive Patients with Dyslipidemia Using Optical Coherence Tomography Angiography: Can A.I. Be Improved?
Reprinted from: *J. Clin. Med.* **2024**, *13*, 7584, https://doi.org/10.3390/jcm13247584 . . . . . **109**

Jinnian Zhang, Weijie Chen, Tanmayee Joshi, Xiaomin Zhang, Po-Ling Loh, Varun Jog, et al.
BAE-ViT: An Efficient Multimodal Vision Transformer for Bone Age Estimation
Reprinted from: *Tomography* **2024**, *10*, 2058–2072, https://doi.org/10.3390/tomography10120146 **132**

Andrea Craus-Miguel, Marc Munar, Gabriel Moyà-Alcover, Ana María Contreras-Nogales, Manuel González-Hidalgo and Juan José Segura-Sampedro
Enhancing Surgical Wound Monitoring: A Paired Cohort Study Evaluating a New AI-Based Application for Automatic Detection of Potential Infections
Reprinted from: *J. Clin. Med.* **2024**, *13*, 7863, https://doi.org/10.3390/jcm13247863 . . . . . **147**

Yujin Eom, Yong-Jin Park, Sumin Lee, Su-Jin Lee, Young-Sil An, Bok-Nam Park and Joon-Kee Yoon
Automated Measurement of Effective Radiation Dose by ^{18}F-Fluorodeoxyglucose Positron Emission Tomography/Computed Tomography
Reprinted from: *Tomography* **2024**, *10*, 2144–2157, https://doi.org/10.3390/tomography10120151 **161**

Joe Hasei, Ryuichi Nakahara, Yujiro Otsuka, Yusuke Nakamura, Kunihiro Ikuta, Shuhei Osaki, et al.
The Three-Class Annotation Method Improves the AI Detection of Early-Stage Osteosarcoma on Plain Radiographs: A Novel Approach for Rare Cancer Diagnosis
Reprinted from: *Cancers* **2025**, *17*, 29, https://doi.org/10.3390/cancers17010029 . . . . . **175**

Matteo Ferrante, Paolo De Marco, Osvaldo Rampado, Laura Gianusso and Daniela Origgi
Effective Dose Estimation in Computed Tomography by Machine Learning
Reprinted from: *Tomography* **2025**, *11*, 2, https://doi.org/10.3390/tomography11010002 . . . . . **190**

Mustafa Koyun, Zeycan Kubra Cevval, Bahadir Reis and Bunyamin Ece
Detection of Intracranial Hemorrhage from Computed Tomography Images: Diagnostic Role and Efficacy of ChatGPT-4o
Reprinted from: *Diagnostics* **2025**, *15*, 143, https://doi.org/10.3390/diagnostics15020143 . . . . . **205**

Ali Yurtseven, Aleksandar Janjic, Mehmet Cayoren, Onur Bugdayci, Mustafa Erkin Aribal and Ibrahim Akduman
XGBoost Enhances the Performance of SAFE: A Novel Microwave Imaging System for Early Detection of Malignant Breast Cancer
Reprinted from: *Cancers* **2025**, *17*, 214, https://doi.org/10.3390/cancers17020214 . . . . . **222**

Mustafa Koyun and Ismail Taskent
Evaluation of Advanced Artificial Intelligence Algorithms' Diagnostic Efficacy in Acute Ischemic Stroke: A Comparative Analysis of ChatGPT-4o and Claude 3.5 Sonnet Models
Reprinted from: *J. Clin. Med.* **2025**, *14*, 571, https://doi.org/10.3390/jcm14020571 . . . . . **242**

Rafał Obuchowicz, Julia Lasek, Marek Wodziński, Adam Piórkowski, Michał Strzelecki and Karolina Nurzynska
Artificial Intelligence-Empowered Radiology—Current Status and Critical Review
Reprinted from: *Diagnostics* **2025**, *15*, 282, https://doi.org/10.3390/diagnostics15030282 . . . . . **257**

Josue Luiz Dalboni da Rocha, Jesyin Lai, Pankaj Pandey, Phyu Sin M. Myat, Zachary Loschinskey, Asim K. Bag and Ranganatha Sitaram
Reply to Keles et al. Comment on "Dalboni da Rocha et al. Artificial Intelligence for Neuroimaging in Pediatric Cancer. *Cancers* 2025, *17*, 622"
Reprinted from: *Cancers* **2025**, *17*, 1778, https://doi.org/10.3390/cancers17111778 . . . . . **285**

Josue Luiz Dalboni da Rocha, Jesyin Lai, Pankaj Pandey, Phyu Sin M. Myat, Zachary Loschinskey, Asim K. Bag and Ranganatha Sitaram
Artificial Intelligence for Neuroimaging in Pediatric Cancer
Reprinted from: *Cancers* **2025**, *17*, 622, https://doi.org/10.3390/cancers17040622 . . . . . **286**

Hreedi Dev, Zhongxiu Hu, Jon D. Blumenfeld, Arman Sharbatdaran, Yelynn Kim, Chenglin Zhu, et al.
The Role of Baseline Total Kidney Volume Growth Rate in Predicting Tolvaptan Efficacy for ADPKD Patients: A Feasibility Study
Reprinted from: *J. Clin. Med.* **2025**, *14*, 1449, https://doi.org/10.3390/jcm14051449 . . . . . **315**

Julia Lasek, Karolina Nurzynska, Adam Piórkowski, Michał Strzelecki and Rafał Obuchowicz
Deep Learning for Ultrasonographic Assessment of Temporomandibular Joint Morphology
Reprinted from: *Tomography* **2025**, *11*, 27, https://doi.org/10.3390/tomography11030027 . . . . . **328**

Muhammed Fikret Yalcinbas, Cengizhan Ozturk, Uzay E. Emir, Onur Ozyurt and Ulas Bagci
Rosette Trajectory MRI Reconstruction with Vision Transformers
Reprinted from: *Tomography* **2025**, *11*, 41, https://doi.org/10.3390/tomography11040041 . . . . . **344**

Elif Keles, Mehmet Numan Colakoglu and Max Bengtsson
Comment on Dalboni da Rocha et al. Artificial Intelligence for Neuroimaging in Pediatric Cancer. *Cancers* 2025, *17*, 622
Reprinted from: *Cancers* **2025**, *17*, 1776, https://doi.org/10.3390/cancers17111776 . . . . . **358**

Emre Haylaz, Ismail Gumussoy, Suayip Burak Duman, Fahrettin Kalabalik, Muhammet Can Eren, Mustafa Sami Demirsoy, et al.
Automatic Segmentation of the Nasolacrimal Canal: Application of the nnU-Net v2 Model in CBCT Imaging
Reprinted from: *J. Clin. Med.* **2025**, *14*, 778, https://doi.org/10.3390/jcm14030778 . . . . . **360**

About the Editors

Karolina Nurzynska

Karolina Nurzyńska obtained a PhD in computer science in 2009 and a post-doctoral degree in 2019. Since then, she has held the position of professor at the Silesian University of Technology. She collaborates with national and international teams, tackling research topics concerning the application of artificial intelligence methods in image understanding. Her main interest involves automating medical image analysis and determining techniques that could simplify and assure repeatability in medical data understanding. However, the design of chatbots dedicated to communication within closed topics has become one of her new interests. In her free time, she loves traveling and spending time in nature.

Michał Strzelecki

Prof. Michał Strzelecki has been employed at the Institute of Electronics, Lodz University of Technology (TUL), since 1988, and since 2007, he has been a Full Professor. In the years 2004–2019, he was the Deputy Director of the Institute for Teaching and Research. He also served as the Vice Dean for Research at the Faculty of Electrical, Electronic, Computer and Control Engineering, TUL, until 2020. His scientific interests include the processing and analysis of biomedical signals and images, data analysis methods, and artificial intelligence. In particular, he is interested in issues related to image texture analysis and the applications and hardware implementations of artificial neural networks. He also works on the development of software aimed at supporting medical imaging diagnostics. Within these interests, Prof. Strzelecki conducts extensive cooperation with domestic and foreign medical universities on the construction and development of systems supporting the diagnostic process. He has participated in numerous international research programs, including COST B21, BM1103, TD1007 (a member of the Management Committee), DGF, 7FP, and projects funded by the National Science Centre (Poland) and the National Centre for Research and Development (Poland). He is the author and co-author of more than 200 scientific publications. In 2006–2008, he was employed at the Jeonbuk National University, Jeonju, Republic of Korea, as a Visiting Professor.

Adam Piórkowski

Adam Piórkowski obtained a PhD in computer science in 2005. In 2015, he obtained a post-doctoral degree in biomedical engineering, and in 2019, he became a professor at AGH in Krakow. He has been involved in medical image processing since 2003, where his particular interests are the segmentation, normalization, and textural analysis of visual data. He conducts scientific projects generally related to medical informatics. He is also an active programmer.

Rafał Obuchowicz

Rafał Obuchowicz currently works in the Department of Radiology of the University Hospital, Krakow, and heads the Imaging Diagnostics department at Lux Med, Krakow. His professional interests are focused on the problems of focal change detection and shape detection in various radiological techniques using classical methods and deep learning machine algorithms. The result of the work undertaken in this direction is the NCBiR project "XR-AI A diagnostic browser for radiology with computer-aided use of artificial intelligence," of which he is the originator and head of the medical team. He also has a professional interest in analyzing image quality using MR and X-ray techniques, resulting in the completed habilitation procedure entitled "Analysis and parametrization

of information in selected techniques of radiological diagnostics." His hobbies include environmental psychology, history, ecology, and urbanization, and he is the husband of Barbra and father of Isabella and Wiktor.

Journal of
Clinical Medicine

MDPI

Article

Are Preoperative CT Findings Useful in Predicting the Duration of Laparoscopic Appendectomy in Pediatric Patients? A Single Center Study

Ismail Taskent [1,*], Bunyamin Ece [1] and Mehmet Ali Narsat [2]

[1] Department of Radiology, Kastamonu University, 37150 Kastamonu, Turkey; bunyaminece@kastamonu.edu.tr
[2] Department of Pediatric Surgery, Kastamonu University, 37150 Kastamonu, Turkey; manarsat@kastamonu.edu.tr
* Correspondence: itaskent@kastamonu.edu.tr

Abstract: Background/Objectives: Preoperative computed tomography (CT) imaging plays a vital role in accurately diagnosing acute appendicitis and assessing the severity of the condition, as well as the complexity of the surgical procedure. CT imaging provides detailed information on the anatomical and pathological aspects of appendicitis, allowing surgeons to anticipate technical challenges and select the most appropriate surgical approach. This retrospective study aimed to investigate the correlation between preoperative CT findings and the duration of laparoscopic appendectomy (LA) in pediatric patients. **Methods:** This retrospective study included 104 pediatric patients diagnosed with acute appendicitis via contrast-enhanced CT who subsequently underwent laparoscopic appendectomy (LA) between November 2021 and February 2024. CT images were meticulously reviewed by two experienced radiologists blinded to the clinical and surgical outcomes. The severity of appendicitis was evaluated using a five-point scale based on the presence of periappendiceal fat, fluid, extraluminal air, and abscesses. **Results:** The average operation time was 51.1 $\pm$ 21.6 min. Correlation analysis revealed significant positive associations between operation time and neutrophil count ($p = 0.014$), C-reactive protein levels ($p = 0.002$), symptom-to-operation time ($p = 0.004$), and appendix diameter ($p = 0.017$). The total CT score also showed a significant correlation with operation time ($p < 0.001$). Multiple regression analysis demonstrated that a symptom duration of more than 2 days ($p = 0.047$), time from CT to surgery ($p = 0.039$), and the presence of a periappendiceal abscess ($p = 0.005$) were independent predictors of prolonged operation time. In the perforated appendicitis group, the presence of a periappendiceal abscess on CT was significantly associated with prolonged operation time ($p = 0.020$). In the non-perforated group, the presence of periappendiceal fluid was significantly related to longer operation times ($p = 0.026$). **Conclusions:** In our study, preoperative CT findings, particularly the presence of a periappendiceal abscess, were significantly associated with prolonged operation times in pediatric patients undergoing laparoscopic appendectomy. Elevated CRP levels, the time between CT imaging and surgery, and a symptom duration of more than 2 days were also found to significantly impact the procedure's duration.

Keywords: acute appendicitis; pediatric; CT findings; laparoscopic surgery; operation time; predictors

1. Introduction

Acute appendicitis, a common cause of abdominal pain in children, often necessitates swift surgical intervention for accurate diagnosis [1]. The utilization of preoperative computed tomography (CT) imaging has notably enhanced diagnostic precision, particularly in cases where clinical and laboratory findings are inconclusive, thus minimizing unnecessary surgeries [2,3]. CT plays a pivotal role in accurately diagnosing acute appendicitis and holds the capability to predict histological severity. This capability facilitates informed decision-making regarding the timing and necessity of operative intervention [4].

Either laparoscopic or open surgery is the preferred surgical approach to treat pediatric patients with acute appendicitis. However, laparoscopic appendectomy (LA) holds several advantages over the open approach. These include shorter recovery times, reduced postoperative pain, a faster return to normal activities, and a more favorable cosmetic outcome [5–9]. However, operation times have been observed to be longer in patients who underwent LA compared to those who had open appendectomy (OA) [5,10].

Operation time is a reliable measure of technical difficulty in laparoscopic procedures, and research shows a link between prolonged operation times and increased complication rates [11–13]. Imaging findings, particularly CT scans, have an important role in predicting surgery difficulty and outcomes. CT can easily detect the presence of periappendiceal fluid, extraluminal air, and abscesses, as well as the diameter and position of the appendix, which are crucial in determining the surgical difficulty. These imaging results may enhance surgical outcomes by allowing surgeons to anticipate difficulties and plan accordingly [14,15]. To the best of our knowledge, there are not enough studies demonstrating the relationship between prolonged operation time and CT findings in LA in the pediatric population.

The purpose of this retrospective study is to investigate the relationship between preoperative CT imaging findings and operation time for LA in pediatric patients.

2. Materials and Methods

2.1. Participants

Pediatric patients who presented to our hospital with acute abdomen symptoms and were suspected of having acute appendicitis but could not be diagnosed with ultrasound, requiring a contrast-enhanced CT, between November 2021 and February 2024, were retrospectively evaluated. All surgeries were performed by a single pediatric surgery specialist at our hospital. During this period, a total of 138 patients were diagnosed with appendicitis via CT, and 127 of them underwent surgical intervention, while the remaining patients were treated with non-surgical methods such as drainage and antibiotics. Out of 386 total appendectomy patients, 127 underwent preoperative CT evaluation. The remaining patients were evaluated using ultrasound imaging. Patients who underwent non-contrast CT (n = 8), those whose CT records were inaccessible (n = 9), and those with incomplete medical records (n = 4) were excluded from this study. Additionally, 2 patients with significant comorbidities that could have affected the operation time were excluded from this study. As a result, a total of 104 patients diagnosed with acute appendicitis on contrast-enhanced CT and underwent LA were included in this study.

Ethical approval was obtained from the institutional review board of our tertiary care center (ethics committee number: KAEK 2022-104, decision date: 19 October 2022) for the retrospective review of laboratory findings and CT scans in patients who underwent LA. This study adhered to the "Declaration of Helsinki". Because this study was retrospective, informed consent was not obtained.

2.2. Data Collection

The data were obtained from the surgical records, anesthesia records, and progress notes from the pediatric surgery clinic at our hospital. Perforated appendicitis was defined as the spillage of appendiceal contents, peritonitis, or abscesses observed at the beginning of surgery. Operation time was defined as the duration from the initiation of the skin incision to the completion of skin closure, based on anesthesia records. Patients were divided into two groups according to operation time: those with operation times $\leq$ 50 min and those with operation times > 50 min. Preoperative parameters included in the analysis were age, gender, body mass index (BMI), white blood cell (WBC) count, neutrophil count (NEU), NEU/WBC ratio, C-reactive protein (CRP) level (normal range in our lab: 0–5 mg/L), heart rate, body temperature, symptom duration (symptom-to-operation time), waiting time (CT to operation time), and findings from CT. To assess the relationship between operation time and outcomes, the complications, readmissions, and length of hospital stay (LOS) for both groups were compared.

2.3. Abdominal CT Assessment

All CT investigations were conducted using 64-detector, 128-slice CT scanners (Revolution EVO; GE Medical Systems, Chicago, IL, USA). All examinations were performed with a low-dose technique utilizing an automated tube current modulation to determine the tube current (mA). All CT scans were conducted at 80–120 kV, with adjustments made automatically based on the patient's physique. The standard section slice thickness was 0.625 mm. The contrast material used (300 mg iodine/mL) amounted to 1 mL/kg. The contrast material was administered automatically. CT examinations were consistently conducted in the portal venous phase.

To obtain a consensus, two radiologists (I.T. with 11 years and B.E. with 10 years of experience) assessed the patients' CT scans using an Advantage Windows workstation (ADW 4.7 Ext. 16 Software, GE Medical Systems, Chicago, IL, USA). The radiologists were not aware of the clinical or surgical results (Figure 1).

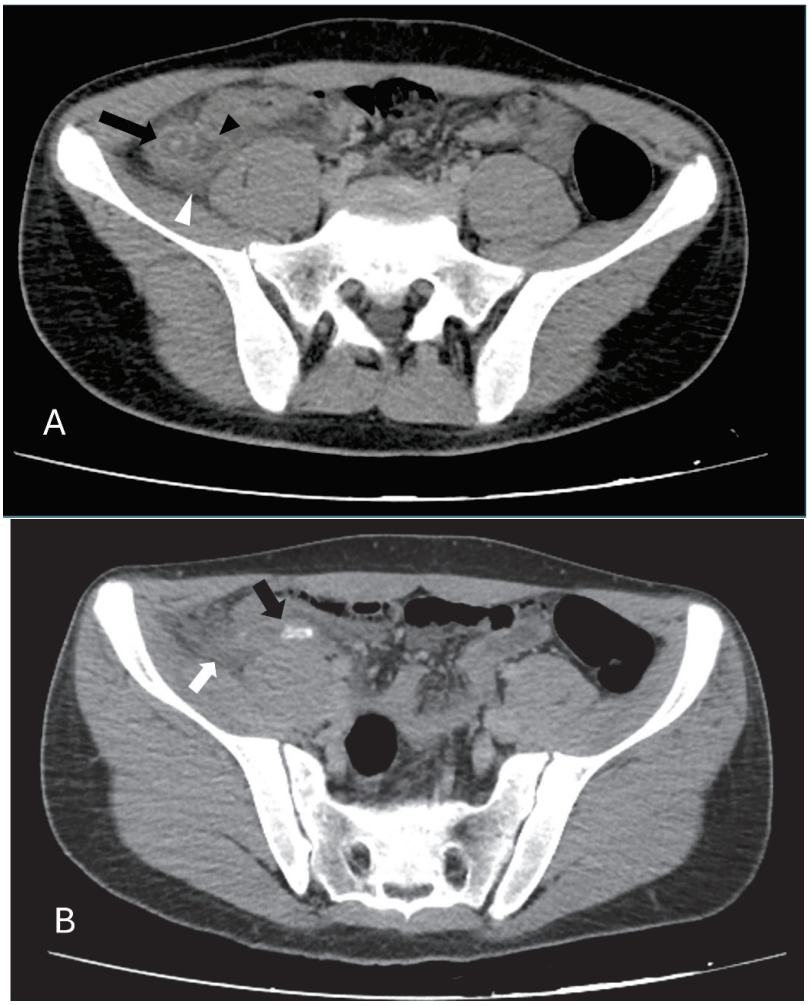


Figure 1. Axial CT images of a 189-month-old male patient. (**A**) The appendix (black arrow) exhibits an increased diameter and a thickened wall, indicative of inflammation. There is a notable collection of periappendiceal fluid (black arrowhead), and the surrounding fat tissue shows increased attenuation (white arrowhead), suggesting inflammatory changes. (**B**) In another section, an appendicolith (black arrow) is clearly visible. Additionally, there is increased attenuation in the periappendiceal fat tissue (white arrow), consistent with inflammatory changes associated with acute appendicitis.

To classify the localizations, three regions were defined in the appendix. An anterior location (where the end of appendix lies anterior to cecum in the large pelvis) was labeled as "Location 1", a retrocecal or retrocolonic location (where the tip of appendix lies behind the

cecum, in the right iliac fossa or reaches the subhepatic area) was labeled as "Location 2", and a pelvic location (where the tip of the appendix lies in the small pelvis) was labeled as "Location 3" [16].

Appendicitis findings were graded based on imaging features observed in the periappendix region. A 5-point scale was devised, considering the presence of an appendicolith, periappendiceal fat stranding, periappendiceal fluid, extraluminal air, and abscess finding. The presence of any of these findings contributed 1 point to the rating, which ranged from 0 to 5 [17].

2.4. Surgical Procedure

All surgeries were performed by the same surgeon with the patient under general anesthesia. A 10 mm optic port was inserted into the umbilicus or subumbilical position, and an additional 5 mm working port was inserted into the left iliac fossa position. A needle grasper was introduced into the abdomen from the right iliac fossa, or a suprapubic or right subcostal 5 mm trocar was used, depending on the location and clinical appearance of the appendix during the initial exploration.

Once a pneumoperitoneum pressure of 8–15 mm Hg was established using CO_2 at a flow rate of 3–5 L per minute, the patient was positioned in a 30° Trendelenburg position with a 15° left tilt. The mesoappendix was dissected using laparoscopic coagulating shears, and the base of the appendix was ligated with endo-clips. After the appendix was divided above the ligated site using laparoscopic coagulating shears, the resected appendix was extracted through the transumbilical port. Subsequently, drainage catheters were inserted through the suprapubic port site into the pelvic cavity or paracolic gutter, depending on the operative findings related to peritoneal contamination from appendiceal perforation and the surgeon's judgment.

No patients required blood transfusion during the perioperative period. For postoperative antibiotic prophylaxis, second-generation cephalosporin and metronidazole were administered.

2.5. Statistical Analysis

The data were analyzed using the Statistical Package for the Social Sciences (SPSS) for Windows version 23 software (IBM SPSS Inc., Chicago, IL, USA). Normal distribution of the data was assessed using the Kolmogorov–Smirnov test. Numerical variables with a normal distribution are presented as mean ± standard deviation (SD) values, while variables without a normal distribution are presented as median (minimum–maximum) values. Categorical variables are reported as the number (n) and percentage (%). The Chi-square test was employed to compare categorical variables. For group comparisons, the independent samples t-test was used for data with a normal distribution, and the Mann–Whitney U test was used for data without a normal distribution. Pearson correlation analysis was utilized for data with a normal distribution, and Spearman correlation analysis was applied for data without a normal distribution. In univariate analysis, simple linear regression was used to evaluate the relationship between the operation time and each independent variable. A multivariable linear regression analysis was performed to assess the impact of various independent variables on operation time. A significance level of $p < 0.05$ was considered statistically significant.

3. Results

This study included 59 (56.7%) males and 45 (43.3%) females, with an average age of 11.7 ± 4.1 years (range: 2.8 to 18 years). The average surgery duration was 51.1 ± 21.6 min. The average time from symptom onset to operation was 32.8 ± 33 h, while from CT scan to operation was 9.4 ± 10.3 h. Appendicoliths were discovered in 55 (52.9%) instances, periappendiceal fat stranding in 86 (82.7%) cases, extraluminal air in 21 (20.2%), ascites in 42 (40.4%) cases, and symptoms of abscess in 15 (14.4%) patients. The average appendix diameter was 11.0 ± 3.2 mm, with females measuring 10.4 ± 3.0 mm and males measuring

11.5 ± 3.2 mm. The appendix vermiformis was classified as follows: just below the anterior abdominal wall (Location 1) in 39 (37.5%) patients, in the retrocecal or retro-ascending colon (Location 2) in 26 (25.0%) patients, and in the pelvis covered by the small intestine (Location 3) in 39 (37.5%) patients. Postoperative complications included wound infections in seven (6.7%) patients and a single case (1.0%) of postoperative abscess. Three patients (2.9%) required readmission. The mean length of stay (LOS) in the hospital was 3.75 ± 1 days, and the mean body temperature was 37.61 ± 0.5 °C. The average heart rate was recorded at 90.42 ± 9.73 beats per minute (BPM) (Table 1).

Table 1. The demographic and clinical data of the patients.

Variables	Value
Number of patients, *n*	104
Sex, *n*, (M/F)	59/45
Age (month)	140.9 ± 49.3
BMI (kg/m^2)	20.7 ± 4.4
WBC (per μL)	15.2 ± 5.3
NEU (per μL)	12.2 ± 5.3
CRP (mg/L)	68.3 ± 78.3
Symptom–operation time (h)	32.8 ± 33.0
CT–operation time (h)	9.4 ± 10.3
Operation time (min)	51.1 ± 21.6
Appendix diameter (mm)	11.0 ± 3.0
BPM	90.4 ± 9.7
Temperature (°C)	37.6 ± 0.5
LOS (days)	3.75 ± 1.00
Readmission, *n* (%)	3 (2.9)
Postop wound infection, *n* (%)	7 (6.7)
Postop abscess, *n* (%)	1 (1)

M: male; F: female; BMI: body mass index; WBC: white blood cell count; NEU: neutrophil; CRP: C-reactive protein; BPM: beats per minute, CT: computed tomography, LOS: length of hospital stay.

The study population had an average age of 140.9 ± 49.3 months. The average BMI was 20.7 ± 4.4. The WBC and NEU counts were 15.2 ± 5.3 and 12.2 ± 5.3 μL, respectively, indicating usual inflammatory responses. Participants had different amounts of inflammation, with an average CRP level of 68.3 ± 78.3 mg/L. The average time between symptom to operation is 32.8 ± 33.0 h, and the average time from CT to operation is 9.5 ± 10.3. The comparison of demographic and clinical data based on gender indicates that no statistically significant differences were observed ($p > 0.05$) (Table 2).

Table 2. The demographic and clinical data of the patients by gender.

Variables	Female (Mean ± SD)	Male (Mean ± SD)	*p*-Value
Age (month)	144.6 ± 48.4	138.1 ± 50.2	0.508
BMI (kg/m^2)	19.7 ± 4.3	21.4 ± 4.4	0.054
WBC (per μL)	15.6 ± 5.8	14.9 ± 5.0	0.511
NEU (per μL)	12.2 ± 5.5	12.2 ± 5.2	0.970
CRP (mg/L)	77.8 ± 93.2	61.1 ± 64.7	0.283
Symptom—operation time (h)	36.8 ± 37.1	29.8 ± 29.5	0.293
CT–operation time (h)	9.9 ± 12.9	9.2 ± 7.9	0.740
Operation time (min)	50.8 ± 26.7	51.2 ± 16.9	0.929
Appendix diameter (mm)	10.3 ± 2.9	11.5 ± 3.2	0.066
BPM	89.6 ± 8.1	91.0 ± 10.8	0.466
Temperature (°C)	37.60 ± 0.46	37.61 ± 0.48	0.942
LOS (days)	3.88 ± 1.07	3.66 ± 0.92	0.247

BMI: body mass index; WBC: white blood cell; NEU: neutrophil; CRP: C-reactive protein; CT: computed tomography; SD: standard deviation, BPM: beats per minute, LOS: length of hospital stay.

According to the correlation analysis results presented in Table 3, the relationships between operation time and continuous variables were evaluated. Analysis results revealed that some variables showed statistically significant effects on operation time. NEU ($p = 0.014$), CRP levels ($p = 0.002$), time from symptom onset to operation ($p = 0.004$), and appendicitis diameter ($p = 0.017$) were positively significant between operation time. On the other hand, other variables such as age, height, weight, BMI, WBC, BPM, temperature, and time from CT to operation did not show a significant relationship with operation time ($p > 0.05$) (Table 3).

Table 3. Correlation analysis of operation time according to continuous variables.

Variables	Correlation Coefficient (*r*)	*p* Value
Age	−0.095	0.336
Height	−0.121	0.222
Weight	−0.034	0.731
BMI	0.007	0.945
WBC	0.106	0.284
NEU	0.240	0.014
CRP	0.295	0.002
Symptom–operation time	0.282	0.004
CT–operation time	0.135	0.173
Appendix diameter	0.234	0.017
BPM	0.031	0.758
Temperature	0.041	0.681
LOS	0.125	0.207
Total score	0.415	<0.001

BMI: body mass index; WBC: white blood cell; NEU: neutrophil; CRP: C-reactive protein; CT: computed tomography, BPM: beats per minute, LOS: length of hospital stay.

The total score, ranging from zero to five, showed a significant and positive correlation with operation time in the correlation analysis (correlation coefficient = 0.415, $p < 0.001$, Figure 2).

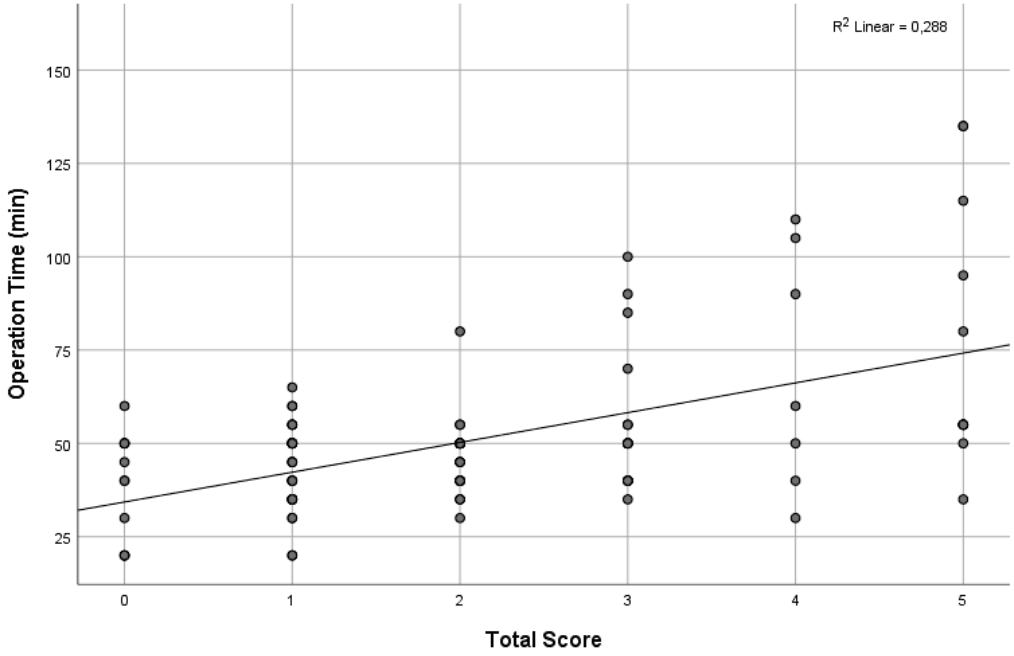


Figure 2. The impact of a scoring system comprising five variables (0 to 5) from CT findings on average operation time (min).

In the analysis, to evaluate the effects of removing specific components from the model on operation time, the results showed that the inclusion or exclusion of certain variables significantly impacted model performance. When all components were included, the model showed an R^2 of 0.359, indicating that 35.9% of the variance in operation time was explained by the predictors, with an F-value of 11.001 ($p < 0.001$). Removing the abscess variable resulted in a drop in R^2 to 0.271 and an increased standard error of the estimate (SEE) to 18.883. Further removals of extraluminal air and periappendicular fluid continued to decrease the model's explanatory power, with the removal of periappendicular fluid lowering R^2 to 0.083 and the F-value to 4.575 ($p = 0.013$). Finally, removing periappendicular fat-stranding reduced R^2 to 0.029, with the model's performance becoming marginally significant ($F = 3.012$, $p = 0.086$). Based on the analysis, the abscess variable is the most important factor in the model, and its removal significantly decreases the model's explanatory power and performance (Table 4).

Table 4. Effect of component removal on model performance for operation time.

Component Removal	R^2	Adj. R^2	SEE	F-Value	*p*-Value
All components included	0.359	0.327	17.739	11.001	<0.001
Abscess removed	0.271	0.241	18.883	9.186	<0.001
Extraluminal air removed	0.224	0.200	19.334	9.601	<0.001
Periappendicular fluid removed	0.083	0.065	20.907	4.575	0.013
Periappendicular fat-stranding removed	0.029	0.019	21.413	3.012	0.086

R^2: coefficient of determination, Adj. R^2: adjusted R^2, SEE: standard error of the estimate.

The relationships between categorical and continuous variables with operation time were evaluated with univariate and multivariable analysis, and the results are presented in Tables 4 and 5. The results revealed that the presence of periappendiceal fat stranding, periappendiceal fluid, extraluminal air, and an abscess had a substantial impact on operation time. However, no significant association was found between appendicolitis, appendix location with operation time. CRP levels, symptom-to-operation time, CT-to-operation time, and appendix diameter showed a significant correlation with operation time. However, other factors such as age, BMI, WBC, and NEU did not have a significant impact on operation time (Tables 5 and 6).

Our analysis of the relationship between symptom duration and operation time revealed that symptom duration does not have a linear effect on operation time. The second-degree polynomial model and LOESS method demonstrated significant fluctuations in operation time as symptom duration increases, indicating a nonlinear relationship (Figure 3). Based on this, we divided the patients into two groups: those with a symptom duration of 2 days or less and those with more than 2 days.

According to the multiple regression analysis results presented in Table 6, the effects of various variables on operation time were examined. The results show that a symptom duration exceeding 2 days ($p = 0.047$) and the time from CT to operation ($p = 0.039$) had a significant effect on operative time. Additionally, the presence of an abscess ($p = 0.005$) also significantly affects the operation time. Other variables, age, appendix diameter, BMI, NEU/WBC, CRP, presence of appendicolith, periappendiceal fat stranding, periappendiceal fluid, extraluminal air, and appendix localization, did not have a significant effect on the operation time ($p > 0.05$). The overall explanatory ratio R^2 value of the model was "0.433", indicating that the independent variables have a certain effect on the operation time (Table 7).

Table 5. Results of the univariate and multivariable analysis of the association between the operation time and categorical variables.

Variables	Operation Time (min) (Mean ± SD)	Univariate Analysis *p*-Value	Multivariable Analysis Coefficient (95% CI)	*p*-Value
Appendicolith				
Absent (*n* = 49)	47.2 ± 18.5	0.165	7.30 (−1.04 to 15.64)	0.086
Present (*n* = 55)	54.5 ± 23.7			
Periappendiceal fat standing				
Absent (*n* = 18)	38.9 ± 13.3	0.007	14.77 (3.98–25.56)	0.008
Present (*n* = 86)	53.7 ± 22.2			
Periappendiceal fluid				
Absent (*n* = 62)	43.2 ± 10.1	<0.001	19.71 (12.01–27.40)	<0.001
Present (*n* = 42)	62.9 ± 28.0			
Extraluminal air				
Absent (*n* = 83)	46.3 ± 14.2	<0.001	23.97 (14.55–33.38)	<0.001
Present (*n* = 21)	70.2 ± 33.2			
Abscess				
Absent (*n* = 89)	46.2 ± 13.7	< 0.001	33.76 (23.72–43.79)	<0.001
Present (*n* = 15)	80.0 ± 34.9			
Appendix localization				
Location 1 (*n* = 39)	48.1 ± 16.7			
Location 2 (*n* = 26)	48.7 ± 22.0	0.235	0.57 (−10.23 to 11.38)	0.916
Location 3 (*n* = 39)	55.8 ± 25.2			

Localization of the appendix was classified using three locations: Location 1: just below the anterior abdominal wall; Location 2: in the retrocecal or retro-ascending colon; Location 3: in the pelvis, on the ventral side covered by the small intestine. (R^2: 0.359).

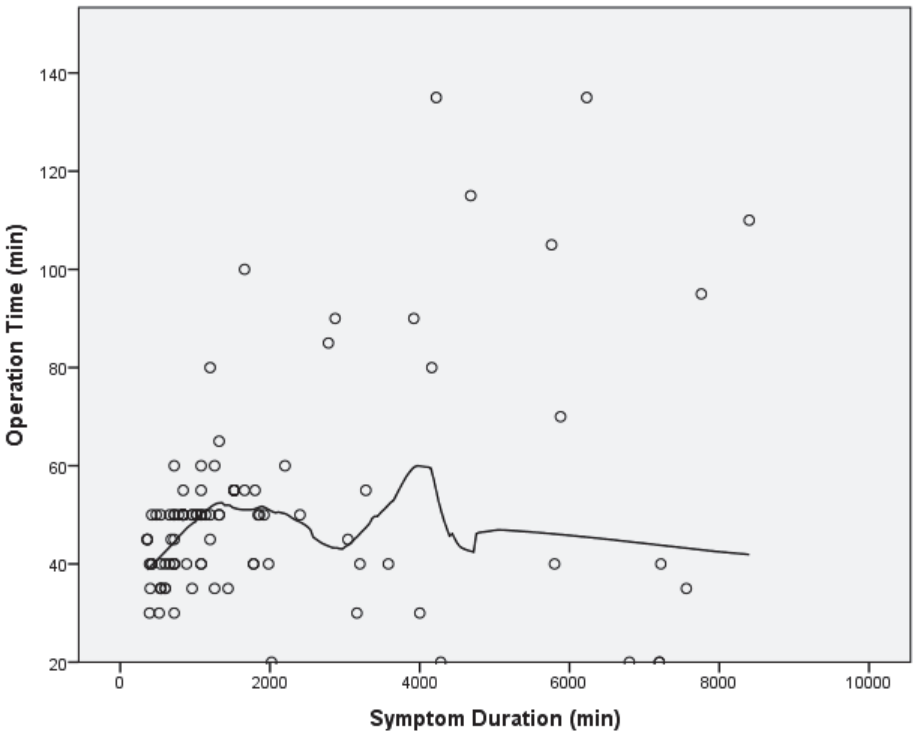


Figure 3. LOESS curve depicting the nonlinear relationship between symptom duration and operation time.

Table 6. Results of the univariate and multivariable analysis of the association between the operation time and continuous variables.

Variables	Value (Mean ± SD)	Univariate Analysis *p*-Value	Multivariable Analysis Coefficient (95% CI)	*p*-Value
Age (month)	140.9 ± 49.3	0.211	−0.02 (−0.117 to 0.065)	0.574
BMI (kg/m^2)	20.7 ± 4.4	0.945	0.31 (−0.66 to 1.28)	0.524
WBC (per μL)	15.2 ± 5.4	0.317	0.12 (−0.59 to 0.83)	0.736
NEU (per μL)	12.2 ± 5.3	0.029	0.51 (−0.21 to 1.24)	0.162
CRP (mg/L)	68.3 ± 78.4	<0.001	0.05 (0.00–0.11)	0.034
Symptom–operation time (h)	32.8 ± 33.0	<0.001	0.25 (0.10–0.40)	0.001
CT–operation time (h)	9.5 ± 10.3	0.946	−0.46 (−0.90 to −0.02)	0.039
Appendix diameter (mm)	11.0 ± 3.0	0.090	1.42 (0.12–2.72)	0.032

BMI: body mass index; WBC: white blood cell; Neu: neutrophil; CRP: C-reactive protein; CT: computed tomography.

Table 7. Results of the multiple regression analysis.

Variables	Coefficient	SD	*t*-Value	CI [0.025–0.975]	*p*-Value
Age	0.03	0.04	0.704	[−0.05 to 0.12]	0.483
Appendix diameter	0.08	0.77	0.108	[−1.45 to 1.62]	0.915
BMI	0.28	0.45	0.623	[−0.61 to 1.18]	0.535
CRP	−0.10	0.02	−0.369	[−0.06 to 0.04]	0.713
NEU/WBC	3.64	20.94	0.174	[−37.95 to 45.24]	0.862
Symptom duration (<2/>2 days)	10.74	0.001	2.351	[0.00 to 0.005]	0.047
CT–operation time	−0.007	0.003	−2.093	[−0.01 to −0.001]	0.039
Appendicolith	0.82	4.16	0.199	[−7.43 to 9.09]	0.843
Periappendiceal fat stranding	6.67	5.20	1.284	[−2.10 to 18.63]	0.203
Periappendiceal fluid	8.54	5.00	1.707	[−1.40 to 18.48]	0.091
Extraluminal air	0.54	7.82	−0.070	[−15.00 to 16.10]	0.944
Abscess	25.31	8.74	2.896	[7.95 to 42.68]	0.005
Localization	3.43	2.11	1.623	[−0.76 to 7.64]	0.108

BMI: body mass index; WBC: white blood cell; Neu: neutrophil; CRP: C-reactive protein; CT: computed tomography. R^2: 0.433.

The median operation time was determined to be 50 min based on our distribution analysis. Using this median as a threshold, operation times exceeding 50 min were classified as long operation time groups. We then compared the clinical and laboratory characteristics between the groups to identify factors influencing prolonged versus shorter operation times. In the long operation time group, CRP levels were significantly higher ($p = 0.002$), symptom-to-operation time was longer ($p < 0.001$), and the duration of symptoms exceeding 2 days was more common ($p = 0.016$). Additionally, periappendiceal fluid ($p < 0.001$), extraluminal air ($p < 0.001$), and the presence of an abscess ($p < 0.001$) were more frequently observed in the long operation time group. No significant differences were found between the groups for other variables such as gender, age, BMI, and appendix diameter (Table 8).

In our study, we also conducted an analysis to examine the factors affecting operation time between two groups classified based on intraoperative findings: perforated and non-perforated appendicitis. In patients with perforated appendicitis, the presence of a periappendiceal abscess on CT was found to be significantly associated with prolonged operative time ($p = 0.020$). However, no significant differences were observed in age, gender, BMI, CRP, or other clinical parameters. In the non-perforated group, the presence of periappendiceal fluid on CT was significantly associated with prolonged operation time ($p = 0.026$), while no significant differences were noted for other variables (Tables 9 and 10).

Table 8. Comparison of clinical and laboratory characteristics between short and long operation time groups.

Variable	Total (n = 104) (Mean ± SD/%)	Short Operation Time Group (n = 75)	Long Operation Time Group (n = 29)	p-Value
Operation time (min)	51.11 ± 21.62	41.53 ± 8.65	75.86 ± 25.32	NA
Sex (F/M)	45/59	32/43	13/16	0.842
Age (month)	140.9 ± 49.3	142.2 ± 50.0	137.7 ± 48.2	0.599
Appendix diameter (mm)	11.1 ± 3.0	10.7 ± 3.2	11.9 ± 2.4	0.106
BMI (kg/m^2)	20.7 ± 4.4	20.6 ± 4.4	21.0 ± 4.6	0.704
CRP (mg/L)	68.4 ± 78.4	54.1 ± 63.9	105.3 ± 99.2	0.002
NEU/WBC	0.73 ± 0.10	0.72 ± 0.10	0.76 ± 0.08	0.118
Symptom–operation time (h)	32.8 ± 33.0	26.9 ± 30.0	48.2 ± 35.0	<0.001
Symptom duration < 2 days	81 (77.9%)	63 (84.0%)	18 (62.1%)	0.016
Symptom duration > 2 days	23 (22.1%)	12 (16.0%)	11 (37.9%)	
CT–operation time (h)	9.5 ± 10.0	9.0 ± 10.0	10.6 ± 9.0	0.059
Appendicolith	55 (52.9%)	37 (49.3%)	18 (62.1%)	0.243
Periappendiceal fat stranding	86 (82.7%)	59 (78.7%)	27 (93.1%)	0.081
Periappendiceal fluid	42 (40.4%)	21 (28.0%)	21 (72.4%)	<0.001
Extraluminal air	21 (20.2%)	8 (10.7%)	13 (44.8%)	<0.001
Abscess	15 (14.4%)	3 (4.0%)	12 (41.4%)	<0.001
Wound infection	7 (6.7%)	3 (4.0%)	4 (13.8%)	0.074
Postop abscess	1 (1%)	0 (0%)	1 (3.4%)	0.106
BPM	90.4 ± 9.7	90.8 ± 9.1	89.4 ± 11.4	0.442
Temperature (°C)	37.61 ± 0.5	37.62 ± 0.5	37.56 ± 0.5	0.497
LOS (days)	3.75 ± 1.0	3.68 ± 0.9	3.96 ± 1.0	0.773
Readmission	3 (2.9%)	2 (2.7%)	1 (3.4%)	0.831

F: female, M: male, BMI: body mass index, CRP: C-reactive protein, WBC: white blood cell; Neu: neutrophil; BPM: beats per minute, LOS: length of hospital stay, SD: standard deviation, NA: not applicable.

Table 9. Distribution and statistical comparison of variables based on short and long operation times in non-perforated appendicitis.

Variable		Total (n = 82)	Short Operation Time (n = 67)	Long Operation Time (n = 15)	p-Value
Age (month)		150.9 ± 44.2	149.2 ± 46.7	158.9 ± 30.5	0.444
Sex	Female	36 (43.9%)	30 (44.8%)	6 (40%)	0.736
	Male	46 (56.1%)	37 (55.2%)	9 (60%)	
BMI (kg/m^2)		21.1 ± 4.2	20.8 ± 4.3	22.6 ± 3.3	0.143
CRP (mg/L)		50.2 ± 58.6	48.4 ± 61.2	58.0 ± 46.3	0.570
NEU/WBC		0.72 ± 0.10	0.71 ± 0.10	0.75 ± 0.08	0.217
BPM		90.9 ± 10.0	91.2 ± 9.2	89.9 ± 13.4	0.663
Temperature (°C)		37.6 ± 0.5	37.6 ± 0.5	37.5 ± 0.5	0.730
Symptom–operation time (h)		27.8 ± 27.8	25.9 ± 28.7	36.4 ± 22.4	0.185
Symptom duration	<2 days	68 (82.9%)	57 (85.1%)	11 (73.3%)	0.275
	>2 days	14(17.1%)	10 (14.9%)	4 (26.7%)	
CT–operation time (h)		9.0 ± 10.2	9.0 ± 11.1	9.1 ± 4.6	0.974
LOS (days)		3.75 ± 1.00	3.68 ± 0.95	4.06 ± 1.16	0.185
Appendix diameter (mm)		10.9 ± 3.2	10.7 ± 3.3	11.8 ± 2.6	0.195
Apendicolith	Absent	43 (52.4%)	35 (52.2%)	8 (53.3%)	0.939
	Present	39 (47.6%)	32 (47.8%)	7 (46.7%)	
Periappendiceal fat standing	Absent	17 (20.7%)	15 (22.4%)	2 (13.3%)	0.434
	Present	65 (79.3%)	52 (77.6%)	13 (86.7%)	
Periappendiceal fluid	Absent	62 (75.6%)	54 (80.6%)	8 (53.3%)	0.026
	Present	20 (24.4%)	13 (19.4%)	7 (46.7%)	
Abscess	Absent	82 (100%)	67 (100%)	15 (100%)	NA
	Present	0 (0%)	0 (0%)	0 (0%)	

BMI: body mass index; CRP: C-reactive protein; LOS: length of hospital stay; BPM: beats per minute; NA: not applicable.

Table 10. Distribution and statistical comparison of variables based on short and long operation times in perforated appendicitis.

Variable		Total (n = 22)	Short Operation Time (n = 8)	Long Operation Time (n = 14)	p-Value
Age (month)		103.5 ± 50.2	83.4 ± 37.4	115.0 ± 54.1	0.161
Sex	Female	9	2 (25%)	7 (50%)	0.251
	Male	13	6 (75%)	7 (50%)	
BMI (kg/m^2)		19.2 ± 5.0	18.8 ± 4.7	19.4 ± 5.4	0.809
CRP (mg/L)		136.1 ± 103.8	101.2 ± 70.8	156.0 ± 116.3	0.243
NEU/WBC		0.78 ± 0.07	0.79 ± 0.06	0.77 ± 0.08	0.600
BPM		88.4 ± 8.4	87.7 ± 7.7	88.8 ± 9.1	0.789
Temperature (°C)		35.6 ± 0.42	37.6 ± 0.5	37.54 ± 0.4	0.631
Symptom–operation time (h)		51.6 ± 43.6	35.5 ± 41.4	60.8 ± 43.6	0.198
Symptom duration	<2 days	13 (59.1%)	6 (75%)	7 (50%)	0.251
	>2 days	9 (40.9%)	2 (25%)	7 (50%)	
CT–operation time (h)		11.1 ± 11	9.30 ± 9.50	12.25 ± 11.98	0.559
LOS (days)		3.77 ± 0.97	3.62 ± 1.06	3.85 ± 0.94	0.602
Appendix diameter (mm)		11.7 ± 2.2	11.4 ± 2.0	11.9 ± 2.4	0.661
Apendicolith	Absent	6 (27.3%)	3 (37.5%)	3 (21.4%)	0.416
	Present	16 (72.7%)	5 (62.5%)	11 (78.6%)	
Periappendiceal fat standing	Absent	1 (4.5%)	1 (12.5%)	0 (0%)	0.176
	Present	21 (95.5%)	7 (87.5%)	14 (100%)	
Periappendiceal fluid	Absent	0 (0%)	0 (0%)	0 (0%)	NA
	Present	22 (100%)	8 (100%)	14 (100%)	
Abscess	Absent	7 (31.8%)	5 (62.5%)	2 (14.3%)	0.020
	Present	15 (68.2%)	3 (37.5%)	12 (85.7%)	

BMI: body mass index; CRP: C-reactive protein, LOS: length of hospital stay; BPM: beats per minute; NA: not applicable.

4. Discussion

In our study, preoperative CT findings were shown to play an important role in predicting the duration of surgery in pediatric patients undergoing laparoscopic appendectomy. Specifically, CT findings such as periappendiceal fluid, extraluminal air, and the presence of a periappendiceal abscess were significantly associated with prolonged operation times. Additionally, elevated CRP levels, the waiting time between the CT imaging and the surgery time, and a symptom duration of more than 2 days were found to be significantly related to prolonged operation times. Furthermore, this study discovered a link between the CT findings and the scoring system developed from these findings and the operation time.

There are several studies with diverse findings on the factors influencing the duration of laparoscopic appendectomy. Siewert et al. [2] demonstrated that preoperative CT findings can be used to predict operative time and complications during laparoscopic appendectomy. Hosokawa et al. [18] discovered that an increase in intra-abdominal fat density and a retrocecal or retro-ascending colon-located appendix prolonged the operation time in pediatric patients. Kohga et al. [19] discovered that the presence of free air was an independent predictor of the development of an intraabdominal abscess. These findings suggest that preoperative CT is a useful tool for predicting postoperative complications and estimating operative time, particularly in complex cases. Our study showed that higher total scores represent more complex surgical cases, and this prolongs the operation time. This finding revealed that as the total score increases, the operation time also increases, and surgeons spend more time dealing with more complex situations. Therefore, evaluating these parameters in the preoperative period may help surgeons predict the duration and degree of difficulty of the operation. However, in our multivariable regression analysis, only periappendiceal abscess was independently associated with operation time. Therefore, while the scoring system can be used as a guide to estimate overall surgical complexity, the independent effect of each component on operation time is limited.

Factors such as being overweight, having high C-reactive protein (CRP) levels, experiencing symptoms for more than 3 days, having an appendix diameter greater than 10 mm, the presence of free air, and the presence of an abscess on CT are findings supporting complicated appendicitis and all independent predictors of prolonged laparoscopic surgery time [12,20]. In cases of complicated appendicitis, additional procedures such as intra-abdominal irrigation and drainage tube placement are often required, which may prolong surgery time [21,22]. In the study conducted by Jeon et al., it was demonstrated that CRP exhibited a significant correlation with operation time, while WBC and NEU showed no correlation [12]. In our study, a significant relationship was found between increasing CRP levels and prolonged operation time in both univariate and multivariable analyses. However, while high WBC and NEU levels showed a positive correlation with surgical times, no significant relationship was observed in multivariable analyses. Furthermore, our study found that the duration of symptoms, CT operation time, and the presence of a periappendiceal abscess were directly related to operative time, with regression analysis confirming these as independent factors.

Although laparoscopic surgery was once considered a relative contraindication in overweight patients, this perception has evolved with increasing surgical experience. Enjoji et al. conducted a study revealing a correlation between an increase in BMI value and the duration of surgery in adult patients undergoing three-port laparoscopic appendectomy [11]. Another study by Enochsson et al. found that being overweight, defined as having a BMI > 26.4, significantly prolonged the duration of surgery in open appendectomy, but this negative impact of being overweight was not observed in laparoscopic appendectomy [16]. Contrarily, Hosokawa et al. demonstrated in their studies that there is no relationship between BMI and the duration of surgery [18,23]. Similarly, in our current study, no correlation was identified between BMI and operation time.

In our study, we also analyzed the factors affecting operation times in both perforated and non-perforated patient groups. In the study by Jeon et al. [12], the appendiceal diameter was associated with prolonged operation time in non-perforated appendicitis, whereas periappendiceal abscess was identified as a more significant factor in perforated cases. Similarly, in our study, the presence of a periappendiceal abscess was the most important factor contributing to prolonged operative time in the perforated group. However, in the non-perforated group, periappendiceal fluid was identified as the main factor influencing operation time. Appendiceal diameter did not have a significant effect in either group.

In rare anatomical positions, laparoscopic appendectomy proves to be a preferable option over the open technique. This preference arises from the surgeon's ability to strategically select trocars and determine their placement once the camera is introduced, and the appendix is properly positioned, contrary to some studies indicating a statistically longer mean operative time in retrocecal and subhepatic groups compared to other anatomical positions [16,18,23]. Based on our study findings, the location of the appendix is not associated with the duration of the operation.

The strengths of our study increase the reliability and accuracy of our results. At first, this study is advantageous due to its substantial sample size, which provides high statistical power and the ability to apply the findings to a larger sample. In addition, the patient cohort demonstrated homogeneity, which minimized variation and enhanced the statistical significance of the results for a specific population group. Furthermore, it is important to note that the surgical procedures were consistently performed by a single surgeon, using the same technique. This conscious approach was carried out to eliminate any potential variations between surgeons and ensure that any variation in surgical outcomes could not be related to differences in surgical expertise or methodology. Together, these strengths enhance the long-term reliability of the findings from our study.

This study has some limitations. First, because this study was retrospective, we were unable to assess additional factors that could influence operation time, considering the fact that all procedures were performed by the same surgeon using the same technique. Second, the utilization of CT scans for patient evaluation, with the resulting radiation exposure,

is an important obstacle. Third, due to the limited sample size, the relationship between prolonged operation time and postoperative outcomes could not be adequately assessed. Future prospective research with larger patient groups will provide further understanding of this topic.

5. Conclusions

In conclusion, our investigations indicate that preoperative CT findings, elevated CRP levels, and a symptom duration of more than 2 days have a significant impact on the duration of laparoscopic appendectomy. In particular, among the CT findings, the presence of a periappendiceal abscess emerged as the strongest independent predictor of operation time. While the five-point CT-based scoring system provides a general framework for assessing surgical complexity, its predictive power for operation time is limited, as not all components significantly influence surgical duration. Future studies with larger patient cohorts and more comprehensive analyses are needed to refine and validate these predictive models.

Author Contributions: I.T.: Conceptualization, Methodology, Writing—Original Draft, Visualization. B.E.: Supervision, Formal Analysis, Writing—Review and Editing. M.A.N.: Investigation, Resources, Data Curation. 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 reviewed and approved by Kastamonu University Clinical Research Ethics Committee (approval code: 2022–KAEK-104 approval date: 19 October 2022).

Informed Consent Statement: Patients were not required to give informed consent to this study because the analysis used anonymous clinical data that were obtained after each patient agreed to treatment by written consent.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

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

References

1. Birnbaum, B.A.; Wilson, S.R. Appendicitis at the millennium. *Radiology* **2000**, *215*, 337–348. [CrossRef] [PubMed]
2. Siewert, B.; Raptopoulos, V.; Liu, S.I.; Hodin, R.A.; Davis, R.B.; Rosen, M.P. CT predictors of failed laparoscopic appendectomy. *Radiology* **2003**, *229*, 415–420. [CrossRef] [PubMed]
3. Gwynn, L.K. The diagnosis of acute appendicitis: Clinical assessment versus computed tomography evaluation. *J. Emerg. Med.* **2001**, *21*, 119–123. [CrossRef] [PubMed]
4. Hansen, A.J.; Young, S.W.; De Petris, G.; Tessier, D.J.; Hernandez, J.L.; Johnson, D.J. Histologic severity of appendicitis can be predicted by computed tomography. *Arch. Surg.* **2004**, *139*, 1304–1308. [CrossRef] [PubMed]
5. Katkhouda, N.; Mason, R.J.; Towfigh, S.; Gevorgyan, A.; Essani, R. Laparoscopic versus open appendectomy: A prospective randomized double-blind study. *Ann. Surg.* **2005**, *242*, 439–448. [CrossRef]
6. Guller, U.; Hervey, S.; Purves, H.; Muhlbaier, L.H.; Peterson, E.D.; Eubanks, S.; Pietrobon, R. Laparoscopic versus open appendectomy: Outcomes comparison based on a large administrative database. *Ann. Surg.* **2004**, *239*, 43–52. [CrossRef]
7. Wei, H.B.; Huang, J.L.; Zheng, Z.H.; Wei, B.; Zheng, F.; Qiu, W.S.; Guo, W.P.; Chen, T.F.; Wang, T.B. Laparoscopic versus open appendectomy: A prospective randomized comparison. *Surg. Endosc.* **2010**, *24*, 266–269. [CrossRef]
8. Wang, X.; Li, Y. Comparison of perioperative outcomes of single-port laparoscopy, three-port laparoscopy and conventional laparotomy in removing giant ovarian cysts larger than 15 cm. *BMC Surg.* **2021**, *21*, 205. [CrossRef]
9. Wang, D.; Dong, T.; Shao, Y.; Gu, T.; Xu, Y.; Jiang, Y. Laparoscopy versus open appendectomy for elderly patients, a meta-analysis and systematic review. *BMC Surg.* **2019**, *19*, 54. [CrossRef]
10. Sauerland, S.; Lefering, R.; Neugebauer, E.A. Laparoscopic versus open surgery for suspected appendicitis. *Cochrane Database Syst. Rev.* **2004**, Cd001546. [CrossRef]
11. Enjoji, M.; Enomoto, N.; Kikuchi, A.; Ueda, Y.; Kato, S.; Ohno, R. Influence of CT-measured appendiceal diameter on operation time of laparoscopic appendectomy. *Surg. Laparosc. Endosc. Percutan Tech.* **2011**, *21*, 20–23. [CrossRef] [PubMed]
12. Jeon, B.G.; Kim, H.J.; Jung, K.H.; Kim, S.W.; Park, J.S.; Kim, K.H.; Kim, I.D.; Lee, S.J. Prolonged operative time in laparoscopic appendectomy: Predictive factors and outcomes. *Int. J. Surg.* **2016**, *36*, 225–232. [CrossRef] [PubMed]

13. Jackson, T.D.; Wannares, J.J.; Lancaster, R.T.; Rattner, D.W.; Hutter, M.M. Does speed matter? The impact of operative time on outcome in laparoscopic surgery. *Surg. Endosc.* **2011**, *25*, 2288–2295. [CrossRef] [PubMed]
14. Jeon, B.G.; Kim, H.J.; Heo, S.C. CT scan findings can predict the safety of delayed appendectomy for acute appendicitis. *J. Gastrointest. Surg.* **2019**, *23*, 1856–1866. [CrossRef] [PubMed]
15. Kim, M.S.; Kim, M.S.; Kim, D.H.; Park, H.W.; Park, H.J.; Hong, H.P.; Kwon, H.J. Pre-operative CT predictors associated with 30-day adverse events in patients with appendiceal inflammatory masses who underwent immediate appendectomies. *Abdom. Imaging* **2015**, *40*, 2263–2271. [CrossRef]
16. Castro, B.A.; Novillo, I.C.; Vázquez, A.G.; Garcia, P.Y.; Herrero, E.F.; Fraile, A.G. Impact of the appendiceal position on the diagnosis and treatment of pediatric appendicitis. *Rev. Paul. Pediatr.* **2019**, *37*, 161–165. [CrossRef]
17. Raptopoulos, V.; Katsou, G.; Rosen, M.P.; Siewert, B.; Goldberg, S.N.; Kruskal, J.B. Acute appendicitis: Effect of increased use of CT on selecting patients earlier. *Radiology* **2003**, *226*, 521–526. [CrossRef]
18. Hosokawa, T.; Tanami, Y.; Sato, Y.; Ishimaru, T.; Kawashima, H.; Oguma, E. Association between the computed tomography findings and operative time for interval appendectomy in children. *Afr. J. Paediatr. Surg.* **2021**, *18*, 73. [CrossRef]
19. Kohga, A.; Yajima, K.; Okumura, T.; Yamashita, K.; Isogaki, J.; Suzuki, K.; Muramatsu, K.; Komiyama, A.; Kawabe, A. Are Preoperative CT Findings Useful for Predicting Postoperative Intraabdominal Abscess in the Patients with Acute Appendicitis? *Medicina* **2019**, *55*, 6. [CrossRef]
20. Byun, J.; Park, S.; Hwang, S.M. Diagnostic Algorithm Based on Machine Learning to Predict Complicated Appendicitis in Children Using CT, Laboratory, and Clinical Features. *Diagnostics* **2023**, *13*, 923. [CrossRef]
21. Schlottmann, F.; Reino, R.; Sadava, E.E.; Campos Arbulú, A.; Rotholtz, N.A. Could an abdominal drainage be avoided in complicated acute appendicitis? Lessons learned after 1300 laparoscopic appendectomies. *Int. J. Surg.* **2016**, *36*, 40–43. [CrossRef] [PubMed]
22. Cho, J.; Park, I.; Lee, D.; Sung, K.; Baek, J.; Lee, J. Risk Factors for Postoperative Intra-Abdominal Abscess after Laparoscopic Appendectomy: Analysis for Consecutive 1,817 Experiences. *Dig. Surg.* **2015**, *32*, 375–381. [CrossRef] [PubMed]
23. Hosokawa, T.; Yamada, Y.; Tanami, Y.; Sato, Y.; Ishimaru, T.; Kawashima, H.; Oguma, E. Associations between sonographic findings and operative time of transumbilical laparoscopic-assisted appendectomy for acute appendicitis in children. *Am. J. Roentgenol.* **2019**, *213*, 191–199. [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.

Journal of
Clinical Medicine

MDPI

Article

Effectiveness of an Artificial Intelligence Software for Limb Radiographic Fracture Recognition in an Emergency Department

Guillaume Herpe [1,2,*], Helena Nelken [1], Tanguy Vendeuvre [3], Jeremy Guenezan [3], Clement Giraud [2], Olivier Mimoz [3], Antoine Feydy [4], Jean-Pierre Tasu [5] and Rémy Guillevin [3,6]

[1] Emergency Radiology Unit, University Hospital Center of Poitiers, 86000 Poitiers, France
[2] Laboratoire de Mathématiques Appliquées LMA, CNRS UMR 7348, 86021 Poitiers, France; clement.giraud@chu-poitiers.fr
[3] Emergency Department, University Hospital Center of Poitiers, 86000 Poitiers, France; tanguy.vendeuvre@chu-poitiers.fr (T.V.); jeremy.guenezan@chu-poitiers.fr (J.G.); olivier.mimoz@chu-poitiers.fr (O.M.); remy.guillevin@chu-poitiers.fr (R.G.)
[4] Department of Musculoskeletal Imaging, Cochin Hospital, AP-HP, 75014 Paris, France; antoine.feydy@aphp.fr
[5] Department of Diagnostic and Interventional Radiology, Poitiers University Hospital, 86000 Poitiers, France; jean-pierre.tasu@chu-poitiers.fr
[6] CHU de Poitiers Service de Radiologie, 86000 Poitiers, France
* Correspondence: herpeguillaume@gmail.com

Abstract: Objectives: To assess the impact of an Artificial Intelligence (AI) limb bone fracture diagnosis software (AIS) on emergency department (ED) workflow and diagnostic accuracy. **Materials and Methods:** A retrospective study was conducted in two phases—without AIS (Period 1: 1 January 2020–30 June 2020) and with AIS (Period 2: 1 January 2021–30 June 2021). **Results:** Among 3720 patients (1780 in Period 1; 1940 in Period 2), the discrepancy rate decreased by 17% ($p = 0.04$) after AIS implementation. Clinically relevant discrepancies showed no significant change (−1.8%, $p = 0.99$). The mean length of stay in the ED was reduced by 9 min ($p = 0.03$), and expert consultation rates decreased by 1% ($p = 0.38$). **Conclusions:** AIS implementation reduced the overall discrepancy rate and slightly decreased ED length of stay, although its impact on clinically relevant discrepancies remains inconclusive. **Key Point:** After AI software deployment, the rate of radiographic discrepancies decreased by 17% ($p = 0.04$) but this was not clinically relevant (−2%, $p = 0.99$). Length of patient stay in the emergency department decreased by 5% with AI ($p = 0.03$). Bone fracture AI software is effective, but its effectiveness remains to be demonstrated.

Keywords: radiology; fractures; bone; AI; artificial intelligence; emergency department; retrospective study; workflow

1. Introduction

Bone fractures have an incidence evaluated between 733 and 4017 per 100,000 patient-years [1]. A missed or delayed diagnosis of fractures on conventional X-rays remains a common problem, ranging from 3% to 10% [2]. Strategies to reduce rates of fracture misdiagnosis are crucial to maintain high standards of patient care, and to limit risks of legal action [3]. Fracture misdiagnosis is a significant concern in clinical settings, particularly in emergency departments where timely and accurate diagnosis is critical. The experience and training of clinicians, especially radiologists, play a crucial role in the accurate identification of fractures. However, several factors contribute to a higher rate of fracture misdiagnosis, including (1) workload and time constraints—in busy emergency departments, clinicians often work under pressure, which can lead to rushed assessments and higher rates of diagnostic errors; (2) complexity of cases—fractures are not always easily identifiable, particularly in areas with complex anatomy or in cases where the fractures are subtle, such as hairline fractures; (3) variability in expertise; and (4) access to expert opinions—while

having access to a specialist, such as an orthopedic surgeon or an experienced radiologist, can significantly reduce diagnostic errors, such access is not always feasible, especially in remote or under-resourced areas.

Given the variability in clinician expertise and the challenges of accessing expert opinions, Artificial Intelligence (AI) has emerged as a valuable tool in assisting with fracture diagnosis. AI-powered tools can (1) support clinicians in diagnosis by identifying fractures with high accuracy, acting as a second set of eyes for clinicians; (2) reduce diagnostic discrepancies, especially in subtle or complex cases; and (3) bridge the expertise gap. In situations where access to a radiology expert is limited, AI can provide decision support, ensuring that even less-experienced clinicians are empowered to make more accurate diagnoses.

These AI tools are particularly valuable in emergency radiology, where speed and accuracy are crucial for optimal patient outcomes [4]. For instance, in some emergency departments, a radiologist on duty is not available 24/7 for radiography reading [5,6].

Artificial Intelligence (AI) including machine learning and deep learning has been used to enable algorithms to learn from data, iteratively improving their own performance without the need for explicit programming. AI has rapidly been put into use in imaging, as a decision aid, as a screening tool, or as a second-reader support for radiologists [3,7].

In bone fracture detection and classification, recent reviews and meta-analyses have reported a high accuracy for AI, with a sensitivity of 92% and 91% and a specificity of 91% on internal and external validation, respectively [8–11]. In a recent meta-analysis of 42 studies, including 37 on radiography, diagnostic performance was comparable for AI and clinicians [11].

In recent years, AI software has been designed as a diagnostic aid with the aim of improving workflow through screening or prioritizing images on worklists and highlighting regions of interest for the clinician [12]. AI could also improve diagnostic certainty by acting as a "second reader" for clinicians when no radiologist is present. However, these advantages remain theoretical, and sometimes commercial, mostly being based on efficacy rather than effectiveness [13].

Given the increasing integration of AI technologies in clinical settings, particularly in emergency departments (EDs), there is a crucial need to evaluate their practical effectiveness. Emergency departments are at the forefront of acute care, where rapid and accurate diagnosis is paramount. This study therefore aims to assess the real-world impact of an AI-based diagnostic tool within this critical environment.

2. Materials and Methods

A consecutive before–after pragmatic retrospective study was conducted in a single French public University Hospital during two periods of time—one before and one after introducing AIS; from 1 January 2020 at 00:01 am to 30 June 2020 at 11:59 pm, called Period 1, no AI was available. During the second period, between 1 January 2021 at 00:01 am and 30 June 2021 at 11:59 pm, called Period 2, the commercially available Bone View (V1., Gleamer, Paris, France) was available for clinicians in the ED.

The data were collected from Poitiers University Hospital, an academic level I trauma center in a 200,000 people referral area with 75,000 visits per year.

This retrospective study was approved by the institutional review board of Comité Ethique et Recherche en Imagerie Médicale (protocol code: CRM-2204-263 and approval date: 23 April 2023). Written informed consent was waived due to retrospective anonymized data collection. The study flowchart is given in Figure 1.

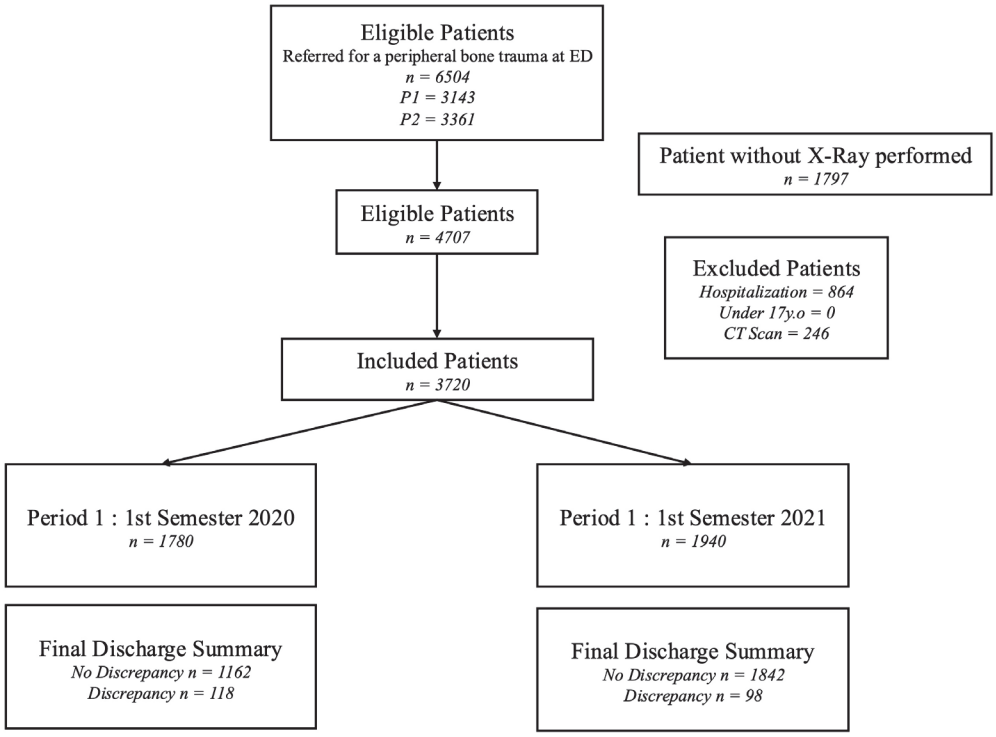


Figure 1. Flowchart of the study.

Two months before Period 2, the AIS was set up (1 November 2020) and a dedicated training session was given to physicians and radiologists. The training consisted of 2 sessions of 1 h (provided by G.H) looking at clinical cases including AIS solution. The AIS validation method and the code were provided in a previously published study [14] and in the Supplementary Materials S1. This version of the AIS was only able to analyze the appendicular skeleton (lower and upper extremities). Spine and pelvic bones were excluded from the analysis.

AIS highlighted a potential fracture in a rectangular box and results were available in less than 1 min on both the PACS system for radiologists (Change Healthcare Radiology v14, Mc Kesson, Vancouver, BC, Canada) and on the emergency physicians' radiographic viewers.

In total, 22 senior emergency physicians and 7 senior radiologists with more than 10 years of experience participated during the 2 periods of the study. ED physicians and radiologists along with their years of experience are listed in the Supplementary Materials S2. All participants have no conflicts of interest.

2.1. Inclusion and Exclusion Criteria

All consecutive patients with limb bone fracture suspicion visiting the ED during Period 1 and Period 2 were included retrospectively from the hospital electronic medical records (Telemaque V14, Poitiers, France).

Inclusion criteria were age (17 years or older), referred to the ED after a recent trauma of less than 72 h, and having undergone a limb radiograph (shoulder, elbow, arm, wrist, hand, hip, knee, leg, ankle, or foot).

To avoid the influence of potential CT scan read back, patients having undergone a CT scan after radiographs were excluded. We also excluded patients for whom radiographs were not read by a radiologist and in-patients for whom radiographs could be analyzed by different physicians in different specialties during hospitalization.

2.2. Gold Standard

The radiologist's reading was considered as the gold standard. The reading was performed without or with the help of AIS, according to the period. In the cases of

discrepancy with an ED physician, a second radiologist reviewed the case to set a final diagnosis. The radiologist's reading (unique or in consensus) was considered as the gold standard.

2.3. Evaluation Criteria

To assess the effectiveness of AIS, the following parameters were recorded:

- Discrepancy rate of limb fracture diagnosis between radiologists and emergency physicians. Discrepancies were defined as differences in the presence or lack of fracture between the final discharge summary performed by the emergency physicians and the radiologist's report. To avoid any mutual influence, emergency physicians established their final diagnosis blinded to the radiologist's report available 24 to 72 h after patient discharge. On the contrary, to alert the clinician to a possible discrepancy, radiologists had knowledge of the final diagnosis of emergency physicians during their readings.
- Length of stay in the ED was defined as the interval between arrival time at the ED and discharge time from the ED and was expressed in minutes and was recorded from the EHR.
- Number of clinician experts' opinions requested whatever the specialty, e.g., orthopedics, radiologists, or resuscitation physicians.
- Any changes in patient management induced by the radiologist's report including changes in treatment (new immobilization with a splint or a plaster, withdrawal of immobilization in the absence of fracture) or follow-up (for instance, new orthopedic appointment) were recorded. All modifications were assessed retrospectively, in consensus by an orthopedic surgeon (TV), an emergency radiologist (GH), and an ED physician (JG), all with more than 10 years of experience, and were defined as being clinically relevant or not. For example, discrepancy was not considered clinically relevant if an elbow fracture was missed but the patient was correctly immobilized with a splint.

2.4. Statistical Analysis

Continuous variables are presented as means and standard deviations (SDs) or medians with ranges, depending on whether they have a normal distribution or not. Categorical variables are presented as numbers with rate. The statistical tests used were Student's *t*-test, the Wilcoxon–Mann–Whitney test and the test of difference in proportions, where its effect size is defined as [15] $h = -\phi_2$, where $= 2\ arcsine\left(\sqrt{P_i}\right)$. An h near 0.2 is a small effect, an h near 0.5 is a medium effect, and an h near 0.8 is a large effect.

We trained a logistic regression model that estimates the probability of discrepancy, presenting the following features: injury location, use of AIS, age, orthopedic appointment within a month, and orthopedic second opinion. The forest plot of this model is displayed in Figure 2 [16–18].

No imputation was made for missing data. Statistical analysis was performed using R software, version 4.0.4.

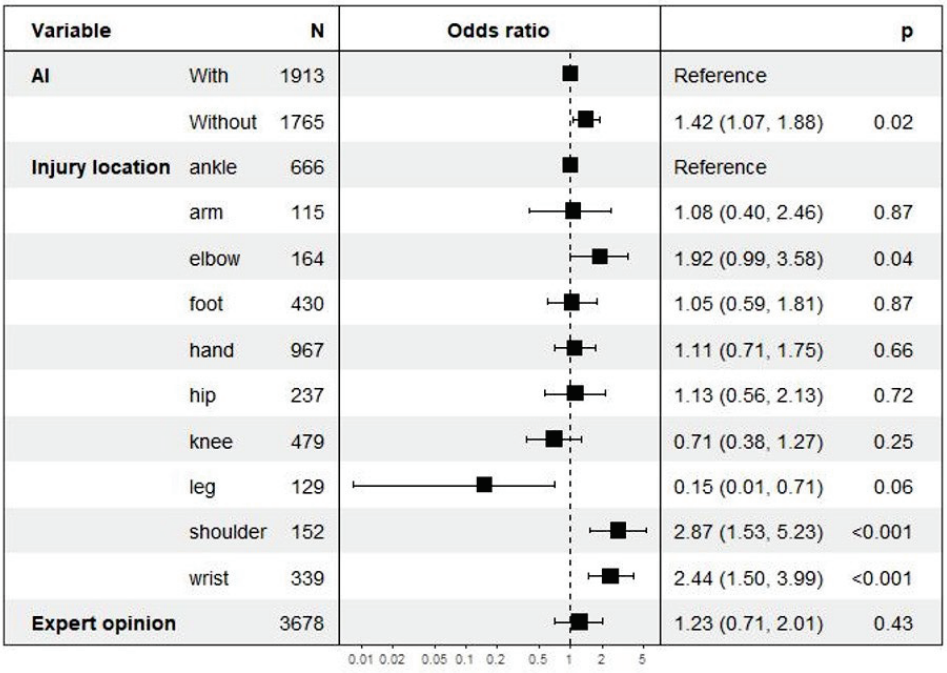


Figure 2. Multivariate risk analysis regarding discrepancies focusing on AI deployment along with fracture location and expert opinion.

3. Results

Among the 41,571 patients visiting the ED during the two study periods (19,991 patients during P1; 21,580 during P2), 16,418 (39%) underwent radiography (7984 [40%] during P1; 8434 [39%] during P2). In total, 4707 (11%) patients fulfilled the inclusion criteria and were therefore included. Of these 4707 patients, 987 (21%) were secondarily excluded because they were later hospitalized (n = 864, 18%) or had a CT scan (246, 5%). Finally, 3720 patients (1780 during Period 1 and 1940 during Period 2) were included in the final analysis.

The clinical characteristics of patients were similar in the two study periods (Table 1). In contrast, there were small differences in the site of the suspected injury between the two periods. There were three times as many shoulder radiographs in Period 2 (Table 1). Other differences were not significant.

Table 1. The demographic characteristics of the study along with the difference among the two subsets and the statistical significant after multiple testing corrections. ** significant p value after multiple testing corrections. SD: standard deviation; AIS: Artificial Intelligence limb bone fracture diagnosis software.

	Without AIS N = 1780	**With AIS N = 1940**	***p* Value**	**Effect Size**
Age (mean ± SD, in years)	42 ± 20	42 ± 21	0.82	
Sex (number, (%))				
Male	974 (54.7)	1102 (56.8)	0.21	0
Injury location (number, (%))				
Hand	449 (25.2)	518 (26.7)	0.3231	0
Ankle	335 (18.8)	331 (17.1)	0.1755	0
Arm	53 (3.0)	62 (3.2)	0.7722	0

Table 1. *Cont.*

	Without AIS N = 1780	With AIS N = 1940	p Value	Effect Size
Foot	207 (11.6)	223 (11.5)	0.9389	
Hip	114 (6.4)	123 (6.3)	0.9896	0
Knee	244 (13.7)	235 (12.1)	0.1611	0
Leg	69 (3.9)	60 (3.1)	0.2243	0
Wrist	172 (9.7)	167 (8.6)	0.2894	0
Shoulder	36 (2.0)	116 (6.0)	<0.0001 **	−0.2
Elbow	86 (4.8)	78 (4.0)	0.2612	0

With the use of AIS, discrepancy rates decreased by 17% from 6.6% in Period 1 to 5.1% in Period 2 ($p = 0.04$) (Table 2). The reduction was significant only for the arm radiographs (Table 3). Discrepancies were more frequent for wrist and elbow ($p = 0.003$) compared to other locations. The discrepancy rate was not associated with either patient age (OR = 1.14, range 0.80–1.62; $p = 0.45$) or orthopedic opinion request (OR = 1.26, range 0.72–2.05, $p = 0.39$) (Figure 2).

Table 2. The difference before and after AIS deployment on the primary and secondary criteria along with the effect size (An h near 0.2 is a small effect, an h near 0.5 is a medium effect, and an h near 0.8 is a large effect). SD: standard deviation. AIS: Artificial Intelligence limb bone fracture diagnosis software.

	Before AIS (N = 1780)	With AIS (n = 1940)	p Value	Effect Size
Discrepancy (number, %)	118 (6.6)	98 (5.1)	0.04	0.1
Clinically relevant discrepancies (number, (%))	65/118 (55)	53/98 (54)	0.99	0
Length of stay (mean ± SD in minutes)	190 ± 132	181 ± 137	0.03	-
Orthopedic opinion referral rate (number, %)	126 (7.1)	123 (6.3)	0.40	0

Mean length of stay in the ED decreased from 190 min (±132 min) for Period 1 to 181 (±137) for Period 2 ($p = 0.03$) (Table 2).

For the two same periods, patient length of stay in the ED did not significantly change from 306 min (±125) to 305 min (±129) ($p = 0.40$).

Expert referral rate did not statistically differ between Period 1 (7.1%) and Period 2 (6.3%) ($p = 0.40$). In the month following patient ED discharge, 810 patients (22%) were consulted by an orthopedic surgeon, with no difference between the two study periods (377 (21) % in Period 1 and 433 (22.3%) in Period 2) ($p = 0.42$).

In total, 118 of 216 discrepancies were clinically relevant with an impact on the patient management (55%). There were no differences in the levels of discrepancy for the two periods (65/118 patients [55%] in Period 1 and 53/98 [54%] patients in Period 2) ($p = 0.99$).

Table 3. Distribution of discrepancies according to the location of the trauma on the overall study period; before and after the trauma along with the effect size. * significant *p* value after multiple testing corrections. AIS: AI limb bone fracture diagnosis software.

Injury Location (Fracture Rate, %)	Overall	Before AIS	With AIS	*p* Value	Effect Size
Foot	22/430 (5.12%)	11/207 (5.31%)	11/223 (4.93%)	1	0
Ankle	33/666 (4.95%)	22/335 (6.57%)	11/331 (3.32%)	0.08	0.2
Elbow	15/164 (9.15%)	11/86 (12.79%)	4/78 (5.13%)	0.15	0.3
Hip	13/237 (5.49%)	9/114 (7.89%)	4/123 (3.25%)	0.19	0.2
Hand	52/967 (5.38%)	24/449 (5.35%)	28/518 (5.41%)	1	0
Wrist	38/339 (11.21%)	20/172 (11.63%)	18/167 (10.78%)	0.93	0
Arm	6/115 (5.22%)	6/53 (11.32%)	0/62 (0%)	0.02 *	0.7
Knee	17/479 (3.55%)	6/244 (2.46%)	11/235 (4.68%)	0.28	−0.1
Shoulder	18/152 (11.84%)	8/36 (22.22%)	10/116 (8.62%)	0.055	0.4
Clavicle	0/28 (0%)	0/8 (0%)	0/20 (0%)	-	-
Leg	1/129 (0.78%)	1/69 (1.45%)	0/60 (0%)	1	0.2
Calcaneus	0/8 (0%)	0/3 (0%)	0/5 (0%)	-	-
Upper limb	0/5 (0%)	0/4 (0%)	0/1 (0%)	-	-
Bone	1/1 (100%)	0/0 (-%)	1/1 (100%)	-	-
Overall	216/3720 (5.81%)	118/1780 (6.63%)	98/1940 (5.05%)	0.04 *	0.1

4. Discussion

Our study aimed to assess how the implementation of AI technology for limb fracture diagnosis affects patient care in the ED. Our results show that AIS has no major effect on clinically relevant discrepancies, despite reducing the rate of total discrepancies by 17%. The need for expert advice remains unchanged and the length of stay in the ED is only minimally affected after AIS setup.

To our knowledge, this study is the first to evaluate the impacts of AIS in real-life conditions in an ED. It should be highlighted that contrary to previous studies, neither the population set nor the readers were selected, and the entire ED physician and emergency radiologist teams participated in the two periods of the study [14,19,20]. Experiences of the teams are provided within the Supplementary Materials.

Our findings are similar to previously reported distributions of fractures among age and anatomical location.

Three points require discussion:

First, the decrease in mean length of stay in the ED is obviously small (9 min on average per patient). However, the number of emergency admissions should be keep in mind; for instance, when there are 100 visits for a potential limb injury per day, this represents a theoretical gain of 15 h. It should be noted that no changes were made to the ED patient workflow after the introduction of AI and it is likely that the addition of

triage [21], prioritization [22] and/or care coordination [23] AI functions could have had a greater impact on this time saving.

Second, the overall reduction in discrepancies of 17% could be considered as interesting and regarded as a significant quality improvement for the patient. Indeed, while the first physician's interpretation facilitates prompt management, reading errors can potentially place the patient at an unnecessary risk of adverse outcomes and the physician at risk of litigation. Discrepancy rates may vary depending on clinicians' expertise. Several studies with controlled data sets have investigated differences in radiographic interpretation between emergency physicians and radiologists, yielding varying levels of agreement ranging from 52% to 99% [24–27].

In our study, AIS has no significant effect on clinically relevant discrepancies. This is in line with a study by Tranovich et al. [5], which found 1044 radiographic discrepancies among 16,111 physician interpretations, with only 0.01% for relevant limb fracture. One should outline that the clinical examination's performance and the frequent expert opinion recourse might act as additional filters to ensure patient safety, preventing treatment based solely on radiographic findings.

Thirdly, the requirement for specialist consultation was not affected by AIS. This is probably due to the necessity for advice on both diagnosis and treatment.

Our study has some limitations, as follows:

First, we did not stratify patient workflow regarding emergency physician experience. This stratification could be interesting, particularly during Period 1. For Period 2, this stratification is less relevant since standalone diagnostic performances of the AI do not rely on ED physicians [14].

Second, the retrospective design could be considered as a weakness. For instance, the number of specialist opinion requests was recorded retrospectively from electronic medical reports, and informal calls were likely omitted. However, without any information being given to the physicians and radiologist on the study, the risk of individual behavior modifications in response to their awareness of being observed was avoided [28].

Third, this is a monocentric study with all the limits due to this design. Our ED might have a specific workflow and the University Hospital cannot reflect the general situation of a healthcare system.

Fourth, we wish to underline that our study was performed during 2020 and 2021, and therefore used the 2021 version of the Boneview algorithm, with the exclusion of spine, ribs, and pelvis. As every release of AIS increases performance and covered anatomical locations, the new version of AIS might have a better performance. This could therefore underestimate the effectiveness of an AIS on the overall common indication of post traumatic radiographs.

Fifth, the observed reduction in hospital length of stay can have different non-controlled causes; several methods have been developed to target a shorter length of stay such as better care coordination, a specific discharge planning program, dedicated case management, a redesigned staffing model, or specialized units for high-risk populations [29]. All of these and other potential management changes in the ED were not controlled for in this study. However, few structural or organizational changes occurred in the emergency department between the two periods. Medical and nursing staff changed only marginally. The number of emergency room visits increased over the period.

Lastly, to assess the financial impact of AI software on a medical institution, a further medico-economic analysis should be conducted.

Despite these limitations, the findings suggest that AIS can contribute to reducing diagnostic discrepancies in busy ED settings, potentially improving patient flow and reducing time-to-diagnosis. However, the lack of significant impact on clinically relevant discrepancies indicates that AIS should be seen as an adjunct to, rather than a replacement for, expert clinical judgment.

5. Conclusions

Clinically relevant discrepancies (those that could have altered patient management) remained stable at around 55% of all discrepancies, despite the overall reduction in discrepancies. This suggests that while AI helped reduce some discrepancies, it did not significantly change the proportion that had clinical relevance, i.e., those that would have added clinical value if detected earlier or differently. Length of stay in the emergency department is only slightly affected by the AIS. These findings underline the importance of distinguishing between general effectiveness (reduced errors) and clinical effectiveness (improved patient outcomes). Further prospective multicentric studies could explore deeper integration to achieve effectiveness rather than efficacy.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jcm13185575/s1, S1. The code underlying this work can be found online at https://github.com/facebookresearch/detectron2 (accessed on 11 August 2024). S2. List of emergency physicians: C.P., D.N., M.N., T.N., T.N., B.A., C.A., D.B., D.M., M.L., M.M., R.E., R.A., R.A., C.R., D.J., G.A., G.J., H.T., K.N., P.S., S.J. List of radiologists (initials): G.H. (10 years of experience), M.M. (25 y.e), J.-C.F. (25 y.e), J.-P.T. (25 y.e), J.V. (10 y.e), N.E. (20 y.e), G.F. (15 y.e).

Author Contributions: Conceptualization, G.H., H.N., T.V., O.M., A.F., J.G., R.G. and J.-P.T.; methodology, G.H. and J.-P.T.; validation, G.H., H.N., T.V., O.M., J.G., R.G. and J.-P.T.; formal analysis, C.G.; investigation, H.N.; resources, H.N. and G.H.; data curation, C.G.; writing—original draft preparation, H.N., G.H., C.G. and T.V.; writing—review and editing, G.H., H.N., T.V., O.M., A.F., J.G., R.G. and J.-P.T.; project administration, G.H., H.N., T.V., O.M., J.G., R.G. and J.-P.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of Comité éthique et Recherche en Imagerie Médicale (protocol code: CRM-2204-263 and approval date: 23 April 2023).

Informed Consent Statement: Patient consent was waived due to the retrospective nature of the data analysis, which involved de-identified imaging data obtained from routine clinical practice. According to institutional and ethical guidelines, studies utilizing anonymized data that do not involve direct interaction with patients and pose no additional risks to their health or privacy may be exempt from obtaining individual consent. Additionally, since this research did not influence clinical care or patient outcomes, and all data handling complied with data protection regulations (such as GDPR or HIPAA), waiving consent was ethically justified. The ethical committee overseeing the study approved this waiver, ensuring the study adhered to relevant legal and ethical standards.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Clinical Relevance Statement: AI in bone fracture is effective on length of stay and radiographic discrepancies, but its effectiveness in clinically relevant conditions remains to be demonstrated.

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

Abbreviations

AI: Artificial Intelligence; ED: emergency department; AIS: AI limb bone fracture diagnosis software.

References

1. Bergh, C.; Wennergren, D.; Möller, M.; Brisby, H. Fracture incidence in adults in relation to age and gender: A study of 27,169 fractures in the Swedish Fracture Register in a well-defined catchment area. *PLoS ONE* **2020**, *15*, e0244291. [CrossRef] [PubMed]
2. Wei, C.-J.; Tsai, W.-C.; Tiu, C.-M.; Wu, H.-T.; Chiou, H.-J.; Chang, C.-Y. Systematic analysis of missed extremity fractures in emergency radiology. *Acta Radiol.* **2006**, *47*, 710–717. [CrossRef] [PubMed]

3. McKinney, S.M.; Sieniek, M.; Godbole, V.; Godwin, J.; Antropova, N.; Ashrafian, H.; Back, T.; Chesus, M.; Corrado, G.S.; Darzi, A.; et al. International evaluation of an AI system for breast cancer screening. *Nature* **2020**, *577*, 89–94. [CrossRef] [PubMed]
4. Williams, S.M.; Connelly, D.J.; Wadsworth, S.; Wilson, D.J. Radiological review of accident and emergency radiographs: A 1-year audit. *Clin. Radiol.* **2000**, *55*, 861–865. [CrossRef]
5. Tranovich, M.J.; Gooch, C.M.; Dougherty, J.M. Radiograph Interpretation Discrepancies in a Community Hospital Emergency Department. *West. J. Emerg. Med.* **2019**, *20*, 626–632. [CrossRef]
6. Fleming, K.A.; Horton, S.; Wilson, M.L.; Atun, R.; DeStigter, K.; Flanigan, J.; Sayed, S.; Adam, P.; Aguilar, B.; Andronikou, S.; et al. The Lancet Commission on diagnostics: Transforming access to diagnostics. *Lancet* **2021**, *398*, 1997–2050. [CrossRef]
7. Canoni-Meynet, L.; Verdot, P.; Danner, A.; Calame, P.; Aubry, S. Added value of an artificial intelligence solution for fracture detection in the radiologist's daily trauma emergencies workflow. *Diagn. Interv. Imaging* **2022**, *103*, 594–600. [CrossRef]
8. Langerhuizen, D.W.G.; Janssen, S.J.; Mallee, W.H.; van den Bekerom, M.P.J.; Ring, D.; Kerkhoffs, G.M.M.J.; Jaarsma, R.L.; Doornberg, J.N. What Are the Applications and Limitations of Artificial Intelligence for Fracture Detection and Classification in Orthopaedic Trauma Imaging? A Systematic Review. *Clin. Orthop. Relat. Res.* **2019**, *477*, 2482–2491. [CrossRef]
9. Kalmet, P.H.S.; Sanduleanu, S.; Primakov, S.; Wu, G.; Jochems, A.; Refaee, T.; Ibrahim, A.; Hulst, L.V.; Lambin, P.; Poeze, M. Deep learning in fracture detection: A narrative review. *Acta Orthop.* **2020**, *91*, 215–220. [CrossRef]
10. Yang, S.; Yin, B.; Cao, W.; Feng, C.; Fan, G.; He, S. Diagnostic accuracy of deep learning in orthopaedic fractures: A systematic review and meta-analysis. *Clin. Radiol.* **2020**, *75*, 713.e17–713.e28. [CrossRef]
11. Kuo, R.Y.L.; Harrison, C.; Curran, T.-A.; Jones, B.; Freethy, A.; Cussons, D.; Stewart, M.; Collins, G.S.; Furniss, D. Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis. *Radiology* **2022**, *304*, 50–62. [CrossRef]
12. Katzman, B.D.; van der Pol, C.B.; Soyer, P.; Patlas, M.N. Artificial intelligence in emergency radiology: A review of applications and possibilities. *Diagn. Interv. Imaging* **2022**, *104*, 6–10. [CrossRef] [PubMed]
13. Herpe, G.; Feydy, A.; D'Assignies, G. Efficacy versus Effectiveness in Clinical Evaluation of Artificial Intelligence Algorithms for Medical Diagnosis: The Award Goes to Effectiveness. *Radiology* **2023**, *307*, e223132. [CrossRef] [PubMed]
14. Duron, L.; Ducarouge, A.; Gillibert, A.; Lainé, J.; Allouche, C.; Cherel, N.; Zhang, Z.; Nitche, N.; Lacave, E.; Pourchot, A.; et al. Assessment of an AI Aid in Detection of Adult Appendicular Skeletal Fractures by Emergency Physicians and Radiologists: A Multicenter Cross-sectional Diagnostic Study. *Radiology* **2021**, *300*, 120–129. [CrossRef] [PubMed]
15. Cohen, J.; Cohen, P.; West, S.G.; Aiken, L.S. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*; Routledge: London, UK, 2013.
16. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016.
17. Wickham, H.; Henry, L. *magrittr: A Forward-Pipe Operator for R*; RStudio: Boston, MA, USA, 2022.
18. Kassambara, A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. 2021. Available online: https://github.com/kassambara/rstatix (accessed on 24 April 2023).
19. 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]
20. Gennaro, G. The "perfect" reader study. *Eur. J. Radiol.* **2018**, *103*, 139–146. [CrossRef] [PubMed]
21. Annarumma, M.; Withey, S.J.; Bakewell, R.J.; Pesce, E.; Goh, V.; Montana, G. Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. *Radiology* **2019**, *291*, 196–202. [CrossRef]
22. O'Neill, T.J.; Xi, Y.; Stehel, E.; Browning, T.; Ng, Y.S.; Baker, C.; Peshock, R.M. Active Reprioritization of the Reading Worklist Using Artificial Intelligence Has a Beneficial Effect on the Turnaround Time for Interpretation of Head CT with Intracranial Hemorrhage. *Radiol. Artif. Intell.* **2021**, *3*, e200024. [CrossRef]
23. Hassan, A.E.; Ringheanu, V.M.; Preston, L.; Tekle, W.G. Artificial Intelligence–Parallel Stroke Workflow Tool Improves Reperfusion Rates and Door-In to Puncture Interval. *Stroke Vasc. Interv. Neurol.* **2022**, *2*, e000224. [CrossRef]
24. Nitowski, L.A.; O'Connor, R.E.; Reese, C.L. The rate of clinically significant plain radiograph misinterpretation by faculty in an emergency medicine residency program. *Acad. Emerg. Med.* **1996**, *3*, 782–789. [CrossRef]
25. Petinaux, B.; Bhat, R.; Boniface, K.; Aristizabal, J. Accuracy of radiographic readings in the emergency department. *Am. J. Emerg. Med.* **2011**, *29*, 18–25. [CrossRef] [PubMed]
26. Al Aseri, Z. Accuracy of chest radiograph interpretation by emergency physicians. *Emerg. Radiol.* **2009**, *16*, 111–114. [CrossRef] [PubMed]
27. Campbell, S.G.; Murray, D.D.; Hawass, A.; Urquhart, D.; Ackroyd-Stolarz, S.; Maxwell, D. Agreement between emergency physician diagnosis and radiologist reports in patients discharged from an emergency department with community-acquired pneumonia. *Emerg. Radiol.* **2005**, *11*, 242–246. [CrossRef] [PubMed]
28. Sedgwick, P.; Greenwood, N. Understanding the Hawthorne effect. *BMJ* **2015**, *351*, h4672. [CrossRef]
29. Tipton, K.; Leas, B.F.; Mull, N.K.; Siddique, S.M.; Greysen, S.R.; Lane-Fall, M.B.; Tsou, A.Y. *Interventions to Decrease Hospital Length of Stay*; Agency for Healthcare Research and Quality (US): Rockville, MD, USA, 2021.

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.

tomography

MDPI

Article

Lightweight MRI Brain Tumor Segmentation Enhanced by Hierarchical Feature Fusion

Lei Zhang [1], Rong Zhang [1], Zhongjie Zhu [1,2,*], Pei Li [1], Yongqiang Bai [1] and Ming Wang [2]

[1] Ningbo Industrial Vision and Industrial Intelligence Lab, Zhejiang Wanli University, Ningbo 315100, China; zhlangl01@hotmail.com (L.Z.); 2022883025@zwu.edu.cn (R.Z.); 2022882023@zwu.edu.cn (P.L.); yongqiangbai@zwu.edu.cn (Y.B.)

[2] Faculty of Information Science and Engineering, Ocean University of China, Qingdao 266100, China; acwarming@stu.ouc.edu.cn

* Correspondence: zhongjiezhu@yeah.net

Abstract: Background: Existing methods for MRI brain tumor segmentation often suffer from excessive model parameters and suboptimal performance in delineating tumor boundaries. Methods: For this issue, a lightweight MRI brain tumor segmentation method, enhanced by hierarchical feature fusion (EHFF), is proposed. This method reduces model parameters while improving segmentation performance by integrating hierarchical features. Initially, a fine-grained feature adjustment network is crafted and guided by global contextual information, leading to the establishment of an adaptive feature learning (AFL) module. This module captures the global features of MRI brain tumor images through macro perception and micro focus, adjusting spatial granularity to enhance feature details and reduce computational complexity. Subsequently, a hierarchical feature weighting (HFW) module is constructed. This module extracts multi-scale refined features through multi-level weighting, enhancing the detailed features of spatial positions and alleviating the lack of attention to local position details in macro perception. Finally, a hierarchical feature retention (HFR) module is designed as a supplementary decoder. This module retains, up-samples, and fuses feature maps from each layer, thereby achieving better detail preservation and reconstruction. Results: Experimental results on the BraTS 2021 dataset demonstrate that the proposed method surpasses existing methods. Dice similarity coefficients (DSC) for the three semantic categories ET, TC, and WT are 88.57%, 91.53%, and 93.09%, respectively.

Keywords: MRI brain tumor segmentation; lightweight; hierarchical feature fusion; macro perception; micro focus

1. Introduction

Brain tumors, posing a significant threat to human health, represent a severe neurological disorder [1,2]. Gliomas, the most prevalent type of brain tumor, are classified into low-grade gliomas (LGG, grades 1–2) and high-grade gliomas (HGG, grades 3–4). HGGs, characterized by rapid infiltration into adjacent tissues, are malignant gliomas that present formidable treatment challenges and lead to reduced survival rates. Early detection, localization, and segmentation of tumors at the LGG stage via medical imaging can facilitate timely surgical interventions and improve patient outcomes. Brain tumor examination and analysis predominantly rely on Magnetic Resonance Imaging (MRI). However, challenges, including complex tumor shapes and blurred edges, persist [3,4]. Thus, a precise MRI brain tumor segmentation method is urgently needed to aid clinical diagnosis and treatment.

Brain tumor segmentation methods primarily encompass two categories: 2D and 3D segmentation methods based on deep learning. Two-dimensional methods dissect MRI images into two-dimensional images, which are then processed using 2D CNNs to extract features and classify pixels. Zhao et al. [5] pioneered the use of CNNS to design an automatic brain tumor segmentation model for 2D sliced MRI images. However, this

approach neglected local detail information and contextual information about the brain tumor. To capture more contextual information, Dong et al. [6] pioneered the application of the U-Net [7] in medical image segmentation. The simple architecture of the U-Net enables the network to consider contextual information. However, the feature extraction capability of the U-Net could be stronger, prompting researchers [8,9] to refine the U-Net. Kong et al. [8] incorporated a feature pyramid module into the U-Net to enhance features by capturing multi-scale semantic information. Chen et al. [9] adjusted the convolution kernel size in each layer of the U-Net encoder to extract features with different receptive fields. Despite the advancement of 2D methods for brain tumors, issues like minor changes between adjacent slices remain. While 2D CNNs process complex brain tumor image slices, they overlook the three-dimensional correlation and often disrupt potential cross-slice continuity, leading to a loss of contextual information.

In contrast, 3D segmentation methods effectively address the abovementioned issues by considering multiple slices simultaneously. They allow for a more comprehensive capture of the three-dimensional shapes and locations of the brain tumors, better meeting the needs of biomedical applications. Inspired by the U-Net, Cicek et al. [10] constructed a 3D U-Net model for volumetric data, converting the input from two-dimensional to three-dimensional images and significantly improving performance. However, this method still could not overcome the inherent essence of the U-Net, namely weak feature extraction capability. To address this problem, Kamnitsas et al. [11] proposed a dual-pathway 3D CNN segmentation network. Its ability to simultaneously process features at multiple scales enhanced its representation capability. However, it did not pay much attention to feature residuals, preventing the network from training deeper and causing information loss. To this end, Alom et al. [12] proposed the R2U-Net based on a recurrent residual network. The network allows the model to train deeper, while the recurrent network makes feature extraction more comprehensive. It achieved better performance, but the recurrent mechanism substantially increased the complexity, necessitating higher hardware requirements. Unlike the recurrent mechanism, Jiang et al. [13] designed a two-stage 3D U-Net model. The first stage performs rough segmentation through a shallow network, while the second stage refines the segmentation through a deeper network. It is trained end-to-end and ranked first in the BraTS 2019 competition, but the parameter volume also demands high hardware requirements.

Recent studies [14–16] demonstrate that transformers can significantly improve the efficiency of brain tumor segmentation. Inspired by the success of transformers in 2D visual tasks, Wang et al. [15] proposed the TransBTS based on the transformer. It captures global contextual information and achieves the best performance at the time. However, this model did not optimize the transformer, resulting in considerable computational complexity. To reduce complexity, Hatamizadeh et al. [16] introduced a sliding window attention mechanism based on the Swin-transformer [17], reducing model complexity. Although transformer-based segmentation methods have achieved notable performance, the quadratic complexity also increases model parameters and computational complexity. To address this, Lee et al. [18] replaced the attention module with a large kernel convolution module. While achieving comparable performance and reducing the parameters, it is merely a module replacement for the Swin-Transformer. It does not consider the local contextual information lost by large kernel convolution.

In summary, existing 3D methods demonstrate a superior ability to capture three-dimensional contextual information. However, certain limitations persist. Three-dimensional MRI images, composed of multiple slices in scans, often exhibit noise across different dimensions, necessitating networks with strong feature extraction capabilities and a keen sensitivity to local details. Contemporary mainstream segmentation models employ transformers as the backbone for feature extraction. While this approach can capture global information correlations, it fails to process local details finely. Secondly, the vast number of parameters inherent in current 3D segmentation models, often an order of magnitude greater than those in 2D models, imposes more stringent hardware requirements.

In response to these challenges, a lightweight MRI brain tumor segmentation method, Enhanced by Hierarchical Feature Fusion (EHFF), is proposed. This method effectively strengthens and merges features while reducing model parameters, achieving superior segmentation performance. The main contributions are as follows:

(1) An adaptive feature learning (AFL) module is constructed as part of a fine-grained feature adjustment network guided by global contextual information. By leveraging macro perception and micro focus, hierarchical features are strengthened, and model parameters are reduced.

(2) To address the lack of attention to local position details, a hierarchical feature weighting (HFW) module is designed. This module extracts refined features across multiple scales through multi-level weighted features, reinforcing the detailed features of spatial positions.

(3) To mitigate the easy loss of up-sampling features in the encoder–decoder structure, a hierarchical feature retention (HFR) module is constructed to retain, up-sample, and fuse features at each layer, thus achieving superior detail retention and reconstruction effects and acting as a supplementary decoder.

2. Proposed Method

The proposed EHFF is a lightweight brain tumor segmentation method that includes an encoder, a decoder, an HFW module, and an HFR module, as depicted in Figure 1. Initially, during the encoding phase, global contextual spatial information is captured through multilevel macro perception and micro focus to guide the enhancement of local fine-grained spatial features. Subsequently, the HFW module refines and weights each hierarchical feature through backpropagation, extracting multi-scale refined features to enhance the detailed features of spatial locations. Ultimately, as an auxiliary decoder, the HFR module collaborates with the primary decoder to perform up-sampling restoration on the refined features, collectively generating the anticipated segmentation image.

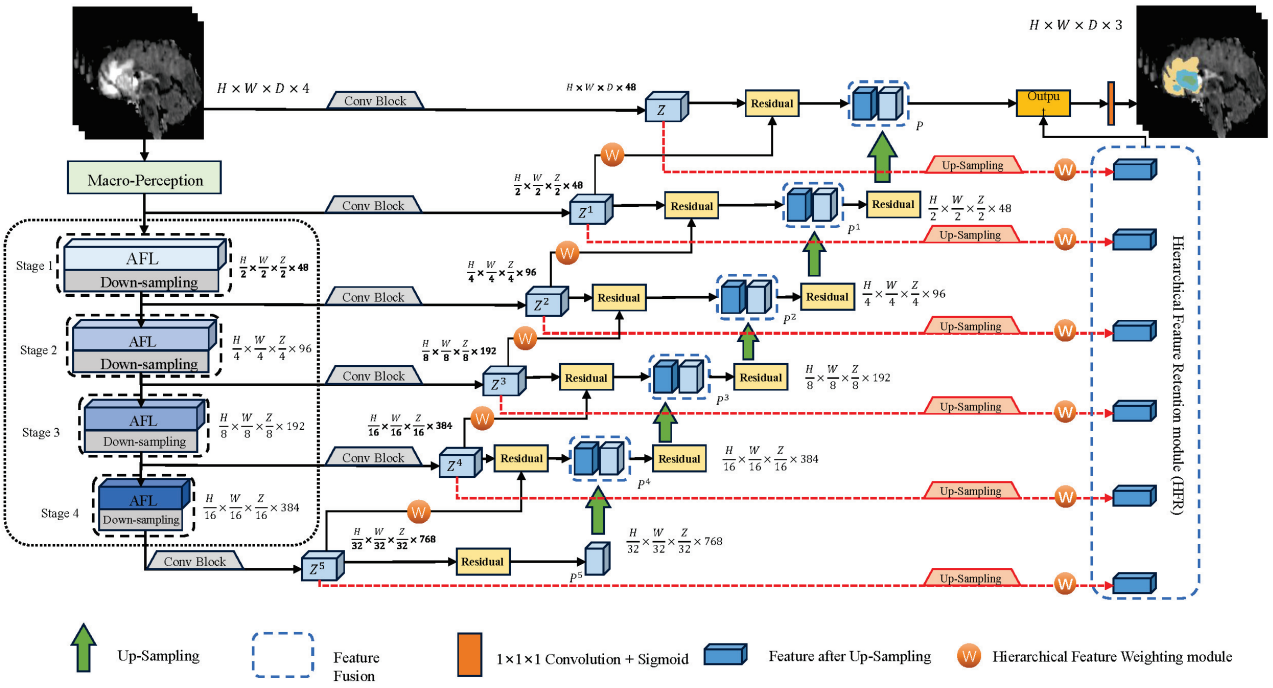


Figure 1. The overall architecture of EHHF.

2.1. Encoder

The encoder is designed to extract high-level features through a multi-level and multi-stage network, as illustrated in Figure 2. It captures the contextual information of brain

tumor images and transmits them via skip connections to the decoder, thereby aiding the model in the segmentation results of the input MRI images.

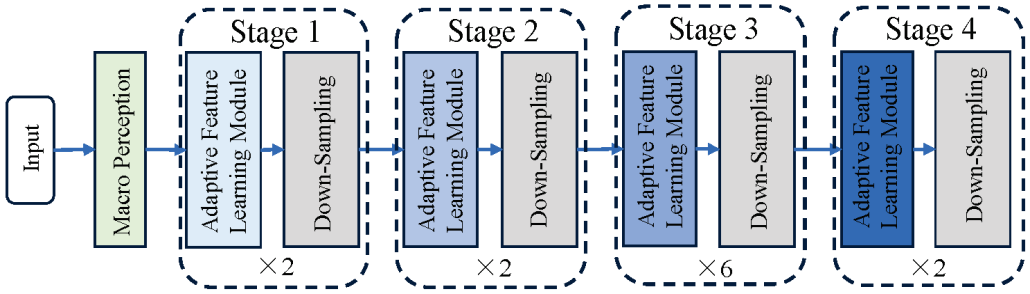


Figure 2. Multistage architecture of the encoder.

For input 3D MRI brain tumor images, a macro-perception module is initially performed on them to generate a feature map with global contextual information, facilitating subsequent feature processing. The macro-perception module comprises several $7 \times 7 \times 7$ large kernel convolutions and corresponding normalization layers. Compared to the mainstream Swin-Transformer-based method, this method of obtaining the global receptive field through large kernel convolutions occupies less memory and reduces model parameters notably. Particularly for the segmentation of MRI brain tumor images, the macro-perception module can preserve local detail features while processing the long-distance spatial position relationship of features.

Recent research indicates that the depth and complexity of the network often result in better performance [19,20]. A single macro-convolution module cannot fully capture the abstract features in complex data, nor can the original features of the MRI image be effectively transmitted. Therefore, a multi-stage deep encoding feature extraction network is designed. This encoding network includes four stages of cascaded AFL feature extraction and down-sampling modules, used to obtain the high-level feature gradually. The AFL module, shown in Figure 3, allows features to continuously learn more complex representation capabilities to better capture the abstract concepts and semantic information in the data and then better fit the data.

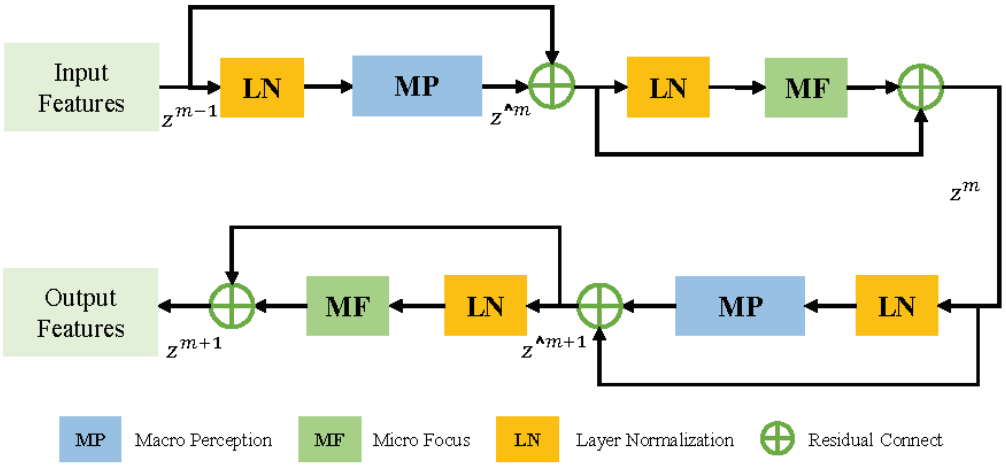


Figure 3. The flowchart of the AFL module.

The AFL module comprises a macro-perception module and a micro-focus module. The macro-perception module is designed to capture the global features and extensive spatial contextual information of MRI brain tumor images. Given the substantial computational demands and parameters of 3D images, a single stack of macro-perception modules may lead to an explosion in model parameters, loss of local details, and compromised

feature representation. Thus, inspired by the MLP layer in Swin-Transformer [17], a micro-focus module is constructed for fine-grained adjustments to enhance the expression and adaptability of spatial detail features. This approach compensates for the drawbacks of stacking large kernel convolutions, thereby improving the effectiveness. The $1 \times 1 \times 1$ convolution in the micro-focus module further refines and boosts the abstract expression of features by introducing non-linear transformations [18], reducing parameter volume and computational complexity.

In the encoding phase, the macro-perception module initially maps the input MRI brain tumor images to acquire the preliminary feature map z^{m-1}. This preliminary feature map, with size $H/2 \times W/2 \times D/2$ and channel dimension 48, is fed into Stage 1 for further feature processing. To be efficient, the preliminary feature map is first normalized and then passed into the AFL module. The macro-perception module in the AFL module captures the deep global spatial contextual feature information to obtain the feature map $z^{\wedge m}$, and the preliminary feature map z^{m-1} is fused through residual connection to gain the global spatial feature representation of different scales. However, relying on residual fusion results in feature redundancy, leading to rich feature information but a loss of attention to local spatial detail. Therefore, the micro-focus module is used to optimize $z^{\wedge m}$ to acquire the refined feature z^m, compressing each channel independently through $1 \times 1 \times 1$ convolution, to decrease feature redundancy by minimizing cross-channel spatial context. This operation adjusts local spatial fine-grain to reduce model complexity, introducing non-linear transformations, enhancing the expressive ability, and enriching feature representation.

The above process is iterated twice in each stage, with the macro-perception module and the micro-focus module processing z^m again. This approach optimizes the spatial detail features while obtaining the global receptive field and acquiring a more complete feature representation. Next, through a $2 \times 2 \times 2$ size, with the stride of 2 down-sampling convolution block, the resolution of the feature map is reduced by a factor of 2, and the refined features are passed into the subsequent stage. Consequently, the output feature map size of Stage 1 is $H/4 \times W/4 \times D/4 \times 96$. Stage 2–4 repeat the operations in Stage 1, and the sizes of the output feature maps are $H/8 \times W/8 \times D/8 \times 192$, $H/16 \times W/16 \times D/16 \times 384$, $H/16 \times W/16 \times D/16 \times 768$, respectively.

The output features of these 4 stages are regarded as hierarchical features, which are applied to the subsequent HFW and HFR modules. They can obtain more comprehensive and rich multi-scale feature representation, acquiring the ability to perceive global spatial context information and local spatial details, which helps capture brain tumor features in MRI images. The above operations are expressed in the following formula:

$$z^{\wedge m} = MP\{LN(z^{m-1})\} + z^{m-1} \tag{1}$$

$$z^m = MF\{LN(z^{\wedge m})\} + z^{\wedge m} \tag{2}$$

$$z^{\wedge m+1} = MP\{LN(z^m)\} + z^m \tag{3}$$

$$z^{m+1} = MF\{LN(z^{\wedge m+1})\} + z^{\wedge m+1} \tag{4}$$

where MP denotes the macro-perception module, MF denotes the micro-focus module, LN denotes the Layer Normalization layer, playing the role of regularization and optimization of training, z^{m-1}, z^m, and z^{m+1}, respectively, represent the input, the output after the first round of iterations, and the final output of the AFL module within each stage, $z^{\wedge m}$ and $z^{\wedge m+1}$ denote the outputs of the first and second MP blocks of the AFL module.

2.2. Decoder

The decoder is designed to progressively up-sample and fuse features of varying scales, originating from the outputs of stages 1 to 4 in the encoder. This process restores the resolution to that of the input image, ultimately generating the final segmentation result. The decoding process begins with up-sampling the highest-level features, which are output

from Stage 4 of the encoder. The output features from the corresponding encoder stage are fused at each up-scaling layer via skip connections [21]. This fusion facilitates the capture of subtle brain tumor features and enhances the precision of brain tumor segmentation.

Specifically, feature maps of varying scales denoted as Z^1–Z^5, output by each encoder stage, are transformed by a convolution block. This block consists of two $3 \times 3 \times 3$ convolution layers that intensify feature representation. The final encoding output of Stage 4, after a $3 \times 3 \times 3$ convolution transformation, is labeled as Z^5 and serves as the input for the first layer of up-sampling. In the subsequent progressive up-sampling process, encoding feature information is introduced into the corresponding decoding stages, and the HFW module is used to apply weights to Z^1–Z^5. These weighted features are then connected to the next layer of the decoder via skip connections, enabling the acquisition of multi-scale refined features and improving feature accuracy. This assists further feature fusion and information transmission, thereby boosting model performance. In the terminal part of the decoding, segmentation output is computed through a $1 \times 1 \times 1$ convolution layer and a sigmoid activation function, realizing pixel-level prediction of brain tumor segmentation. The above operations are expressed in the following formula:

$$Z^1 = Conv_{3\times3}\{MP(input)\} \tag{5}$$

$$Z^i = Conv_{3\times3}(Stage_{i-1}), i \in [2,5] \tag{6}$$

$$P^5 = Up(Z^5) \tag{7}$$

$$P^i = Up\left\{Z^i, P^{i-1}, HFW\left(Z^{i+1}\right)\right\}, i \in [1,4] \tag{8}$$

$$P = Up\{input, P^1, HFW(Z^1)\} \tag{9}$$

where Z^1 denotes the features mapped by the first layer macro-perception module, $Conv_{3\times3}$ denotes the Conv Block consisting of two $3 \times 3 \times 3$ convolution kernels, Z^i denotes the features mapped by the output of i-th Stage through Conv Block module, $input$ denotes the input image, Up denotes the up-sampling, HFW denotes the hierarchical feature weighting module, P^i denotes the output of the i-th up-sampling, and P denotes the final output of the decoder.

2.3. Hierarchical Feature Weighting Module

The hierarchical feature weighting (HFW) module is designed to refine feature information through backpropagation, aiming to minimize information redundancy and enhance the feature representation, thereby augmenting model performance. Traditional encoder–decoder architectures typically obtain high-level features through down-sampling and pooling. However, a semantic gap often exists between the encoding and decoding processes, which hinders the integration of high-level and low-level features, resulting in blurred prediction boundaries. To address this issue, the HFW module was developed to retain hierarchical feature information from different encoder stages. By weighting features across various levels, the module facilitates a more effective fusion of high-level and low-level features. This approach reduces information loss during up-sampling and strengthens the precision of spatial features. The framework of the HFW module, as shown in Figure 4, incorporates a learnable weight parameter optimized through backpropagation. This optimization is based on the weighted learning of features Z^1–Z^5 from stages 1 to 4 through backpropagation, which determines the contribution of the output in each stage. This process extracts multi-scale refined features, enriching the detail of spatial locations.

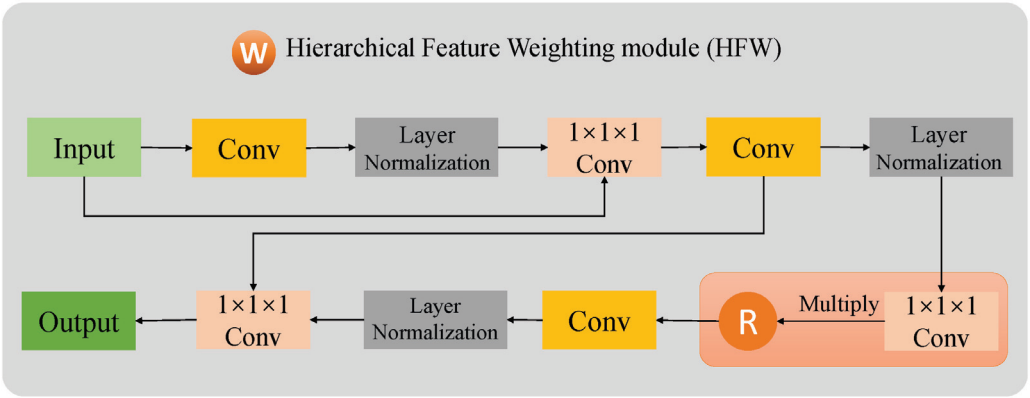


Figure 4. The flowchart of the HFW module.

Traditional methods of residual connection fusion, which involve simply adding or concatenating features, provide comprehensive feature representation but also introduce excessive feature redundancy and irrelevant information. To address this, a weighted fusion approach is employed, offering a more nuanced method of feature integration. This approach discards superfluous feature information, giving the model flexibility in information fusion and potentially reducing the risk of overfitting [22].

The HFW module takes the hierarchical features Z^1–Z^5, output by each stage of the encoder [23], as input, targeting the acquisition of local spatial details under a global receptive field. Initially, local spatial details of each hierarchical feature are captured through a convolutional block composed of multiple $3 \times 3 \times 3$ convolutions and normalization layers. Subsequently, feature information is compressed and refined through several $1 \times 1 \times 1$ and $3 \times 3 \times 3$ convolutional kernels, yielding compressed refined features E. A learnable random variable R is then introduced, multiplied by the refined features E, and processed through another series of $3 \times 3 \times 3$ convolutions, normalization layer, and $1 \times 1 \times 1$ convolution to obtain the final weighted feature. Given the problem of information loss due to network depth, multiple residual connections are designed within the HFW module to ensure that the acquisition of refined local spatial detail features does not lead to gradient vanishing or feature loss. The above operations are expressed in the following formula:

$$E = Conv_{1\times1\times1}\{LN[Conv(input)]\}_2 \tag{10}$$

$$Out = Conv_{1\times1\times1}\{LN[Conv(E \times R)]\} \tag{11}$$

where $Conv$ denotes multiple $3 \times 3 \times 3$ convolutional layer, $Conv_{1\times1}$denotes $1 \times 1 \times 1$ convolutional layer, R denotes random parameter, E denotes compression refined feature, and subscript $_2$ denotes that the operation loops twice.

2.4. Hierarchical Feature Retention Module

The hierarchical feature retention module is designed to function as a supplementary decoder, preserving hierarchical features Z^1–Z^5 and executing hierarchical up-sampling fusion. In the final stage, it collaborates with the primary decoder to decode and output the final predicted segmentation of brain tumor images. Contrary to the primary decoder, which integrates encoder and decoder features through skip connections, the HFR module focuses on the direct outputs of various stages. It directly performs hierarchical up-sampling on the preserved multi-scale encoding features, enabling the model to retain spatial information and relationships across different resolution feature maps. For 3D data of brain tumors, this spatial correspondence is crucial for capturing the connections between pixels and spatial details, significantly aiding in accurate measurement and analysis across multiple slices. This capability is instrumental in assisting the model to determine the volume and contour segmentation of brain tumors.

Regarding the retained information Z^1–Z^5 from the encoder, they are upscaled to produce five outputs with size H × W × D × 48, indirectly capturing different scale feature information at various spatial resolutions. Subsequently, these features are fused proportionally with the final output from the primary decoder to yield the ultimate segmentation prediction image. The above operations are expressed in the following formula:

$$Q = \sum_{i=1}^{5} Up\left(Z^i\right) \tag{12}$$

$$Output = Sigmoid\{Conv_{1\times1\times1}[P \times \partial + Q \times (1 - \partial)]\} \tag{13}$$

where Q denotes the output of the HFR module, Up denotes up-sampling, $Sigmoid$ denotes the activation function, and ∂ denotes the fusion coefficient, which is set to 0.8 according to the empirical value.

3. Experimental Results and Analysis

Experiments were conducted on the BraTS 2021 dataset [24], and the results were compared with other mainstream methods. The experiments were carried out using the PyTorch 1.10 deep learning framework in a Python 3.8 environment. The hardware setup consisted of a single 24GB NVIDIA GeForce RTX 3090. And the initial learning rate was set to 0.001. The model was trained for 300 epochs with a batch size of 1, utilizing the Adam optimizer for end-to-end optimization.

3.1. Dataset

The BraTS 2021 dataset comprises 1251 cases in the training set, 219 cases in the validation set, and 570 cases in the test set. Each MRI scan includes 3D images from four modalities. Within the training set, each case's scan consists of 3D images from t1, t1ce, t2, flair modalities, and one shared label. The shared labels comprise four classification categories: 0, 1, 2, and 4. To gain a visual understanding of the brain tumor segmentation annotations and labels across different modalities of MRI images, a case was selected and visualized using the 3D Slicer medical image analysis software , as depicted in Figure 5. The final task of the BraTS 2021 Challenge necessitates the identification of three specific sub-regions: the enhanced tumor ET (label 4), the tumor core TC (labels 1 and 4), and the entire tumor WT (labels 1, 2, and 4). Since the organizers of this dataset did not make the test data publicly available, comparisons were made using a five-fold cross-validation method on the training data only.

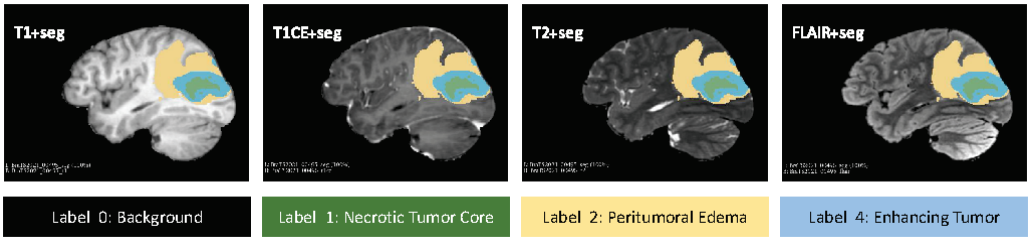


Figure 5. Visualization of 3D MRI images and brain tumor annotation.

3.2. Evaluation Metric

In terms of MRI brain tumor segmentation, the evaluation of segmentation performance commonly employs the Dice Similarity Coefficient (DSC) as the standard metric. The DSC serves as a widely used measure to quantify the similarity between segmentation results and reference standards. The formula is as follows:

$$DSC = \frac{2|S \cap G|}{|S| + |G|} \tag{14}$$

where $|S \cap G|$ denotes the number of intersection pixels between the segmentation result and the standard label, while $|S|$ and $|G|$ represent the number of pixels in the segmentation

result and the standard label, respectively. The DSC value ranges between 0 and 1, wherein a value closer to 1 indicates a higher degree of similarity between the segmentation result and the standard label.

3.3. Comparative Experiment

We conducted five-fold cross-validations using five subsets of the BraTS 2021 dataset. The proposed method was compared with other mainstream methods, including nn-Former [25], TransBTS [15], UNETR [14], and Swin UNETR [16], as illustrated in Tables 1–5. The visualization of these segmentation results is presented in Figure 6. Figure 7 presents a bar chart comparison, facilitating a more intuitive appreciation of the superiority inherent in our method. And the experimental results in Table 6 (including 3 additional methods: 3D-UX-Net [18], TransUNet [26], and MedNeXt-S [27]) revealed that, across three semantic categories (ET, TC, and WT) in five subsets, the DSC achieved by the proposed method surpassed those of competing methods, demonstrating considerable competitiveness. Specifically, the five-fold cross-validation results for ET, TC, and WT achieved by the proposed method were 88.57%, 91.53%, and 93.09%, respectively, with an average five-fold cross-validation result of 91.06%. This represents an average improvement of 0.83% of the BraTS 2021 dataset when compared to the advanced method, Swin UNETR. At the same time, the proposed method's standard deviations for segmentation accuracy are 0.445, 0.477, and 0.139, respectively, which are at relatively low levels, indicating our stability.

Further, Table 7 provided a comparison of the parameter number and inference time across different methods. It is evident that the parameter count and the inference time for the proposed method are significantly lower than that of other methods, with a reduction of approximately 29% compared to Swin UNETR. This demonstrates the relative simplicity of the proposed method and further substantiates the lightweight nature of the EHFF.

Table 1. Performance comparison on Fold 0 in the BraTS 2021 dataset.

	nn-Former	**TransBTS**	**UNETR**	**Swin UNETR**	**EHFF**
ET	72.34	85.67	85.23	85.33	**87.63**
TC	82.46	89.64	88.64	89.54	**90.39**
WT	86.45	91.12	91.61	92.32	**92.45**
Avg	80.42	88.81	88.49	89.06	**90.16**

Note: the bold denotes the best result.

Table 2. Performance comparison on Fold 1 in the BraTS 2021 dataset.

	nn-Former	**TransBTS**	**UNETR**	**Swin UNETR**	**EHFF**
ET	76.31	88.41	89.13	89.41	**89.68**
TC	83.54	90.32	89.77	91.64	**92.45**
WT	87.88	91.10	92.49	93.31	**93.47**
Avg	82.58	89.94	90.46	91.45	**91.87**

Table 3. Performance comparison on Fold 2 in the BraTS 2021 dataset.

	nn-Former	**TransBTS**	**UNETR**	**Swin UNETR**	**EHFF**
ET	77.91	86.41	87.11	87.12	**88.29**
TC	85.15	89.75	89.54	91.21	**91.24**
WT	87.65	90.24	92.12	92.45	**92.98**
Avg	83.57	88.80	89.59	90.26	**90.84**

Table 4. Performance comparison on Fold 3 in the BraTS 2021 dataset.

	nn-Former	TransBTS	UNETR	Swin UNETR	EHFF
ET	74.12	86.74	87.39	87.42	**88.52**
TC	83.02	90.11	90.23	90.69	**91.64**
WT	86.63	91.13	92.14	92.54	**93.12**
Avg	81.26	89.33	89.92	90.22	**91.09**

Table 5. Performance comparison on Fold 4 in the BraTS 2021 dataset.

	nn-Former	TransBTS	UNETR	Swin UNETR	EHFF
ET	75.36	86.72	86.78	86.87	**88.74**
TC	82.71	89.21	90.24	90.44	**91.91**
WT	87.82	91.41	92.36	93.12	**93.45**
Avg	81.96	89.11	89.79	90.14	**91.37**

Table 6. Average DSC performance comparison on Fold 0-4 in Brats2021 dataset.

	nn-Former	TransBTS	UNETR	Swin UNETR	3D-UX-Net *	TransUNet *	MedNeXt-S *	EHFF	Std
ET	75.21	86.79	87.13	87.23	-	-	-	**88.57**	0.445
TC	83.38	89.81	89.68	90.70	-	-	-	**91.53**	0.477
WT	87.29	91.00	92.14	92.75	-	-	-	**93.09**	0.139
Avg	81.96	89.20	89.65	90.23	90.63	89.17	**91.27**	90.85	0.361

Note: - indicates missing data, * indicates data obtained directly from the original paper, and the Std denotes the standard deviation of the proposed EHFF.

Table 7. Parameters and inference time comparison.

Method	Parameters	Inference Time (s)
TransBTS	**32,987,588**	-
nn-Former	149,478,267	-
UNETR	125,018,259	-
Swin UNETR	62,187,296	0.96
EHFF (Ours)	44,053,394	**0.93**

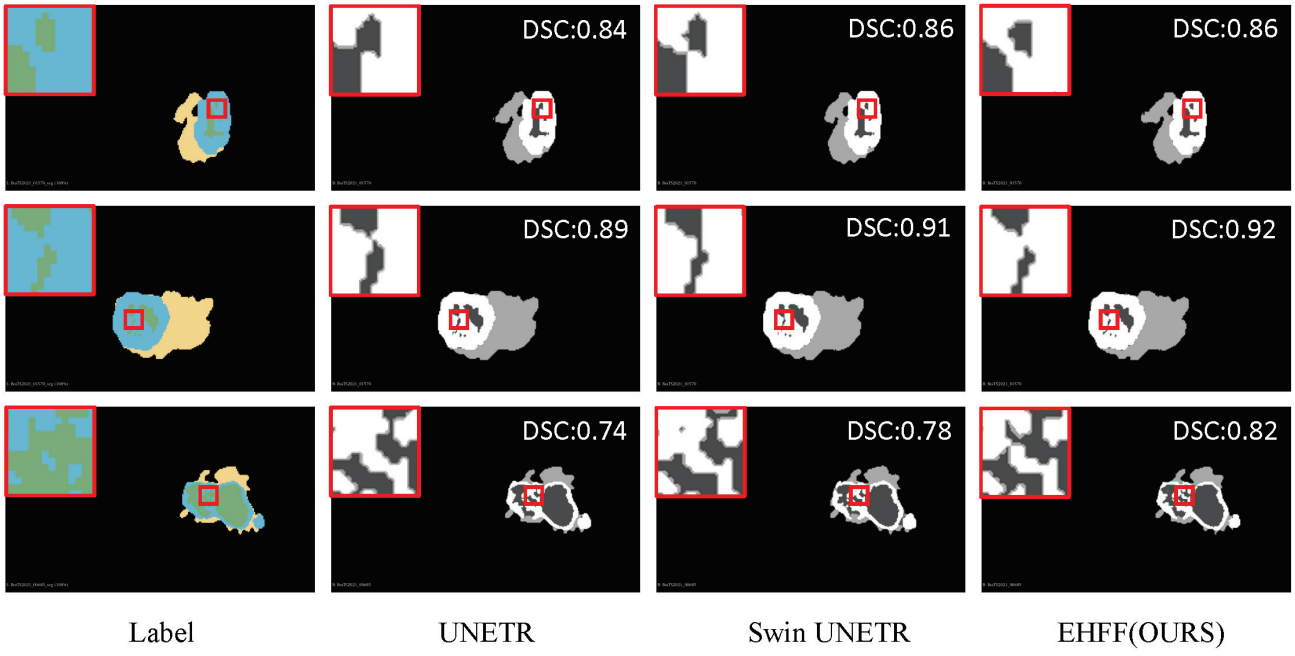


Figure 6. Segmentation comparison by various methods.

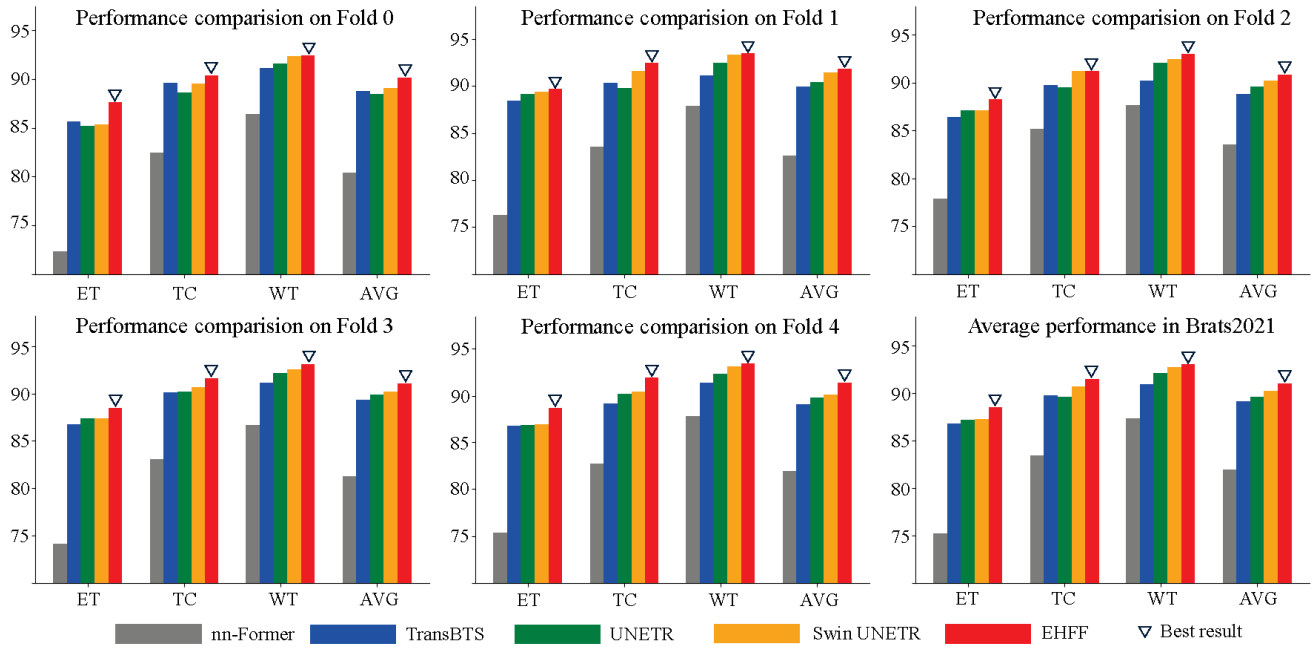


Figure 7. The visual comparison of the performance.

Compared to methods like nn-Former, UNETR, Swin UNETR, and TransUNet, which use the transformer for global context modeling, our method offers significant advantages. The high computational complexity of the transformer leads to large model parameters and inefficiencies, especially when handling long sequences. In contrast, our proposed method uses a macro-perception module built with large kernel convolutions to model global context, significantly reducing the number of model parameters. Combined with a micro-focus module, EHFF effectively captures image features. And the HFW module is employed to refine these features, enabling precise brain tumor segmentation.

Although our method does not have fewer parameters than TransBTS, it performs better overall. In the five-fold cross-validation on the BraTS 2021 dataset, EHFF surpassed TransBTS in all 15 metrics, especially achieving nearly 2% higher accuracy in the WT index. This distinction highlights the superior effectiveness of our method, even with a slightly larger parameter count.

When compared to 3D UX-NET, both methods abandon the transformer. However, 3D UX-NET focuses too much on global features, neglecting the finer details of brain tumors. Our method improves on this by preserving important detail information through hierarchical feature weighting, ensuring precise segmentation by enhancing attention to tumor details across multiple scales.

Compared to the latest method, MedNeXt-S, our method has certain limitations. MedNeXt-S makes deeper use of CNNs, extracting higher-level information through deeper convolutional modules. This further enhances the representation of tumor details, improving segmentation performance. This design has inspired us, and in future research, we plan to explore deeper CNN architectures to achieve even better performance while maintaining or reducing model complexity.

3.4. Ablation Experiment

To validate the efficacy of the modules within the proposed EHFF, ablation experiments were conducted on two key components: the HFW module and the HFR module. In order to efficiently showcase the effectiveness of these modules, the experiments were exclusively performed on a subset of the BraTS 2021 dataset, specifically fold 4. The experimental results clearly demonstrate that the EHFF achieves its optimal performance when both the HFW module and the HFR module are employed together in Table 8a. In addition, ablation

experiments were carried out on the fusion coefficient, as indicated in Table 8b. The results reveal that the optimal performance is achieved when the fusion coefficient is set to 0.8.

Table 8. Abalation experiments.

(a) Module Abalation Comparison.			(b) Fusion Coefficient.	
HFW	**HFR**	**DSC**	**Fusion Coefficient**	**DSC**
×	×	89.87	0.6	90.88
✓	×	90.63	0.7	91.18
×	✓	90.74	0.8	**91.37**
✓	✓	**91.37**	0.9	90.77

3.5. Limitations

Brain tumor segmentation is a task that involves multi-scale feature processing. The proposed method uses large kernel convolutions to extract global features. However, the use of large kernel convolutions can reduce the ability to capture fine details. In this study, we enhanced tumor detail features through a micro-focus module. Despite this, the method is still not detailed enough and may result in the loss of some tumor details. This issue is also common in other current brain tumor segmentation methods. In the future, we plan to design a more refined local feature processing module to work alongside large kernel convolutions, aiming to achieve a more lightweight and precise brain tumor segmentation.

4. Conclusions

In this paper, we propose a lightweight MRI brain tumor segmentation method enhanced by hierarchical feature fusion to achieve precise segmentation of MRI brain tumors. Initially, the AFL module is constructed to thoroughly capture the global features of MRI brain tumor images by embracing macro perception and micro focus. This module subsequently refines these features at a granular level, thereby reducing computational complexity while enhancing the spatial information. Secondly, the HFW module is developed to extract and augment multi-scale features through multi-level weighted processing. This module can enhance the detail features of spatial locations and mitigate the shortfall in local detail attention. Finally, hierarchical features are fused and up-sampled through the HFR module, which co-predicts the accurately segmented brain tumor images in conjunction with the decoder. The five-fold cross-validation results on the BraTS 2021 dataset show that the proposed method achieves standard deviations of 0.445, 0.477, and 0.139 in the ET, TC, and WT categories, respectively. These low standard deviations demonstrate the stability of the proposed method. Additionally, the average segmentation accuracy of the proposed method, measured by DSC, reaches 88.27%, 91.31%, and 92.96% for these categories, which are 1.34%, 0.83%, and 0.34% higher than those of Swin UNETR, respectively. These results demonstrate that the proposed method can effectively improve the quality of segmented images and provide better clinical support for brain tumor segmentation imaging.

In general, while our method demonstrates obvious advantages over others, it still faces limitations in practical clinical applications. For instance, deep learning-based brain tumor segmentation requires significant computational resources and costly annotated datasets, which increases the overall cost. Looking ahead, our future work will involve using cloud computing and distributed learning to reduce computational costs, thereby achieving more lightweight brain tumor segmentation.

Author Contributions: Conceptualization, L.Z. and Z.Z.; methodology, L.Z. and Z.Z.; validation, L.Z. and M.W.; data curation, L.Z.; writing—original draft preparation, L.Z. and R.Z.; writing—review and editing, L.Z., Z.Z., Y.B., P.L. and M.W.; visualization, L.Z. and R.Z.; supervision, L.Z. and R.Z.; resources, Z.Z.; project administration, Z.Z.; funding acquisition, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Natural Science Foundation of China (61671412); the Zhejiang Provincial Natural Science Foundation of China (LZ24F010004, LY21F010014); the Yongjiang Sci-Tech Innovation 2035 (2023-23); and the Ningbo Municipal Major Project of Science and Technology Innovation 2025 (2022Z076).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

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

References

1. Bauer, S.; Wiest, R.; Nolte, L.P.; Reyes, M. A survey of MRI-based medical image analysis for brain tumor studies. *Phys. Med. Biol.* **2013**, *58*, R97. [CrossRef] [PubMed]
2. Mabray, M.C.; Barajas, R.F.; Cha, S. Modern brain tumor imaging. *Brain Tumor Res. Treat.* **2015**, *3*, 8–23. [CrossRef] [PubMed]
3. Wadhwa, A.; Bhardwaj, A.; Verma, V.S. A review on brain tumor segmentation of MRI images. *Magn. Reson. Imaging* **2019**, *61*, 247–259. [CrossRef] [PubMed]
4. Zhu, Z.; He, X.; Qi, G.; Li, Y.; Cong, B.; Liu, Y. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. *Inf. Fusion* **2023**, *91*, 376–387. [CrossRef]
5. Zhao, L.; Jia, K. Multiscale CNNs for brain tumor segmentation and diagnosis. *Comput. Math. Methods Med.* **2016**, *2016*, 8356294. [CrossRef] [PubMed]
6. Dong, H.; Yang, G.; Liu, F.; Mo, Y.; Guo, Y. Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks. In *Medical Image Understanding and Analysis: 21st Annual Conference, MIUA 2017, Edinburgh, UK, 11–13 July 2017*; Springer International Publishing: Cham, Switzerland, 2017; Proceedings 21; pp. 506–517.
7. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; Springer International Publishing: Cham, Switzerland, 2015; Proceedings, Part III 18; pp. 234–241.
8. Kong, X.; Sun, G.; Wu, Q.; Liu, J.; Lin, F. Hybrid pyramid u-net model for brain tumor segmentation. In *Intelligent Information Processing IX: 10th IFIP TC 12 International Conference, IIP 2018, Nanning, China, 19–22 October 2018*; Springer International Publishing: Cham, Switzerland, 2018; Proceedings 10; pp. 346–355.
9. Chen, Y.; Cao, Z.; Cao, C.; Yang, J.; Zhang, J. A modified U-Net for brain Mr image segmentation. In *Cloud Computing and Security: 4th International Conference, ICCCS 2018, Haikou, China, 8–10 June 2018*; Springer International Publishing: Cham, Switzerland, 2018; Revised Selected Papers, Part VI 4; pp. 233–242.
10. Çiçek, Ö; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, 17–21 October 2016*; Springer International Publishing: Cham, Switzerland, 2016; Proceedings, Part II 19; pp. 424–432.
11. Kamnitsas, K.; Ledig, C.; Newcombe, V.F.J.; Simpson, J.P.; Kane, A.D.; Menon, D.K.; Rueckert, D.; Glocker, B. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* **2017**, *36*, 61–78. [CrossRef]
12. Alom, M.Z.; Hasan, M.; Yakopcic, C.; Taha, T.M.; Asari, V.K. Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation. *arXiv* **2018**, arXiv:1802.06955.
13. Jiang, Z.; Ding, C.; Liu, M.; Tao, D. Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, 17 October 2019*; Springer International Publishing: Cham, Switzerland, 2020; Revised Selected Papers, Part I 5; pp. 231–241.
14. Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H.R.; Xu, D. Unetr: Transformers for 3d medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 574–584.
15. Wang, W.; Chen, C.; Ding, M.; Li, J.; Yu, H.; Zha, S. Transbts: Multimodal brain tumor segmentation using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 109–119.
16. Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Roth, H.R.; Xu, D. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*; Springer International Publishing: Cham, Switzerland, 2021; pp. 272–284.
17. 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, BC, Canada, 11–17 October 2021; pp. 10012–10022
18. Lee, H.H.; Bao, S.; Huo, Y.; Landman, B.A. 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.

19. 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.
20. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
21. Oyedotun, O.K.; Al Ismaeil, K.; Aouada, D. Why is everyone training very deep neural network with skip connections? *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *34*, 5961–5975. [CrossRef] [PubMed]
22. Ma, S.; Zhao, B.; Hou, Z.; Yu, W.; Pu, L.; Zhang, L. Robust visual object tracking based on feature channel weighting and game theory. *Int. J. Intell. Syst.* **2023**, *2023*, 6731717. [CrossRef]
23. Zhu, Z.; Zhang, L.; Bai, Y.; Wang, Y.; Li, P. Scene text image super-resolution through multi-scale interaction of structural and semantic priors. *IEEE Trans. Artif. Intell.* **2024**, *5*, 3653–3663. [CrossRef]
24. Baid, U.; Ghodasara, S.; Mohan, S.; Bilello, M.; Calabrese, E.; Colak, E.; Farahani, K.; Kalpathy-Cramer, J.; Kitamura, F.C.; Pati, S.; et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv* **2021**, arXiv:2107.02314.
25. Zhou, H.-Y.; Guo, J.; Zhang, Y.; Han, X.; Yu, L.; Wang, L.; Yu, Y. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Trans. Image Process.* **2023**, *32*, 4036–4045. [CrossRef] [PubMed]
26. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Alan, L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
27. Saikat, R.; Gregor, K.; Constantin, U.; Michael, B.; Jens, P.; Fabian, I.; Paul, J.; Klaus, M. Mednext: transformer-driven scaling of convnets for medical imagesegmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Vancouver, BC, Canada, 8–12 October 2023; pp. 405–415.

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.

Journal of
Clinical Medicine

MDPI

Article

Fully Automated Detection of the Appendix Using U-Net Deep Learning Architecture in CT Scans

Betül Tiryaki Baştuğ [1,*], Gürkan Güneri [2], Mehmet Süleyman Yıldırım [3], Kadir Çorbacı [4] and Emre Dandıl [5,*]

[1] Department of Radiology, Medical Faculty, Bilecik Şeyh Edebali University, Bilecik 11230, Türkiye
[2] Department of General Surgery, Medical Faculty, Bilecik Şeyh Edebali University, Bilecik 11230, Türkiye; gurkanguneri@gmail.com
[3] Department of Sogut Vocational School, Computer Technology, Bilecik Şeyh Edebali University, Bilecik 11600, Türkiye; mehmets.yildirim@bilecik.edu.tr
[4] Department of General Surgery, Bilecik Osmaneli Mustafa Selahattin Çetintaş Hospital, Bilecik 11500, Türkiye; dr.kadircorbaci@gmail.com
[5] Department of Computer Engineering, Faculty of Engineering, Bilecik Seyh Edebali University, Bilecik 11230, Türkiye
* Correspondence: betultryak@yahoo.com (B.T.B.); emre.dandil@bilecik.edu.tr (E.D.)

Abstract: Background: The accurate segmentation of the appendix with well-defined boundaries is critical for diagnosing conditions such as acute appendicitis. The manual identification of the appendix is time-consuming and highly dependent on the expertise of the radiologist. **Method:** In this study, we propose a fully automated approach to the detection of the appendix using deep learning architecture based on the U-Net with specific training parameters in CT scans. The proposed U-Net architecture is trained on an annotated original dataset of abdominal CT scans to segment the appendix efficiently and with high performance. In addition, to extend the training set, data augmentation techniques are applied for the created dataset. **Results:** In experimental studies, the proposed U-Net model is implemented using hyperparameter optimization and the performance of the model is evaluated using key metrics to measure diagnostic reliability. The trained U-Net model achieved the segmentation performance for the detection of the appendix in CT slices with a Dice Similarity Coefficient (DSC), Volumetric Overlap Error (VOE), Average Symmetric Surface Distance (ASSD), Hausdorff Distance 95 (HD95), Precision (PRE) and Recall (REC) of 85.94%, 23.29%, 1.24 mm, 5.43 mm, 86.83% and 86.62%, respectively. Moreover, our model outperforms other methods by leveraging the U-Net's ability to capture spatial context through encoder–decoder structures and skip connections, providing a correct segmentation output. **Conclusions:** The proposed U-Net model showed reliable performance in segmenting the appendix region, with some limitations in cases where the appendix was close to other structures. These improvements highlight the potential of deep learning to significantly improve clinical outcomes in appendix detection.

Keywords: appendix detection; deep learning; medical imaging; segmentation; U-Net architecture

1. Introduction

The integration of artificial intelligence (AI) into medical diagnostics has transformed disease detection and treatment planning, improving both accuracy and efficiency. AI technologies, particularly deep learning algorithms, such as convolutional neural networks (CNN), have demonstrated remarkable potential in interpreting complex medical images and analyzing extensive clinical data. These advances have been crucial in several medical fields, including oncology, cardiology, and neurology, where AI systems have matched or exceeded human diagnostic performance [1,2].

Currently, AI-based approaches proposed for medical imaging can be divided into two subgroups: machine learning-based methods and advanced neural network-based

methods such as deep learning [3]. AI, particularly deep learning models, has been instrumental in transforming diagnostics in many medical fields [4]. On the other hand, the development and application of Explainable AI (XAI) in medical diagnostics has gained considerable momentum in recent years. By providing interpretable results, XAI systems have been instrumental in building the trust and acceptance of AI tools among healthcare professionals. In addition, AI-based and XAI-based decision support systems play a critical role in these systems by providing clinicians with transparent and interpretable decision paths [5–8]. With the increasing integration of XAI, the adoption of AI systems in clinical practice would continue to grow, driven by both improved accuracy and interpretability of AI predictions [9]. Incorporating these advanced AI tools, such as deep learning-based architectures for appendix detection, into diagnostic workflows holds great promise for improving the speed, accuracy, and reliability of medical diagnoses.

Appendicitis, a common and potentially life-threatening condition, is one area where AI-driven diagnostic tools can make a significant impact. The timely and accurate diagnosis of appendicitis is crucial, as delays can lead to severe complications such as perforation and sepsis. While clinical examination and imaging techniques such as computed tomography (CT) scans are traditionally used for diagnosis, these methods are often subjective and highly dependent on clinician expertise [10].

Deep learning has increasingly become a critical tool in the medical field, offering advanced solutions for diagnosis, treatment planning, and patient care [4]. In recent years, several studies have shown that AI models, including machine learning and deep learning techniques, can improve diagnostic accuracy and support clinical decision-making [11–13]. AI, particularly deep learning, has shown promise in appendicitis diagnosis, with several studies highlighting the advantages of CNN-based models in analyzing medical images. Recent studies have underscored the effectiveness of AI in enhancing appendix diagnosis [14,15]. A systematic review by Issaiy et al. [16] highlighted the use of different AI algorithms, including deep learning models such as CNN, in diagnosing acute appendicitis. These models often outperform traditional diagnostic methods in identifying appendix from medical images like CT scans. In a recent study, Liang et al. [17] developed a deep learning model combined with radiomics to differentiate between complicated and uncomplicated acute appendicitis using pelvic CT scans. Their model outperformed radiologists, suggesting improved diagnostic performance. In another study, Marcinkevics et al. [18] focused on pediatric appendicitis and used machine learning models such as logistic regression, random forests, and generalized boosted regression models to predict the diagnosis and severity of appendicitis. These models were trained and validated using clinical data, demonstrating that AI can effectively aid in the diagnosis and management of pediatric appendicitis. Rajpurkar et al. conducted a study using a 3D deep learning model to diagnose appendicitis. They used an approach that exploits volumetric data from CT scans to improve diagnostic accuracy. The model demonstrated high sensitivity and specificity, outperforming traditional diagnostic methods and representing a significant advance in automated appendicitis detection [19]. These studies and applications highlight the significant advances and contributions of AI in the detection and diagnosis of appendicitis. However, there remains a gap in the previous works regarding fully automated segmentation of the appendix, particularly using deep learning architectures like U-Net.

Deep learning is extensively used in medical imaging and diagnostics, where CNNs analyze medical images [20]. For instance, deep learning models such as U-Net are used for image segmentation tasks in radiology, significantly enhancing diagnostic accuracy [19]. The U-Net deep learning architecture has many applications, including the detection of brain tumors [21], segmentation of retinal layers in optical coherence tomography images [22], breast cancer detection [23], and liver segmentation [24]. These examples highlight the versatility and effectiveness of the U-Net model in various medical imaging tasks, underscoring its importance in enhancing diagnostic accuracy and improving patient care. Despite these promising results, several challenges persist in optimizing deep learning models and AI methods for appendix detection. Factors such as data quality, model

complexity, and the integration of multi-modal data significantly impact the performance of deep learning methods. Addressing these challenges through strategies like data augmentation, advanced model architectures, and rigorous model optimization is essential for maximizing the potential of deep learning in medical diagnostics [25,26].

In this study, we propose a U-Net-based architecture specifically designed for appendix segmentation in CT scans. The U-Net model is widely known for its encoder–decoder structure, which allows the precise localization and segmentation of complex structures in images. In the U-Net architecture, skip connections between the encoder and decoder layers help preserve spatial information and improve segmentation accuracy [27], which is essential for distinguishing the appendix from surrounding tissue. Our U-Net model has been designed with specific training parameters and data augmentation techniques to handle variability in CT scan quality and anatomy. These features are designed to optimize the model's performance for the challenging task of appendix segmentation, which is often complicated by the small size and unclear boundaries of the appendix in medical images. To the best of our knowledge, there are almost no studies on appendix segmentation from CT scans. This study proposes U-Net deep learning architecture using the original dataset in appendix segmentation and discusses potential improvements to enhance diagnostic accuracy and clinical outcomes. This study provides several important contributions to the field of automated appendix detection and segmentation:

1. We have developed a U-Net model specifically tailored for appendix segmentation in CT scans, addressing a significant gap in the state-of-the-art scans. The architecture builds on the strengths of U-Net, using special training parameters and data enhancement techniques to cope with the variations in image quality and anatomy complexity;
2. The proposed model is trained on an original annotated dataset of abdominal CT scans and evaluated using key metrics such as DSC, VOE, ASSD, and HD95. These metrics demonstrate the reliable performance of the model in accurately segmenting the appendix, with a particular focus on minimizing false positives and false negatives;
3. This study employs hyperparameter optimization techniques to fine-tune the U-net architecture to ensure the highest possible segmentation performance. In addition, data augmentation strategies are applied to expand the training set and improve the model's ability to generalize across different CT scan conditions;
4. While the model demonstrates high segmentation performance, we discuss potential limitations, particularly in cases where the appendix is close to other anatomical structures. We suggest directions for future improvement to enhance the diagnostic accuracy and clinical utility of the system.

The subsequent sections of the paper are organized as follows. The details of the original dataset generated for this study, the data augmentation procedures, and the methodology of the proposed system are described in Section 2. Section 3 deals with the results of the experimental analyses. The evaluation of the obtained results, the comparison with previous studies and the comparison with state-of-the-art methodology are discussed in Section 4. The study is concluded with inferences and focused future work in the last section.

2. Materials and Methods

The proposed model and the general structure of the study are shown in the block diagram in Figure 1. The methodology of the proposed approach consists of three main phases. The first phase is data collection and dataset preparation. In this phase, the necessary data are collected and a dataset for the proposed U-Net model is created. The second phase involves the structuring of the proposed U-Net deep learning model and the preparation of the working environment. In this phase, the architecture of the model is determined and the necessary software and hardware infrastructure is set up. The third phase is the automatic segmentation of the appendix region using the best weights obtained from the training processes. In this phase, the model is tested, the weights that give the

best performance are determined, and the success of the model in the segmentation of the appendix region with these weights is evaluated.

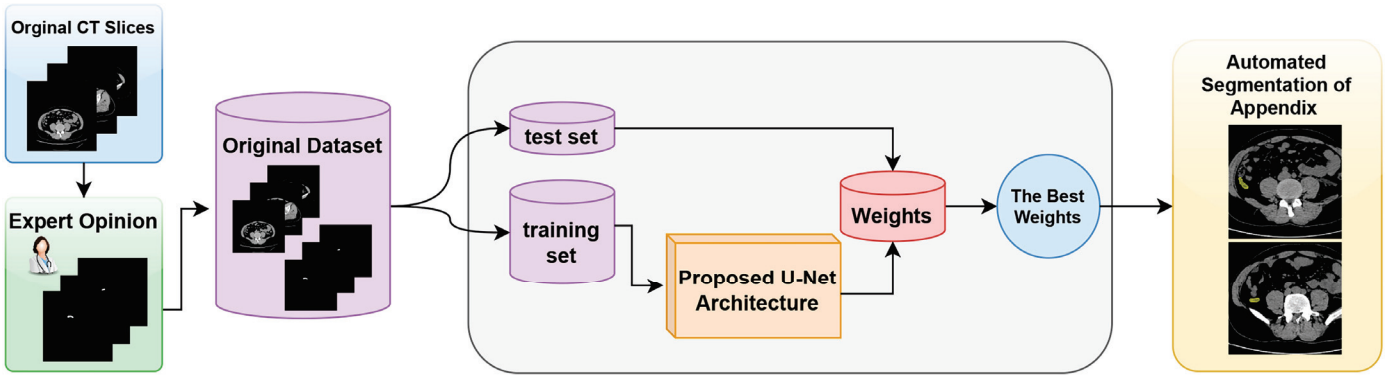


Figure 1. Block diagram of the proposed detection system for the fully automated detection of appendix region in CT scans.

A. Dataset

In this study, an original dataset was prepared for appendix detection and segmentation from CT scans using the proposed U-Net-based deep learning architecture. CT examinations were performed using a 64-MDCT scanner (multi-slice CT Aquillion 64; Toshiba). The imaging protocol included the following parameters: a slice thickness of 2.5 mm, a reconstruction interval of 0.777 mm, a gantry rotation time of 0.6 s, a tube voltage of 120 kV, and a tube current of 200 mA. The field of view was 40 to 50 cm. During CT imaging, slices were obtained from the upper abdomen to the lower abdomen with the patient in the supine position.

The original dataset consisted of CT scans of 299 patients who presented to the Bilecik Training and Research Hospital with abdominal pain and suspected acute appendicitis. These patients were assessed for demographic and laboratory characteristics and the presence or absence of appendicitis. As a result of the evaluations, 140 of these patients had a normal appendix, while the remaining 159 patients were diagnosed with acute appendicitis. Of the patients included in the dataset, 190 were male and the remaining 109 were female, with ages ranging from 18 to 91 years. The patients' radiological images were obtained between July 2021 and April 2024. In addition, it was confirmed that there was no ethical problem in conducting this study with Decision Number 13 of the 8th meeting of Bilecik Seyh Edebali University Non-Interventional Clinical Research Ethics Committee on 5 December 2023. In addition, patients who have been treated at Bilecik Training and Research Hospital have agreed to provide their data for research purposes.

In the dataset, expert physicians annotated the appendix regions with ground truth (GT) masks in these CT scans, using ITK-SNAP software [28], and the annotations were stored in NIfTI format. During the data preparation phase, axial slices identified by the expert physicians as containing the appendix were extracted. This process resulted in 672 slices from patients with appendicitis and 748 slices from healthy individuals, giving a total of 1420 slices extracted. These annotated slices formed the basis of the dataset used to train and evaluate the proposed U-Net deep learning model. Figure 2 shows sample slices from the dataset, indicating the annotated appendix regions with GT masks. This dataset, with its detailed annotation and diverse sample population, provided a robust approach for training the U-Net model to accurately segment the appendix in CT images.

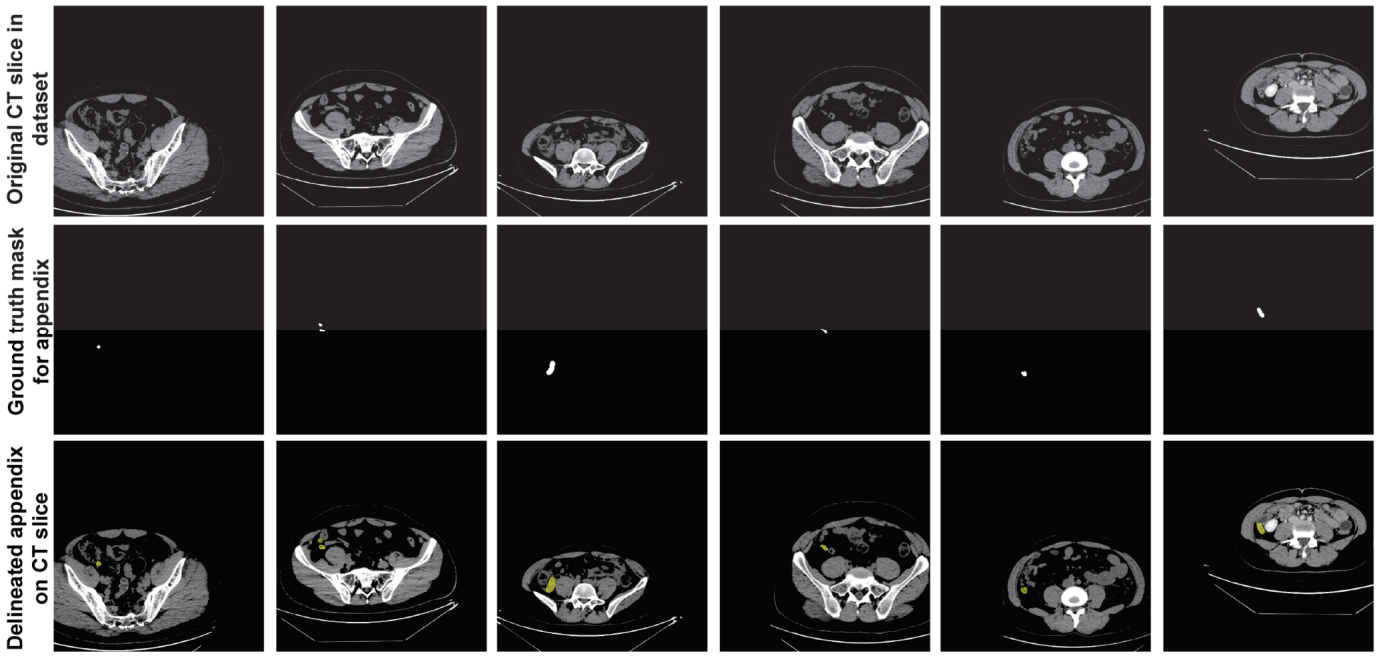


Figure 2. Sample slices from the dataset indicating the annotated appendix regions with GT masks.

B. Proposed Methodology

In this study, a deep learning model based on the U-Net architecture was proposed for detecting appendix areas on axial CT scans. The U-Net model is widely used in medical image processing, especially for biomedical image segmentation [29–31]. It has demonstrated its effectiveness in various medical applications by accurately classifying each pixel in an input image to delineate different tissues and structures. The U-Net architecture consists of two parts. The structure of the U-Net can be summarized for each key component in the contracting path (encoder) and the expansive path (decoder). The first part is the contracting path, which serves as the encoder [19]. In the encoder part, convolution, padding, and pooling operations are performed to compress the data into feature representations [32]. For each layer *Le* in the encoder, the output O_{Le}^{en} is given Equation (1). The encoder extracts features and reduces spatial resolution while segmenting the image.

$$O_{Le}^{en} = \mathrm{ReLU}\left(\mathrm{Conv}\left(O_{Le-1}^{en}\right)\right) \tag{1}$$

where O_{Le}^{en} represents the feature maps after layer *Le* in the encoder.

The second part is the expanding path, which serves as the decoder. In the decoder part, deconvolution, padding, and pooling operations are performed to expand the data, and the output from each corresponding level of the encoder is added at each layer to bring the data back to its original input size [33]. On the other hand, for each layer *Ld*, the output O_{Ld}^{de} is shown in Equation (2). In the expansive path, the feature maps are upsampled using deconvolution (transposed convolution).

$$O_{Ld}^{de} = \mathrm{ReLU}\left(\mathrm{DeConv}\left(O_{Ld-1}^{de}\right)\right) \tag{2}$$

where O_{Ld}^{de} represents the feature maps after layer *Ld* in the decoder.

Skip connections are a key feature of the U-Net architecture, allowing information to flow from the encoder to the decoder at corresponding levels [34]. This is a concatenation operation between the feature maps from the encoder and the decoder at each layer [27]. The skip connection is shown in Equation (3).

$$O_{Ld}^{de} = \mathrm{concat}\left(O_{Ld}^{de},\ O_{nL-Le}^{en}\right) \tag{3}$$

where nL is the number of layers in the encoder.

In U-Net architecture, the final layer performs a convolution operation to obtain the segmented data. The final layer $O_{final}(x, y)$ applies a 1×1 convolution to map the feature maps to the desired number of output classes, as seen in Equation (4).

$$O_{final}(x, y) = \text{sigmoid}\left(Conv_{1\times1}\left(O_{Ld}^{de}\right)\right) \quad (4)$$

In this study, the proposed U-Net architecture for automatic appendix detection and segmentation from CT scans is shown in Figure 3. The proposed U-Net model for appendix segmentation differs significantly from the models commonly used in the past, mainly because it uses sigmoid output instead of softmax in the final layer. In addition, the contracting path was repeated four times in the proposed U-Net model. The specific architecture of the U-Net, designed for segmentation, features a symmetric encoder–decoder structure with skip connections, which is particularly effective at capturing the fine details needed to segment small anatomical structures such as the appendix [35]. In contrast, state-of-the-art models are not as well-adapted to this task and often require additional layers or modules to achieve a similar performance [36,37].

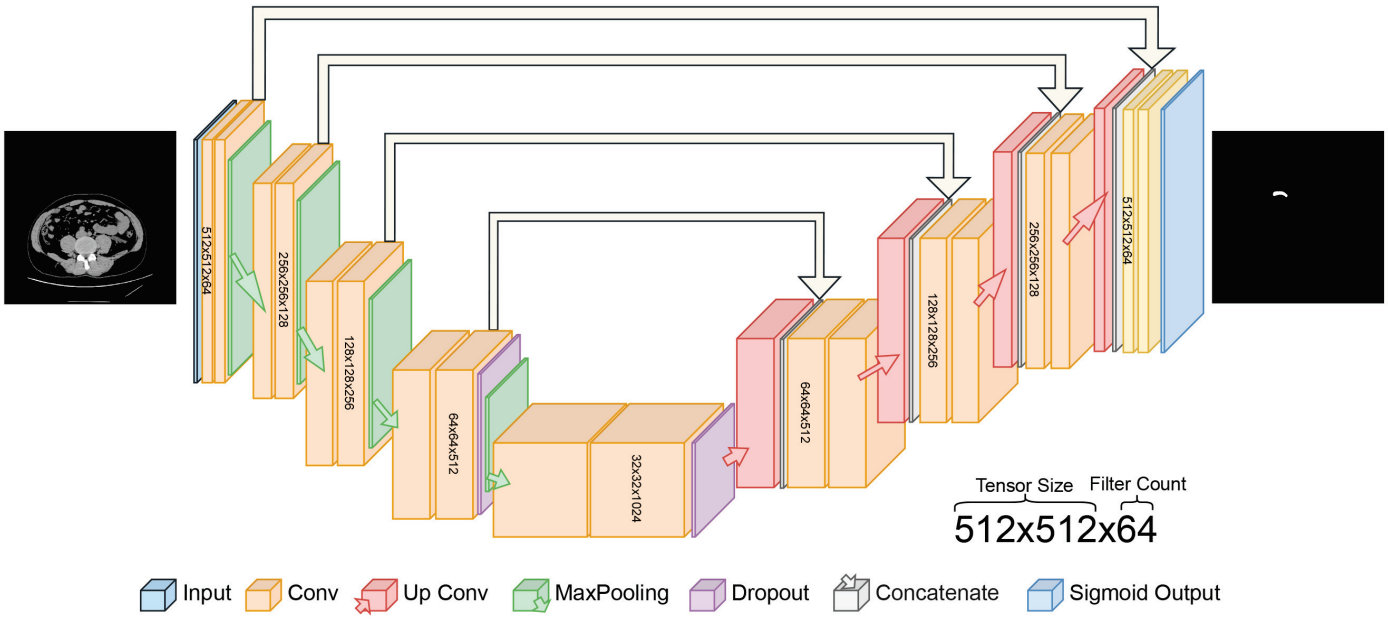


Figure 3. Proposed U-Net deep learning architecture for automated detection of appendix.

C. Key Performance Metrics

In this study, the performance of the proposed deep learning model for automatic appendix segmentation is evaluated using a number of important key metrics. The performance of the proposed model was measured using overlap-based, distance-based and pixel-based key metrics. Furthermore, these metrics were used to measure the similarity between the segmented areas (A_S) by the proposed U-Net-based deep learning model and the expert annotations (A_E) as a ground truth mask for the appendix. The Dice Similarity Coefficient (DSC), in Equation (5), is commonly used to compare the predicted segmentation result by the proposed method with the ground truth mask by the expert. On the other hand, the Volumetric Overlap Error (VOE), in Equation (6), is used to evaluate the dissimilarity between two volumes and measure how much the predicted segmentation deviates from the ground truth segmentation. The Average Symmetric Surface Distance (ASSD), a distance-based key metric, measures the average distance between the surfaces of the predicted segmentation and the ground truth and is given in Equation (7). The Hausdorff Distance (HD), given in Equations (8)–(10), is a measure of the distance between the

predicted segmentation boundary and the ground truth segmentation boundary. Hausdorff Distance 95 (HD95) uses this by taking the 95th percentile of the distances between boundary points rather than the maximum distance. Precision (PRE), in Equation (11), refers to the ratio of correctly predicted positive pixels (true positive, TP) to the total number of pixels predicted as positive (true positive + false positive, TP + FP) by the model in the segmentation tasks. Recall (REC) represents the proportion of correctly predicted positive pixels (TP) out of all true positive pixels (TP + false negative, FN) in the ground truth and is shown in in Equation (12).

$$\mathrm{DSC}\ (A_E, A_S) = \frac{2|A_E \cap A_S|}{|A_E| \cup |A_S|} \times 100 \tag{5}$$

$$\mathrm{VOE}\ (A_E, A_S) = \left(1 - \frac{|A_S \cap A_E|}{|A_S| + |A_E| - |A_S \cup A_E|}\right) \times 100 \tag{6}$$

$$\mathrm{ASSD}\ (A_E, A_S) = \frac{1}{|A_E| + |A_S|} \times \left(\sum_{x \in A_S} \min_{y \in A_E} d(x,y) + \sum_{y \in A_S} \min_{x \in A_E} d(y,x)\right) \tag{7}$$

$$\mathrm{hd}(A_S, A_E) = \max_{x \epsilon A_S} \min_{y \epsilon A_E} \|x - y\|_2 \tag{8}$$

$$\mathrm{hd}(A_E, A_S) = \max_{x \epsilon A_E} \min_{y \epsilon A_S} \|x - y\|_2 \tag{9}$$

$$\mathrm{HD}(A_S, A_E) = \max(\mathrm{hd}(A_S, A_E), \mathrm{hd}(A_E, A_S)) \tag{10}$$

$$\mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times 100 \tag{11}$$

$$\mathrm{REC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100 \tag{12}$$

3. Results

In this study, comprehensive experiments are conducted to evaluate the performance of the proposed U-Net-based deep learning model for appendix segmentation on CT scans from an original dataset. The experimental analyses reveal the performance and challenges of the proposed model in appendix segmentation. The experimental analyses were performed on a workstation equipped with an NVIDIA RTX 3060 GPU, 32 GB RAM, and an Intel i5-13400T CPU. The operating system used was Ubuntu 22 and the scripts were developed using Python with TensorFlow and Keras frameworks.

In this study, data augmentation was performed to address several key challenges in training deep learning models for appendix segmentation in CT scans. Data augmentation was applied to artificially increase the size and variability of the prepared dataset. This helps to avoid overfitting [38], where the model memorizes the training data rather than learning generalizable features. In addition, the CT scans in the dataset can vary significantly between patients in terms of anatomical differences, scan quality, patient position and image noise. The data augmentation enables the proposed U-Net deep learning architecture to learn more robust and discriminative features. This improves the model's ability to generalize to unseen data by simulating variations in real CT scans. In this study, in the original dataset prepared for appendix segmentation, the number of slices in the training set was increased from 1199 to 2399, the number of CT slices in the test set was increased from 221 to 441, and the total number of slices was increased from 1420 to 2840 using data augmentation strategies. The data augmentation procedures used to increase the size of the dataset were width_shift_range = 0.2 (WS), height_shift_range = 0.2 (HS), rotation_range = 2 (R) and zoom_range = 0.05 (Z). The data augmentation procedures for the dataset were randomly generated on both training and test sets separately or together. Figure 4 shows the effect of the data augmentation procedures on some CT scans and GT masks in the dataset.

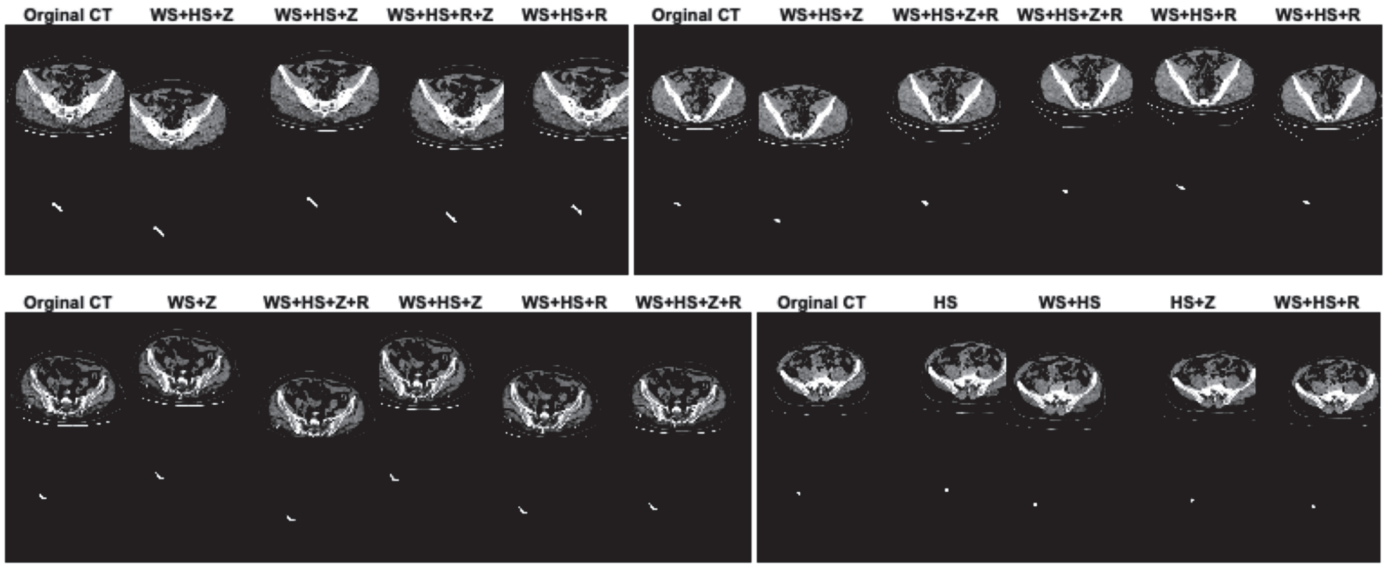


Figure 4. The effect of the data augmentation procedures on some CT scans and GT masks in the dataset.

The proposed U-Net model for appendix segmentation was subjected to training processes in the working environment using the original dataset obtained after data augmentation. In the original version of the dataset, a total of 1420 slices were collected from CT scans. Using the data augmentation procedures, the total number of CT slices in the dataset was doubled, resulting in a total of 2840 CT slices. For appendix segmentation, 2399 (~84%) of the 2840 images were allocated to the training phase and the remaining 441 (~16%) to the test phase.

In this study, the results of the proposed U-Net architecture for appendix detection from CT scans are also compared with the results of the state-of-the-art DenseNet [39] and Residual U-Net (Res U-Net) [40] architectures. The most prominent feature of DenseNet is that each layer of the network is directly connected to the outputs of all the previous layers and is widely used in segmentation problems. Res U-Net is a variation of the classical U-Net model where residual links are used as a basis to increase the depth of the model.

In experimental studies of this study, some hyperparameters in the U-Net architecture proposed for appendix segmentation are optimized and their optimal values for high segmentation performance are determined. First, ReLU is used as the activation function in the U-Net architecture. ReLU allows the U-Net to learn and model complex, non-linear mappings between input images and output segmentations. On the other hand, the Adam function was chosen as the optimization function. The Adam optimizer is well fitted to train the U-Net as it handles the complex nature of the deep layers of the architecture with adaptive learning rates and impulse-like behaviour. The learning rate was set to 0.001 for training the U-Net network. The learning rate plays a significant role in how quickly or slowly the U-Net model converges to an optimal solution as it learns from the data, especially for image segmentation tasks such as appendix detection in CT scans. Training was performed with these parameters for 100 epochs. The development of the loss during the training process is shown in Figure 5a. Looking at this figure, it is clear that the training process achieved a continuous improvement. During the training process, the weights after 100 epochs were obtained and subjected to the test phase. The DSC evolution for the 100 epochs can be seen in Figure 5b, which illustrates the performance during testing. It can be seen that the DSC value reaches an effective level after the first 10 epochs.

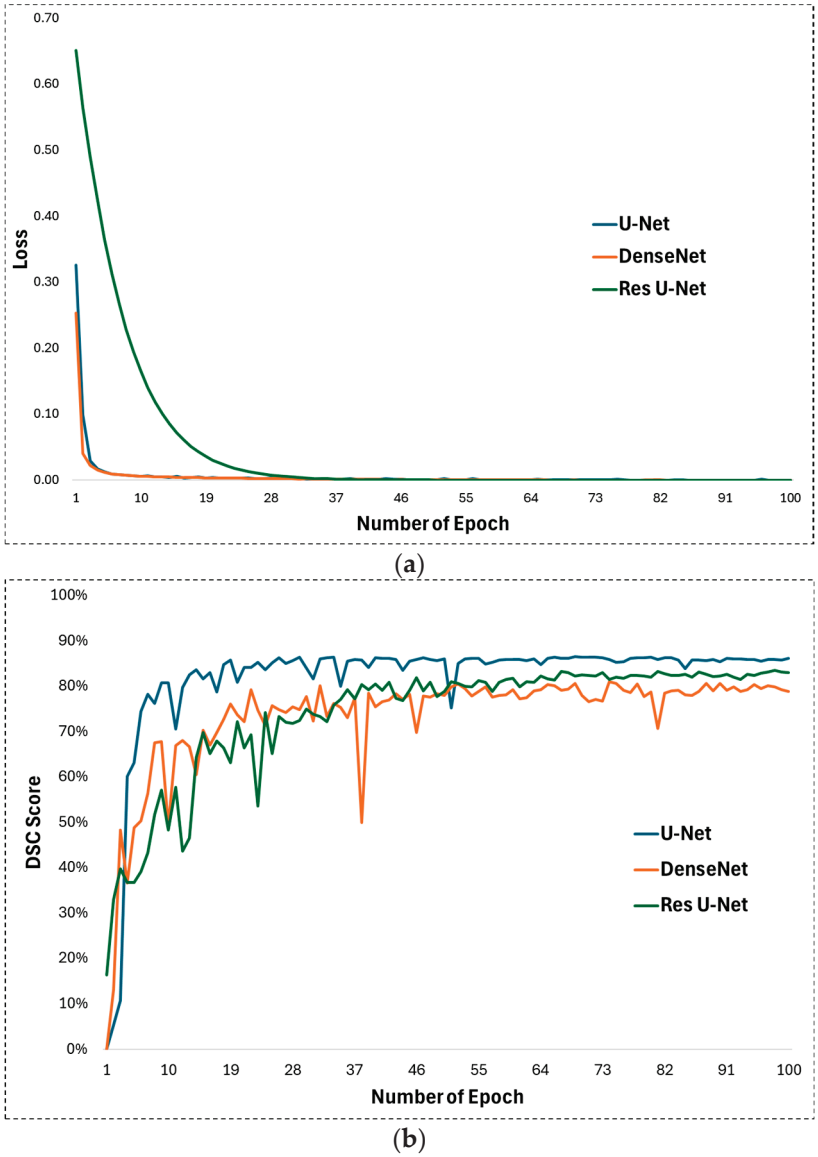


Figure 5. (**a**) Training loss and (**b**) DSC development during test phase for U-Net, DenseNet, and Res U-Net methods.

The appendix regions successfully detected and segmented on CT slices using the proposed U-Net deep learning architecture during the experimental studies are shown in Figure 6. While the mask of an original CT slice in Figure 6a containing the appendix region delineated by the expert is shown in Figure 6b, the segmentation of the appendix by the proposed U-Net deep learning architecture is shown in Figure 6c along with the DSC score. On the other hand, the overlap of the expert mask and the U-Net segmentation on the CT slice is also shown in Figure 6d and its zoomed version is shown in Figure 6e. As can be seen, the proposed U-Net-based deep learning architecture is very successful in segmenting the appendix CT scans and is very close to the expert GTs. Some examples of the segments that the proposed model has difficulty in detecting are shown in Figure 7. When analysing Figure 7, it is clear that the model cannot successfully detect some slices. The lack of success in these slices can be attributed to the fact that the appendix region is close to or adjacent to other areas, or the boundaries of the region are similar to those of the neighbouring areas. Such cases stood out as instances where the boundaries of the appendix were blurred and the model was unstable.

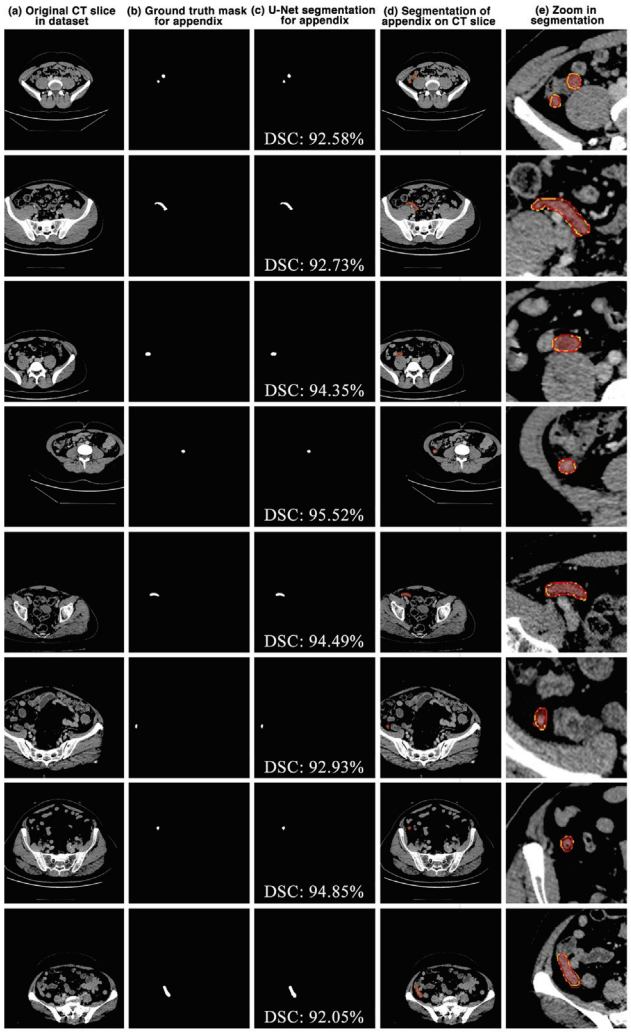


Figure 6. The appendix regions successfully detected and segmented on CT slices using the proposed U-Net deep learning architecture during the experimental studies. Red: ground truth mask for appendix, yellow: U-Net segmentation for appendix.

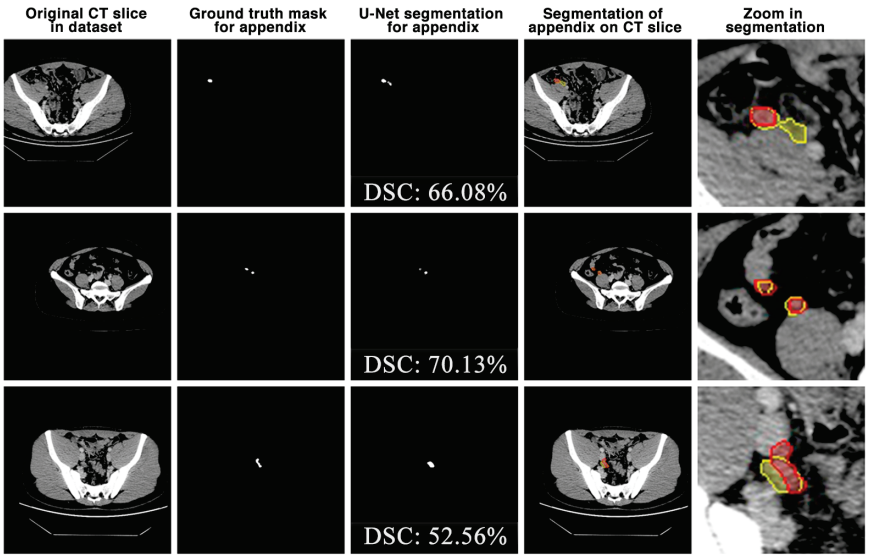


Figure 7. Some examples of unsuccessful appendix detection and segmentation by the proposed U-Net model. Red: ground truth mask for appendix, yellow: U-Net segmentation for appendix.

4. Discussion

The performance metrics results obtained using the proposed U-Net deep learning architecture and other state-of-the-art models for appendix segmentation are shown in Table 1. The proposed U-Net model is compared with other previous deep learning-based methods. Our model is particularly successful in accurately segmenting small and ambiguous patch boundaries. Comparisons with other methods show that U-Net provides better performance due to its encoder–decoder structure and the ability of skip connections to preserve spatial information. For DSC, which is the most important key metric for appendix segmentation performance, the proposed U-Net architecture achieves a score of 86.58%, while the Res U-Net and DenseNet architectures achieves 83.53% and 80.64% for the same metric, respectively. For the other key metrics with 22.99% for VOE, 1.08 mm for ASSD, 3.87 mm for HD95, and 87.08% for REC, it can be concluded that the U-Net architecture is more successful than Res U-Net and DenseNet. In addition, it is seen that DenseNet architecture is more successful with 88.56% only for PRE key metric.

Table 1. The comparison of the results of the proposed U-Net architecture with the results of the state-of-the-art DenseNet and Res U-Net architectures for the detection of the appendix from CT scans.

Method	DSC [%]	VOE [%]	ASSD [mm]	HD95 [mm]	PRE [%]	REC [%]
DenseNet	80.64	30.12	1.49	5.84	88.56	77.49
Res U-Net	83.53	26.99	1.73	6.93	88.06	81.51
Proposed U-Net	86.58	22.99	1.08	3.87	87.08	87.08

In Figure 8, the average scores achieved on the test set of CT slices for the proposed U-Net and other state-of-the-art methods such as DenseNet and Res U-Net are represented by box plots in terms of DSC in Figure 8a, VOE in Figure 8b, ASSD in Figure 8c, HD95 in Figure 8d, PRE in Figure 8e and REC in Figure 8f key metrics. From these results, it can be concluded that the proposed U-Net-based deep learning model is highly successful in detecting the appendix. Furthermore, our dataset, which was annotated by expert physicians and included CT scans from both healthy individuals and patients, provided high-quality and diverse training data, which enhanced the generalization of the U-Net model. We used hyper-parameter optimization with the ReLu activation function and the ADAM optimizer, which are known for efficient training and convergence [41,42]. These choices, combined with comprehensive evaluation metrics such as DSC, VOE, ASSD and HD95, ensured detailed evaluation and high performance. Such differences in architecture, data quality, augmentation, and evaluation strategies are likely to contribute to U-Net model's competitive DSC of 86.58% in appendix detection, highlighting the importance of these factors in achieving superior diagnostic performance.

Due to the absence of previous studies specifically focusing on appendix segmentation, it was not possible to compare our study's results directly with existing studies on this topic. However, comparisons can be drawn from similar tasks in the appendix such as segmentation, diagnosis, and classification, as shown in Table 2. In previous studies, classification research has generally been carried out using different CNN architectures in the diagnosis of appendicitis [15,43–45]. However, it is very important to determine the location of the appendix from the scans in cases. Therefore, in this study, the segmentation of the appendix is first achieved with a high score, and an important step in the diagnosis is completed. These comparative analyses help highlight the potential and robustness of our proposed model for appendix segmentation, suggesting its promising application in this area.

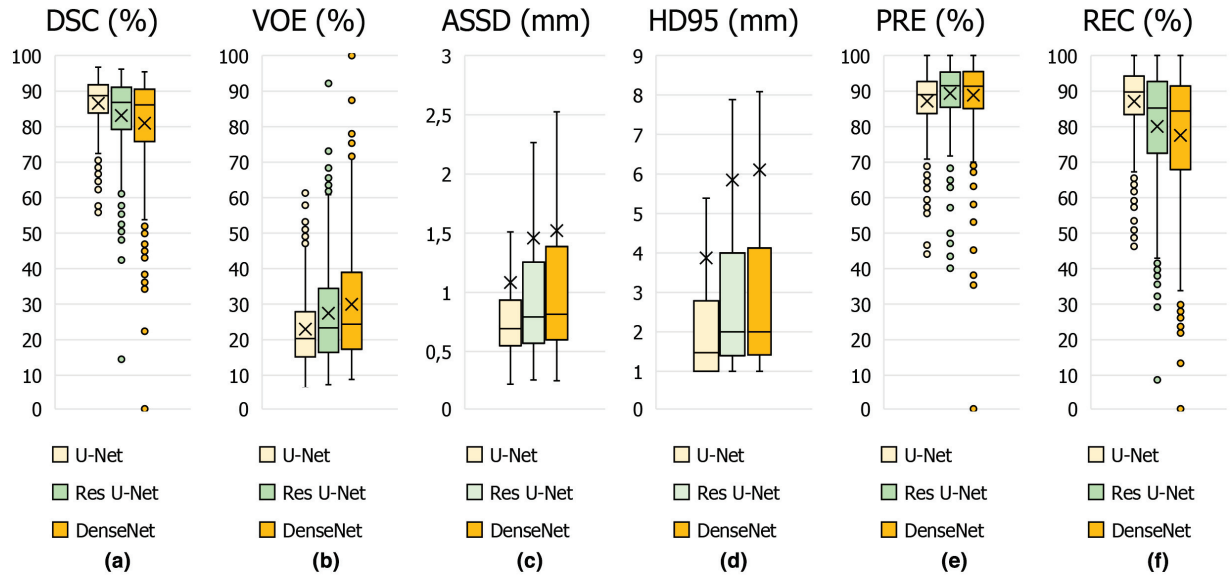


Figure 8. Boxplot showing the performance metrics for appendix segmentation obtained using the proposed U-Net deep learning architecture and other models in the study.

Table 2. Comparison of previous studies for automatic segmentation of appendix and classification of appendicitis.

Study	Year	Number of Images/Subjects	Research Topic	Methodology	Key Metric Evaluation (%)
Al et al. [43]	2019	319 CT examinations	Classification of acute appendicitis	Reinforcement Learning and CNN	AUC = 96.1
Rajpurgar et al. [15]	2020	646 CT examinations	Classification of appendicitis	3D CNN	AUC = 82.6
Park et al. [44]	2020	667 CT images	Classification of appendicitis	3D CNN	AUC = 96.0
Park et al. [45]	2023	4078 CT images	Classification of acute appendicitis, diverticulitis, and normal appendix	CNN (EfficientNet)	AUC = 95.1 (acute appendicitis) AUC = 97.2 (acute diverticulitis) AUC = 97.9 (normal appendix)
Ours	2024	940 CT images	Segmentation of appendix	U-Net-based deep learning architecture with hyperparameter optimization	DSC = 86.58

Automated appendix detection using U-Net deep learning architecture has some significant advantages and some disadvantages. On the positive side, these systems are much faster than manual methods, saving valuable time and enabling faster diagnoses. They offer high accuracy and consistency because they are trained on large datasets, minimizing human error. In addition, deep learning models can be customized for different patient populations and imaging modalities, resulting in more personalized and accurate analyses. These systems can process large amounts of data, improving their performance over time, and are able to detect small abnormalities in the early stages of appendicitis, allowing for earlier treatment. However, there are challenges to consider. The performance of deep learning models is highly dependent on the quality and diversity of the training data. Poor quality or insufficient datasets can lead to inaccurate results. Developing and implementing these systems can be expensive, both initially and in terms of ongoing maintenance and updates. The complexity of building and training these models requires

specialized expertise, which may require the hiring of skilled personnel. There is also a risk of misleading results, particularly in rare cases or situations not represented in the training data, which can lead to false positives or false negatives. Finally, ensuring the privacy and security of medical data is critical, and robust protocols must be in place to protect this sensitive information. In conclusion, while the fully automated segmentation of the appendix region using deep learning models can significantly improve the speed and accuracy of medical diagnoses, its success depends largely on the quality of the data and the continuous improvement of the system.

Figure 9 shows the comparison of the segmentation performances of the proposed U-Net deep learning model and other state-of-the-art DenseNet and Res U-Net architectures on the same CT slices in the dataset in terms of the DSC key metric. Although the performances of the architectures are close to each other for the segmentation results, it is seen that the U-Net architecture is slightly more successful.

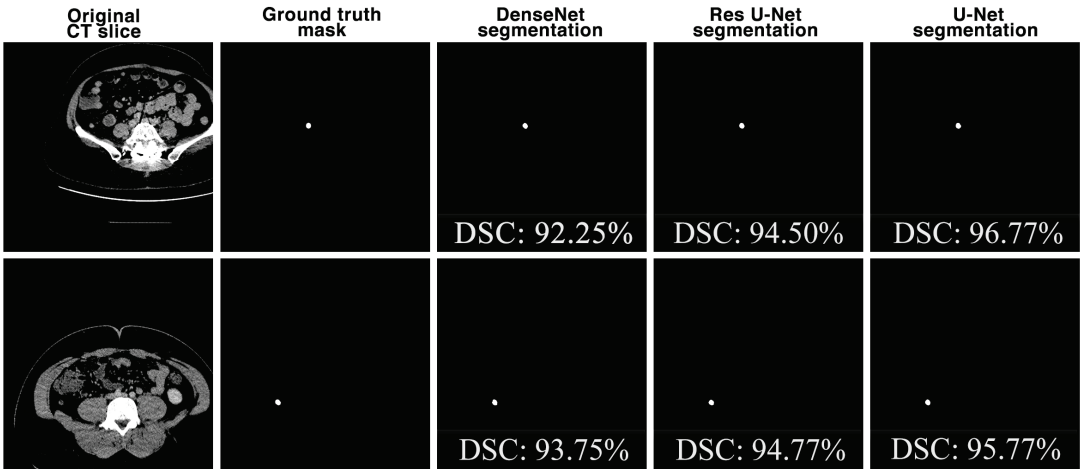


Figure 9. The comparison of the appendix segmentation performances of the proposed U-Net deep learning model and other state-of-the-art DenseNet and Res U-Net architectures on the same CT slices.

5. Conclusions

Our study showed the significant potential of using a U-Net-based deep learning model for the accurate segmentation and detection of the appendix in CT images. With an achieved DSC score of 86.58%, the model shows promise in aiding clinical diagnosis and potentially improving patient outcomes. Detailed annotation by expert physicians and extensive data augmentation techniques contributed to the robustness and reliability of the model. The results of this study, demonstrating high performance in the detection of appendicitis using a U-Net-based deep learning model, have significant potential clinical implications. Improved diagnostic accuracy can reduce misdiagnosis and prevent complications such as perforation and sepsis, while the reduction in diagnostic time enables faster treatment decisions. The model's ability to provide consistent and standardized diagnoses reduces clinician variability and human error. In addition, the model's rapid and accurate diagnosis can optimize healthcare resources by minimizing unnecessary imaging and surgery. As a decision support tool, the model can assist radiologists by highlighting areas of interest and providing second opinions, which is particularly beneficial for less experienced practitioners. Early and accurate detection allows for timely intervention, improving patient outcomes. Integrating the U-Net model into clinical workflows can streamline the diagnostic process, and further research could extend its applications by refining the model and incorporating multimodal data to improve its accuracy. Overall, the robust performance of the model highlights the transformative potential of deep learning in medical diagnostics, particularly in the timely and accurate diagnosis of appendicitis.

To further improve our appendix detection model, we can focus on expanding and diversifying the dataset by including more CT scans from a varied patient population and multiple medical centres to improve generalization. Advanced data augmentation techniques, such as 3D augmentation and the use of generative adversarial networks for synthetic data generation, can introduce additional variability. Improving the model architecture by integrating hybrid models, using transfer learning, and applying ensemble methods can also improve performance. Automated hyperparameter tuning and advanced regularization techniques can optimize model parameters and prevent overfitting. The integration of clinical and imaging data through multimodal fusion techniques can provide a more comprehensive input for improved accuracy. Finally, continuous evaluation in real-world clinical settings and regular model updates with new data can ensure that the model remains accurate and relevant, ultimately leading to better diagnostic outcomes and improved patient care. To further improve model performance, future work should focus on expanding the dataset with more diverse and comprehensive samples, employing advanced augmentation and hybrid modelling techniques, and integrating multimodal data. In addition, the incorporation of explicable AI methods may improve the interpretability and trustworthiness of the model. Continuous evaluation in clinical settings and regular updates with new data can ensure that the model remains accurate and relevant. These advances can solidify the role of AI in medical diagnostics, ultimately leading to more precise and timely treatments for patients.

Author Contributions: Conceptualization: B.T.B. and E.D.; methodology: B.T.B., E.D., and M.S.Y.; software: E.D. and M.S.Y.; validation: E.D., B.T.B., K.Ç., and G.G.; formal analysis: E.D., B.T.B., M.S.Y., K.Ç., and G.G.; investigation: E.D., B.T.B., M.S.Y., K.Ç., and G.G.; data curation: B.T.B., K.Ç., and G.G.; writing—original draft preparation: B.T.B., E.D., and M.S.Y.; writing—review and editing: B.T.B., K.Ç., E.D., and M.S.Y.; visualization: E.D. and M.S.Y.; supervision: B.T.B. and E.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: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board According to Decision No. 13 of the 8th meeting of the Ethics Committee of Non-Interventional Clinical Research of Bilecik Şeyh Edebali University dated 5 December 2023.

Informed Consent Statement: Informed consent was obtained from all the subjects involved in the study.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: We would like to thank Bilecik Training and Research Hospital for providing the dataset and all the patients who volunteered to participate. All individuals have agreed to this consent. Furthermore, we dedicate this study to the 100th anniversary of the foundation of the Republic of Türkiye.

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

References

1. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]
2. McKinney, S.M.; Sieniek, M.; Godbole, V.; Godwin, J.; Antropova, N.; Ashrafian, H.; Back, T.; Chesus, M.; Corrado, G.S.; Darzi, A. International evaluation of an AI system for breast cancer screening. *Nature* **2020**, *577*, 89–94. [CrossRef] [PubMed]
3. Obuchowicz, R.; Strzelecki, M.; Piórkowski, A. Clinical Applications of Artificial Intelligence in Medical Imaging and Image Processing—A Review. *Cancers* **2024**, *16*, 1870. [CrossRef] [PubMed]
4. Aggarwal, R.; Sounderajah, V.; Martin, G.; Ting, D.S.; Karthikesalingam, A.; King, D.; Ashrafian, H.; Darzi, A. Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. *NPJ Digit. Med.* **2021**, *4*, 65. [CrossRef] [PubMed]
5. Kim, S.Y.; Kim, D.H.; Kim, M.J.; Ko, H.J.; Jeong, O.R. XAI-Based Clinical Decision Support Systems: A Systematic Review. *Appl. Sci.* **2024**, *14*, 6638. [CrossRef]

6. Kostopoulos, G.; Davrazos, G.; Kotsiantis, S. Explainable Artificial Intelligence-Based Decision Support Systems: A Recent Review. *Electronics* **2024**, *13*, 2842. [CrossRef]
7. Ouanes, K.; Farhah, N. Effectiveness of Artificial Intelligence (AI) in Clinical Decision Support Systems and Care Delivery. *J. Med. Syst.* **2024**, *48*, 74. [CrossRef]
8. Bozyel, S.; Şimşek, E.; Koçyiğit, D.; Güler, A.; Korkmaz, Y.; Şeker, M.; Ertürk, M.; Keser, N. Artificial intelligence-based clinical decision support systems in cardiovascular diseases. *Anatol. J. Cardiol.* **2024**, *28*, 74. [CrossRef]
9. Metta, C.; Beretta, A.; Pellungrini, R.; Rinzivillo, S.; Giannotti, F. Towards Transparent Healthcare: Advancing Local Explanation Methods in Explainable Artificial Intelligence. *Bioengineering* **2024**, *11*, 369. [CrossRef]
10. Birnbaum, B.A.; Wilson, S.R. Appendicitis at the millennium. *Radiology* **2000**, *215*, 337–348. [CrossRef]
11. Khalifa, M.; Albadawy, M. AI in diagnostic imaging: Revolutionising accuracy and efficiency. *Comput. Methods Programs Biomed. Update* **2024**, *5*, 100146. [CrossRef]
12. Zeltzer, D.; Herzog, L.; Pickman, Y.; Steuerman, Y.; Ber, R.I.; Kugler, Z.; Shaul, R.; Ebbert, J.O. Diagnostic accuracy of artificial intelligence in virtual primary care. *Mayo Clin. Proc. Digit. Health* **2023**, *1*, 480–489. [CrossRef]
13. Yuan, R.; Janzen, I.; Devnath, L.; Khattra, S.; Myers, R.; Lam, S.; MacAulay, C. MA19. 11 Predicting Future Lung Cancer Risk with Low-Dose Screening CT Using an Artificial Intelligence Model. *J. Thorac. Oncol.* **2023**, *18*, S174. [CrossRef]
14. Mijwil, M.M.; Aggarwal, K. A diagnostic testing for people with appendicitis using machine learning techniques. *Multimed. Tools Appl.* **2022**, *81*, 7011–7023. [CrossRef]
15. Rajpurkar, P.; Park, A.; Irvin, J.; Chute, C.; Bereket, M.; Mastrodicasa, D.; Langlotz, C.P.; Lungren, M.P.; Ng, A.Y.; Patel, B.N. AppendiXNet: Deep learning for diagnosis of appendicitis from a small dataset of CT exams using video pretraining. *Sci. Rep.* **2020**, *10*, 3958. [CrossRef]
16. Issaiy, M.; Zarei, D.; Saghazadeh, A. Artificial intelligence and acute appendicitis: A systematic review of diagnostic and prognostic models. *World J. Emerg. Surg.* **2023**, *18*, 59. [CrossRef]
17. Liang, D.; Fan, Y.; Zeng, Y.; Zhou, H.; Zhou, H.; Li, G.; Liang, Y.; Zhong, Z.; Chen, D.; Chen, A. Development and validation of a deep learning and radiomics combined model for differentiating complicated from uncomplicated acute appendicitis. *Acad. Radiol.* **2024**, *31*, 1344–1354. [CrossRef]
18. Marcinkevics, R.; Reis Wolfertstetter, P.; Wellmann, S.; Knorr, C.; Vogt, J.E. Using machine learning to predict the diagnosis, management and severity of pediatric appendicitis. *Front. Pediatr.* **2021**, *9*, 662183. [CrossRef]
19. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K. Chexnet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv* **2017**, arXiv:1711.05225.
20. Sarvamangala, D.; Kulkarni, R.V. Convolutional neural networks in medical image understanding: A survey. *Evol. Intell.* **2022**, *15*, 1–22. [CrossRef]
21. Aboussaleh, I.; Riffi, J.; Fazazy, K.E.; Mahraz, M.A.; Tairi, H. Efficient U-Net architecture with multiple encoders and attention mechanism decoders for brain tumor segmentation. *Diagnostics* **2023**, *13*, 872. [CrossRef] [PubMed]
22. Ma, D.; Lu, D.; Chen, S.; Heisler, M.; Dabiri, S.; Lee, S.; Lee, H.; Ding, G.W.; Sarunic, M.V.; Beg, M.F. Lf-unet–a novel anatomical-aware dual-branch cascaded deep neural network for segmentation of retinal layers and fluid from optical coherence tomography images. *Comput. Med. Imaging Graph.* **2021**, *94*, 101988. [CrossRef] [PubMed]
23. Platania, R.; Shams, S.; Yang, S.; Zhang, J.; Lee, K.; Park, S.-J. Automated breast cancer diagnosis using deep learning and region of interest detection (bc-droid). In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Boston, MA, USA, 20–23 August 2017; pp. 536–543.
24. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, 17–21 October 2016; Proceedings, Part II 19. pp. 424–432.
25. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
26. Shorten, C.; Khoshgoftaar, T.M.; Furht, B. Text data augmentation for deep learning. *J. Big Data* **2021**, *8*, 101. [CrossRef] [PubMed]
27. Zhang, X.; Yang, S.; Jiang, Y.; Chen, Y.; Sun, F. FAFS-UNet: Redesigning skip connections in UNet with feature aggregation and feature selection. *Comput. Biol. Med.* **2024**, *170*, 108009. [CrossRef]
28. Yushkevich, P.A.; Piven, J.; Hazlett, H.C.; Smith, R.G.; Ho, S.; Gee, J.C.; Gerig, G. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage* **2006**, *31*, 1116–1128. [CrossRef]
29. 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.
30. Micallef, N.; Seychell, D.; Bajada, C.J. Exploring the u-net++ model for automatic brain tumor segmentation. *IEEE Access* **2021**, *9*, 125523–125539. [CrossRef]
31. Ryu, J.Y.; Hong, H.K.; Cho, H.G.; Lee, J.S.; Yoo, B.C.; Choi, M.H.; Chung, H.Y. Deep learning for the automatic segmentation of extracranial venous malformations of the head and neck from MR Images Using 3D U-Net. *J. Clin. Med.* **2022**, *11*, 5593. [CrossRef]

32. Polattimur, R.; Dandıl, E.; Yıldırım, M.S.; Şenol, A.U.; Tezel, Z.E.; Selvi, A.O.; Kabay, S.C. Fully automated axial plane segmentation of cervical spinal cord using U-Net in MR scans. In Proceedings of the 2023 7th International Symposium on Innovative Approaches in Smart Technologies (ISAS), İstanbul, Türkiye, 23–25 November 2023; pp. 1–7.
33. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
34. Wu, J.; Zhang, Y.; Wang, K.; Tang, X. Skip connection U-Net for white matter hyperintensities segmentation from MRI. *IEEE Access* **2019**, *7*, 155194–155202. [CrossRef]
35. 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. pp. 234–241.
36. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
37. 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.
38. Salehi, A.W.; Khan, S.; Gupta, G.; Alabduallah, B.I.; Almjally, A.; Alsolai, H.; Siddiqui, T.; Mellit, A. A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope. *Sustainability* **2023**, *15*, 5930. [CrossRef]
39. Iandola, F.; Moskewicz, M.; Karayev, S.; Girshick, R.; Darrell, T.; Keutzer, K. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv* **2014**, arXiv:1404.1869.
40. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
41. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 807–814.
42. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
43. Al, W.A.; Yun, I.D.; Lee, K.J. Reinforcement learning-based automatic diagnosis of acute appendicitis in abdominal ct. *arXiv* **2019**, arXiv:1909.00617.
44. Park, J.J.; Kim, K.A.; Nam, Y.; Choi, M.H.; Choi, S.Y.; Rhie, J. Convolutional-neural-network-based diagnosis of appendicitis via CT scans in patients with acute abdominal pain presenting in the emergency department. *Sci. Rep.* **2020**, *10*, 9556. [CrossRef]
45. Park, S.H.; Kim, Y.J.; Kim, K.G.; Chung, J.-W.; Kim, H.C.; Choi, I.Y.; You, M.-W.; Lee, G.P.; Hwang, J.H. Comparison between single and serial computed tomography images in classification of acute appendicitis, acute right-sided diverticulitis, and normal appendix using EfficientNet. *PLoS ONE* **2023**, *18*, e0281498. [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.

Journal of
Clinical Medicine

MDPI

Article

Opportunistic Screening for Low Bone Mineral Density in Adults with Cystic Fibrosis Using Low-Dose Computed Tomography of the Chest with Artificial Intelligence

Matthias Welsner [1,*], Henning Navel [2,3], Rene Hosch [3], Peter Rathsmann [4], Florian Stehling [5], Annie Mathew [6], Sivagurunathan Sutharsan [1], Svenja Strassburg [1], Dirk Westhölter [1], Christian Taube [1], Sebastian Zensen [7], Benedikt M. Schaarschmidt [7], Michael Forsting [7], Felix Nensa [3], Mathias Holtkamp [3,7], Johannes Haubold [3,7], Luca Salhöfer [3,7,†] and Marcel Opitz [7,†]

[1] Department of Pulmonary Medicine, Adult Cystic Fibrosis Center, University Hospital Essen-Ruhrlandklinik, University of Duisburg-Essen, 45239 Essen, Germany
[2] Department of Electrical Engineering and Applied Natural Sciences, Westphalian University of Applied Sciences, 45897 Gelsenkirchen, Germany
[3] Institute for Artificial Intelligence in Medicine, University Medicine Essen, 45147 Essen, Germany
[4] Department of Radiology, St. Josef Hospital Werden, University Medicine Essen, 45239 Essen, Germany
[5] Pediatric Pulmonology and Sleep Medicine, Cystic Fibrosis Center, Children's Hospital, University of Duisburg-Essen, 45147 Essen, Germany
[6] Department of Endocrinology, Diabetes and Metabolism, Division of Laboratory Research, University Hospital Essen, 45147 Essen, Germany
[7] Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, University of Duisburg-Essen, 45147 Essen, Germany
* Correspondence: matthias.welsner@rlk.uk-essen.de
† These authors contributed equally to this work.

Abstract: Background: Cystic fibrosis bone disease (CFBD) is a common comorbidity in adult people with cystic fibrosis (pwCF), resulting in an increased risk of bone fractures. This study evaluated the capacity of artificial intelligence (AI)-assisted low-dose chest CT (LDCT) opportunistic screening for detecting low bone mineral density (BMD) in adult pwCF. **Methods:** In this retrospective single-center study, 65 adult pwCF (mean age 30.1 ± 7.5 years) underwent dual-energy X-ray absorptiometry (DXA) of the lumbar vertebrae L1 to L4 to determine BMD and corresponding z-scores and completed LDCTs of the chest within three months as part of routine clinical care. A fully automated CT-based AI algorithm measured the attenuation values (Hounsfield units [HU]) of the thoracic vertebrae Th9–Th12 and first lumbar vertebra L1. The ability of the algorithm to diagnose CFBD was assessed using receiver operating characteristic (ROC) curves. **Results:** HU values of Th9 to L1 and DXA-derived BMD and the corresponding z-scores of L1 to L4 showed a strong correlation (all $p < 0.05$). The area under the curve (AUC) for diagnosing low BMD was highest for L1 (0.796; $p = 0.001$) and Th11 (0.835; $p < 0.001$), resulting in a specificity of 84.9% at a sensitivity level of 75%. The HU threshold values for distinguishing normal from low BMD were <197 (L1) and <212 (Th11), respectively. **Conclusions:** Routine LDCT of the chest with the fully automated AI-guided determination of thoracic and lumbar vertebral attenuation values is a valuable tool for predicting low BMD in adult pwCF, with the best results for Th11 and L1. However, further studies are required to define clear threshold values.

Keywords: cystic fibrosis; adults; bone disease; osteoporosis; computed tomography; DXA; artificial intelligence; opportunistic screening; bone mineral density

1. Introduction

Cystic fibrosis (CF) is a rare autosomal recessive disease with serious chronically debilitating morbidities. CF is caused by a decreased quantity and/or function of the

cystic fibrosis transmembrane regulator (CFTR) protein [1]. The CFTR is an ion channel that regulates the flow of chloride and other ions across the epithelia in various tissues, including the lungs, pancreas, other gastrointestinal organs, and sweat glands. Decreased CFTR quantity or function results in failure to regulate ion transport in these tissues, leading to a multisystem pathology associated with CF [2].

First described in 1979, bone disease is a well-recognized comorbidity in people with cystic fibrosis (pwCF), and its prevalence is increasing in the aging CF population [3,4]. Low bone mineral density (BMD) can increase the risk of low-trauma bone fractures and may result in the development of postural deviations, such as hyperkyphosis and scoliosis, which can have negative consequences for cough and airway clearance and can also limit the effectiveness of chest physiotherapy, contributing to the development of pulmonary exacerbations [5,6]. Therefore, it is crucial to diagnose bone demineralization at an early stage to ensure the appropriate treatment, management, and prevention of low-trauma fractures.

Dual-energy X-ray absorptiometry (DXA) is considered the gold standard for the diagnosis of cystic fibrosis-related bone disease (CFBD) [7]. CFBD is defined as the presence of low BMD, expressed as height- and age-adjusted z-scores of less than -2.0 standard deviations (SDs) determined by DXA [8]. European cystic fibrosis bone mineralization guidelines recommend routine bone density scans every five years if the BMD z-score is >-1, every two years if the z-score is between -1 and -2, and every year if the z-score is below -2 in adult pwCF younger than 50 years of age [8]. Therefore, osteoporosis in children, adolescents, and adults with CF is defined as having a BMD Z score of <-2 and a significant (low-trauma) fracture history [8].

The etiology of CFBD is multifactorial and encompasses a wide range of factors, including exocrine pancreatic insufficiency, diabetes mellitus, chronic inflammation, sex hormone deficiency, vitamin D and vitamin K deficiency, calcium malabsorption, malnutrition, delayed puberty, use of exogenous glucocorticoids, physical inactivity, and the direct effect of CFTR dysfunction on bone cell activity [9,10]. Even with normal BMD as measured with DXA, pwCF exhibits imbalanced bone turnover that favors bone resorption, abnormal bone microstructure, and increased risk of fractures compared to people without CF [11–13].

Despite the known high prevalence of osteoporosis (23.5%) and osteopenia (38%) in adult pwCF, current screening methods for CFBD, such as DXA, are often limited due to factors, such as accessibility, cost, and the potential risk of additional radiation exposure [6]. The significance of these limitations becomes particularly relevant in the context of CF, considering its growing life expectancy and the subsequent increase in the prevalence of osteopenia and osteoporosis with advancing age [14]. Therefore, it is encouraging to explore and establish alternative screening procedures that can overcome these limitations and improve screening efficiency for the early detection and management of low BMD and prevention of low-trauma bone fractures.

Low-dose computed tomography (LDCT) of the chest is a widely used imaging technique for detecting structural lung damage and evaluating the progression of pulmonary disease in adult pwCF [15]. LDCT scans of the chest with and without artificial intelligence (AI) techniques have been explored as an opportunistic approach for assessing low BMD in various studies outside of CF [16–21]. Opportunistic screening (OS) describes a method that enables the extraction of useful information from pre-existing images or image stacks such as chest CT scans, which were originally obtained for purposes unrelated to the targeted information [22]. These studies explored the feasibility of using CT attenuation values (Hounsfield Units [HU]) as surrogate markers for BMD measured using DXA and showed their remarkable potential for opportunistic low BMD screening. Despite the promising potential of LDCT with AI for OS of low BMD, its application in pwCF remains relatively unexplored, particularly in the context of establishing dependable HU threshold values for this unique population. Research studies that focus on the CF population, for whom the early detection of low BMD could significantly impact management approaches and

outcomes, are scarce, thereby revealing a significant gap in the current research and clinical practice landscape. The present study aims to investigate the feasibility of implementing a fully automated AI-assisted approach for screening CFBD using LDCT scans of the chest. This method aims to automatically measure the attenuation values of the thoracic (Th9–Th12) and the first lumbar vertebra (L1) during routine clinical practice and assess their usefulness in identifying CFBD in adult pwCF.

2. Methods

2.1. Study Design and Participants

This single-center retrospective study identified 71 pwCF aged 18–49 years from the Adult Cystic Fibrosis Center of Ruhrlandklinik Essen, Germany, who underwent DXA and LDCT of the chest within a 3-month period between April 2019 and May 2024, as part of routine clinical care. All subjects had a confirmed diagnosis of CF based on two disease-defining variants in their CFTR gene. Of the 71 pwCF, five patients were excluded from the study because of incorrect recognition of the region of interest (ROI) by AI (Figure 1), and one patient was excluded because of metal artifacts in the lumbar vertebrae that prevented DXA from being performed. After exclusion, 65 pwCF samples were available for analysis. Clinically relevant data, including pulmonary function test results, body mass index (BMI), 25-hydroxyvitamin D levels, and medication, were obtained from the electronic records nearest to DXA. None of the participants had a history of significant bone fracture. The duration between DXA and CT scan measurements was 6 ± 16 days.

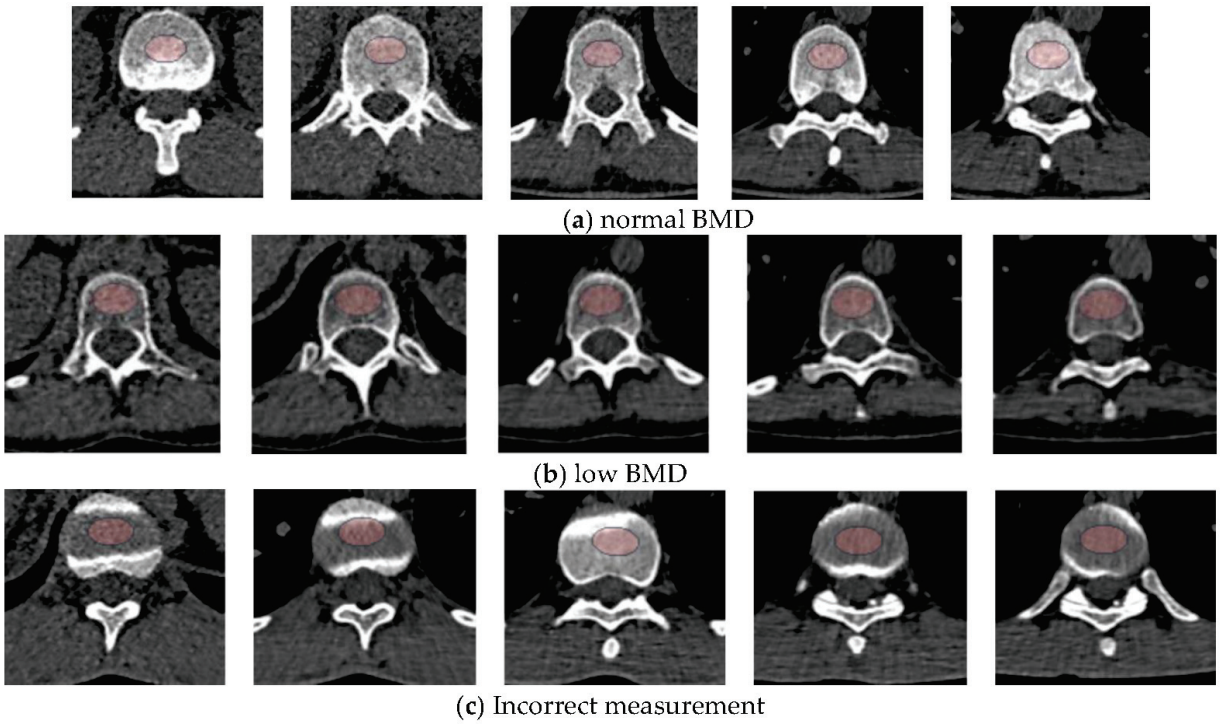


Figure 1. Examples of AI-assisted determinations of attenuation values for vertebral bodies L1 to Th9 using low-dose computed tomography (LDCT). From left to right: L1–Th9. (**a**) Normal findings. The ROI placed by the AI was appropriately centered on the vertebral body to be measured. The measured HU values ranged from 223 to 364. (**b**) Low BMD: The ROI is centered correctly by the AI on the vertebral bodies for the assessment of attenuation values. This results in HU values ranging between 89 and 117. (**c**) Incorrect measurements: The ROIs for L1, Th12, Th10, and Th9 were placed within the intervertebral disc space, resulting in underscored HU measurements for L1 (89), Th12 (122), Th10 (88), and Th9 (94). Conversely, the ROI for Th11 was positioned centrally within the vertebral body, with an HU value of 255. AI, artificial intelligence; ROI, region of interest; HU, Hounsfield unit; BMD, bone mass density.

This retrospective single-center study was approved by the local review board of the Essen University Hospital (no. 24-11977-BO) and followed the ethical principles of the Declaration of Helsinki for medical research involving human subjects.

2.2. Image Acquisition

2.2.1. DXA Scanning

BMD was measured using a Horizon Ci scanner (Hologic Inc., Marlborough, MA, USA, software version 5.6) and a Lunar Prodigy scanner (GE Healthcare Technologies, Chicago, IL, USA). Routine lumbar anteroposterior scans were conducted with the BMD region of interest comprising the 1st to 4th lumbar vertebrae. Manufacturer reference values were used to determine the z-scores. A z-score of <-2 indicated low BMD.

2.2.2. Computer Tomography and AI

All patients included in this study received non-contrast-enhanced LDCT scans of the chest using a 64-detector row single-source CT scanner (SOMATOM Definition AS, Siemens Healthcare GmbH, Forchheim, Germany) with a gantry rotation time of 300 ms (collimation: 64 $\times$ 0.6 mm; slice thickness: 1.0 mm; tube current-time product: 25 mAs; tube voltage: 120 kV). To optimize radiation protection, automatic tube current modulation (CARE Dose4D; Siemens Healthcare GmbH, Forchheim, Germany) and automatic tube voltage selection (CARE kV algorithm; Siemens Healthcare GmbH, Forchheim, Germany) were applied. Convolution kernels B31F and B70F were used for image reconstruction. The calibration of the CT system was performed quarterly using a standard water phantom provided by the manufacturer. In addition, daily calibration of the CT system was conducted using air.

Patients were scanned in an inspiratory breath-hold and head-first supine position with their arms elevated. For further image processing, reconstruction in the soft tissue window B31F with a 1 mm slice thickness was used for open-source Body and Organ Analysis (BOA), a comprehensive CT image segmentation algorithm designed for seamless workflow integration [23].

The open-source BOA integrates various segmentation algorithms with an emphasis on workflow integration through DICOM node integration, providing comprehensive body segmentation in CT images with extensive body volume coverage. The BOA incorporates two segmentation algorithms: Body Composition Analysis (BCA) and TotalSegmentator [24]. In this study, only BCA, which was trained using the nnU-Net framework, was used, as described by Haubold et al. [23].

The CT data were fully blinded, and BCA was used to automatically determine BMD from the available dataset. Based on HU, the following parameters for vertebrae Th9 to L1 were calculated: mean attenuation values with standard deviation and 95% confidence interval (CI). All the generated datasets were checked for quality. Inadequate ROI placement resulted in exclusion from the study (Figure 1).

2.3. Pulmonary Function Testing and Body Mass Index

BMI calculations and pulmonary function tests (PFTs) were obtained from the electronic records nearest to the DXA scans. A JAEGER MasterScreen body (CareFusion, Hoechberg, Germany) was used to measure forced expiratory volume in 1 s (FEV1). The Global Lung Function Initiative (GLI) reference values were used [25].

2.4. Statistical Analysis

Statistical analyses were performed using GraphPad Prism, version 10.2 (GraphPad Software, San Diego, CA, USA). Data are presented as the mean $\pm$ standard deviation. Normal distribution was tested using the Shapiro–Wilk test. Student's *t*-test or Mann–Whitney U-test was used to assess differences between groups, as appropriate. The chi-square test (χ^2) or Fisher's exact test was used to assess the differences between categorical variables. Pearson's correlation coefficient was used to measure the linear correlation

between two sets of metric data. Receiver operating characteristic (ROC) curve analysis was performed to identify predictors of low BMD using AI-determined HU values of vertebrae Th9 to L1. ROC curve plotting and ROC analysis, including the determination of the area under the ROC curve (AUC), specificity, and sensitivity for each variable, and determination of thresholds, were calculated. Statistical significance was set at $p < 0.05$.

3. Results

3.1. Study Population

A total of 65 pwCF were included in this study, of which there were 25 females (38%). The majority of the study population was homozygous for F508del, accounting for 46% of the total, with an average age of 30.1 ± 7.5 years. The mean ppFEV1 was 58.8 ± 21.7, and the BMI ranged from 14.0 to 39.0 kg/m^2, with a mean of 21.8 ± 4.0. Exocrine pancreatic insufficiency was identified in 53 of 65 pwCF (82%), whereas CF-related diabetes mellitus was present in 16 patients (25%). The 25-hydroxyvitamin D levels ranged from 4.3 to 54.2 ng/mL, with 77% receiving oral cholecalciferol replacement therapy. Table 1 presents the demographic and clinical characteristics of the study population.

Table 1. Patient demographics and characteristics.

Characteristics	Overall (*n* = 65)	Normal (*n* = 53)	CFBD (*n* = 12)	*p*-Value (normal vs. CFBD)
Age, years	30.1 ± 7.5 (18–49)	30.2 ± 7.8 (18–49)	29.8 ± 6.5 (20–42)	0.820
Female sex, n (%)	25 (38)			0.754
Genotype, n (%)				0.598
F508del homozygous	30 (46)	24 (45)	6 (50)	
F508del heterozygous	23 (36)	18 (34)	5 (42)	
Other	12 (18)	11 (21)	1 (8)	
CFTR modulator therapy, n (%)				0.470
Elexacaftor/tezacaftor/ivacaftor	18 (28)	13 (25)	5 (42)	
Tezacaftor/ivacaftor	6 (9)	4 (8)	2 (17)	
Lumacaftor/ivacaftor	3 (5)	3 (6)	0	
Ivacaftor	2 (3)	2 (4)	0	
None	36 (55)	31 (58)	5 (42)	
ppFEV1	58.8 ± 21.7 (21–120)	61.4 ± 22.6 (21–120)	47.3 ± 12.6 (29–68)	0.006
FEV1 [L]	2.4 ± 1.1 (0.7–4.9)	2.5 ± 1.1 (0.7–4.9)	1.9 ± 0.5 (1.2–3.0)	0.006
BMI [kg/m^2]	21.8 ± 4.0 (14.0–39.0)	22.4 ± 4.0 (14.9–39.0)	19.4 ± 3.3 (14.0–24.6)	0.020
Pancreatic insufficiency, n (%)	53 (82)	42 (80)	11 (92)	0.317
Cystic fibrosis-related diabetes, n (%)	16 (25)	15 (28)	1 (8)	0.147
25-hydroxyvitamin D levels [ng/mL] #	22.0 ± 9.5 (4.3–54.2)	21.6 ± 7.6 (6.9–38.8)	23.7 ± 15.7 (4.3–54.2)	0.660
Cholecalciferol supplementation, n (%)	50 (77)	39 (74)	11 (92)	0.179
Steroid therapy, n (%)	5 (8)	4 (8)	1 (8)	0.926

Values are the mean ± standard deviation, range, or number of patients (%). CFTR, cystic fibrosis transmembrane conductance regulator; ppFEV1, percent-predicted forced expiratory volume in 1 s; BMI, body mass index; CFBD, cystic fibrosis-related bone disease. # n = 64.

3.2. Dual-Energy X-ray Absorptiometry (DXA)

Twelve participants (18%) exhibited low BMD for age on the DXA scan (L1-L4) with a z-score of <−2, as shown in Table 2. The overall mean bone mineral density was 0.97 ± 0.15 g/cm^2, corresponding to a z-score of −1.0 ± 1.3. Statistically significant differences were observed between pwCF with normal BMD and those with low BMD in terms of ppFEV1 ($p = 0.019$) and BMI ($p = 0.020$). No other clinical characteristics were significantly different between the two groups (Table 1).

Table 2. Results of DXA and CT-attenuation values measurement.

Characteristics	Overall (n = 65)	Normal (n = 53)	CFBD (n = 12)	p-Value (Normal vs. CFBD)
DXA z-score L1-L4	−1.0 ± 1.3 [−1.35–−0.73]	0.6 ± 0.8 [−0.81–−0.35]	−3.0 ± 0.6 [−3.43–−2.67]	<0.001
DXA BMD L1-L4 [g/cm^2]	0.97 ± 0.15 [0.93–1.01]	1.02 ± 0.11 [0.99–1.05]	0.73 ± 0.06 [0.69–0.77]	<0.001
DXA z-score L1	−1.5 ± 1.2 [−1.77–−1.19]	−1.1 ± 0.9 [−1.36–−0.86]	−3.1 ± 0.7 [−3.59–−2.66]	<0.001
DXA BMD L1 [g/cm^2]	0.89 ± 0.14 [0.86–0.92]	0.93 ± 0.11 [0.90–0.96]	0.70 ± 0.08 [0.65–0.75]	<0.001
HU Th9	247.1 ± 73.9 [228.79–265.40]	260.3 ± 68.3 [241.44–279.10]	188.9 ± 72.1 [143.09–234.75]	0.002
HU Th10	234.0 ± 58.7 [219.41–248.50]	242.4 ± 55.6 [227.07–257.76]	196.6 ± 60.0 [158.75–234.46]	0.013
HU Th11	237.7 ± 54.2 [224.30–251.18]	250.4 ± 45.6 [237.79–262.93]	182.0 ± 56.0 [146.36–217.56]	<0.001
HU Th12	237.0 ± 45.8 [225.71–248.42]	244.5 ± 37.9 [234.0–254.91]	204.4 ± 63.2 [164.26–244.58]	0.055
HU L1	230.4 ± 57.6 [216.15–244.71]	242.2 ± 49.5 [228.54–255.87]	178.4 ± 64.0 [137.74–219.07]	<0.001
HU Th9–L1	237.3 ± 49.7 [224.94–249.57]	247.1 ± 43.4 [235.13–259.01]	193.9 ± 54.8 [159.0–228.76]	0.001

Values are presented as the mean ± standard deviation. Brackets indicate the 95% CI. CI, confidence interval; DXA, dual-energy X-ray absorptiometry; BMD, bone mineral density; HU, Hounsfield units; CFBD, cystic fibrosis-related bone disease.

3.3. CT-Derived Bone Density

Table 2 shows AI-calculated CT-derived attenuation values for Th9 to L1, with L1 consistently exhibiting the lowest HU values among the examined vertebral bodies. Patients with DXA-determined low BMD demonstrated significantly lower HU values in all examined vertebral bodies, except for Th12 (p = 0.055), compared to those without bone disease (all p <0.05).

3.4. Correlation between DXA (Z-Scores and BMD) and CT-Derived Bone Density

The findings of the correlation analysis are presented in Figures 2 and 3, respectively. The analysis showed a positive association between the AI-estimated HU values and the DXA-derived BMD and z-scores for L1–L4. The strongest correlations were identified between L1 and BMD (r = 0.52; $p < 0.001$) and z-scores (r = 0.58; $p < 0.001$). A correlation analysis conducted independently on L1, as determined by LDCT and DXA, revealed a significant association for both BMD (r = 0.56; $p < 0.001$) and z-score (r = 0.57; $p < 0.001$). However, the mean HU values of Th9 to L1 did not exhibit improved results when compared with DXA (r = 0.43, $p < 0.001$ for BMD and r = 0.48, $p < 0.001$ for z-score).

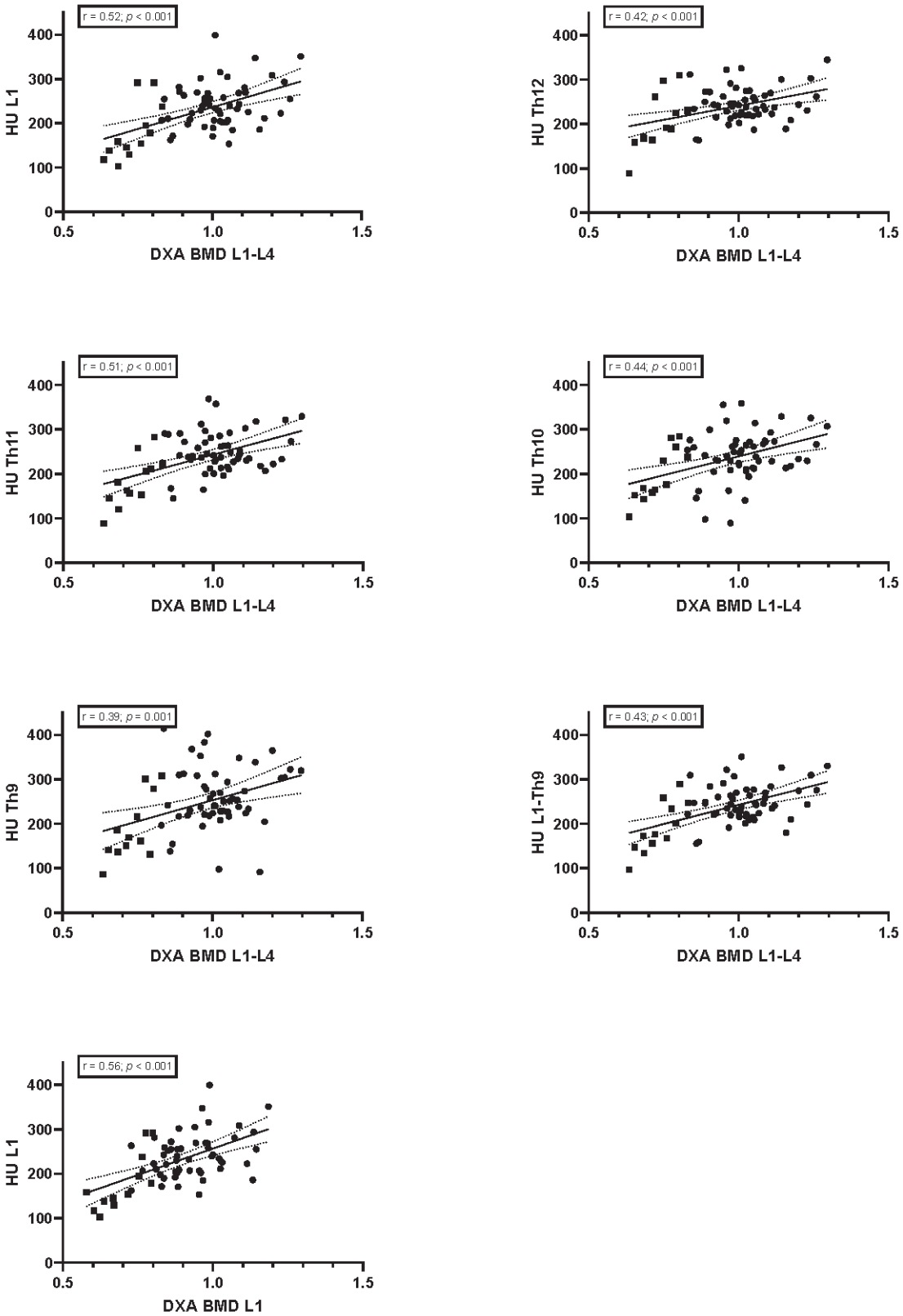


Figure 2. Scatter plot of CT-attenuation values and DXA-derived BMD. ●, normal BMD; ■, CFBD. CFBD, cystic fibrosis-related bone disease; HU, Hounsfield units; DXA, dual-energy X-ray absorptiometry; BMD, bone mineral density; r, correlation coefficient.

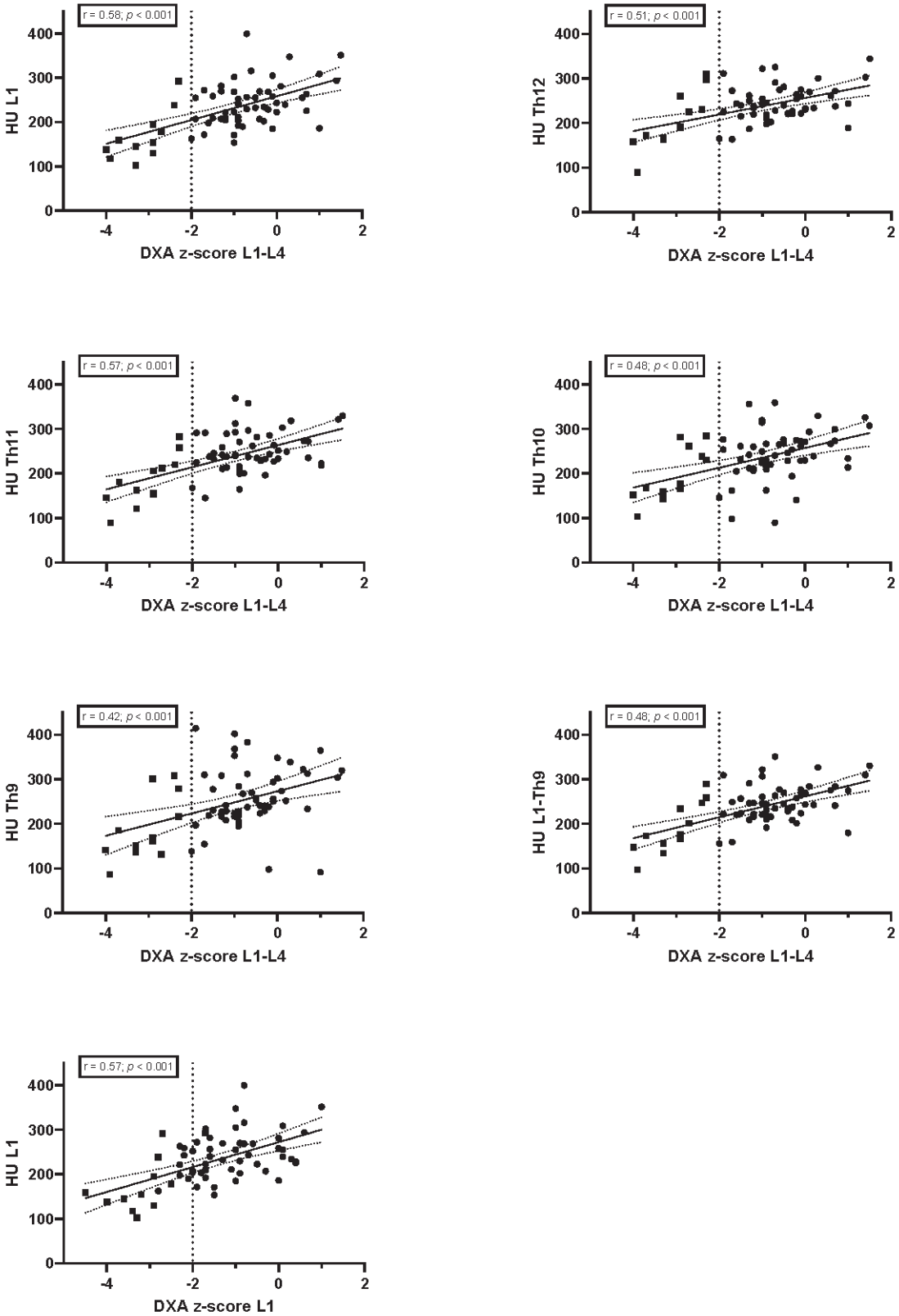


Figure 3. Scatter plot of CT-attenuation values and DXA-derived z-scores. ●, normal BMD; ■, CFBD CFBD, cystic fibrosis-related bone disease; HU, Hounsfield units; DXA, dual-energy X-ray absorptiometry; BMD, bone mineral density; r, correlation coefficient.

3.5. ROC Analysis

The results of the ROC analysis of AI-estimated attenuation values revealed that Th11 and L1 demonstrated the highest AUC values for distinguishing between the low BMD and normal group. Th11 had an AUC value of 0.835 ($p < 0.001$), whereas L1 had an AUC value of 0.796 ($p = 0.001$) (Figure 4). Table 3 shows the HU thresholds at a sensitivity level of 75% for each vertebra. Th11 and L1 exhibited the highest specificity for differentiating between pwCF with low BMD and those with normal BMD at a sensitivity level of 75%, with thresholds of <212 HU and <197 HU, respectively, and a specificity of 84.9% for both.

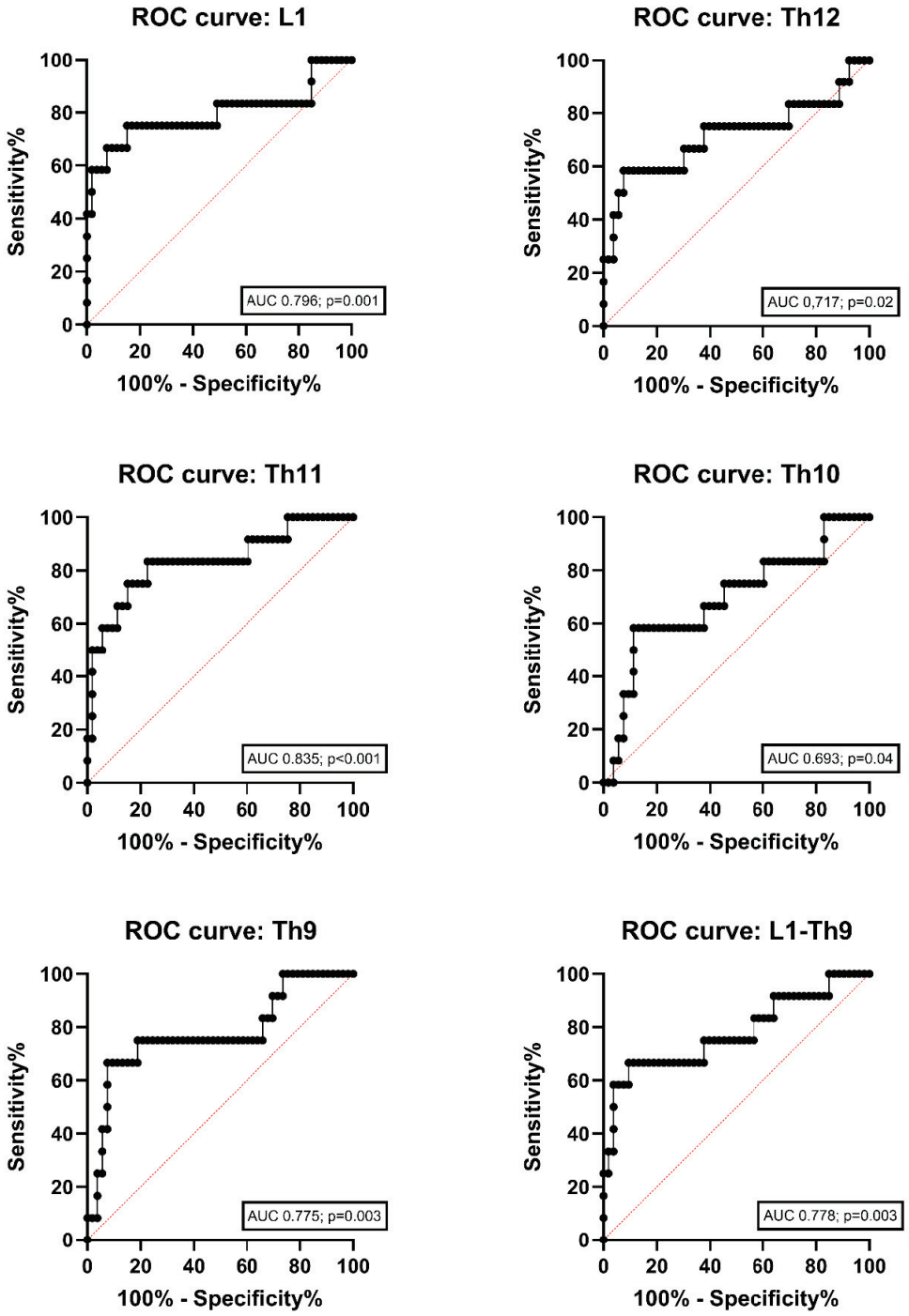


Figure 4. Receiver operating characteristic (ROC) curves of CT-attentuation values for predicting low BMD. AUC, area under the curve; BMD, bone mineral density; ROC, receiver operating characteristic.

Table 3. Sensitivity and specificity and likelihood ratio of different density thresholds for Th9 to L1 indicating low BMD.

HU Threshold		Sensitivity (%)	95% CI	Specificity%	95% CI	Likelihood Ratio
<190	Th9	66.7	39.1–86.2	92.5	82.1–97.0	8.8
<217	Th9	75.0	46.8–91.1	81.1	68.6–89.4	4.0
<185	Th10	58.3	32.0–80.7	88.7	77.4–94.7	5.2
<238	Th10	75.0	46.8–91.1	54.7	41.5–67.3	1.7
<163	Th11	50.0	25.4–74.6	98.1	90.1–99.9	26.5
<212	Th11	75.0	46.8–91.1	84.9	72.9–92.1	5.0
<179	Th12	41.7	19.3–68.0	96.2	87.2–99.3	11.0
<232	Th12	75.0	46.8–91.1	62.3	48.8–74.1	2.0
<160	L1	58.3	19.3–68.0	98.1	90.1–99.9	30.9
<197	L1	75.0	46.8–91.1	84.9	72.9–91.1	5.0
<178	Th9 -L1	58.3	32.0–80.7	96.2	87.2 to 99.3%	15.5
<234	Th9 -L1	75.0	46.8–91.1	62.3	48.8 to 74.1%	2.0

BMD, bone mineral density; CI, confidence interval; HU, Hounsfield units.

4. Discussion

Our study indicates that the LDCT of the chest using a fully automated AI-guided determination of Th9 to L1 trabecular attenuation values has promising potential for opportunistic screening of low BMD in adult pwCF. Our findings revealed a positive correlation between the HU values of Th9 to L1 and both DXA-derived z-scores and BMD of L1 to L4. Moreover, we established threshold values for Th9 to L1 with acceptable specificity and sensitivity levels for diagnosing low BMD, with the best results achieved for Th11 and L1.

The growing significance of the early diagnosis of comorbidities in pwCF is noteworthy, considering the increasing life expectancy and subsequent increase in comorbidities [3]. In addition to CF, AI has been investigated for its potential to enhance OS for low BMD during routine non-enhanced LDCT scans of the chest by leveraging existing imaging resources to identify individuals at risk of osteoporosis [20]. This method has the potential to improve screening rates, reduce costs and radiation dose, and facilitate early intervention, particularly in settings where access to DXA is limited or in high-risk populations such as CF, which may benefit from routine BMD assessment [26,27].

It is difficult to compare our data with those of existing studies. The outcomes of AI-assisted LDCT scanning for low BMD screening vary considerably based on factors such as the scanned region, the patient population being examined, and the specific CT scanner and AI software used. A meta-analysis conducted by Ong et al. revealed a sensitivity of 41.0–100% and a specificity of 31.0–100%, which reflects the great heterogeneity of the AI techniques used and the examined populations for differentiating between normal and low BMD [28]. With respect to the disease, our study cohort was much younger than those used in other studies for assessing opportunistic low BMD screening with LDCT. These studies predominantly involved older adults, often postmenopausal women or patients aged >50 years, with varying disease conditions, reflecting the higher prevalence and clinical concern of low BMD in these groups [17]. Normative attenuation values for L1 and Th9-12 in young adults can be inferred from the data provided by Jang et al. [29] and Patel and Lee [30]. Looking at L1, Jang et al. reported a mean L1 attenuation of 226 $\pm$ 44 HU in patients younger than 30 years and 192 $\pm$ 39 HU in patients 45–49 years of age. These results are in line with ours, showing overall HU values for L1 of 230 $\pm$ 58. Patel and Lee suggested that young adults typically exhibit higher HU values at every spine level than older age groups, where the mean attenuation values of the spine decrease linearly with age [30].

Therefore, determining HU thresholds to differentiate normal from low BMD is challenging and varies considerably across the literature. Research conducted by Yang et al. established cutoff values for distinguishing between normal and low BMD, with values ranging from 147.5 (L1) to 168.5 (Th9) [20]. The current study reported a high diagnostic efficiency for AI-based screening, with AUC values ranging from 0.772 to 0.793 (Th9–L1), which is consistent with the AUC values in our study (0.693–0.835). In a literature review by Zaidi et al. an L1–HU value of 135 was defined as the limit for determining low BMD [31]. However, our study identified higher threshold values for single vertebrae to differentiate between normal and low BMD, mainly because of the younger age of the studied population (Table 3). This highlights the complexity involved in setting the threshold values and emphasizes the dependence of these values on the patient population being studied.

Although the use of fully automated AI-guided LDCT for the opportunistic screening of low BMD is increasing, several factors should be considered. Automated segmentation plays a key role in the accurate identification of bone and vertebral levels [32]. In many CT scans of the chest, the vertebra of interest is not fully captured and may be at the edge of the field of view, leading to partial inclusion, which can lead to the over- or underestimation of the HU values. However, spinal deformities such as scoliosis or kyphosis, which are prevalent in pwCF, can complicate the automatic detection of vertebral bodies [5]. The accurate identification of spinal levels is essential because research indicates that BMD varies among different spinal levels and may lead to the incorrect determination of attenuation values [30].

DXA is recognized as a precise method for assessing BMD yet presents certain limitations. Inaccurate positioning, false scan mode, artifacts, and spinal deformations may affect the accuracy of BMD determination [33].

Limitations

First, the retrospective design of this study, which involved individuals who underwent paired chest CT scans and DXA examinations at a three-month interval, may have resulted in selection bias and could potentially limit the generalizability of the findings. Furthermore, the small number of patients, particularly those with confirmed osteoporosis, may have influenced the determination of the reference values and led to over- or underestimation. Additionally, the AI-supported recognition of vertebral bodies for determining HU may be faulty (Figure 1), which could result in the under-sharpening or over-sharpening of the measurements, directly affecting the diagnostic outcome. It must also be noted that the definition of low BMD or osteoporosis in pwCF differs from the World Health Organization (WHO)'s definition for assessing bone density, which is employed in the majority of studies on opportunistic CT screening [34]. This variance may introduce potential bias when comparing our findings with those of other researchers. The accuracy of AI models is high; however, they require validation against large, diverse datasets before integration into routine clinical workflows. Finally, our data do not provide information regarding the potential fracture risk in pwCF DXA-measured BMD as it does not correlate well with fracture risk in pwCF [11]. Additional longitudinal data are necessary to determine whether the overall prediction of fracture risk using LDCT of the chest with and without AI will be able to address this limitation in adults with CF.

5. Conclusions

Fully automated AI-supported opportunistic screening for low BMD using LDCT scans of the chest shows great potential for use in adult pwCF. However, future research is needed to determine the diagnostic accuracy of this method and to establish clear thresholds to effectively differentiate between normal and low BMD for the diagnosis of CFDB and assessment of potential bone fracture risk in adults with CF.

Author Contributions: Conceptualization, M.W. and M.O.; Data curation, H.N., R.H., P.R., A.M., S.S. (Sivagurunathan Sutharsan), S.S. (Svenja Strassburg), B.M.S., F.N., J.H., L.S. and M.O.; Formal analysis, M.W., F.S. and D.W.; Investigation, M.W., H.N., R.H., P.R., J.H. and L.S.; Methodology,

M.W., S.Z. and M.O.; Project administration, S.S. (Sivagurunathan Sutharsan), C.T., M.F., F.N. and M.M.; Resources, R.H., C.T., M.F. and J.H.; Software, S.Z., B.M.S., F.N., M.M., J.H., L.S. and M.O.; Supervision, C.T., M.F. and M.O.; Validation, H.N., A.M. and B.M.S.; Writing—original draft, M.W. and M.O.; Writing—review and editing, R.H., P.R., F.S., A.M., S.S. (Sivagurunathan Sutharsan), S.S. (Svenja Strassburg), D.W., C.T., S.Z., B.M.S., M.F., F.N., M.M., J.H. and L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study did not receive any specific grants from funding agencies in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: This study was approved by the local ethics committee of the University Hospital Essen (24-11977-BO) and followed the Declaration of Helsinki Ethical Principles for Medical Research Involving Human Subjects. Informed consent was waived because of the retrospective nature of the study and the analysis used anonymous clinical data.

Data Availability Statement: Data used to support the current findings are available from the corresponding author upon request.

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

Abbreviations

AI	artificial intelligence
AUC	area under the curve
BCA	body composition analysis
BMI	body mass index
BOA	body organ analysis
CF	cystic fibrosis
CFBD	cystic fibrosis-related bone disease
CFTR	cystic fibrosis transmembrane conductance regulator
CT	computed tomography
DXA	dual-energy X-ray absorptiometry
LDCT	low-dose computed tomography
PFT	pulmonary function testing
$ppFEV_1$	percent predicted forced expiratory volume in 1 s
pwCF	people with cystic fibrosis
ROC	receiver operating characteristic

References

1. Mall, M.A.; Burgel, P.R.; Castellani, C.; Davies, J.C.; Salathe, M.; Taylor-Cousar, J.L. Cystic fibrosis. *Nat. Rev. Dis. Primers* **2024**, *10*, 53. [CrossRef] [PubMed]
2. Ong, T.; Ramsey, B.W. Cystic Fibrosis: A Review. *JAMA* **2023**, *329*, 1859–1871. [CrossRef] [PubMed]
3. Ticona, J.H.; Lapinel, N.; Wang, J. Future Comorbidities in an Aging Cystic Fibrosis Population. *Life* **2023**, *13*, 1305. [CrossRef] [PubMed]
4. Mischler, E.H.; Chesney, P.J.; Chesney, R.W.; Mazess, R.B. Demineralization in cystic fibrosis detected by direct photon absorptiometry. *Am. J. Dis. Child.* **1979**, *133*, 632–635. [CrossRef]
5. Fainardi, V.; Nora, M.; Salghetti, A.; Petraglia, F.; Fanzaghi, P.; Anelli, G.; Cattabiani, M.; Scopelliti, G.; Deolmi, M.; Ferrari, I.; et al. Prevalence of Scoliosis in Children and Adolescents with Cystic Fibrosis. *Children* **2024**, *11*, 321. [CrossRef]
6. Paccou, J.; Zeboulon, N.; Combescure, C.; Gossec, L.; Cortet, B. The prevalence of osteoporosis, osteopenia, and fractures among adults with cystic fibrosis: A systematic literature review with meta-analysis. *Calcif. Tissue Int.* **2010**, *86*, 1–7. [CrossRef]
7. Putman, M.S.; Anabtawi, A.; Le, T.; Tangpricha, V.; Sermet-Gaudelus, I. Cystic fibrosis bone disease treatment: Current knowledge and future directions. *J. Cyst. Fibros.* **2019**, *18* (Suppl. S2), S56–S65. [CrossRef]
8. Sermet-Gaudelus, I.; Bianchi, M.L.; Garabédian, M.; Aris, R.M.; Morton, A.; Hardin, D.S.; Elkin, S.L.; Compston, J.E.; Conway, S.P.; Castanet, M.; et al. European cystic fibrosis bone mineralisation guidelines. *J. Cyst. Fibros.* **2011**, *10*, S16–S23. [CrossRef] [PubMed]
9. Jacquot, J.; Delion, M.; Gangloff, S.; Braux, J.; Velard, F. Bone disease in cystic fibrosis: New pathogenic insights opening novel therapies. *Osteoporos. Int.* **2016**, *27*, 1401–1412. [CrossRef]
10. Mora Vallellano, J.; Delgado Pecellin, C.; Delgado Pecellin, I.; Quintana Gallego, E.; Lopez-Campos, J.L. Evaluation of bone metabolism in children with cystic fibrosis. *Bone* **2021**, *147*, 115929. [CrossRef]
11. Stephenson, A.; Jamal, S.; Dowdell, T.; Pearce, D.; Corey, M.; Tullis, E. Prevalence of vertebral fractures in adults with cystic fibrosis and their relationship to bone mineral density. *Chest* **2006**, *130*, 539–544. [CrossRef] [PubMed]

12. Baroncelli, G.I.; De Luca, F.; Magazzu, G.; Arrigo, T.; Sferlazzas, C.; Catena, C.; Bertelloni, S.; Saggese, G. Bone demineralization in cystic fibrosis: Evidence of imbalance between bone formation and degradation. *Pediatr. Res.* **1997**, *41*, 397–403. [CrossRef] [PubMed]
13. Putman, M.S.; Milliren, C.E.; Derrico, N.; Uluer, A.; Sicilian, L.; Lapey, A.; Sawicki, G.; Gordon, C.M.; Bouxsein, M.L.; Finkelstein, J.S. Compromised bone microarchitecture and estimated bone strength in young adults with cystic fibrosis. *J. Clin. Endocrinol. Metab.* **2014**, *99*, 3399–3407. [CrossRef]
14. Boyle, R.L.; Psoter, K.J.; Merlo, C.A.; Sidhaye, A.R.; Lechtzin, N.; Patel, S.; Montemayor, K.; Horne, A.; West, N.E. Prevalence and Risk Factors for Low Bone Mineral Density in Adults with Cystic Fibrosis. *JBMR Plus* **2022**, *6*, e10666. [CrossRef]
15. Ciet, P.; Bertolo, S.; Ros, M.; Casciaro, R.; Cipolli, M.; Colagrande, S.; Costa, S.; Galici, V.; Gramegna, A.; Lanza, C.; et al. State-of-the-art review of lung imaging in cystic fibrosis with recommendations for pulmonologists and radiologists from the "iMAging managEment of cySTic fibROsis" (MAESTRO) consortium. *Eur. Respir. Rev.* **2022**, *31*, 210173. [CrossRef]
16. He, Y.; Lin, J.; Zhu, S.; Zhu, J.; Xu, Z. Deep learning in the radiologic diagnosis of osteoporosis: A literature review. *J. Int. Med. Res.* **2024**, *52*, 3000605241244754. [CrossRef] [PubMed]
17. Liu, R.W.; Ong, W.; Makmur, A.; Kumar, N.; Low, X.Z.; Shuliang, G.; Liang, T.Y.; Ting, D.F.K.; Tan, J.H.; Hallinan, J. Application of Artificial Intelligence Methods on Osteoporosis Classification with Radiographs-A Systematic Review. *Bioengineering* **2024**, *11*, 484. [CrossRef]
18. Hu, N.; Wang, M.; Yang, M.; Chen, X.; Wang, J.; Xie, C.; Zhang, B.; Wang, Z.; Chen, X. Bone mineral density in lower thoracic vertebra for osteoporosis diagnosis in older adults during CT lung cancer screening. *BMC Geriatr.* **2024**, *24*, 237. [CrossRef]
19. Sollmann, N.; Loffler, M.T.; El Husseini, M.; Sekuboyina, A.; Dieckmeyer, M.; Ruhling, S.; Zimmer, C.; Menze, B.; Joseph, G.B.; Baum, T.; et al. Automated Opportunistic Osteoporosis Screening in Routine Computed Tomography of the Spine: Comparison with Dedicated Quantitative CT. *J. Bone Miner. Res.* **2022**, *37*, 1287–1296. [CrossRef]
20. Yang, J.; Liao, M.; Wang, Y.; Chen, L.; He, L.; Ji, Y.; Xiao, Y.; Lu, Y.; Fan, W.; Nie, Z.; et al. Opportunistic osteoporosis screening using chest CT with artificial intelligence. *Osteoporos. Int.* **2022**, *33*, 2547–2561. [CrossRef]
21. Fang, Y.; Li, W.; Chen, X.; Chen, K.; Kang, H.; Yu, P.; Zhang, R.; Liao, J.; Hong, G.; Li, S. Opportunistic osteoporosis screening in multi-detector CT images using deep convolutional neural networks. *Eur. Radiol.* **2021**, *31*, 1831–1842. [CrossRef]
22. Pickhardt, P.J.; Summers, R.M.; Garrett, J.W.; Krishnaraj, A.; Agarwal, S.; Dreyer, K.J.; Nicola, G.N. Opportunistic Screening: Radiology Scientific Expert Panel. *Radiology* **2023**, *307*, e222044. [CrossRef] [PubMed]
23. Haubold, J.; Baldini, G.; Parmar, V.; Schaarschmidt, B.M.; Koitka, S.; Kroll, L.; van Landeghem, N.; Umutlu, L.; Forsting, M.; Nensa, F.; et al. BOA: A CT-Based Body and Organ Analysis for Radiologists at the Point of Care. *Investig. Radiol.* **2024**, *59*, 433–441. [CrossRef] [PubMed]
24. Wasserthal, J.; Breit, H.C.; Meyer, M.T.; Pradella, M.; Hinck, D.; Sauter, A.W.; Heye, T.; Boll, D.T.; Cyriac, J.; Yang, S.; et al. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiol. Artif. Intell.* **2023**, *5*, e230024. [CrossRef]
25. GLI Calculator. Available online: http://gli-calculator.ersnet.org/ (accessed on 24 June 2024).
26. Engelke, K.; Chaudry, O.; Bartenschlager, S. Opportunistic Screening Techniques for Analysis of CT Scans. *Curr. Osteoporos. Rep.* **2023**, *21*, 65–76. [CrossRef] [PubMed]
27. Guenoun, D.; Champsaur, P. Opportunistic Computed Tomography Screening for Osteoporosis and Fracture. *Semin. Musculoskelet. Radiol.* **2023**, *27*, 451–456. [CrossRef]
28. Ong, W.; Liu, R.W.; Makmur, A.; Low, X.Z.; Sng, W.J.; Tan, J.H.; Kumar, N.; Hallinan, J. Artificial Intelligence Applications for Osteoporosis Classification Using Computed Tomography. *Bioengineering* **2023**, *10*, 1364. [CrossRef]
29. Jang, S.; Graffy, P.M.; Ziemlewicz, T.J.; Lee, S.J.; Summers, R.M.; Pickhardt, P.J. Opportunistic Osteoporosis Screening at Routine Abdominal and Thoracic CT: Normative L1 Trabecular Attenuation Values in More than 20 000 Adults. *Radiology* **2019**, *291*, 360–367. [CrossRef]
30. Patel, S.P.; Lee, J.J.; Hecht, G.G.; Holcombe, S.A.; Goulet, J.A. Normative Vertebral Hounsfield Unit Values and Correlation with Bone Mineral Density. *J. Clin. Exp. Orthop.* **2016**, *2*, 14. [CrossRef]
31. Zaidi, Q.; Danisa, O.A.; Cheng, W. Measurement Techniques and Utility of Hounsfield Unit Values for Assessment of Bone Quality Prior to Spinal Instrumentation: A Review of Current Literature. *Spine* **2019**, *44*, E239–E244. [CrossRef]
32. Klinder, T.; Ostermann, J.; Ehm, M.; Franz, A.; Kneser, R.; Lorenz, C. Automated model-based vertebra detection, identification, and segmentation in CT images. *Med. Image Anal.* **2009**, *13*, 471–482. [CrossRef] [PubMed]
33. Watts, N.B. Fundamentals and pitfalls of bone densitometry using dual-energy X-ray absorptiometry (DXA). *Osteoporos. Int.* **2004**, *15*, 847–854. [CrossRef]
34. Chadha, M.; Chaddha, R.; Divakar, H.; Kalyan, H.; Seth, S.; Shah, P. Osteoporosis: Epidemiology, Pathogenesis, Evaluation and Treatment. *Open J. Orthop.* **2022**, *12*, 153–182. [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.

tomography

MDPI

Article

Pulmonary Fissure Segmentation in CT Images Using Image Filtering and Machine Learning

Mikhail Fufin *, Vladimir Makarov, Vadim I. Alfimov, Vladislav V. Ananev and Anna Ananeva

Medical Informatics Laboratory, Yaroslav-the-Wise Novgorod State University, 41 B. St. Petersburgskaya, Veliky Novgorod 173003, Russia; vladimir.makarov@novsu.ru (V.M.); vadim.alfimov@novsu.ru (V.I.A.); vladislav.ananev@novsu.ru (V.V.A.); s235373@std.novsu.ru (A.A.)

* Correspondence: s234029@std.novsu.ru

Abstract: Background: Both lung lobe segmentation and lung fissure segmentation are useful in the clinical diagnosis and evaluation of lung disease. It is often of clinical interest to quantify each lobe separately because many diseases are associated with specific lobes. Fissure segmentation is important for a significant proportion of lung lobe segmentation methods, as well as for assessing fissure completeness, since there is an increasing requirement for the quantification of fissure integrity. Methods: We propose a method for the fully automatic segmentation of pulmonary fissures on lung computed tomography (CT) based on U-Net and PAN models using a Derivative of Stick (DoS) filter for data preprocessing. Model ensembling is also used to improve prediction accuracy. Results: Our method achieved an F1 score of 0.916 for right-lung fissures and 0.933 for left-lung fissures, which are significantly higher than the standalone DoS results (0.724 and 0.666, respectively). We also performed lung lobe segmentation using fissure segmentation. The lobe segmentation algorithm shows results close to those of state-of-the-art methods, with an average Dice score of 0.989. Conclusions: The proposed method segments pulmonary fissures efficiently and have low memory requirements, which makes it suitable for further research in this field involving rapid experimentation.

Keywords: lung; fissure; segmentation; computed tomography; machine learning; CNN; stick derivative

1. Introduction

The human lungs are divided by fissures into anatomically independent lobes. In clinical practice, the segmentation of lung lobes is useful for the diagnosis and evaluation of lung diseases [1]. It is often of clinical interest to quantify each lobe separately because many diseases are associated with specific lobes. For example, pulmonary lobe segmentation methods can be used to assess the infection severity of COVID-19 by individual lobe [2]. In addition, the segmentation of lobes is extremely important in the surgical treatment of lung diseases [3,4]. For example, the surgical treatment of non-small cell lung cancer often includes lobectomy [5], which is the removal of the diseased lobe.

Therefore, analyzing an affected lung region at the individual lobe level can provide valuable insights for the purpose of the diagnosis and evaluation of a variety of medical conditions.

One of the most popular approaches to lobe segmentation estimates lobar boundaries based on information from fissures, airways, and vessels [6–8]. In addition, there is an increasing requirement for fissure integrity or completeness quantification, which is highly relevant to lung disease characterization [9]. Some studies have proposed methods for the automated analysis of pulmonary fissure integrity [10,11].

Fissure segmentation is important for a significant proportion of lung lobe segmentation methods as well as for assessing fissure integrity. However, there are a lot of challenges that make automatic fissure segmentation a difficult task, namely, imperfections in computed tomography (CT) technology, the presence of image noise, inhomogeneous image

intensity, natural variability in lung anatomy, and variability in lobe shape due to the influence of lung diseases.

A fissure is a double layer of connective tissue formed by invagination of the outer pleural membrane of the lung [12]. The left lung consists of two lobes separated by a left oblique fissure. The right lung consists of three lobes separated by a right oblique fissure and a right horizontal fissure.

The lobe boundaries formed by pulmonary fissures are often either partially invisible on CT images or difficult to distinguish from the adjacent vessels, bronchi, and pathological structures. On cross-sectional CT images, fissures show up as thin curve-like structures (less than 1 mm thick), slightly denser than the surrounding lung parenchyma. Often, fissures are only partially visible or even absent on one or more cross-sections. Moreover, due to lung diseases, serious deformations of the shape of individual lobes can happen. Pathologies such as fibrosis or emphysema may locally resemble fissures or obstruct their shape and appearance [13].

Even CT scans of healthy patients show high anatomical variability, which in itself poses a great challenge in the task of segmenting pulmonary fissures and/or lobes and requires consideration of the larger context surrounding the object of interest. Another factor to consider is the quality of the image. Fissures can be indistinguishable on low-resolution CT scans (with slice thickness greater than 1 mm) so that even a human cannot recognize them clearly. In this case, even creating a dataset to train or validate the algorithm can be difficult, because the worse the image quality, the more disagreement between the experts. The presence of noise has similar issues to low resolution.

The accurate segmentation of lobes requires both local and global information to be taken into account. Local information here means the intensity of voxels located in the vicinity of the fissure within a radius of at least one order of magnitude less than the image size. The presence of vessels, bronchi, airways, and ribs of the thorax is an example of global information, which helps in narrowing down the fissure detection area. This information is especially useful when a fissure in the image does not have enough contrast, is missing, or looks sparse. Because of these issues, local information is not sufficient to identify fissures and/or lung lobes with enough confidence.

Convolutional neural networks (CNNs), as a class of deep learning (DL) models, are effective in many computer vision tasks. They can extract complex and non-obvious correlations from data without relying on models invented for the specific problem. Instead, each model architecture typically solves a broad class of problems, and for a particular problem, it is trained on domain-specific data in a process known as machine learning (ML). Deep learning is a type of machine learning that involves training neural networks with hidden layers, such as convolutional neural networks. Since we do not use any other type of ML in this paper, hereafter, the terms ML and DL will be used interchangeably.

ML is applied for tasks like fissure and lung lobe segmentation as well. Depending on the particular kind of neural network architecture applied in the machine learning model, either individual CT slices or 3D segments may be used as its input. Based on the input type, fissure and lobe segmentation methods can be divided into volumetric and non-volumetric. The advantages of non-volumetric segmentation neural networks are low memory consumption and lower dataset size requirements. The advantage of volumetric neural networks is that such models do not lose relationships between slices, as the input is a 3D image rather than separate slices. This allows for performing fissure and/or lobe segmentation even when the fissure cannot be recognized on some cross-sections.

Before the widespread adoption of neural networks, non-ML fissure and lobe segmentation algorithms served as the basic approach to the task. Their advantage is that they require much less data for the parameter search and validation.

Early attempts to segment pulmonary fissures and lobes were based on various methods, such as the watershed transform [14,15], Voronoi division [16], adaptive sweeping [17], and minimal path [18]. Many of them were based on the Derivative of Stick (DoS) filter proposed by Xiao et al. [6]. The DoS filter enhances fissures in the scan by processing

individual slices of a certain cross-section. Filtering is performed in three cross-sections orthogonal to each other, and then the results are combined. Next, thresholding is performed using multiple thresholds, and these results are also combined. To eliminate false positives, a postprocessing pipeline based on a 3D connected component analysis is used. Based on this work, Peng et al. [7] introduced a new framework based on lung anatomy knowledge and airway and pulmonary fissure segmentation using ODoS [19], an improved version of the DoS method, to segment lobes. Zhao et al. [20] proposed an anisotropic differential operator called the directional derivative of plate (DDoP) filter, which is a 3D version of the DoS filter. Chen et al. [21] segmented pulmonary lobes by applying a multistage spline surface fitting method to the masks obtained with the DDoP filter. Ross et al. [22] also used a thin-plate spline [23] surface fitting method to segment pulmonary lobes with lobar fissure masks as input data, although they employed a particle system rather than a DoS-based algorithm to segment fissures.

Non-ML algorithms are less efficient in taking global information into account and are more sensitive to changes in data sampling. It is an extremely difficult task to devise an algorithm that utilizes enough global information to produce reliable results. The reason is that the fissure position and shape depend on global information in a non-obvious way due to natural anatomic variability. Because of this, many modern approaches to this problem make use of CNNs.

For example, Gerard et al. [24] used a model called FissureNet that is composed of two Seg3DNet networks. One Seg3DNet model finds the region of interest around pulmonary fissures, and another model of the same type refines the prediction of the previous network. FissureNet is trained separately for each lung, with CT scans split into 3D chunks of $64 \times 200 \times 200$ voxels. Gerard and Reinhardt [25] proposed the LobeNet model, which is a FissureNet extension for lobe segmentation. LobeNet consists of four Seg3DNet networks. The first two networks correspond to the two Seg3DNet networks of FissureNet and segment pulmonary fissures. The third Seg3DNet network uses CT scan data and lung fissure segmentation masks to coarsely estimate lobe boundaries. The fourth network refines the results of the previous one. Later, Gerard, Herrmann, et al. [26] used LobeNet to develop a segmentation algorithm that predicts left and right lung regions in humans with diffuse opacification and consolidation.

As ML methods have evolved, an increasing number of authors have proposed approaches that segment lung lobes using grayscale information directly, avoiding the fissure segmentation step. Park et al. [27] used 3D U-Net to segment pulmonary lobes on CT scans. Wang et al. [28] used V-Net for the same segmentation task with 3D chunks of CT scans as input. To better account for global and positional information, Wang et al. [28] added a CoordConv layer to the network. CoordConv is a simple extension of a regular convolutional layer to incorporate positional information by including additional channels for voxel coordinates.

Some researchers have chosen non-volumetric CNNs for pulmonary fissure and lobe segmentation tasks. Chen et al. [29] proposed a scheme called LLASN (Lung Lobes Adversarial based Segmentation Network) in which U-Net is used to generate segmentation results, and a discriminator network is used to discriminate the generated segmentation results from ground-truth labels. Dadras et al. [30] employed multiple ML techniques (self-supervision, attention, and augmentation) to train a lung lobe segmentation model based on 2D U-Net.

We propose a method for the fully automatic segmentation of pulmonary fissures on lung CT based on a DoS filter and non-volumetric CNNs. The advantage of our method over ordinary segmentation networks working with individual slices is better performance due to adding a DoS filter as a preprocessing step to account for the volumetric information of the input image. Model ensembling is also used to improve prediction accuracy.

The advantage of the proposed method over methods based on volumetric CNNs such as 3D U-Net and V-Net is that it requires less memory. This can be easily demonstrated by the fact that a single slice of a 3D image is an edge case of a 3D chunk, and while non-

volumetric CNNs can be trained on batches as small as a singular slice, volumetric CNNs should use wider chunks to have a measurable advantage over non-volumetric CNNs.

The main contributions of this paper are the following:

- This paper proposes a novel method for pulmonary fissure segmentation on lung CT using 2D CNNs and the DoS filter.
- We suggest a pulmonary lobe segmentation method using a fissure detection algorithm and an interpolation technique known as thin-plate splines.
- We draw more attention to the problem of the automatic segmentation of objects such as parenchyma, fissures, lobes, vessels, and airways, which can be helpful in diagnosis and surgical planning.

This paper is structured as follows. Section 2 describes the segmentation pipeline, including model training, training loss design, and the DoS filter used to preprocess data for model training and validation, as well as for inference. This section also provides a classification of the segmentation errors used in this paper and describes the data used for the experiments. Section 3 shows the experimental results, including a comparison of the proposed method with the DoS method, cross-validation results, and lobe segmentation experiment results. In Section 4, we evaluate the performance of the proposed method, discuss its advantages as well as limitations and weaknesses, and suggest how the proposed approach can be improved and/or be utilized in the future.

2. Materials and Methods

2.1. Derivative-of-Stick Filter

The key idea of the DoS filter is to use stick filters of $L \times L$ size, where L is the length of the stick. The image is filtered with $2(L-1)$ versions of the filter, one for each possible filter orientation. The lung parenchyma tissue has the lowest density in areas immediately adjacent to the fissures. To take advantage of that, Xiao et al. [6] use three parallel sticks spaced S pixels apart from each other instead of a single stick.

Two nonlinear derivatives for fissure enhancement are introduced, $\ell_{\max}$ and $\ell_{\min}$:

$$\ell_{\max}^{S,\theta}(x) = \max(\mu_M - \mu_L, \mu_M - \mu_R) - \kappa \cdot \sqrt{E\left[I_j^2\right] - \left(E\left[I_j\right]\right)^2}, \tag{1}$$

$$\ell_{\min}^{S,\theta}(x) = \min(\mu_M - \mu_L, \mu_M - \mu_R) - \kappa \cdot \sqrt{E\left[I_j^2\right] - \left(E\left[I_j\right]\right)^2}. \tag{2}$$

Here, θ is the orientation angle of the stick, x is the spatial position, and κ is a positive coefficient.

μ_M, μ_L, and μ_R are the mean intensity values along the middle, left, and right sticks, respectively:

$$\mu = \frac{1}{L}\sum_{j=1}^{L} I_j. \tag{3}$$

Here, E is the expected value operator, and I_j is the intensity of the j-th pixel. The second term is introduced to suppress blob-like structures.

The stick responses of different orientations can be integrated as follows:

$$F_{\max}(x) = \max\left(\max_{1\le i\le 2(L-1)}\left(\ell_{\max}^{S,\theta_i}\right), \quad 0\right) \tag{4}$$

$$F_{\min}(x) = \max\left(\max_{1\le i\le 2(L-1)}\left(\ell_{\min}^{S,\theta_i}\right), \quad 0\right) \tag{5}$$

Here, θ_i denotes the discrete angle of the i-th stick. Since we are looking for bright objects, only non-negative response values are considered.

A fissure with a step-like appearance or a thickened fissure can be converted into a standard thin curvilinear structure using the F_{max} operation. Normal thin fissures are not affected by F_{max}. When F_{min} is applied to the result of F_{max}, both normal and pathological fissures will be enhanced equally. The combined filter can be described as follows:

$$F_\circ = F_{\max} \circ F_{\min}. \tag{6}$$

Because a fissure can be barely visible or interrupted on slices in one cross-section and clearly visible in another, Xiao et al. [6] integrate responses from all three perpendicular cross-sections:

$$F^{3D}(x) = \left(F_\circ^A + F_\circ^S + F_\circ^C\right) \cdot \frac{\text{median}\left(F_\circ^A, F_\circ^S, F_\circ^C\right)}{\max(F_\circ^A, F_\circ^S, F_\circ^C)}. \tag{7}$$

Here, $F_\circ^A$, $F_\circ^S$, and $F_\circ^C$ denote the response of the DoS $F_\circ$ filter in the axial, sagittal, and coronal cross-sections, respectively.

2.2. The Proposed Method

We introduce two new segmentation methods in addition to the baseline method, in which a neural network is applied without preprocessing data with a DoS filter. By comparing different methods, we tried to answer two questions: which cross-section gives the best result and whether preprocessing a CT scan with a DoS filter improves the result. To answer these questions, three image preprocessing approaches were chosen. The first performs histogram equalization and intensity normalization to the $0 \ldots 255$ range. The second also involves histogram equalization and normalization, but the image is first processed by a DoS filter. The third approach combines the outputs of the previous two methods by creating an additional data dimension.

The pipeline of the image segmentation application is shown in Figure 1. First, a lung mask is applied to the image in such a way that everything outside the mask is filled with zeros. The image is also cropped to a region of interest that is equal to the bounding box of the lung mask. Then, the image is filtered and/or normalized depending on the preprocessing approach of choice. The 3D scan is then sliced. As the previous step produces images of variable size due to the natural variability in lung anatomy and differences in CT scan resolutions, slices are resized to 512×512 pixels. This step is performed because the model deals with images of the same size.

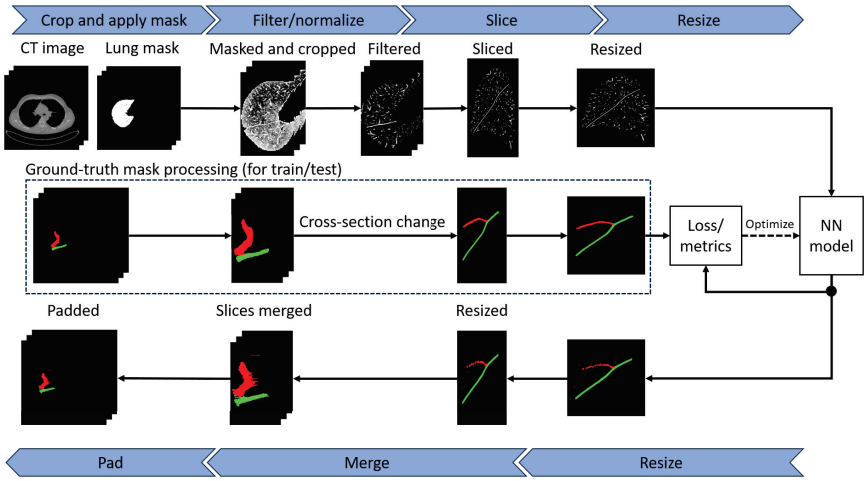


Figure 1. The segmentation pipeline overview. First, the image is masked with the lung mask and then cropped to the mask's bounding box. Next, the image is normalized and optionally filtered by a DoS filter. Then, individual slices are fed as input to the neural network. Finally, several inverse operations are performed to transform the output and retrieve the fissure segmentation. Green and red colors indicate voxel values that depend on the object class, and black is the background.

The segmentation model expects individual slices as input. The neural network outputs a mask with integer values from 0 to 2, where 0 is the background, 1 is an oblique fissure, and 2 is a horizontal fissure. All of the preprocessing steps, except for DoS filtering, have an inverse postprocessing step. The mask then gets scaled back to the cropped image resolution. The individual slices are merged into a 3D image, which is then padded to the size of the original CT image.

Figure 2 depicts a pipeline of the combined preprocessing approach. Firstly, the image is processed with the DoS filter, and then histogram equalization and normalization to 0…255 are applied. The output image has three channels of the same size as the input. All channels except for the second contain the same image processed by the DoS filter. The second channel contains the image prepared in a similar way except for the DoS filtration. The result is then sliced and saved into separate images for further use in model training and testing.

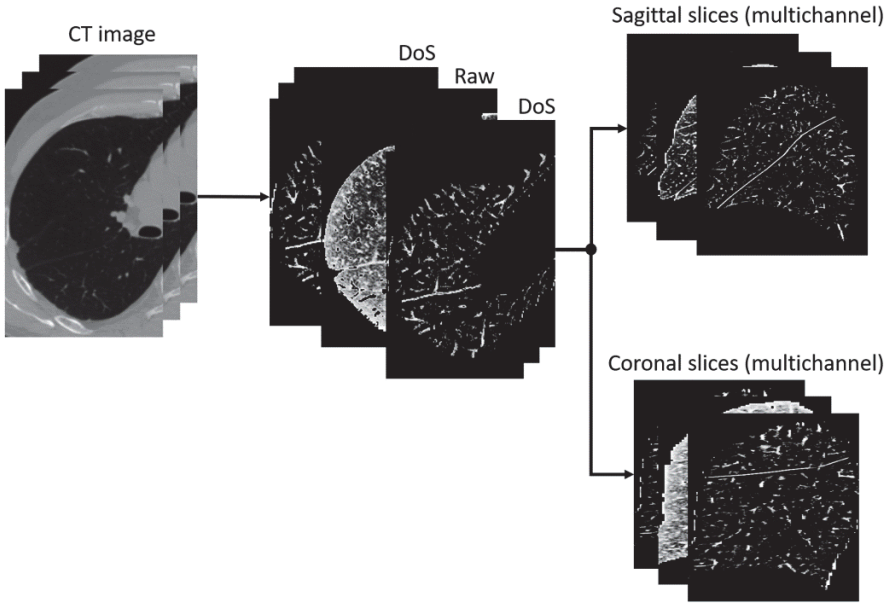


Figure 2. The combined preprocessing approach pipeline. Only one image channel contains the normalized image not processed by the filter.

Six models were trained for each lung, one for each combination of preprocessing approach and cross-section (either sagittal or coronal). There are twelve models in total, not including the standalone DoS method, ensembles, and the models trained for the cross-validation. For two of them, PAN architecture and the Focal loss were used, and for the rest, U-Net and the Dice loss were used, which is explained in Section 3. The axial cross-section is not used because the horizontal fissure of the right lung appears only on a few slices due to its orientation (hence its name). Therefore, a training dataset composed only of axial slices would be very unbalanced. Though usually an issue like that is solved by applying augmentations, there is no clear way to solve this issue in this particular case.

To further improve the performance of the models, we apply model ensembling. The pipeline is shown in Figure 3. The one-hot function is applied to each of the segmentation masks generated by the models that make up the ensemble, and then the argmax function is applied to the sum of the results. To ensure the reliability of the results, we also performed 10-fold cross-validation.

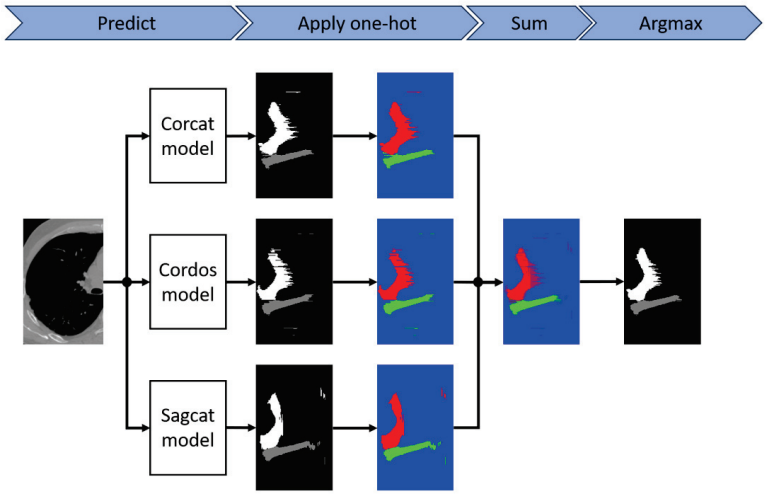


Figure 3. Model ensembling pipeline. One-hot is applied to the output of each model, and then argmax is applied to the sum of the results. Here, black-and-white images represent the standard coding of classes, where for each class a particular value is used. Here, black and white images represent standard class encoding, where a specific value is used for each class. Color images represent one-hot class encoding, each color refers to one of the channels (blue—background, red and green—fissures).

2.3. Error Types

Fissure segmentation errors can be divided into the following categories: false negatives (type II errors), misclassification, oversegmentation, and false fissure segmentation (or simply false segmentation). The last two are examples of false positives (or type I errors). Examples are shown in Figure 4. Oversegmentation (see *a*) occurs when the predicted mask overlaps with the true mask along the entire fissure cross-section, but the Jaccard measure (the overlap area divided by the union area) is less than 70%. False segmentation (see *c*) is the incorrect classification of structures' voxels as fissures. Misclassification (see *d*) happens when the model attributes a voxel to the wrong fissure type.

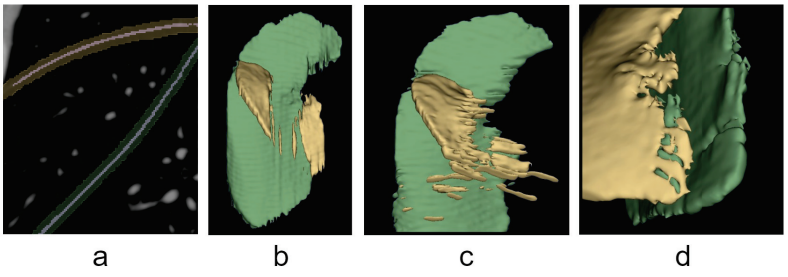


Figure 4. Error types shown in the right lung segmentation example: (**a**) oversegmentation; (**b**) false negatives; (**c**) false segmentation; (**d**) misclassification. Green indicates the oblique fissure and yellow indicates the horizontal fissure.

There are methods to mitigate errors of each type. The postprocessing stage of the DoS method is an example of a false positive reduction algorithm. In this paper, we use model ensembling to eliminate false segmentation and to increase the overall accuracy of the segmentation as well.

To reduce the misclassification error rate, we propose an algorithm based on the *k*-nearest neighbor (KNN) approach. The idea is to keep the two largest components, one for each of the two classes (for the right lung), and train the KNN method on the coordinates of these voxels. This allows falsely classified voxels to be assigned to a more appropriate class based on their proximity to a particular fissure. For this purpose, all voxel coordinates of the original mask are passed to the trained model, resulting in a new class being assigned

to them. Although an ensemble model is capable of reducing a large proportion of such errors by itself, curvilinear approximation algorithms that derive the lobe segmentation are very sensitive to misclassification. Therefore, the proposed algorithm can be used as a precautionary measure.

Oversegmentation errors are usually not critical, because in both the predicted segmentation and the true segmentation, the fissure masks are wider than real fissures, which look like very thin lines in sagittal and coronal sections. Since we consider fissure segmentation as an intermediate step in the lobe segmentation task, usually followed by curvilinear approximation (e.g., using the thin-plate spline method [23]) or an additional lobe segmentation model [25], the mask width does not play a major role as long as it does not interfere with the approximation step.

Moreover, oversegmentation is preferable to type II errors or false segmentation because people usually highlight not the fissure itself but the approximate area around it, and the width varies not only in annotations made by different people but also in annotations made by the same person. But, since standard measures such as F_1, the Jaccard coefficient, and the Dice coefficient do not take this into account, they encourage accuracy instead of completeness.

The simplest solution would be to use the measures that allow for configuring the importance of false positives over false negatives, such as F_{beta}. However, such measures would equally encourage oversegmentation and false segmentation. It is possible to modify a measure so that oversegmentation will have less effect on its value than false segmentation. In this paper, we use the measures proposed by Xiao et al. [6], which satisfy this requirement.

The changed *Precision*, *Recall*, and F_1 measures are based on counting true positives as well as type I and type II errors. Xiao et al. [6] use a 3 mm margin to define true positive values: i.e., voxels that are no more than 3 mm away from the ground-truth mask are considered true positive values for TP_1, and all other voxels of the predicted mask contribute to type I errors or false positives FP [6]. In a similar way, the voxels of the ground-truth mask are divided into TP_2 and type II errors FN, for which the same intersection criterion is used. In general, TP_1 and TP_2 are not equal. Then, *Precision* and *Recall* measures are defined as $TP_1/(TP_1 + FP)$ and $TP_2/(TP_2 + FN)$. The F_1 measure is defined by the same equation: $F_1 = 2 \cdot Precision \cdot Recall/(Precision + Recall)$.

2.4. Loss Function

In this paper, we use both the Dice loss function [31] and the Focal loss function. Both the Dice loss and Focal loss perform well in tasks with high class imbalance, like fissure segmentation. The Dice loss is based on the Dice score, while the Focal loss is an improved version of the Cross-Entropy loss. The Focal loss has been described in detail in [32]. The Dice loss function is defined as follows:

$$\mathrm{D} = \sum_c D^c = \sum_c \left(- \frac{2 \sum_i p^c(i) g^c(i)}{\sum_i p^c(i) + \sum_i g^c(i) + \gamma} \right) \tag{8}$$

where $p^c(i)$ and $g^c(i)$ represent the predicted value and ground truth at position i, respectively; c denotes the class number; and γ is a very small positive value (e.g., 1×10^{-5}) to avoid division by 0.

2.5. Data

Ninety-nine chest CT scans were collected from the Novgorod Regional Clinical Hospital. The patients who had undergone the scans had been diagnosed with lung cancer or had symptoms of lung cancer. The scans were picked according to the following criterion: the pulmonary fissures are clearly visible to a physician over most of the area of the respective lung.

Scans were produced by different scanners, namely, Aquilion manufactured by Toshiba in Shimoishigami, Otawara-shi, Tochigi, Japan, Ingenuity manufactured by Philips Health-

care in Cleveland, Ohio, USA, Optima CT660 manufactured by GE BE Private in Whitefield, Bangalore, India, BrightSpeed manufactured by GE Hangwei Medical Systems in Beijing, China, as well as SOMATOM Definition AS, Sensation, Perspective, and Emotion all manufactured by Siemens in Erlangen, Germany. The slice thickness ranged from 0.3 mm to 2 mm, with 86 out of 99 scans (87%) having a slice thickness of 1 mm or less.

Our access to the data was facilitated through the Cooperation Agreement between the Ministry of Health of the Novgorod Region and the Novgorod State University (Agreement number 20230609, signed on 9 June 2023). The study was conducted in accordance with the Declaration of Helsinki, and protocol #5 was approved by the Novgorod State University Ethics Board (approval date: 22 January 2024).

Nine scans were randomly selected for the test set, and the remaining ninety scans were used as training data. Pulmonary fissures on the scans were manually annotated by a physician using 3D Slicer [33]. Lung masks were obtained automatically using the Chest Imaging Platform software (version 5.2.2) based on 3D Slicer (version 5.5.0-2023-11-24) [34].

3. Results

3.1. Fissure Segmentation

The segmentation models were trained using Pytorch version 2.1.0 on an NVIDIA Quadro RTX 8000 graphics card with 48 GB of memory. The following parameters were used for the DoS filter: $L = 11$, $gap = 3$, and $k = 0.5$. The values of L and gap were taken from [6], and the value $k = 0.5$ was chosen experimentally. The rotate, translate, scale, and shear augmentations from the Albumentations package [35] were used in training. The U-Net and Pyramid Attention Network (PAN) architectures, along with the Focal and Dice loss functions, were used. The specific implementations of these architectures and loss functions were taken from the Segmentation Models Pytorch Python package [36].

There are two models per data preprocessing method, one trained on the slices obtained in the sagittal cross-section and one trained on the coronal slices, which gives us six models for each lung and twelve models in total. For brevity, the models were given the following designations. Models trained on sagittal slices have the prefix *sag*, and models trained on coronal slices have the prefix *cor*. Models trained on data preprocessed by the DoS filter have the suffix *dos*. Models trained on data not processed by the filter have the suffix *raw*. Models working on hybrid-type data (as shown in Figure 2) have the prefix *cat* (short for concatenation, since it concatenates the results of the other two approaches). The names of the models included in the ensemble are abbreviated to the first letters of the prefix and suffix (e.g., *cd* is the same as *cordos*, i.e., the model trained on coronal slices using the DoS filter). The abbreviation reference can be found in Table 1.

Table 1. Abbreviation reference.

Model Name	Description	In-Ensemble Alias
cordos	A model trained on the slices retrieved in the coronal cross-section processed by the DoS filter.	*cd*
corcat	Coronal cross-section, combined processing.	*cc*
corraw	Coronal cross-section, without the DoS filter.	*cr*
sagdos	Sagittal cross-section, with the DoS filter.	*sd*
sagcat	Sagittal, combined.	*sc*
sagraw	Sagittal, without the DoS filter.	*sr*
cd + *cr* + *sc* + *sd*	An ensemble of the *cordos*, *corraw*, *sagcat*, and *sagdos* models.	-

The combination of the U-Net architecture, the ResNet 34 encoder, and the Dice loss function was used as the baseline method. The PAN and DeepLabV3+ models with the ResNet 101, ResNet 152, EfficientNet B3, and EfficientNet B4 encoders were also tested. With the exception of the *sagcat* models (hybrid processing, sagittal cross-section), no significant improvement over the baseline was observed during the validation phase.

For all other models, the baseline was chosen. The *sagcat* models are PANs with the ResNet 34 encoder and were trained using the Focal loss function.

For the right lung, the models were trained on slices of 90 volumetric CT images, and test data consisting of slices of 9 CT images were used for evaluation. The resulting comparison of the models with one of the ensembles is shown in Table 2. The ensemble schematic can be seen in Figure 3. In addition to the *Precision*, *Recall*, and F_1 scores, their alternative versions, described earlier in this paper and named P_{DoS}, R_{DoS}, and F_{DoS}, were also used. These scores are less sensitive to oversegmentation than the standard *Precision*, *Recall*, and F_1 scores. With the exception of *Recall* and R_{DoS}, the ensemble consistently performs better in all other measures, since ensembles tend to sacrifice Recall for Precision.

Table 2. Testing models on the right lung.

Method	F_1	*Precision*	*Recall*	F_{DoS}	P_{DoS}	R_{DoS}
Ensemble (all models)	0.683	0.571	0.878	0.914	0.898	0.943
sagcat	0.652	0.522	0.897	0.890	0.850	0.948
sagraw	0.620	0.482	0.907	0.876	0.828	0.945
cordos	0.653	0.542	0.855	0.875	0.833	0.939
sagdos	0.636	0.508	0.884	0.873	0.831	0.934
corcat	0.642	0.531	0.846	0.872	0.825	0.944
corraw	0.623	0.497	0.873	0.872	0.817	0.952

The comparison of the six-model ensemble with the original DoS algorithm of [6], which includes the 3D connected component analysis pipeline, is shown in Table 3. The visual comparison of the masks obtained using the ensemble and using DoS is shown in Figure 5. Since DoS does not distinguish between oblique and horizontal fissures of the right lung, all fissures had the same class in the comparison. We can conclude that the DoS algorithm is much less effective than neural network-based methods.

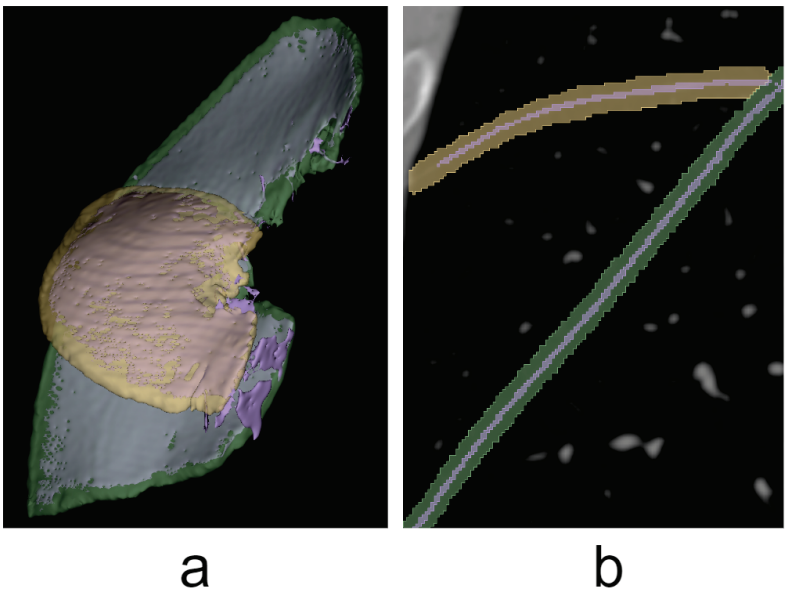


Figure 5. A comparison of the ensemble of neural network models and the DoS method using one of the test images. The mask obtained with DoS is highlighted in purple. $F_{DoS} = 0.908$ for the DoS result, and $F_{DoS} = 0.967$ for the ensemble result (measured on this scan alone). (**a**) A 3D model. (**b**) A sagittal slice.

Table 3. DoS and ensemble comparison using the right lung.

Method	F_1	*Precision*	*Recall*	F_{DoS}	P_{DoS}	R_{DoS}
Ensemble (all models)	0.687	0.572	0.884	0.916	0.898	0.946
Standalone DoS	0.305	0.733	0.201	0.724	0.763	0.708

For both the right and left lungs, we combined the 6 models to create 57 ensembles. The top five models for the right lung based on the F_{DoS} score are shown in Table 4.

Table 4. The five best ensembles for the right-lung fissure segmentation.

Method	F_1	*Precision*	*Recall*	F_{DoS}	P_{DoS}	R_{DoS}
All 6 models	0.683	0.571	0.878	0.914	0.898	0.943
cd + cr + sc + sd	0.686	0.581	0.867	0.914	0.901	0.938
cc + cr + sc + sd	0.684	0.578	0.867	0.913	0.900	0.938
cd + cr + sc + sr	0.680	0.570	0.873	0.913	0.899	0.939
cc + cd + sc + sr	0.689	0.587	0.860	0.913	0.903	0.934

The F_1 and F_{DoS} scores for each of the right-lung fissures are presented in Table 5.

For the left-lung fissure segmentation, the same approach was used as for the right-lung fissure segmentation. Specifically, models were trained on slices obtained in two cross-sections with different preprocessing techniques (L-models), and then model ensembles were created. A training set of 68 images and a test set of 9 images were used for the left-lung fissure segmentation, with encoders' weights initialized with the weights from the respective right-lung fissure segmentation model encoders before training. For example, the left-lung *sagcat* model encoder weights were initialized with the weights of the same model of the right lung. All but the largest connected component of the oblique fissure were also removed before testing.

Table 5. F_1 and F_{DoS} scores depending on right-lung fissure and model.

Method	$F_1^{Oblique}$	$F_1^{Horizontal}$	$F_{DoS}^{Oblique}$	$F_{DoS}^{Horizontal}$
cc + cd + cr + sc + sd + sr	0.678	0.682	0.909	0.915
sagcat	0.643	0.653	0.884	0.887
sagraw	0.615	0.617	0.873	0.867
cordos	0.652	0.638	0.877	0.856
corraw	0.630	0.590	0.873	0.855
sagdos	0.630	0.632	0.871	0.855
corcat	0.652	0.608	0.880	0.844

Table 6 shows the comparison of the best single model, the two best ensembles, and the standalone DoS method. We also used models trained for right-lung fissure segmentation (R-models) to segment left-lung fissures. The comparison results can be seen in the same table, namely, the best single model and the best ensemble. In the task of left-lung fissure segmentation, the F_{DoS} of R-models is just 2.6% lower than that of L-models, which shows that R-models have achieved significant generalization ability.

Table 6. Our four best ensembles, the best model, and the DoS method on the left lung. *sagcat R* and *sc+sd+sr R* are the best model and the best ensemble trained on the right-lung slices.

Method	F_1	*Precision*	*Recall*	F_{DoS}	P_{DoS}	R_{DoS}
sagcat	0.725	0.753	0.724	0.926	0.933	0.922
cd + cr + sc + sd + sr	0.737	0.745	0.747	0.933	0.944	0.923
cd + sc + sr	0.738	0.743	0.753	0.932	0.940	0.927
sagcat R	0.721	0.658	0.826	0.907	0.933	0.886
sc + sd + sr R	0.710	0.626	0.846	0.907	0.933	0.886
Standalone DoS	0.285	0.702	0.188	0.666	0.767	0.617

3.2. Cross-Validating Fissure Segmentation Models

Since the training and test datasets we used for training and evaluation are relatively small, we performed 10-fold cross-validation to ensure that the results are reliable. For this purpose, we combined the training and test datasets into two sets of 99 and 77 scans each for the right and left lungs, respectively. We then divided the first set into 10 folds of 9 or 10 CT scans and the second set into 10 folds of 7 or 8 scans for the left lung. The second-best ensemble for the right lung and the second-best ensemble for the left lung were also re-evaluated. This choice is explained by the fact that these ensembles include fewer models than the top ones but perform no more than 0.2% worse according to Tables 4 and 6. For each fold, the models included in the selected ensembles were re-trained and re-evaluated.

Table 7 shows the overall cross-validated performance of the right-lung models. The values to the left of the brackets are the mean values across folds, while the values in the brackets are their standard deviations. The cross-validated performance of the single best model (*sagcat*) is 0.4% higher than the non-cross-validated result (see Table 4), while the top performance of the ensemble dropped by 4.4%. Although the best ensemble previously showed better performance than any single model by a margin of 2.4%, after cross-validation, *sagcat* showed superior performance, incidentally also by a margin of 2.4%. Despite this, it cannot be concluded that ensembling is useless, since the ensemble still has higher Precision than the single best model, almost preserving the old margin (4.8% then vs. 4.2% now).

Table 7. Right-lung models' overall cross-validated performance.

Method	F_1	*Precision*	*Recall*	F_{DoS}	P_{DoS}	R_{DoS}
cd + cr + sc + sd	0.550 (0.051)	0.452 (0.053)	0.786 (0.064)	0.870 (0.055)	0.900 (0.048)	0.858 (0.061)
sagcat	0.559 (0.059)	0.425 (0.076)	0.904 (0.068)	0.894 (0.055)	0.858 (0.057)	0.942 (0.058)
corraw	0.522 (0.049)	0.394 (0.047)	0.833 (0.053)	0.861 (0.047)	0.824 (0.049)	0.912 (0.046)
cordos	0.507 (0.045)	0.399 (0.050)	0.768 (0.058)	0.839 (0.048)	0.828 (0.053)	0.865 (0.051)
sagdos	0.495 (0.060)	0.384 (0.062)	0.781 (0.071)	0.810 (0.073)	0.819 (0.085)	0.823 (0.073)

Table 8 demonstrates cross-validated by-fissure performance for the right-lung models. It can be seen that both before and after cross-validation, the models detected the horizontal fissures better than the oblique fissures. Although it was previously more evident with the ensemble, now the single best model performs similarly.

Table 8. Right-lung models' by-fissure performance with 10-fold cross-validation.

Method	$F_1^{Oblique}$	$F_1^{Horizontal}$	$F_{DoS}^{Oblique}$	$F_{DoS}^{Horizontal}$
cd + cr + sc + sd	0.541 (0.050)	0.565 (0.061)	0.881 (0.051)	0.839 (0.072)
sagcat	0.529 (0.066)	0.622 (0.051)	0.887 (0.054)	0.903 (0.062)
corraw	0.518 (0.045)	0.517 (0.064)	0.871 (0.040)	0.816 (0.074)
cordos	0.510 (0.045)	0.482 (0.056)	0.857 (0.042)	0.774 (0.067)
sagdos	0.486 (0.063)	0.506 (0.057)	0.825 (0.074)	0.761 (0.080)

Figures 6 and 7 contain box-with-whisker plots for the oblique and horizontal fissures. The box extends from the first quartile (Q1) to the third quartile (Q3) of the data, with a line at the median. The whiskers extend from the box to the farthest data point lying within 1.5× the inter-quartile range (IQR) from the box. Flier points are those past the ends of the whiskers.

Table 9 demonstrates cross-validated performance for the left-lung models. After cross-validation, the performance drops by 3.4%, but unlike the right-lung models, the ensemble still performs better, albeit by a smaller margin of 0.4% (compared to 0.7%). The boxplot for the left lung is shown in Figure 8.

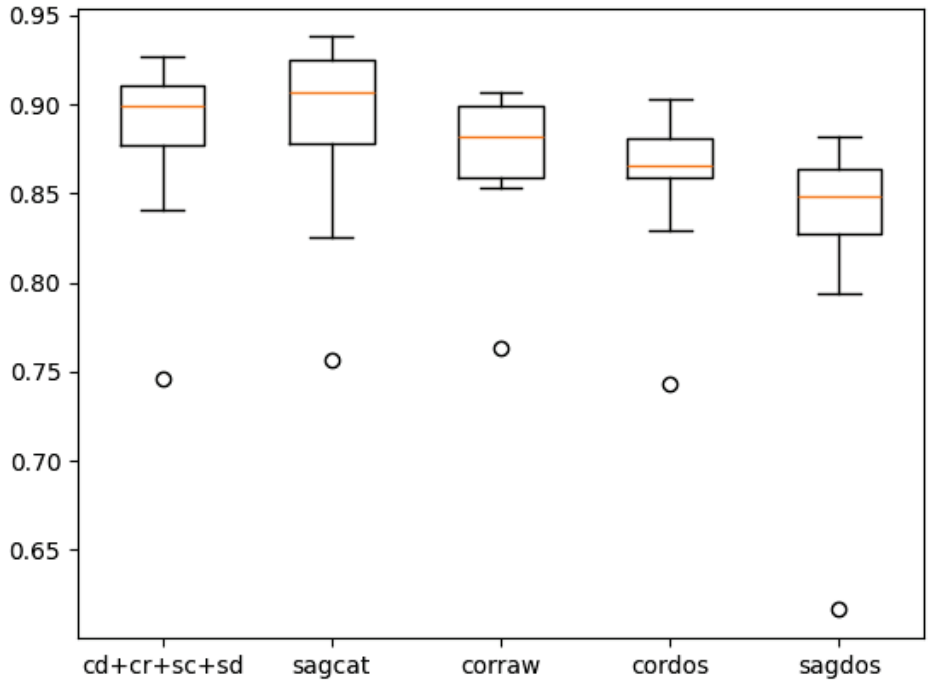


Figure 6. Plot for the right-lung models' cross-validated performance on the oblique fissure.

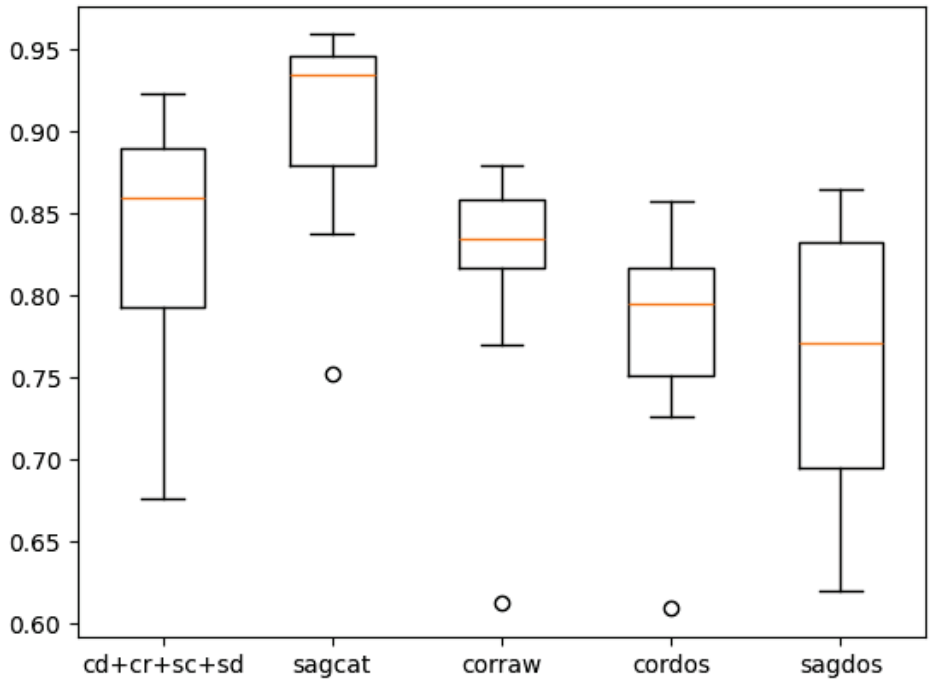


Figure 7. Plot for the right-lung models' cross-validated performance on the horizontal fissure.

Table 9. Left-lung models' performance with 10-fold cross-validation.

Method	F_1	*Precision*	*Recall*	F_{DoS}	P_{DoS}	R_{DoS}
cd + sc + sr	0.726 (0.061)	0.787 (0.067)	0.690 (0.056)	0.899 (0.065)	0.950 (0.064)	0.869 (0.065)
sagcat	0.707 (0.059)	0.782 (0.064)	0.664 (0.056)	0.895 (0.065)	0.945 (0.054)	0.869 (0.065)
sagraw	0.722 (0.059)	0.738 (0.072)	0.722 (0.057)	0.894 (0.062)	0.926 (0.066)	0.873 (0.058)
cordos	0.672 (0.072)	0.752 (0.063)	0.621 (0.079)	0.847 (0.085)	0.929 (0.063)	0.794 (0.099)

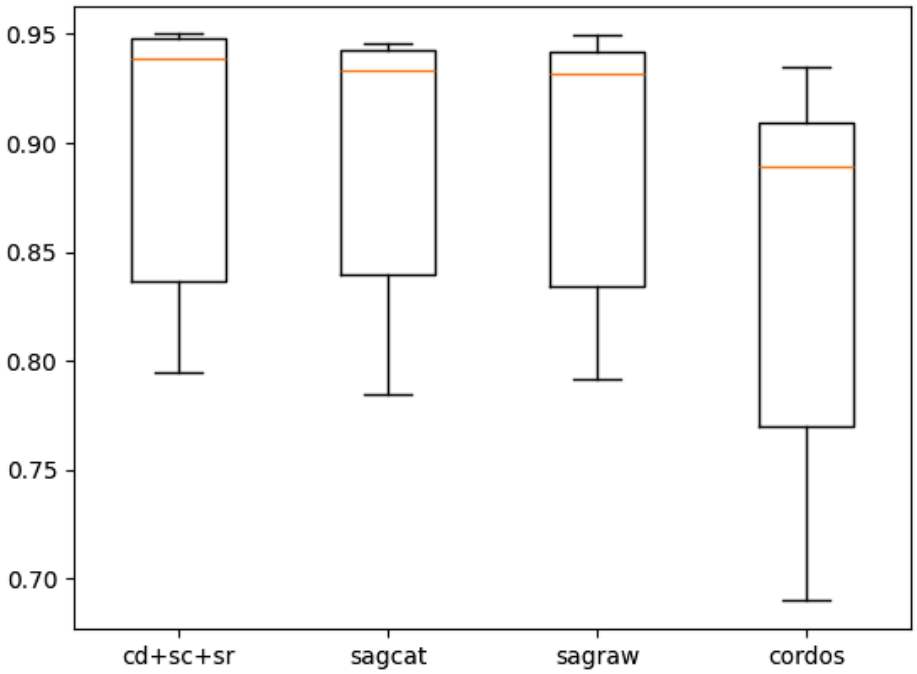


Figure 8. Plot for the left-lung models' cross-validated performance.

3.3. Pulmonary Lobe Segmentation

Fissure masks can be used to obtain lung lobe masks by applying curvilinear approximation algorithms such as thin-plate splines [23]. Such lobe segmentation will not be accurate because not only fissures but also airways have a large influence on the appearance of lung lobes; for example, they occlude lung fissures' visibility on a CT image in certain areas of the lung. Consequently, meaningful segmentation of the lobes requires either segmenting the airways [7] or segmenting the lobes directly [28]. Both of these approaches are beyond the scope of this paper.

Nevertheless, it is possible to perform an additional indirect assessment of the fissure segmentation quality using the lobe segmentation obtained using just the lung mask and fissure masks. For this purpose, the lobes found using ground-truth masks were compared with those found using predicted masks.

Before applying thin-plate splines, all components with sizes less than 500 voxels were removed from the right-lung fissure masks, and all components except the largest one were removed from the left-lung fissure mask. To eliminate misclassification, KNN was applied as described above. The lung mask was also used to establish the outer boundaries of the lobes.

The *cd* + *cr* + *sc* + *sd* + *sr* ensemble was used to segment the left-lung lobes, and the ensemble of all models (*cc* + *cd* + *cr* + *sc* + *sd* + *sr*) was used to segment the right-lung lobes, since these are the best ensembles for the left and right lung lobes, respectively. The *sagcat* model was also used to segment the lobes of both lungs. The comparison is shown in Table 10.

Table 10. The Dice scores of the results of lobe segmentation by two ensembles and the *sagcat* model with the approximation algorithm applied. UR, MR, LR, UL, and LL stand for upper right, middle right, lower right, upper left, and lower left, respectively.

Method	UR	MR	LR	Right avg	UL	LL	Left avg	All avg
All-models ensemble *R*	0.991	0.977	0.991	0.986	-	-	-	0.989
cd + *cr* + *sc* + *sd* + *sr L*	-	-	-	-	0.993	0.993	0.993	0.989
sagcat L	0.980	0.952	0.986	0.973	0.994	0.994	0.994	0.981

Here, UR, MR, LR, UL, and LL are the upper-right, middle-right, left-right, upper-left, and lower-left lobes, respectively; right avg, left avg, and all avg are the average scores for the right lung, left lung, and both lung lobes, respectively. In the first two rows, all avg is the average for the two ensembles. R or L in the index designates either the right or left lung, respectively, for the model.

Ensembles of models perform better than any single model because they are less prone to type I errors. The results of left-lung lobe segmentation with ensembles are shown in Figure 9. Ensembles show a higher average result on both lungs, 0.989. In comparison, Wang et al. [28] achieved a median Dice score of 0.993, and Gerard et al. [25] achieved a median Dice score of 0.947.

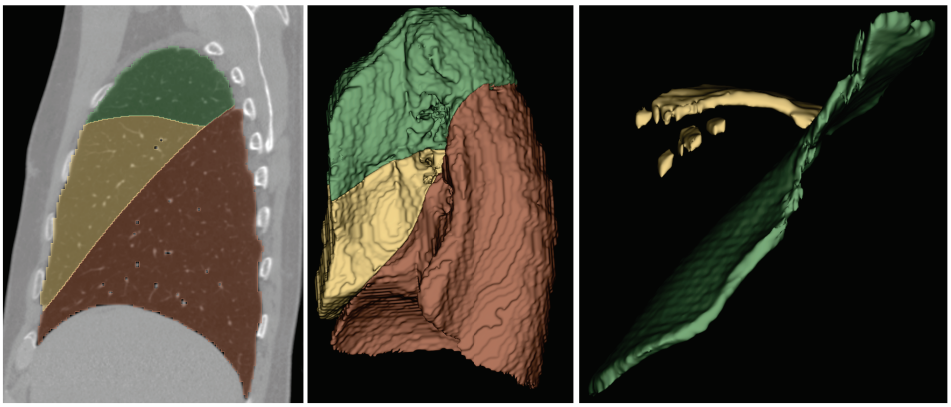


Figure 9. Right-lung lobes on one of the test images. (**a**) Lobes in the sagittal cross-section. Green, yellow and red colors are for the upper, middle, and red lobes, respectively. (**b**) A 3D model of lobes. (**c**) A 3D model of fissures. Green and yellow colors are for the oblique and horizontal fissures, respectively.

Figure 10 shows an example of a discrepancy between the predicted and true masks. The greatest discrepancy is observed in the regions of the lung adjacent to the mediastinum. In these regions, the fissures are the least segmented in both training and test samples as a result of various factors, such as other organs overlapping with the fissures.

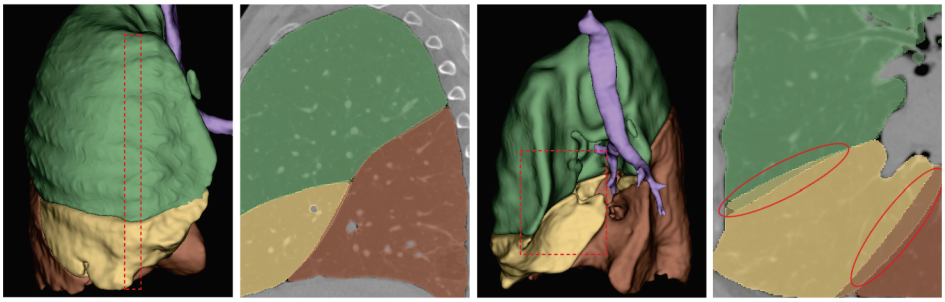


Figure 10. A comparison of true and predicted lobe masks of the right lung on one of the test images. Colors have the same meanings as in the Figure 9, except for purple, which denotes airways. Here, the values of F_1 are 0.992 for the upper lobe; 0.952 for the middle lobe; and 0.986 for the lower lobe. (**a**) A region on a 3D lobe model on which the section corresponding to (**b**) is highlighted with red. (**b**) A sagittal slice with both masks superimposed. (**c**) The 3D lobe model with masks superimposed in the area of the right lung adjacent to the mediastinum; the areas of discrepancy are circled in red. (**d**) The same area highlighted on the 3D model.

4. Discussion

We have proposed a new method for pulmonary fissure segmentation on CT images based on the combination of DoS filtering, CNNs, and model ensembling techniques. Our method shows better results than the DoS method. Segmentation networks were trained on individual slices of sagittal and coronal cross-sections of preprocessed DoS-filtered CT images. Several models were also trained on slices of non-filtered images.

For the right lung, the best result is shown by the ensemble of models, with a modified F1 score of 0.916, compared to the standalone DoS method [6] score of 0.724. For the left lung, the ensemble of five models shows the best result, with a score of 0.933 (the standalone DoS method shows a score of 0.666).

While our method relies on 2D CNN models, it still accounts for volumetric information, firstly because a DoS filter is used as a preprocessing step and secondly since different components of the ensembles are trained on different cross-sections. Ensembles of networks trained on both coronal and sagittal slices give more accurate predictions than individual models or ensembles of models trained on only one of the cross-sections, which was confirmed experimentally. An axial cross-section was not used because the right horizontal fissure is almost perpendicular to the axis in most cases, making the data for training and evaluation very limited. Assuming that 32-bit floating point (FP) encoding is used, each model requires memory for at least one slice of 512×512 per inference pass, which is 1MB for a grayscale image and 3MB for an RGB image, which may be the original slice, a synthesized feature such as a DoS-filtered slice, or both. For comparison, Gerard et al. [24] used fixed-size image crops of $64 \times 200 \times 200$, which is 5 MB for a 32-bit FP-encoded grayscale volume [24]. There is a trade-off between inference speed and memory consumption, as using very small batches or single samples slows down inference but reduces memory requirements.

Most modern computers have gigabytes of RAM. CPUs are able to segment fissures with CNN models by storing data and weights in the RAM rather than the video memory, although they will be much slower than GPUs. In addition to input and output data, there are model weights and intermediate outputs of hidden layers, the memory for which may not be reused in the same pass because the model can have skip connections. Assuming that segmenting the fissures in the scan will require no more than hundreds of times more memory than the input data themselves, the task can be accomplished on most modern computers, even with a 3D model.

The task of fissure segmentation is far from being completely solved and has many issues. Therefore, approaches that allow rapid experimentation are needed. In the inference and evaluation stage, the batch size only affects the execution time. In the training stage,

the batch size is an important hyperparameter that has a huge impact on the learning process. For instance, a larger batch size gives a more accurate estimate of the true gradient, which makes the whole learning process more stable, especially at high learning rates. Two-dimensional CNNs give more freedom to choose the batch size because they are less constrained by memory resources. For this reason, approaches based on 2D CNNs, such as the proposed solution, have an advantage in research, even though this advantage is less significant in the inference process.

We performed cross-validation to ensure that the results are reliable, as the dataset was relatively small. Although on the right lung, the ensemble performed worse, according to the F_1 score, than the single model in cross-validation, it still had a higher *Precision* score, which is more important for applications such as lung lobe segmentation, as they are very sensitive to false positives. Models trained on data preprocessed with the DoS filter performed better both before and after cross-validation; therefore, the combination of the DoS filter and the CNN model gives more reliable prediction results than the CNN model trained on raw data.

The scope of the proposed method is limited to CT scans in which fissures are clearly visible. If a person cannot see fissures on a particular CT scan, they need additional data to do so, for example, a higher-resolution scan of the same patient taken under the same conditions. Patients rarely undergo CT scans twice in a row due to health risks. However, a dataset of high-resolution scans can be collected that can be algorithmically processed to mimic the performance of a low-resolution scanner. In this case, the segmentation masks for training and evaluation can be obtained from high-resolution scans, and the model can be trained and estimated on low-resolution scans. However, we then need a way to evaluate the accuracy of an algorithm that imitates low-resolution scanning.

In addition, in many cases, fissures remain partially invisible despite high scanning quality. Getting more people to annotate the same data and finding a consensus among them partially solves this problem, but the time and effort required increase proportionately. Another possible approach is to use other types of scanning, such as magnetic resonance imaging (MRI), along with CT. Although cases in which both types of imaging are performed on the same patient in a short period of time are rare, there are studies that address such cases [37,38].

Another weakness of our approach is that cases where the disease (more precisely, cancer or nodules mimicking cancer) obscures the lung fissures were rare in our dataset, while these are the most interesting cases, since one possible application of lobe detection is the diagnosis of lung cancer, which largely depends on whether the tumor has reached the inner boundaries of the lobe, which are defined by the lung fissures. Lobectomy, surgery to remove the diseased lobe of the lung, is more effective if the tumor does not cross the lobe boundary [5].

In addition, lung lobe segmentation was performed using fissure segmentation. Although the only purpose of lobe segmentation was to evaluate the quality of the fissure segmentation method, the lobe segmentation algorithm still shows results close to those of state-of-the-art methods, with an average Dice coefficient value of 0.989, which suggests the high efficiency of the proposed fissure segmentation method. In comparison, Wang et al. [28] achieved a median Dice score of 0.993, and Gerard et al. [25] achieved a median Dice score of 0.947.

To improve the fissure segmentation algorithm, we can collect more data, especially those where the appearance of fissures is affected by a disease, which will make the model more robust in edge cases. In addition, in future work, the performance of the proposed method and state-of-the-art 3D CNN models can be compared on the same data. Alternatively, future work can include the segmentation of airways and vessels to further improve the lobe segmentation algorithm. Also, the fissure segmentation algorithm can be used in the task of fissure integrity assessment.

5. Conclusions

We proposed a new method for pulmonary fissure segmentation on CT images using CNNs, image filtering, and model ensembles. The proposed method shows a modified F1 score of 0.916 for right-lung fissures and 0.933 for left-lung fissures, with a cross-validated performance of 0.894 and 0.899, respectively. We also developed a lung lobe segmentation method using the fissure segmentation method and thin-plate spline algorithm. It shows an average Dice coefficient of 0.989, so we conclude that the underlying fissure segmentation algorithm is effective. Future work may include making the fissure segmentation algorithm more robust in edge cases, particularly by utilizing a larger and more balanced dataset or developing an accurate lobe segmentation method. The fissure segmentation method can also be part of a fissure integrity assessment algorithm.

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

Funding: This work was financed by the Ministry of Science and Higher Education of the Russian Federation within the framework of state support for the creation and development of World-Class Research Centers "Digital Biodesign and Personalized Healthcare", project # 075-15-2022-306.

Institutional Review Board Statement: Our access to data was facilitated through the Cooperation Agreement between the Ministry of Health of the Novgorod Region and the Novgorod State University (Agreement number 20230609, signed on 9 June 2023).

Informed Consent Statement: Not applicable.

Data Availability Statement: The imaging data that were used in this study can be shared upon request.

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

References

1. Cederlund, K.; Tylén, U.; Jorfeldt, L.; Aspelin, P. Classification of Emphysema in Candidates for Lung Volume Reduction Surgery: A New Objective and Surgically Oriented Model for Describing CT Severity and Heterogeneity. *Chest* **2002**, *122*, 590–596. [CrossRef] [PubMed]
2. Yousefzadeh, M.; Hasanpour, M.; Zolghadri, M.; Salimi, F.; Yektaeian Vaziri, A.; Mahmoudi Aqeel Abadi, A.; Jafari, R.; Esfahanian, P.; Nazem-Zadeh, M.R. Deep learning framework for prediction of infection severity of COVID-19. *Front. Med.* **2022**, *9*, 940960. [CrossRef] [PubMed]
3. Nakamura, H.; Taniguchi, Y.; Haruki, T. Securing and Treatment of Interlobar Pulmonary Artery in Severe Insufficient Interlobar Fissure of the Lung. *Kyobu Geka. Jpn. J. Thorac. Surg.* **2019**, *72*, 829–833.
4. Nene, A.R.; Gajendra, K.S.; Sarma, M.V.R. Lung lobes and fissures: A morphological study. *J. Exp. Clin. Anat.* **2011**, *5*, 30–38. [CrossRef]
5. Deslauriers, J.; Greógoire, J. Surgical Therapy of Early Non-Small Cell Lung Cancer. *Chest* **2000**, *117*, 104S–109S. [CrossRef]
6. Xiao, C.; Stoel, B.C.; Bakker, M.E.; Peng, Y.; Stolk, J.; Staring, M. Pulmonary Fissure Detection in CT Images Using a Derivative of Stick Filter. *IEEE Trans. Med. Imaging* **2016**, *35*, 1488–1500. [CrossRef]
7. Peng, Y.; Zhong, H.; Xu, Z.; Tu, H.; Li, X.; Peng, L. Pulmonary Lobe Segmentation in CT Images Based on Lung Anatomy Knowledge. *Math. Probl. Eng.* **2021**, *2021*, 5588629. [CrossRef]
8. van Rikxoort, E.M.; Prokop, M.; de Hoop, B.; Viergever, M.A.; Pluim, J.P.W.; van Ginneken, B. Automatic Segmentation of Pulmonary Lobes Robust Against Incomplete Fissures. *IEEE Trans. Med. Imaging* **2010**, *29*, 1286–1296. [CrossRef]
9. Pu, J.; Wang, Z.; Gu, S.; Fuhrman, C.; Leader, J.; Meng, X.; Tedrow, J.; Sciurba, F. Pulmonary Fissure Integrity and Collateral Ventilation in COPD Patients. *PLoS ONE* **2014**, *9*, e96631. [CrossRef]
10. Althof, Z.; Gerard, S.; Pan, Y.; Christensen, G.; Hoffman, E.; Reinhardt, J. Automatic Quantification of Pulmonary Fissure Integrity: A Repeatability Analysis. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020; pp. 581–585. [CrossRef]
11. Althof, Z.; Gerard, S.; Eskandari, A.; Galizia, M.; Hoffman, E.; Reinhardt, J. Attention U-net for automated pulmonary fissure integrity analysis in lung computed tomography images. *Sci. Rep.* **2023**, *13*, 14135. [CrossRef]
12. Aziz, A.; Ashizawa, K.; Nagaoki, K.; Hayashi, K. High resolution CT anatomy of the pulmonary fissures. *J. Thorac. Imaging* **2004**, *19*, 186–191. [CrossRef] [PubMed]

13. Li, Q.; Chen, L.; Li, X.; Xia, S.; Kang, Y. An improved random forests approach for interactive lobar segmentation on emphysema detection. *Granul. Comput.* **2020**, *5*, 503–512. [CrossRef]
14. Kuhnigk, J.M.; Hahn, H.; Hindennach, M.; Dicken, V.; Krass, S.; Peitgen, H.O. Lung lobe segmentation by anatomy-guided 3D watershed transform. *Proc. SPIE-Int. Soc. Opt. Eng.* **2003**, *5032*, 1482–1490. [CrossRef]
15. Ukil, S.; Reinhardt, J. Anatomy-Guided Lung Lobe Segmentation in X-Ray CT Images. *IEEE Trans. Med Imaging* **2009**, *28*, 202–214. [CrossRef] [PubMed]
16. Zhou, X.; Hayashi, T.; Hara, T.; Fujita, H.; Yokoyama, R.; Kiryu, T.; Hoshi, H. Automatic segmentation and recognition of anatomical lung structures from high-resolution chest CT images. *Comput. Med. Imaging Graph.* **2006**, *30*, 299–313. [CrossRef]
17. Wei, Q.; Hu, Y.; Gelfand, G.; Macgregor, J. Segmentation of Lung Lobes in High-Resolution Isotropic CT Images. *Biomed. Eng. IEEE Trans.* **2009**, *56*, 1383–1393. [CrossRef]
18. Appia, V.; Patil, U.; Das, B. Lung Fissure detection in CT Images using Global Minimal Paths. *Proc. SPIE* **2010**, *7623*, 76231. [CrossRef]
19. Peng, Y.; Xiao, C. An oriented derivative of stick filter and post-processing segmentation algorithms for pulmonary fissure detection in CT images. *Biomed. Signal Process. Control.* **2018**, *43*, 278–288. [CrossRef]
20. Hong, Z.; Stoel, B.; Staring, M.; Bakker, E.; Stolk, J.; Zhou, P.; Xiao, C. A framework for pulmonary fissure segmentation in 3D CT images using a directional derivative of plate filter. *Signal Process.* **2020**, *173*, 107602. [CrossRef]
21. Hong, Z. Lung Lobe Segmentation Based on Lung Fissure Surface Classification Using a Point Cloud Region Growing Approach. *Algorithms* **2020**, *13*, 263. [CrossRef]
22. Ross, J.; Kindlmann, G.; Okajima, Y.; Hatabu, H.; Diaz, A.; Silverman, E.; Washko, G.; Dy, J.; Estépar, R. Pulmonary lobe segmentation based on ridge surface sampling and shape model fitting. *Med. Phys.* **2013**, *40*, 121903. [CrossRef] [PubMed]
23. Bookstein, F. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 567–585. [CrossRef]
24. Gerard, S.; Patton, T.; Christensen, G.; Bayouth, J.; Reinhardt, J. FissureNet: A Deep Learning Approach For Pulmonary Fissure Detection in CT Images. *IEEE Trans. Med. Imaging* **2018**, *38*, 156–166. [CrossRef]
25. Gerard, S.E.; Reinhardt, J.M. Pulmonary Lobe Segmentation Using A Sequence of Convolutional Neural Networks For Marginal Learning. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 1207–1211. [CrossRef]
26. Gerard, S.; Herrmann, J.; Xin, Y.; Martin, K.; Rezoagli, E.; Ippolito, D.; Bellani, G.; Cereda, M.; Guo, J.; Hoffman, E.; et al. CT image segmentation for inflamed and fibrotic lungs using a multi-resolution convolutional neural network. *Sci. Rep.* **2021**, *11*, 1455. [CrossRef] [PubMed]
27. Park, J.; Yun, J.; Kim, N.; Park, B.; Cho, Y.; Park, H.; Song, M.; Lee, M.; Seo, J.B. Fully Automated Lung Lobe Segmentation in Volumetric Chest CT with 3D U-Net: Validation with Intra- and Extra-Datasets. *J. Digit. Imaging* **2019**, *33*, 221–230. [CrossRef]
28. Wang, W.; Chen, J.; Zhao, J.; Chi, Y.; Xie, X.; Zhang, L.; Hua, X. Automated Segmentation Of Pulmonary Lobes Using Coordination-Guided Deep Neural Networks. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 1353–1357. [CrossRef]
29. Chen, Q.; Tang, H.; Dillenseger, J.L. LLASN: Lung Lobes Segmentation that Uses Adversarial Network. In Proceedings of the 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 22–24 October 2021; pp. 240–244. [CrossRef]
30. Dadras, A.; Jaziri, A.; Frodl, E.; Vogl, T.; Dietz, J.; Bucher, A. Lightweight Techniques to Improve Generalization and Robustness of U-Net Based Networks for Pulmonary Lobe Segmentation. *Bioengineering* **2023**, *11*, 21. [CrossRef]
31. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Jorge Cardoso, M. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Cardoso, M.J., Arbel, T., Carneiro, G., Syeda-Mahmood, T., Tavares, J.M.R., Moradi, M., Bradley, A., Greenspan, H., Papa, J.P., Madabhushi, A., et al., Eds.; Springer: Cham, Switzerland, 2017; pp. 240–248.
32. Yeung, M.; Sala, E.; Schönlieb, C.B.; Rundo, L. Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Comput. Med. Imaging Graph.* **2022**, *95*, 102026. [CrossRef] [PubMed]
33. 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]
34. Estépar, R.S.J.; Ross, J.C.; Harmouche, R.; Onieva, J.; Diaz, A.A.; Washko, G.R. Chest Imaging Platform: An Open-Source Library and Workstation for Quantitative Chest Imaging. In *C66. LUNG IMAGING II: NEW PROBES AND EMERGING TECHNOLOGIES*; American Thoracic Society: New York, NY, USA , 2015; pp. A4975–A4975. Available online: https://www.atsjournals.org/doi/abs/10.1164/ajrccm-conference.2015.191.1_MeetingAbstracts.A4975 (accessed on 29 September 2024).
35. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and Flexible Image Augmentations. *Information* **2020**, *11*, 125. [CrossRef]
36. Iakubovskii, P. Segmentation Models Pytorch-Python Library with Neural Networks for Image Segmentation Based on PyTorch. 2019. Available online: https://github.com/qubvel/segmentation_models.pytorch (accessed on 29 September 2024).

37. Chalela, J.; Kidwell, C.; Nentwich, L.; Luby, M.; Butman, J.; Demchuk, A.; Hill, M.; Patronas, N.; Latour, L.; Warach, S. Magnetic resonance imaging and computed tomography in emergency assessment of patients with suspected acute stroke: A prospective comparison. *Lancet* **2007**, *369*, 293–298. [CrossRef]
38. Ardekani, M.; Redaei, M.; Ghasemipour, S.; Ameri Ahmad, Z. Comparative study of CT-scan modality with MRI modality findings in patients suspected of COVID-19. *Egypt. J. Radiol. Nucl. Med.* **2023**, *54*, 75. [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

MaskAppendix: Backbone-Enriched Mask R-CNN Based on Grad-CAM for Automatic Appendix Segmentation

Emre Dandıl [1,*], Betül Tiryaki Baştuğ [2], Mehmet Süleyman Yıldırım [3], Kadir Çorbacı [4] and Gürkan Güneri [5]

[1] Department of Computer Engineering, Faculty of Engineering, Bilecik Seyh Edebali University, 11230 Bilecik, Türkiye
[2] Radiology Department, Faculty of Medicine, Bilecik Şeyh Edebali University, 11230 Bilecik, Türkiye; betultryak@yahoo.com
[3] Department of Söğüt Vocational School, Computer Technology, Bilecik Şeyh Edebali University, Söğüt, 11600 Bilecik, Türkiye; mehmets.yildirim@bilecik.edu.tr
[4] General Surgery Department, Bilecik Osmaneli Mustafa Selahattin Çetintaş Hospital, 11500 Bilecik, Türkiye; dr.kadircorbaci@gmail.com
[5] General Surgery Department, Faculty of Medicine, Bilecik Şeyh Edebali University, 11230 Bilecik, Türkiye; gurkanguneri@gmail.com
* Correspondence: emre.dandil@bilecik.edu.tr

Abstract: Background: A leading cause of emergency abdominal surgery, appendicitis is a common condition affecting millions of people worldwide. Automatic and accurate segmentation of the appendix from medical imaging is a challenging task, due to its small size, variability in shape, and proximity to other anatomical structures. Methods: In this study, we propose a backbone-enriched Mask R-CNN architecture (MaskAppendix) on the Detectron platform, enhanced with Gradient-weighted Class Activation Mapping (Grad-CAM), for precise appendix segmentation on computed tomography (CT) scans. In the proposed MaskAppendix deep learning model, ResNet101 network is used as the backbone. By integrating Grad-CAM into the MaskAppendix network, our model improves feature localization, allowing it to better capture subtle variations in appendix morphology. Results: We conduct extensive experiments on a dataset of abdominal CT scans, demonstrating that our method achieves state-of-the-art performance in appendix segmentation, outperforming traditional segmentation techniques in terms of both accuracy and robustness. In the automatic segmentation of the appendix region in CT slices, a DSC score of 87.17% was achieved with the proposed approach, and the results obtained have the potential to improve clinical diagnostic accuracy. Conclusions: This framework provides an effective tool for aiding clinicians in the diagnosis of appendicitis and other related conditions, reducing the potential for diagnostic errors and enhancing clinical workflow efficiency.

Keywords: appendix segmentation; deep learning; CT imaging; mask R-CNN; grad-CAM; Detectron

1. Introduction

The appendix, a small, tube-like structure, extends from the cecum, the beginning of the large intestine, and is typically located in the lower right abdomen. Specifically, the appendix is found in the right iliac fossa, a region in the lower right part of the abdomen. Anatomically, it is a blind-ended pouch with a length that varies significantly among individuals, ranging from 5 to 35 cm [1]. As in the structure of the intestinal wall of the colon, the appendix's wall histologically consists of mucosa, submucosa, muscularis externa, and serosa [2]. Physiologically, the appendix is thought to have a role in maintaining gut flora and responding to infections. However, its precise function remains debated, with theories ranging from a vestigial organ to a site for immune system regulation [3].

Appendicitis is an inflammation of the appendix [4]. It is a common surgical emergency characterized by acute or chronic inflammation, often leading to severe abdominal

pain and, if untreated, potentially serious complications. The most common cause of abdominal pain is acute appendicitis [5]. Acute appendicitis is an important health problem that affects approximately 17 million people worldwide every year [6]. The most common cause of appendicitis is obstruction of the appendiceal lumen, which can be due to fecaliths (hardened stool), lymphoid hyperplasia, or foreign bodies. This obstruction impairs the drainage of mucus and other secretions, leading to increased intraluminal pressure, reduced blood flow, and subsequent bacterial overgrowth and inflammation [7–11]. Symptoms typically include right lower abdominal pain, nausea, vomiting, and fever. The condition can progress rapidly, and if the appendix becomes necrotic or ruptures, it can lead to peritonitis, a widespread abdominal infection, or abscess formation [12,13].

Treatment for appendicitis usually involves surgical removal of the appendix, known as an appendectomy, which can be performed via open surgery or laparoscopic technique [14]. A meta-analysis showed that appendectomy was significantly more effective in the treatment of patients diagnosed with acute appendicitis than in patients who received antibiotics alone [15]. This situation has shown us once again the importance of making the correct diagnosis of acute appendicitis. Diagnosis of acute appendicitis is difficult in emergency departments because diagnosis based solely on clinical evaluation has low sensitivity and low specificity [16]. In a study of 800,000 patients, in-hospital mortality was found to be 0.12% [17]. In some perioperative cases, appendectomy alone cannot be performed or is not sufficient. Complicated surgical procedures such as ileocecal resection, right hemicolectomy, and end ileostomy increase morbidity [18]. On the other hand, misdiagnosis of appendicitis may result in negative appendectomies. A negative appendectomy can cause long-term problems in addition to the risks of anesthesia and surgery. The increased incidence of inflammatory bowel disease and clostridium difficile infection in patients who have had appendectomy suggests that the appendix plays an important immunologic role [19]. Therefore, correctly identifying acute appendicitis and correctly grading it radiologically can allow the surgeon to prepare for possible morbidities and provide the opportunity to reduce morbidity. Early diagnosis and intervention are crucial to prevent complications and ensure favorable outcomes.

Computed tomography (CT) is a pivotal imaging modality in the assessment of appendiceal diseases, providing high-resolution cross-sectional images that aid in both diagnosis and management. Diagnosis of appendicitis is primarily clinical but can be supported by imaging such as ultrasound or CT to confirm the diagnosis and assess for complications [20–22]. When performed with oral or intravenous contrast, CT enhances the visualization of the appendix and surrounding structures, allowing for accurate differentiation between appendicitis and other abdominal pathologies. Considering the sensitivity and specificity of CT, it is a very successful radiological examination method in the diagnosis and management of acute appendicitis. However, the urgency of the disease increases the workload of radiologists in the evaluation of these CTs and causes loss of time when planning treatment for patients [23]. Therefore, the development of technologies that interpret CTs at least as accurately as radiologists and faster than them is an important need for clinicians. A better segmentation of the appendix significantly improves diagnostic accuracy, accelerates decision making and allows for more effective treatment planning. Precise segmentation may allow clinicians to visualize the appendix more clearly on CT scans, making it easier to identify signs of inflammation, perforation, or abscess development. By accurately delineating the appendix from surrounding structures, the risk of missing important diagnostic details can be reduced. In emergency situations, faster and more reliable identification of appendicitis can be life-saving. Automated segmentation of the appendix simplifies the radiology workflow by providing instant and accurate contours of the appendix, allowing radiologists to make faster and more assured decisions [24].

Because the appendix is a small, finger-shaped organ in the lower right abdomen, it is sometimes difficult to see, especially on medical scans. The appendix usually causes no symptoms or discomfort in everyday life. However, if appendicitis develops as a result of inflammation or infection of the appendix, the appendix must be removed urgently as

it can lead to serious complications. In addition, the appendix can be located in many different places in the abdominal cavity [25]. This variability makes the diagnosis difficult, as the inflammation can be different depending on the location. On the other hand, misdiagnosing appendicitis as another condition can delay necessary surgery and increase the risk of complications such as perforation and peritonitis [26]. Well-defined segmentation delineates the appendix boundaries, ensuring radiologists can confidently identify the appendix. It is therefore very important to segment the appendix accurately with precise boundaries. Previous studies on appendix segmentation are scarce or very limited. In particular, it can be seen that the proposed studies on the appendix mostly focus on the classification of normal appendix and appendicitis and the prediction of appendicitis. In one of these studies, Marcinkevics et al. [27] applied prediction of appendicitis in children based on comprehensive medical history, clinical assessment, laboratory parameters, and abdominal ultrasound. The study used machine learning approaches including logistic regression, random forests and boosting machines. In another study proposed by Byun et al. [28], CT image review, laboratory tests, and clinical features were used with a decision tree-based machine learning approach to predict appendicitis in children.

In medical imaging and image processing, artificial intelligence (AI) is emerging as a leading field [29]. AI is a powerful tool for the transformation of raw data into meaningful insights [30]. The emergence of deep learning has transformed the field of AI and has significantly changed the field of data analytics [31]. Significant contributions have been provided to improve clinical decision support systems by integrating deep learning into clinical applications [32]. In particular, the quality of image segmentation has been significantly improved through the use of deep learning methods [33]. Deep learning approaches also provide significant benefits to medical image segmentation, particularly in terms of feature learning, automatization, and improved accuracy. Deep learning-based methods automatically learn structured features from image data, detecting both low-level features such as edges and textures and high-level features such as shape and structure, without the need for manual feature extraction [34]. Deep learning approaches have achieved state-of-the-art performance in medical image segmentation tasks [35]. In recent years, many deep learning-based methods have been proposed in evaluation of appendiceal diseases. These computer-aided methods have enhanced diagnostic accuracy, streamlined workflows, and offered novel insights. In the context of appendiceal diseases, deep learning models can assist in detecting appendicitis, identifying appendiceal abnormalities, and predicting disease progression. For instance, these models can be trained to segment the appendix accurately and to classify images based on the presence of inflammation, perforation, or other complications. In addition, deep learning-based systems can assist radiologists by providing additional diagnostic insights and reducing the burden of image analysis [36]. In one of the proposed studies based on deep learning, using a reinforcement learning based on CNN, Al et al. [37] proposed an approach that can detect acute appendicitis in abdominal CT images. Lee et al. [38] focused on the distinction between acute appendicitis and diverticulitis using 2D CNN in abdominal CT scans. Liang et al. [39] presented an approach using combined deep learning and radiomics to differentiate between complicated and uncomplicated cases of acute appendicitis. Park et al. [40] proposed a CNN method for the classification of appendicitis using CT scans of patients with abdominal pain presenting to the emergency department. Rajpurkar et al. [41] presented a 3D deep learning model (AppendiXNet) for classification of appendicitis from CT scans. In another study, Park et al. [36] proposed the classification of normal appendix, diverticulitis, and acute appendicitis using EfficientNet, a deep learning model.

Previous studies on the appendix have mostly focused on predicting appendicitis using machine learning with different findings and classifying appendicitis from CT scans. Especially before the diagnosis of appendicitis, it is important to accurately segment the appendix organ with its real boundaries. At this stage, the appearance of borders can make it difficult to make a correct diagnosis or cause complications in the treatment process. Appendicitis is diagnosed using a combination of the doctor's clinical experience,

physical examination findings, imaging techniques, and laboratory tests. However, false positive or false negative results may be obtained due to the limitations mentioned above. Therefore, in patients with suspected appendicitis, the boundaries of the appendix must be precisely delineated. The main reason for our study is this need for automatic evaluation of appendix. Based on this, we aimed to develop a model that can evaluate the appendix on CT images. In this study, we propose the MaskAppendix for appendix segmentation, examining its performance in terms of accuracy, computational efficiency, and robustness in CT imaging modality. Furthermore, we compare the effectiveness of MaskAppendix with the state-of-the-art methods to highlight the strengths and limitations of each model. Our results demonstrate the potential of MaskAppendix to transform the landscape of appendix segmentation, paving the way for more streamlined and reliable diagnostic workflows in detection of appendix. Applying MaskAppendix to appendix segmentation leverages its ability to precisely delineate the appendix from surrounding structures in complex CT images. This capability is particularly valuable given the appendix's small size and the variability in its anatomical location and appearance across different patients. By automating the segmentation process, MaskAppendix can potentially reduce the workload of radiologists and improve the consistency of appendix identification in clinical practice. The highlights of this study are:

- Automatic segmentation of the appendix is provided on CT scans;
- Segmentation of the appendix is applied with high performance using MaskAppendix on the Detectron platform;
- An appendix CT image dataset is created for this study;
- The precise appendix segmentation on CT scans with localization is enhanced using Grad-CAM;
- This framework provides an effective tool for aiding clinicians in the diagnosis of appendicitis;
- The MaskAppendix method achieves state-of-the-art performance in appendix segmentation.

The following sections of this study are organized as follows. In Section 2, the details of the appendix dataset prepared for this study are presented and the infrastructure of the proposed MaskAppendix is explained in detail. In Section 3, the results of the experimental studies are presented. An Ablation Study is also conducted in this section. The findings and evaluations are discussed in Section 4 and the results are compared with state-of-the-art methods. Section 5 contains conclusions and future work.

2. Materials and Methods

The general methodology of the MaskAppendix model proposed in this study for the automatic segmentation of the appendix is shown in Figure 1. In the methodology of MaskAppendix, in the first step, the dataset consisting of CT scans was prepared and divided into two categories as training and test data. In the second step, the MaskAppendix deep learning model was trained on the prepared training set and the optimal network weights were obtained. In the last step, the appendix was segmented on the test set using the best weights obtained in the training phase for the MaskAppendix model and the appendix segmentation results of the proposed model were evaluated. The model exploits the strengths of each component: Mask R-CNN's segmentation precision, ResNet's powerful feature extraction, Grad-CAM's transparency and Detectron's modularity. In the proposed MaskAppendix model, the ResNet50 and ResNet101 architectures serve as the backbone of Mask R-CNN, providing the deep and rich features that are critical for detecting and segmenting small and complex structures such as the appendix. By integrating Grad-CAM into this architecture, it is possible for clinicians to visualize which features or areas of the CT scan influenced the model's decision. The combination of Mask R-CNN's instance segmentation, ResNet's deep feature extraction, and Grad-CAM's explainability ensures that the model not only provides highly accurate segmentations, but also provides insight into how those decisions were achieved. In addition, Detectron

simplifies the deployment of the Mask R-CNN model with a ResNet backbone. It enables efficient training, testing, and fine-tuning of the segmentation models, saving time and computing resources.

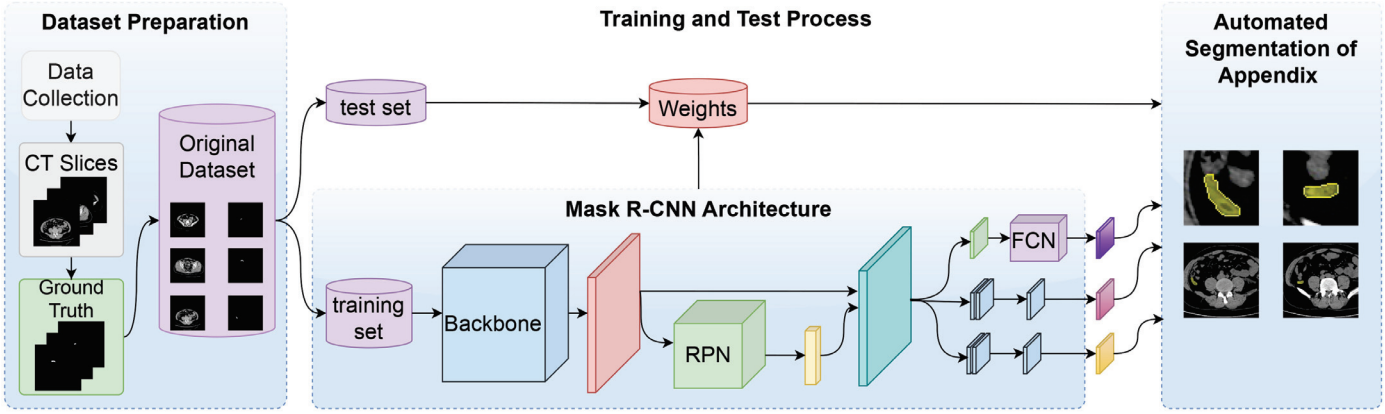


Figure 1. Overall architecture of the proposed MaskAppendix model for automatic segmentation of the appendix.

2.1. Dataset

In this study, an original dataset of CT scans was created for automatic segmentation of appendix. Abdominal CT images of different patients in the dataset were collected at Bilecik Training and Research Hospital in the presence of physicians from the Radiology and General Surgery Departments. The abdominal CT scans were obtained from 235 patients who presented to the emergency department with abdominal pain and suspected acute appendicitis. A total of 153 of the patients were male and the remaining 82 were female, aged between 18 and 91 years. Patients were evaluated in the ward in terms of demographic and laboratory characteristics as well as the presence or absence of appendicitis. Therefore, 129 of the patients were diagnosed with acute appendicitis by physicians and the remaining 106 were healthy with a normal appendix. Patients were placed in the supine position during the CT scan, and slices were acquired from the upper abdomen to the lower abdomen. CT scans were performed between March 2022 and April 2024. The CT scanner and some important acquisition parameters are listed in Table 1. The study was ethically approved by Non-Interventional Clinical Research Ethics Committee of Bilecik Şeyh Edebali University according to decision number 13 of the 8th meeting on 5 December 2023. In addition, the patients whose data were included in the dataset created as part of the study agreed to the use of their data for research purposes.

Table 1. The used CT scanner and some important acquisition parameters.

CT scanner	64-MDCT (multi-slice CT Aquillion 64; Toshiba)
Slice thickness	2.5 mm
Reconstruction interval	0.777 mm
Gantry rotation time	0.6 s
Tube voltage	120 kV
Tube current	200 mA (in average)
The field of view range	From 40 to 50 cm
Image size	512 × 512

In the dataset created for the appendix segmentation in the study, the patients' CT scans were obtained in DICOM format. During data preparation, axial slices from each patient's CT scan that were considered to contain the appendix were identified by specialized physicians. For the dataset, a total of 940 slices, 459 slices from scanned patients with appendicitis and 481 slices from healthy individuals, were identified as suitable for appendix segmentation. The appendix regions in these CT slices were extracted by expert

physicians using the ITK-SNAP v4.0.2 software [42,43], as shown in Figure 2. For each original CT slice in the dataset shown in Figure 2a, the ground truth mask delineated by the expert is shown in Figure 2b and the appendix region extracted on the CT slice is shown in Figure 2c. In addition, the ground truth masks of the appendix regions for the dataset were saved in NIfTI format. Image pre-processing techniques were then used to enhance the CT images by histogram equalization, noise removal and line sharpening.

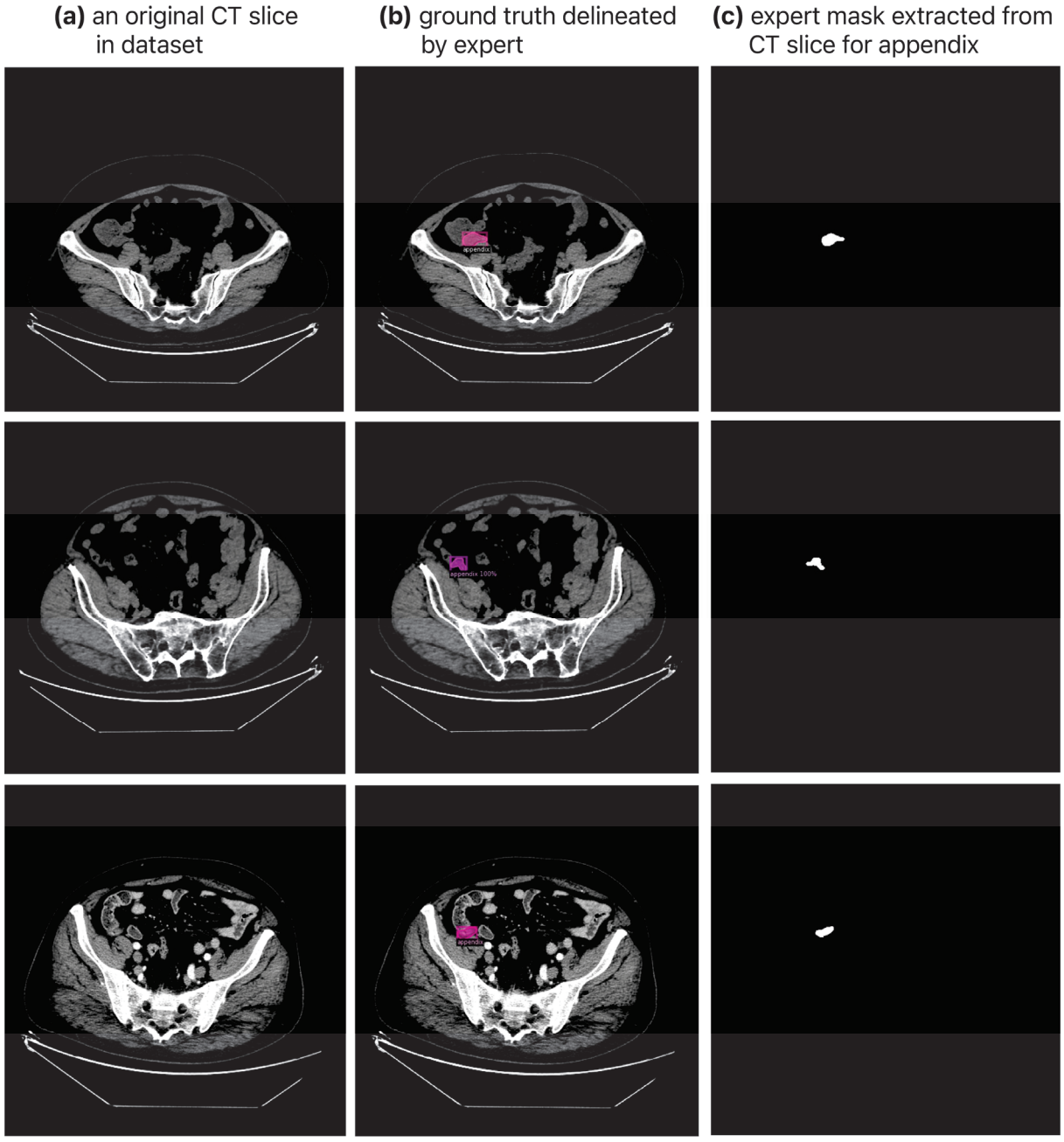


Figure 2. Original CT in dataset slices and the appendix region masks in these CT slices extracted by expert physicians.

2.2. Proposed Mask R-CNN Model

Mask R-CNN is one of the current deep learning methods widely used for object recognition from images and accurate segmentation of objects in images [44]. Mask R-CNN is also a heuristic extension of faster R-CNN [45], another widely used method for object recognition, and has a very easy to train architecture. A typical Mask R-CNN architecture

consists of three stages [46,47]. In the first stage, as in the faster R-CNN, the input image is processed by a backbone network and passed to a Region Proposal Network (RPN) to extract features and generate the feature map [47]. In the second stage of Mask R-CNN, RPN is applied to the associated candidate regions. Here, inspired by fast R-CNN [48], region of interest alignment (RoIAlign) is performed by applying a RoI pooling layer to the candidate bounding boxes to make each candidate region the same size. In the final stage of Mask R-CNN, the bounding boxes of the objects are classified by passing them through a Fully Connected Layer (FCL). Unlike Faster R-CNN, Mask R-CNN uses a Fully Convolutional Network (FCN) to create a mask in parallel with each bounding box [49]. In this study, a deep learning model based on the Mask R-CNN architecture shown in Figure 3 is proposed for appendix detection on axial CT scans.

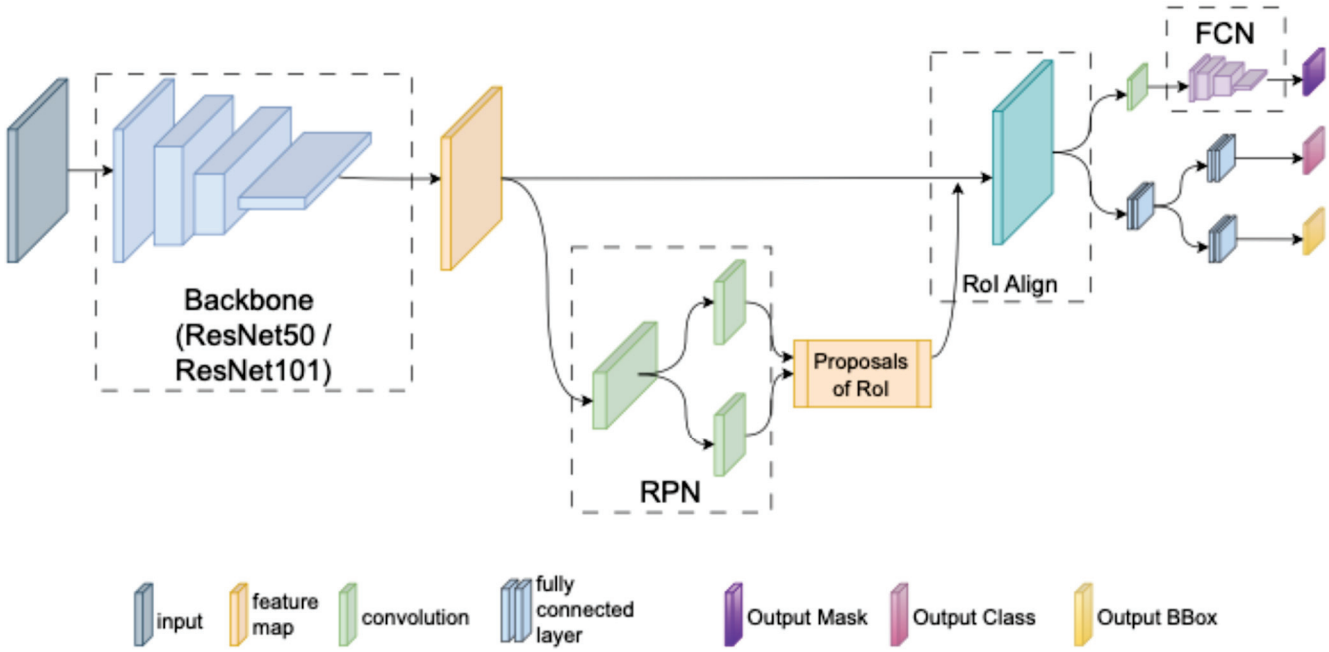


Figure 3. The proposed Mask R-CNN framework for automatic segmentation of the appendix on CT images.

In the masked R-CNN, the loss function given in Equation (1) is used for the mask part (head). In this equation, L_{mask} calculates the pixel-wise cross-entropy between the target mask and the estimated mask. $L_{Mask\ R-CNN}$ denotes a multitasking loss function based on the loss function used for fast R-CNN [48,50]. Here, the classification loss L_{class} and the bounding box loss (L_{bbox}) are the same as in fast R-CNN [48]. L_{mask} is the average binary cross-entropy loss for the RoI associated with the ground truth class [49].

$$L_{Mask\ R-CNN} = L_{class} + L_{bbox} + L_{mask} \quad (1)$$

2.3. ResNet

In the Mask R-CNN architecture, the ResNet [51] architecture is widely used as the backbone to perform feature extraction from images differently. The architecture of the ResNet network, which is an improved combination of CNN structures, consists of block structures [52]. The vanishing gradient problem, which arises due to the large number of layers in deep learning models, is solved by using the residual blocks introduced by the ResNet network. Residual blocks allow an input to be directly summed with the output, creating short-cut connections. In the ResNet architecture, residual links can be created at different depths. In this study, ResNet50 and ResNet101 networks with a depth of 50 and 101 layers, respectively, are used as the backbone in the first stage of the Mask R-CNN architecture for feature extraction and feature map generation from images. Since the

ResNet50 and ResNet101 architectures are deeper ResNet networks, bottleneck blocks are used in their infrastructure. In bottleneck blocks, 1×1 convolution is used to reduce the number of channels, 3×3 convolution and 1×1 convolution are used to restore the number of channels. In this way, high accuracy and speed benefits can be achieved in experimental studies. First, for each abdominal CT image, the features are extracted using the ResNet backbone and the window shifting method is used to determine the bounding box recommendations associated with the feature map. The internal structure of the ResNet backbone network is shown in Figure 4a. On the other hand, Figure 4b shows the depth generation with residual block in ResNet50 and ResNet101.

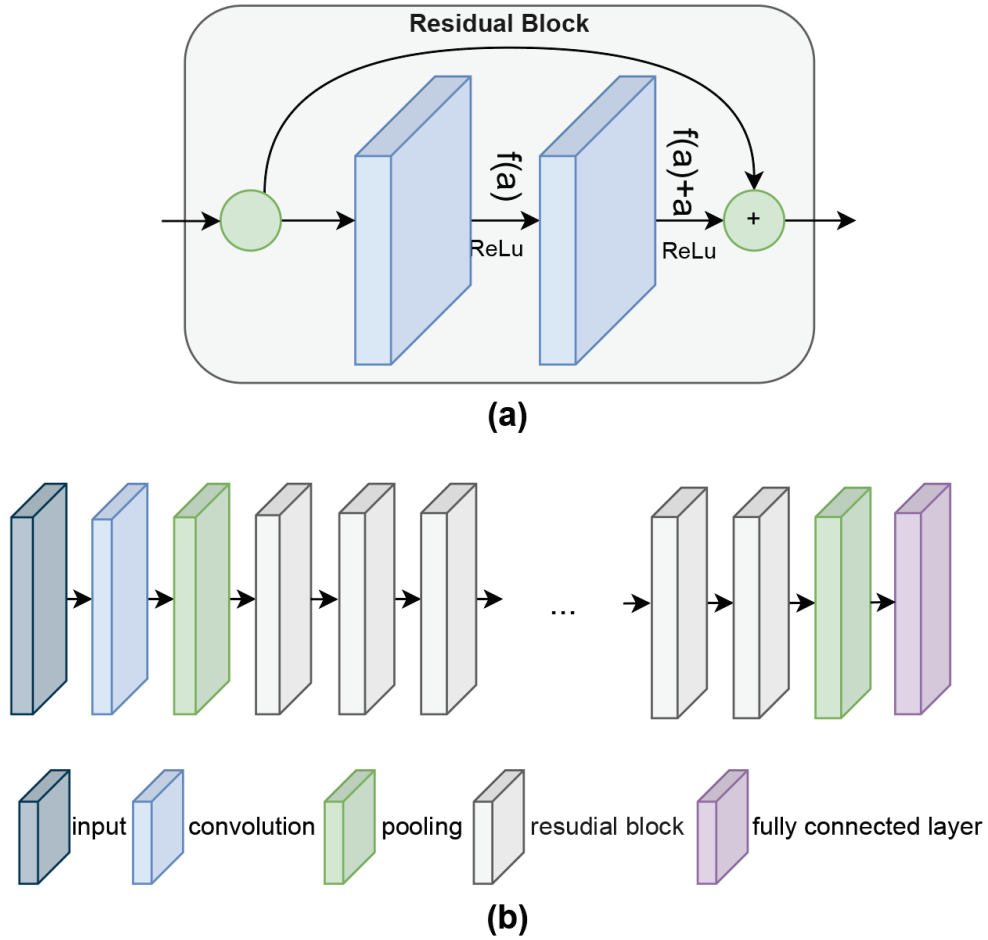


Figure 4. (**a**) Internal structure of the ResNet backbone network, (**b**) depth generation with residual block in ResNet50 and ResNet101.

2.4. Detectron2 Platform

The Detectron infrastructure is a tool that provides advanced object recognition and segmentation applications, in particular the Mask R-CNN architecture. Detectron2 [53] is a state-of-the-art open-source library developed by Facebook AI Research and compiled entirely in PyTorch. Detectron2 also offers a dynamic architecture, faster training, and easier model configuration. In addition, the PyTorch-based Detectron2 infrastructure includes different configurations for Mask R-CNN and can be used directly with fine-tuned parameters. Because of these features, the MaskAppendix model proposed in this study was run on the Detectron2 platform.

3. Experimental Studies

In this study, several experimental analyses were performed to verify the performance of the proposed Mask R-CNN based model with two different backbones running on the

Detectron2 library for appendix detection from CT images, and the results and findings were evaluated. All implementations of the proposed MaskAppendix were done using the Python programming language. A workstation computer, whose hardware components are listed in Table 2, was used for the experimental studies. Two GPUs were used to increase the processing speed and to obtain more effective segmentation results. The model used is based on the Mask R-CNN framework on the open source Detectron2 platform.

Table 2. Hardware characteristics of the computer used for experimental studies.

Hardware	Characteristic
Computer	Workstation
Central Processor (CPU)	Intel Core i9-9900 K @ 5 GHz (8 Core/16 Thread)
Memory (RAM) (×2)	16 GB DDR4 2666 MHz
Mainboard	WS Z390 Pro DDR4
GPU (×2)	NVIDIA GeForce GTX 1080Ti 11 GB
Hard disk drive	500 GB SSD + 3 TB SATA 6 Gb 3.5″

In this study, several improvements were performed for test evaluation and training using the model developed on the open-source Mask R-CNN framework on the Detectron2 platform. Hyperparameter selection and optimization are critical to the success of deep learning models such as MaskAppendix. In this study, by systematically experimenting with different hyperparameters, performing ablation studies, and using grid search methods, the best performing configuration was identified. This process ensures that the model achieves high segmentation accuracy, robustness, and computational efficiency, making it a valuable support for automatic appendix segmentation in CT scans. The optimization of the hyperparameters used for the MaskAppendix model proposed in this study for automatic appendix segmentation is presented in Table 3. These parameters have been identified with the best results based on the tests carried out in the experimental studies. ResNet50 or ResNet101 was used as the backbone for ablation in the Mask R-CNN network. In addition, it was noted that faster results were obtained using two GPUs on the computer where the experimental analysis was performed. On the other hand, the number of worker threads or processes used for data loading and the number of images processed in each training batch were 16 and 2, respectively. The learning rate in the network was chosen to be 0.001 for better learning of details. Furthermore, the optimal CHECKPOINT_PERIOD for training the network is 1000 and MAX_ITER is 100,000, which was also verified by the performance on the test set. For the sizes of the anchors for object detection, the values 16, 32, 64, 128, and 256 were applied as a list, and the number of RoI was set to 256.

Table 3. Optimization of the hyperparameters used for the Mask R-CNN architecture run on the Detectron2 platform.

Hyperparameter	Value	Identification
BACKBONE	ResNet50 and ResNet101	Network model for feature extraction and feature mapping
GPU_COUNT	2	Number of GPUs on the computer on which the network runs
NUM_WORKERS	16	The number of worker threads or processes used for data loading
IMS_PER_BATCH	2	The number of images processed in each training batch
BASE_LR	0.001	Learning rate of the network
CHECKPOINT_PERIOD	1000	How often the state of the model is recorded during training
MAX_ITER	100,000	Maximum number of iterations required to complete the training
ANCHOR_GENERATOR.SIZES	(16, 32, 64, 128, 256)	The sizes of the anchors for object detection
NUM_CLASSES	1	Number of classes for CT slices
BATCH_SIZE_PER_IMAGE	256	The number of RoI

In this study, the appendix segmentation performance of the proposed MaskAppendix architecture was evaluated using both distance-based and overlap-based metrics. For appendix segmentation from abdominal CT images, the segmentation results obtained using the proposed MaskAppendix model and the ground truth masks delineated by expert physicians were compared with the key metrics Dice Similarity Coefficient (DSC), Jaccard Similarity Index (JSI), Volumetric Overlap Error (VOE), Average Symmetric Surface Distance (ASD), and Hausdorff Distance 95 (HD95) to demonstrate the accuracy of the achieved segmentation. These metrics are given in Equation (2), Equation (3), Equation (4), Equation (5), and Equations (6), (7), and (8) respectively. In the equations given for the criteria used, the area of the ground truth masks delineated by the reference expert physicians is denoted as A_{GT} and the area masked by the proposed model is denoted as A_{MA}. DSC calculates the overlap ratio of the areas of the ground truth masks delineated by the experts and the area masked by the proposed method [54]. On the other hand, by taking into account the intersection of the ground truth mask areas and the masked areas using the proposed method, the overlap ratio can be calculated from a different perspective with the JSI metric [55]. While the VOE key metric is used to determine the overlap error for the ground truth mask regions and the masked regions using the proposed method, the difference that physically occurs between the ground truth mask regions and the masked regions using the proposed method is calculated by ASD. In the calculation of the ASD metric, d represents the minimum distance between two segmentation masks. HD95 is used in the metric space to show how far apart the ground truth mask areas and the areas masked using the proposed method are from each other.

$$\mathrm{DSC}(A_{MA}, A_{GT}) = \frac{2|A_{MA} \cap A_{GT}|}{|A_{MA}| \cup |A_{GT}|} \times 100 \tag{2}$$

$$\mathrm{JSI}(A_{MA}, A_{GT}) = \frac{|A_{MA} \cap A_{GT}|}{|A_{MA}| + |A_{GT}| - |A_{MA} \cap A_{GT}|} \times 100 \tag{3}$$

$$\mathrm{VOE}(A_{MA}, A_{GT}) = 100 \times \left(1 - \frac{|A_{MA} \cap A_{GT}|}{|A_{MA} \cup A_{GT}|}\right) \tag{4}$$

$$\mathrm{ASD}(A_{MA}, A_{GT}) = \frac{1}{|S(A_{MA})| + |S(A_{GT})|} \begin{pmatrix} \sum_{S_{A_{MA}} \in S(A_{MA})} d(SA_{MA}, S(A_{GT})) + \\ \sum_{S_{A_{ref}} \in S(A_{GT})} d(SA_{GT}, S(A_{MA})) \end{pmatrix} \tag{5}$$

$$hd(A_{MA}, A_{GT}) = \max_{x \epsilon A_{MA}} \min_{y \epsilon A_{GT}} \|x - y\|_2 \tag{6}$$

$$hd(A_{GT}, A_{MA}) = \max_{x \epsilon A_{GT}} \min_{y \epsilon A_{MA}} \|x - y\|_2 \tag{7}$$

$$\mathrm{HD95}(A_{GT}, A_{MA}) = \max\left(hd(A_{MA}, A_{GT}), hd(A_{GT}, A_{MA})\right) \tag{8}$$

In the dataset prepared within the scope of this study, suitable slices of the CT abdominal scans obtained from the patients, in which the appendix region is visible, were selected in the presence of expert physicians. The effectiveness of the proposed method was evaluated by means of various analyses, using the results of the experimental studies carried out on these slices for appendix segmentation. From a total of 940 CT images selected from appendicitis and normal patients, 752 (~80%) were used for the training phase of the MaskAppendix model and the remaining 188 (~20%) were used for the test phase. The training set contains 378 normal appendix images and 374 appendicitis images, while 103 of the test set are normal appendix images and the remaining 85 are appendicitis images. In addition, for the MaskAppendix model proposed in this study, a training phase was performed separately for ResNet50 and ResNet101 backbone networks on the training set. Time and performance evaluations of the training phase were performed. Figure 5a shows the total loss values obtained according to the number of iterations for each backbone in the training phase using ResNet50 and ResNet101 networks as the backbone in the Mask R-CNN architecture. Although the loss values obtained for the training phase of the Mask R-CNN network for both backbones were quite close to each other, a slightly lower loss value was achieved for the Mask R-CNN with

ResNet101. On the other hand, in the test phase of the proposed MaskAppendix model, measurements were made separately for the check point in each period using the weights obtained in training. Figure 5b shows the progression of the DSC key metric as a function of the number of iterations for Mask R-CNN with ResNet50 and Mask R-CNN with ResNet101 networks on the test set. For DSC, it can be said that the segmentation results for both networks are quite similar, and again a slightly higher DSC is obtained for Mask R-CNN with ResNet101. Here, each iteration number represents a checkpoint period.

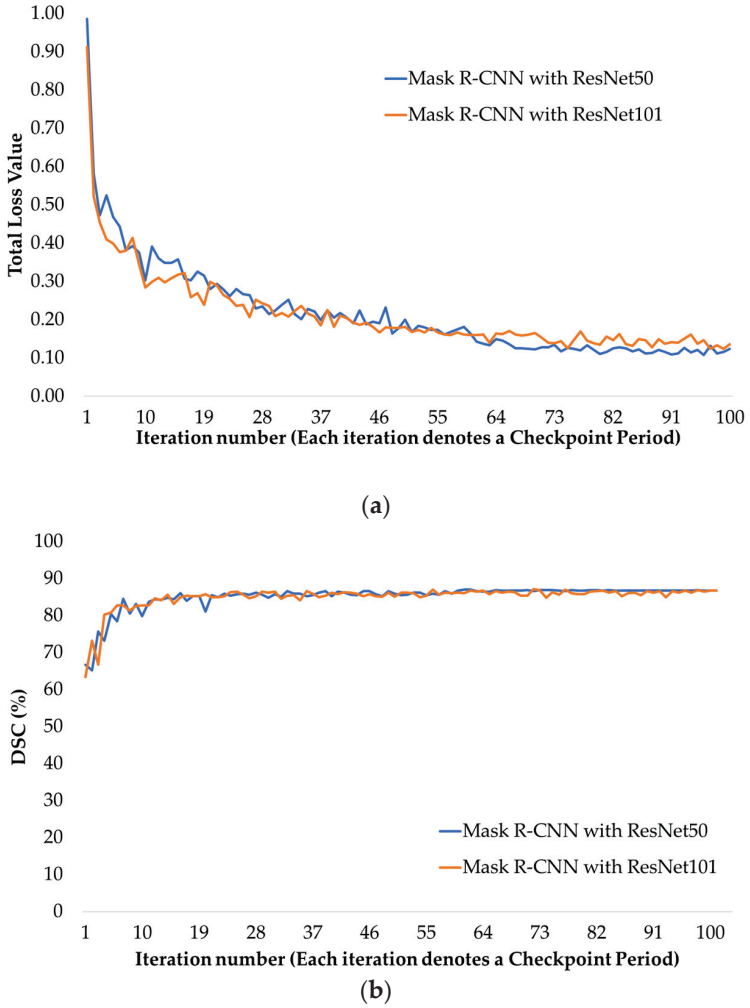


Figure 5. (**a**) Progression of total loss values during training using ResNet50 and ResNet101 backbone in Mask R-CNN network, (**b**) DSC progression using Mask R-CNN with ResNet50 and Mask R-CNN with ResNet101 on the test set.

3.1. Results

In the experimental studies conducted with the proposed MaskAppendix model, automatic segmentation of the appendix from CT images was successfully performed, as shown in Figure 6. Although the appendix covers a small area on the abdominal CT, its pattern is similar to other tissues, and its boundaries are not clear in some CT slices, the proposed MaskAppendix model successfully masked the appendix. The ground truth delineated by the expert raters for the appendix in an original abdominal CT scan in the dataset in Figure 6a is shown in Figure 6b, and the ground truth only mask showing the appendix extracted from the whole CT slice is shown in Figure 6c. Figure 6d shows the automatic segmentation of the appendix in the CT image using the proposed MaskAppendix, while Figure 6e shows the zoomed segmented appendix region with the overlap ratio and the achieved DSC score. It can be seen that the segmentation results of the expert physician

proposed MaskAppendix model with the delineated ground truth masks have very high overlap rates and the segmentation is successful with high DSC achieved.

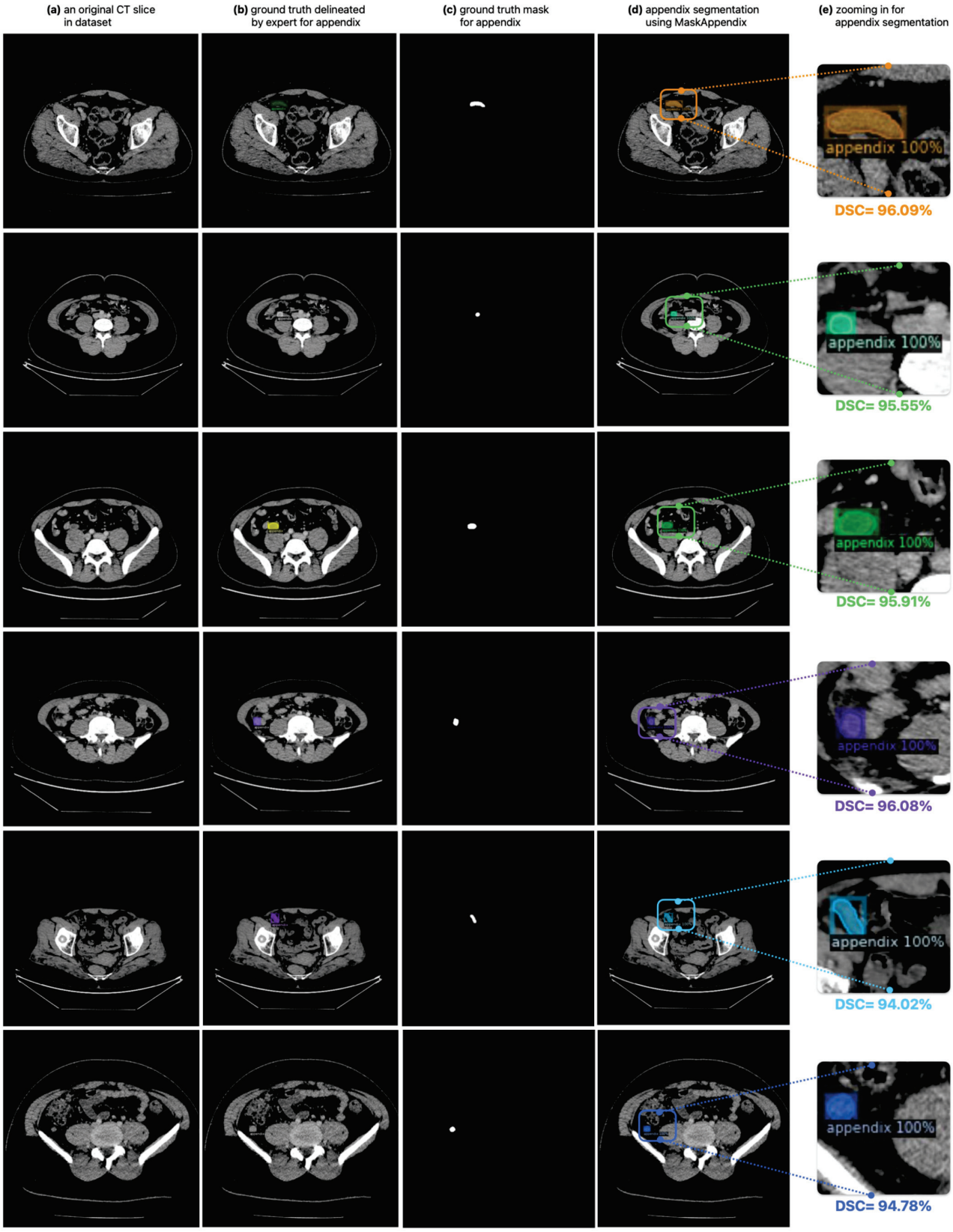


Figure 6. The successful automatic segmentation of the appendix in CT slices using the proposed MaskAppendix model.

In this study, the average results of the CT scans in the test set for the proposed MaskAppendix model on the dataset originally prepared for appendix segmentation are given in Table 4, based on the DSC, JSI, VOE, ASD, and HD95 key metrics. In addition, the segmentation results obtained with MaskAppendix for appendix segmentation are compared with the segmentation results obtained with the state-of-the-art methods U-Net [56] and DenseNet [57]. While U-Net is a deep learning architecture based on encoders and decoders widely used in medical image segmentation, DenseNet is a method based on connecting each layer in the network to the outputs of all previous layers. As can be seen from this table, the results of the DSC, JSI, VOE, ASD, and HD95 metrics used to measure segmentation performance in the appendix show that the proposed MaskAppendix model is more successful. In terms of the DSC key metric, 79.88%, 85.94%, and 87.17% were obtained for DenseNet, U-Net, and MaskAppendix (Mask R-CNN with ResNet101), respectively. Moreover, the proposed method is more successful for JSI, another key metric used for overlap accuracy, and VOE, a metric used for overlap error measurement. On the other hand, for ASD and HD95, key metrics calculated according to the smallest distance between two clusters, results with smaller distances than U-Net and DenseNet architectures were obtained using Mask R-CNN with ResNet101, 0.47 mm and 3.70 mm, respectively. The MaskAppendix model is more successful than U-Net and DenseNet in appendix segmentation due to its ability to perform instance segmentation with Mask R-CNN, enhanced feature localization with Grad-CAM, and deeper feature extraction with ResNet101. It is particularly well-suited to the small, variable and complex nature of the appendix due to its region proposal network, anchor boxes, and two-stage segmentation process, resulting in more accurate and reliable segmentation results.

Table 4. Average results achieved on CT scans in the test set for MaskAppendix and other state-of-the-art methods based on DSC, JSI, VOE, ASD, and HD95 key metrics.

Method	DSC [%]	JSI [%]	VOE [%]	ASD [mm]	HD95 [mm]
DenseNet	79.88	70.61	29.39	1.67	6.89
U-Net	85.94	76.70	23.29	1.24	5.43
MaskAppendix	87.17	78.11	21.89	0.47	3.70

3.2. Ablation Study

In this study, the results of the ablation experiments were also evaluated to verify the effect of the ResNet50 and ResNet101 modules on the segmentation performance of the MaskAppendix architecture, proposed as the backbone for the segmentation of the appendix region on CT slices. As shown in Table 5, the results obtained by optimizing the same network hyperparameters using ResNet50 and ResNet101 separately in the Mask-R-CNN architecture are compared for the key metrics DSC, JSI, VOE, ASD, and HD95. When ResNet101 is used as the backbone in the Mask R-CNN architecture (MaskAppendix), the appendix segmentation results are slightly better than Mask R-CNN with ResNet50. For the DSC and JSI metrics used for overlap accuracy, Mask R-CNN with Res-Net101 showed higher accuracy with values of 87.17% and 78.11%, respectively. In addition, for the VOE metric, which indicates the overlap error, Mask R-CNN with ResNet101 (MaskAppendix) has a lower error than Mask R-CNN with ResNet50. On the other hand, it can be said that Mask R-CNN with ResNet101 is more successful with a lower distance of 0.47 mm for the ASD metric used to evaluate the accuracy in the Appendix segmentation results, and Mask R-CNN with ResNet50 is more successful with 2.68 mm for the HD95 metric.

Table 5. Comparison of ablation results for ResNet50 and ResNet101 backbone networks in Mask R-CNN deep learning architecture for appendix segmentation from CT scans.

Method	Backbone	DSC [%]	JSI [%]	VOE [%]	ASD [mm]	HD95 [mm]
Mask R-CNN	ResNet50	87.07	77.79	22.21	0.48	2.68
Mask R-CNN	ResNet101	87.17	78.11	21.88	0.47	3.70

The boxplot evaluation of the results of the ablation experiments performed to verify the effect of the ResNet50 and ResNet101 modules, proposed as the backbone in the mask R-CNN architecture, on the segmentation performance is presented in Figure 7. Figure 7a shows the box plots verifying the segmentation results of the DSC metric, Figure 7b the JSI metric, Figure 7c the VOE metric, Figure 7d the ASD metric, and Figure 7e the HD95 metric for the Mask R-CNN with ResNet50 and Mask R-CNN with ResNet101 models. It can be seen that the results of the boxplot averages for Mask R-CNN with ResNet50 and Mask R-CNN with ResNet101 for each metric are close to each other.

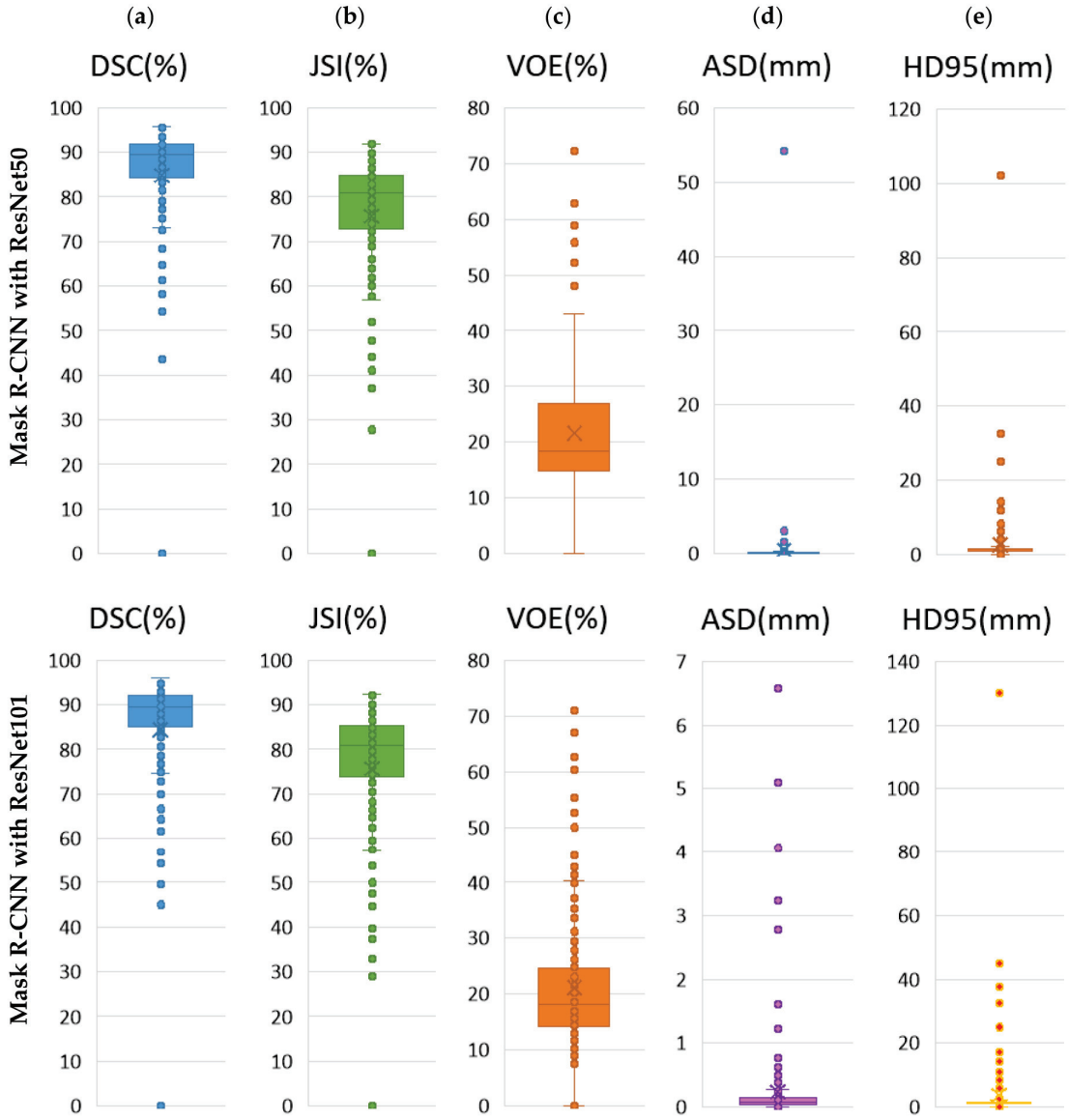


Figure 7. Boxplot evaluation of the results of ablation experiments carried out to verify the effect of ResNet50 and ResNet101 modules proposed as backbone in the Mask R-CNN architecture on segmentation performance. Boxplot for (**a**) DSC, (**b**) JSI, (**c**) VOE, (**d**) ASD, and (**e**) HD95.

4. Discussion

To the best of our knowledge, the studies proposed for appendix segmentation on CT scans are very limited, if not non-existent. It seems that the studies proposed for appendix detection are mostly proposed for the classification of normal appendix and appendicitis [36,37,40,41]. Compared to the segmentations obtained with the state-of-the-art U-Net and DenseNet methods, in our study we achieved a segmentation performance for the appendix with 87.17% DSC using the proposed MaskAppendix deep learning model. On the other hand, in our previous study [24], the appendix was segmented with a DSC score of 85.94% using the U-Net architecture, while in this study a DSC score of 87.17% was achieved using MaskAppendix. Comparing the appendix segmentation results in this study with our previous study, we can see that the results for appendix segmentation are more successful.

Evaluating the results of the experimental studies conducted for the automatic segmentation of the appendix region in the appendix abdomen CT slice dataset created for this study, it can be clearly stated that the segmentation results of the MaskAppendix model (Mask R-CNN with ResNet101) are more successful than the segmentation results obtained using state-of-the-art U-Net and DenseNet architectures. Although it has been verified that the proposed method achieves high segmentation scores with key metric measurements on the slices in the dataset, some appendix regions segmented with lower performance by the proposed MaskAppendix model are shown in Figure 8. Although the segmentation results on the images are reasonable, there may be many different reasons for this and similar problems. The low segmentation results for the appendix may be due to the fact that in some CT slices the appendix region is close to or connected to other areas, the patterns are similar, and the boundaries of the region are similar to those of the neighboring areas. On the other hand, the presence of noise in the relevant images, the interweaving of the textural colors of the appendix region and other organs, the fact that the appendix has more than one part, errors made by the expert in marking, and the number of iterations in the training of the MaskAppendix network, which was insufficient to detect the appendix regions in these images, could also have reduced the segmentation success.

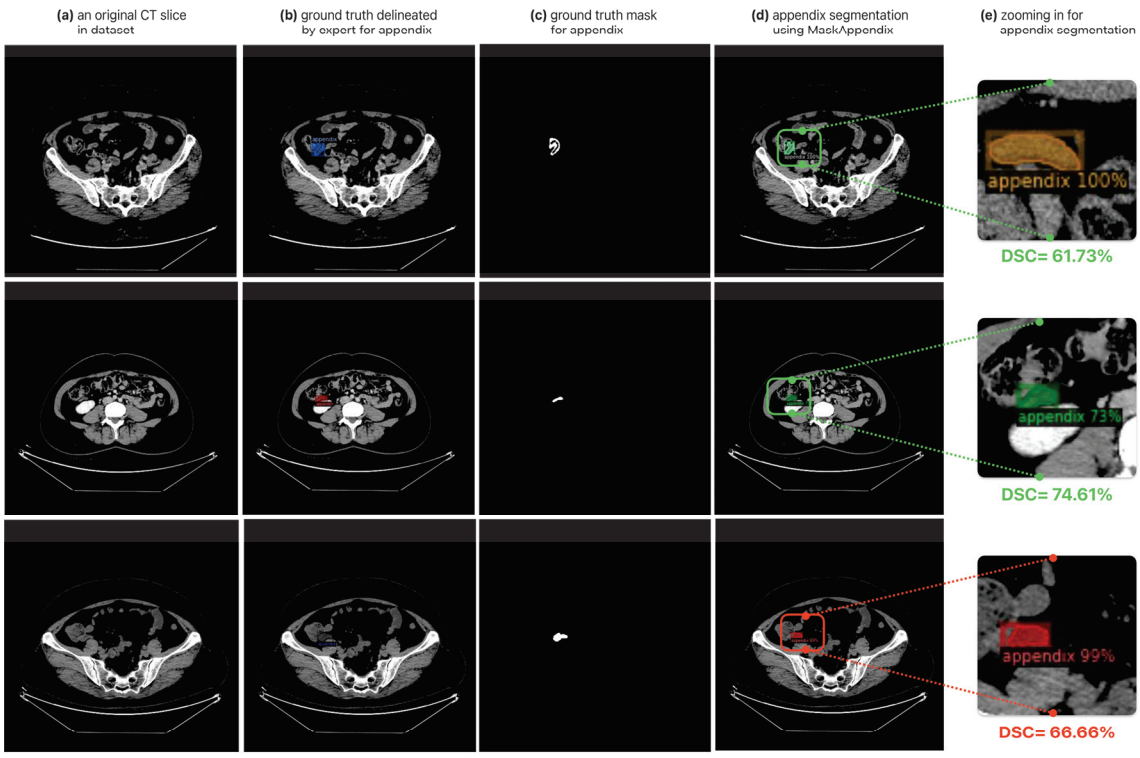


Figure 8. Appendix regions segmented with lower performance using the proposed MaskAppendix model in some CT slices.

The combination of Mask R-CNN and Grad-CAM improves segmentation accuracy by providing complementary strengths in object detection, feature localization and explainability. Grad-CAM provides a mechanism to highlight the key regions of the image that contribute to predicting a particular class [58]. In the context of appendix segmentation, the combination of Mask R-CNN and Grad-CAM helps Mask R-CNN to focus on the most relevant parts of the image where the appendix is probably located. Rather than being distracted by surrounding anatomical structures, Grad-CAM prompts the model to increase its attention to regions that contribute most to the correct identification of the appendix. In this study, Gradient-weighted Class Activation Mapping (Grad-CAM), a discriminative localization technique, was also used to evaluate whether the proposed MaskAppendix model (Mask R-CNN with ResNet101 backbone) correctly segmented the location of the appendix. Grad-CAM can be applied to any CNN model on images and can produce a more transparent heatmap by generating visual descriptions [59]. In this study, the feature map output from the final convolutional layer of the MaskAppendix model was used to generate Grad-CAM on an input CT slice. Figure 9a shows the expertly delineated ground truth mask of the appendix and its position on an original CT slice is shown in Figure 9b. On the other hand, the automatic segmentation of the appendix in this CT slice with the proposed MaskAppendix model is shown in Figure 9c. Figure 9d shows the Grad-CAM evaluation of the correct segmentation and position of the appendix. It can be seen that the MaskAppendix segmentation result and the visual results of the heatmap features output with Grad-CAM are compatible with each other. By integrating Grad-CAM into the Mask R-CNN framework, the proposed model gained the ability to better understand and highlight the fine details that distinguish the appendix from neighboring anatomical structures. As a result, clinical outcomes can be improved using Grad-CAM as the appendix can be more clearly visualized and diseases such as appendicitis can be diagnosed more quickly and precisely.

Table 6 also evaluates the total runtime of the Mask R-CNN with ResNet50 and Mask R-CNN with ResNet101 (MaskAppendix) models for training and test sets. There is a significant difference in depth between ResNet50 and ResNet101 backbone networks. This is expected to be reflected in the feature training times of the networks. Total time for Mask R-CNN with ResNet50 for training and the time for one Checkpoint Period were 528.6 and 5.29 min, respectively. On the other hand, since the backbone network used for MaskAppendix is deeper, the same times were obtained as 723.6 and 7.24 min, respectively. The test time was calculated as 29.3 s for Mask R-CNN with ResNet50 and 36.5 s for MaskAppendix. As a result, it can be seen that the training and test times obtained according to the depth of the networks used as backbone in the infrastructure of the Mask R-CNN architecture are reasonable and compatible.

Table 6. Total runtimes of the Mask R-CNN with ResNet50 and Mask R-CNN with ResNet101 (MaskAppendix) models for training and test sets.

Method	Backbone	Total Time for Training (min)	The Time for a Checkpoint Period (min)	Total Time for Test (s)
Mask R-CNN	ResNet50	528.6	5.29	29.3
Mask R-CNN	ResNet101	723.6	7.24	36.5

In this study, for appendix segmentation, the ground truth masks for the appendix provided by expert physicians in the dataset created with CT scans of healthy individuals and appendicitis patients were important for the generalization of the proposed model. Nevertheless, there are several limitations to the proposed MaskAppendix model, although it shows promising results in appendix segmentation. A larger, more diverse dataset including different imaging conditions would be required to fully assess the robustness and generalizability of the model. In addition, the model has yet to undergo extensive clinical validation, despite performing well in a controlled experimental setting. The impact of

the model in real-time clinical settings, including its integration into radiology workflows, remains to be investigated, as does how radiologists interact with its output. In addition, widespread adoption of the proposed model may be challenged by the computational cost of training and deploying the model in real-world environments.

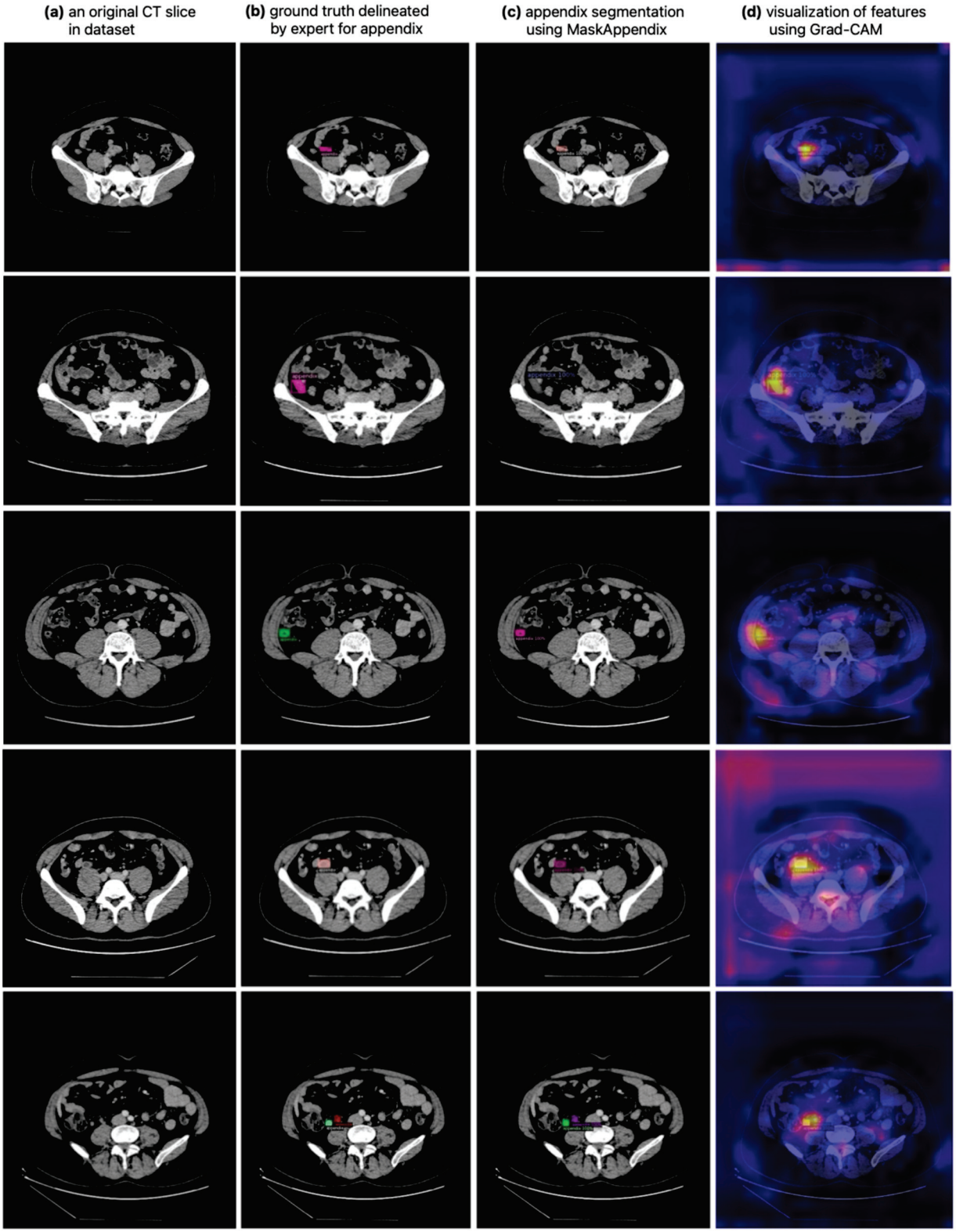


Figure 9. Evaluation of visual results of heatmap feature outputs with Grad-CAM with MaskAppendix segmentation result.

5. Conclusions

The contribution of computer-aided tools to assist physicians in the diagnosis of disease from medical images has increased in recent years. It is known that segmentation of these images by manual selection of the RoI by physicians is prone to human failure, requires more time, and consists of too many procedural dependencies. Therefore, automated segmentation of the appendix from CT images can make a significant contribution in terms of both time and cost. This study presents the potential of the proposed MaskAppendix deep learning model on the Detectron2 platform for accurate appendix segmentation from CT images. In the automatic segmentation of the appendix region in CT slices, the results of 87.17% DSC, 78.11% JSI, 21.88% VOE, 0.47 mm ASD, and 3.70 mm HD95 obtained with the proposed approach are more successful than the state-of-the-art U-Net and DenseNet deep learning architectures. Therefore, the obtained results have the potential to improve the accuracy of clinical diagnosis. On the other hand, in the proposed MaskAppendix model, two different networks, ResNet50 and ResNet101, were used as the backbone and their effects on appendix segmentation from CT slices were evaluated by ablation experiments. The ablation studies confirm that higher segmentation performance can be achieved when ResNet101 is used as the backbone in the Mask R-CNN architecture. In addition, the verification of the correct segmentation of the appendix location in the proposed MaskAppendix model was evaluated using a heatmap generated by generating visual annotations on the images using Grad-CAM.

The proposed method can have a significant improvement in the diagnosis of appendicitis and patient care using the MaskAppendix model, which integrates Mask R-CNN with Grad-CAM for automated appendix segmentation. First of all, the precise segmentation provided by the MaskAppendix model can give surgeons a clear understanding of the location, size, and surrounding structures of the appendix. Therefore, the proposed MaskAppendix model can overcome the challenges of accurately and automatically segmenting the appendix on CT scans, allowing radiologists to visualize the appendix with greater clarity. In addition, the method can be particularly valuable in cases where the appendix is not easily distinguishable, such as in patients with abnormal anatomy or conditions. By using Grad-CAM to improve feature localization, the model can identify subtle of appendicitis, improving diagnostic confidence. Moreover, clinicians can provide personalized care based on each patient's specific anatomy and condition by automating and improving the accuracy of appendix segmentation.

In future studies, the dataset can be expanded by increasing the patient population with CT scans obtained from different resolution, size, and scanner types. In addition, it is planned to improve the existing model and increase the segmentation performance by developing hybrid deep learning architectures and automatic hyperparameter optimization.

Author Contributions: Conceptualization: E.D.; methodology: B.T.B., E.D. and M.S.Y. software: E.D. and M.S.Y.; validation: E.D., B.T.B., K.Ç. and G.G.; formal analysis: E.D., B.T.B., M.S.Y., K.Ç. and G.G.; investigation: E.D., B.T.B., M.S.Y., K.Ç. and G.G.; data curation: B.T.B., K.Ç. and G.G.; writing—original draft preparation: E.D. writing—review and editing: B.T.B., K.Ç. and E.D., visualization: E.D. and M.S.Y.; supervision: E.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: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board According to Decision No. 13 (protocol code) of the 8th meeting of the Ethics Committee of Non-Interventional Clinical Research of Bilecik Şeyh Edebali University, dated 5 December 2023.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: We would like to thank Bilecik Training and Research Hospital for providing the dataset and all the patients who volunteered to participate. All individuals have agreed to this consent.

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

References

1. Schumpelick, V.; Dreuw, B.; Ophoff, K.; Prescher, A. Appendix and cecum: Embryology, anatomy, and surgical applications. *Surg. Clin.* **2000**, *80*, 295–318.
2. Kooij, I.; Sahami, S.; Meijer, S.; Buskens, C.; Te Velde, A. The immunology of the vermiform appendix: A review of the literature. *Clin. Exp. Immunol.* **2016**, *186*, 1–9. [CrossRef] [PubMed]
3. Bollinger, R.R.; Barbas, A.S.; Bush, E.L.; Lin, S.S.; Parker, W. Biofilms in the large bowel suggest an apparent function of the human vermiform appendix. *J. Theor. Biol.* **2007**, *249*, 826–831. [CrossRef] [PubMed]
4. D'souza, N.; Nugent, K. Appendicitis. *Am. Fam. Physician* **2016**, *93*, 142–143. [PubMed]
5. Zachos, K.; Kolonitsiou, F.; Panagidis, A.; Gkentzi, D.; Fouzas, S.; Alexopoulos, V.; Kostopoulou, E.; Roupakias, S.; Vervenioti, A.; Dassios, T. Association of the bacteria of the vermiform appendix and the peritoneal cavity with complicated acute appendicitis in children. *Diagnostics* **2023**, *13*, 1839. [CrossRef] [PubMed]
6. Lee, M.S.; Purcell, R.; McCombie, A.; Frizelle, F.; Eglinton, T. Retrospective cohort study of the impact of faecoliths on the natural history of acute appendicitis. *World J. Emerg. Surg.* **2023**, *18*, 18. [CrossRef]
7. Bhangu, A.; Søreide, K.; Di Saverio, S.; Assarsson, J.H.; Drake, F.T. Acute appendicitis: Modern understanding of pathogenesis, diagnosis, and management. *Lancet* **2015**, *386*, 1278–1287. [CrossRef]
8. Dimberg, J.; Rubér, M.; Skarstedt, M.; Andersson, M.; Andersson, R.E. Genetic polymorphism patterns suggest a genetic driven inflammatory response as pathogenesis in appendicitis. *Int. J. Color. Dis.* **2020**, *35*, 277–284. [CrossRef]
9. Lamps, L.W. Beyond acute inflammation: A review of appendicitis and infections of the appendix. *Diagn. Histopathol.* **2008**, *14*, 68–77. [CrossRef]
10. Rogers, M.B.; Brower-Sinning, R.; Firek, B.; Zhong, D.; Morowitz, M.J. Acute appendicitis in children is associated with a local expansion of fusobacteria. *Clin. Infect. Dis.* **2016**, *63*, 71–78. [CrossRef]
11. Wagner, M.; Tubre, D.J.; Asensio, J.A. Evolution and current trends in the management of acute appendicitis. *Surg. Clin.* **2018**, *98*, 1005–1023. [CrossRef] [PubMed]
12. Dinç, T.; Sapmaz, A.; Erkuş, Y.; Yavuz, Z. Complicated or non-complicated appendicitis? That is the question. *Turk. J. Trauma Emerg. Surg.* **2022**, *28*, 390. [CrossRef]
13. Li, Z.; Li, Z.; Zhao, L.; Cheng, Y.; Cheng, N.; Deng, Y. Abdominal drainage to prevent intra-peritoneal abscess after appendectomy for complicated appendicitis. *Cochrane Database Syst. Rev.* **2021**, *8*, CD010168. [CrossRef] [PubMed]
14. Shiihara, M.; Sudo, Y.; Matsushita, N.; Kubota, T.; Hibi, Y.; Osugi, H.; Inoue, T. Therapeutic strategy for acute appendicitis based on laparoscopic surgery. *BMC Surg.* **2023**, *23*, 161. [CrossRef] [PubMed]
15. Zagales, I.; Sauder, M.; Selvakumar, S.; Spardy, J.; Santos, R.G.; Cruz, J.; Bilski, T.; Elkbuli, A. Comparing outcomes of appendectomy versus non-operative antibiotic therapy for acute appendicitis: A systematic review and meta-analysis of randomized clinical trials. *Am. Surg.* **2023**, *89*, 2644–2655. [CrossRef]
16. Cho, S.U.; Oh, S.K. Accuracy of ultrasound for the diagnosis of acute appendicitis in the emergency department: A systematic review. *Medicine* **2023**, *102*, e33397. [CrossRef]
17. Stöß, C.; Nitsche, U.; Neumann, P.-A.; Kehl, V.; Wilhelm, D.; Busse, R.; Friess, H.; Nimptsch, U. Acute Appendicitis: Trends in Surgical Treatment: A Population-Based Study of Over 800,000 Patients. *Dtsch. Aerzteblatt Int.* **2021**, *118*, 244.
18. Caballero-Alvarado, J.; Torres, V.L.; Peralta, K.L.; Corvera, C.Z. Powikłane ostre zapalenie wyrostka robaczkowego z zajęciem podstawy wyrostka robaczkowego: Przegląd strategii leczenia operacyjnego. *Pol. Przegląd Chir.* **2023**, *96*, 65–70. [CrossRef]
19. Clanton, J.; Subichin, M.; Drolshagen, K.; Daley, T.; Firstenberg, M.S. Fulminant Clostridium difficile infection: An association with prior appendectomy? *World J. Gastrointest. Surg.* **2013**, *5*, 233. [CrossRef]
20. Sippola, S.; Virtanen, J.; Tammilehto, V.; Grönroos, J.; Hurme, S.; Niiniviita, H.; Lietzen, E.; Salminen, P. The accuracy of low-dose computed tomography protocol in patients with suspected acute appendicitis: The OPTICAP study. *Ann. Surg.* **2020**, *271*, 332–338. [CrossRef]
21. Song, H.; Lee, S.; Park, J.H.; Kim, H.Y.; Min, H.D.; Jeon, J.-J.; Lee, K.H.; Group, L. Can patient triaging with clinical scoring systems reduce CT use in adolescents and young adults suspected of having appendicitis? *Radiology* **2021**, *300*, 350–358. [CrossRef] [PubMed]
22. Soundappan, S.S.; Karpelowsky, J.; Lam, A.; Lam, L.; Cass, D. Diagnostic accuracy of surgeon performed ultrasound (SPU) for appendicitis in children. *J. Pediatr. Surg.* **2018**, *53*, 2023–2027. [CrossRef]
23. Wang, D.C.; Parry, C.R.; Feldman, M.; Tomlinson, G.; Sarrazin, J.; Glanc, P. Acute abdomen in the emergency department: Is CT a time-limiting factor? *Am. J. Roentgenol.* **2015**, *205*, 1222–1229. [CrossRef] [PubMed]
24. Başţuğ, B.T.; Güneri, G.; Yıldırım, M.S.; Çorbacı, K.; Dandıl, E. Fully Automated Detection of the Appendix Using U-Net Deep Learning Architecture in CT Scans. *J. Clin. Med.* **2024**, *13*, 5893. [CrossRef]

25. Zacharzewska-Gondek, A.; Szczurowska, A.; Guziński, M.; Sąsiadek, M.; Bladowska, J. A pictorial essay of the most atypical variants of the vermiform appendix position in computed tomography with their possible clinical implications. *Pol. J. Radiol.* **2019**, *84*, e1. [CrossRef]
26. Weinberger, H.; Zeina, A.-R.; Ashkenazi, I. Misdiagnosis of acute appendicitis in the emergency department: Prevalence, associated factors, and outcomes according to the patients' disposition. *Ochsner J.* **2023**, *23*, 271–276. [CrossRef] [PubMed]
27. Marcinkevics, R.; Reis Wolfertstetter, P.; Wellmann, S.; Knorr, C.; Vogt, J.E. Using machine learning to predict the diagnosis, management and severity of pediatric appendicitis. *Front. Pediatr.* **2021**, *9*, 662183. [CrossRef]
28. Byun, J.; Park, S.; Hwang, S.M. Diagnostic Algorithm Based on Machine Learning to Predict Complicated Appendicitis in Children Using CT, Laboratory, and Clinical Features. *Diagnostics* **2023**, *13*, 923. [CrossRef]
29. Obuchowicz, R.; Strzelecki, M.; Piórkowski, A. Clinical Applications of Artificial Intelligence in Medical Imaging and Image Processing—A Review. *Cancers* **2024**, *16*, 1870. [CrossRef]
30. Abbaoui, W.; Retal, S.; El Bhiri, B.; Kharmoum, N.; Ziti, S. Towards revolutionizing precision healthcare: A systematic literature review of artificial intelligence methods in precision medicine. *Inform. Med. Unlocked* **2024**, *46*, 101475. [CrossRef]
31. Sharma, A.; Lysenko, A.; Jia, S.; Boroevich, K.A.; Tsunoda, T. Advances in AI and machine learning for predictive medicine. *J. Hum. Genet.* **2024**, *69*, 487–497. [CrossRef] [PubMed]
32. Roy, S.; Pal, D.; Meena, T. Explainable artificial intelligence to increase transparency for revolutionizing healthcare ecosystem and the road ahead. *Netw. Model. Anal. Health Inform. Bioinform.* **2023**, *13*, 4. [CrossRef]
33. Rayed, M.E.; Islam, S.S.; Niha, S.I.; Jim, J.R.; Kabir, M.M.; Mridha, M. Deep learning for medical image segmentation: State-of-the-art advancements and challenges. *Inform. Med. Unlocked* **2024**, *47*, 101504. [CrossRef]
34. Archana, R.; Jeevaraj, P.E. Deep learning models for digital image processing: A review. *Artif. Intell. Rev.* **2024**, *57*, 11. [CrossRef]
35. Liu, X.; Song, L.; Liu, S.; Zhang, Y. A review of deep-learning-based medical image segmentation methods. *Sustainability* **2021**, *13*, 1224. [CrossRef]
36. Park, S.H.; Kim, Y.J.; Kim, K.G.; Chung, J.-W.; Kim, H.C.; Choi, I.Y.; You, M.-W.; Lee, G.P.; Hwang, J.H. Comparison between single and serial computed tomography images in classification of acute appendicitis, acute right-sided diverticulitis, and normal appendix using EfficientNet. *PLoS ONE* **2023**, *18*, e0281498. [CrossRef]
37. Al, W.A.; Yun, I.D.; Lee, K.J. Reinforcement learning-based automatic diagnosis of acute appendicitis in abdominal ct. *arXiv* **2019**, arXiv:1909.00617.
38. Lee, G.P.; Park, S.H.; Kim, Y.J.; Chung, J.-W.; Kim, K.G. Enhancing disease classification in abdominal CT scans through RGB superposition methods and 2D convolutional neural networks: A study of appendicitis and diverticulitis. *Comput. Math. Methods Med.* **2023**, *2023*, 7714483. [CrossRef] [PubMed]
39. Liang, D.; Fan, Y.; Zeng, Y.; Zhou, H.; Zhou, H.; Li, G.; Liang, Y.; Zhong, Z.; Chen, D.; Chen, A. Development and validation of a deep learning and radiomics combined model for differentiating complicated from uncomplicated acute appendicitis. *Acad. Radiol.* **2024**, *31*, 1344–1354. [CrossRef]
40. Park, J.J.; Kim, K.A.; Nam, Y.; Choi, M.H.; Choi, S.Y.; Rhie, J. Convolutional-neural-network-based diagnosis of appendicitis via CT scans in patients with acute abdominal pain presenting in the emergency department. *Sci. Rep.* **2020**, *10*, 9556. [CrossRef]
41. Rajpurkar, P.; Park, A.; Irvin, J.; Chute, C.; Bereket, M.; Mastrodicasa, D.; Langlotz, C.P.; Lungren, M.P.; Ng, A.Y.; Patel, B.N. AppendiXNet: Deep learning for diagnosis of appendicitis from a small dataset of CT exams using video pretraining. *Sci. Rep.* **2020**, *10*, 3958. [CrossRef] [PubMed]
42. Yushkevich, P.A.; Piven, J.; Hazlett, H.C.; Smith, R.G.; Ho, S.; Gee, J.C.; Gerig, G. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage* **2006**, *31*, 1116–1128. [CrossRef]
43. ITKSNAP. Available online: www.itksnap.org/ (accessed on 15 December 2023).
44. Jia, W.; Tian, Y.; Luo, R.; Zhang, Z.; Lian, J.; Zheng, Y. Detection and segmentation of overlapped fruits based on optimized mask R-CNN application in apple harvesting robot. *Comput. Electron. Agric.* **2020**, *172*, 105380. [CrossRef]
45. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
46. Yu, Y.; Zhang, K.; Yang, L.; Zhang, D. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Comput. Electron. Agric.* **2019**, *163*, 104846. [CrossRef]
47. Dandıl, E.; Yıldırım, M.S.; Selvi, A.O.; Uzun, S. Automated liver segmentation using Mask R-CNN on computed tomography scans. *Gazi Üniversitesi Mühendislik Mimar. Fakültesi Derg.* **2021**, *37*, 29–46. [CrossRef]
48. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
49. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
50. Zimmermann, R.S.; Siems, J.N. Faster training of Mask R-CNN by focusing on instance boundaries. *Comput. Vis. Image Underst.* **2019**, *188*, 102795. [CrossRef]
51. 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.
52. Targ, S.; Almeida, D.; Lyman, K. Resnet in resnet: Generalizing residual architectures. *arXiv* **2016**, arXiv:1603.08029.
53. Detectron2. Available online: https://github.com/facebookresearch/detectron2 (accessed on 11 August 2024).

54. Dice, L.R. Measures of the amount of ecologic association between species. *Ecology* **1945**, *26*, 297–302. [CrossRef]
55. Yuan, Y.; Chao, M.; Lo, Y.-C. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. *IEEE Trans. Med. Imaging* **2017**, *36*, 1876–1886. [CrossRef] [PubMed]
56. 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. pp. 234–241.
57. Iandola, F.; Moskewicz, M.; Karayev, S.; Girshick, R.; Darrell, T.; Keutzer, K. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv* **2014**, arXiv:1404.1869.
58. Tang, D.; Chen, J.; Ren, L.; Wang, X.; Li, D.; Zhang, H. Reviewing CAM-Based Deep Explainable Methods in Healthcare. *Appl. Sci.* **2024**, *14*, 4124. [CrossRef]
59. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [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.

Journal of
Clinical Medicine

MDPI

Article

Monitoring Progression in Hypertensive Patients with Dyslipidemia Using Optical Coherence Tomography Angiography: Can A.I. Be Improved?

Irina Cristina Barca [1,*], Vasile Potop [1] and Stefan Sorin Arama [2]

[1] Ophthalmology Department, "Carol Davila" University of Medicine and Pharmacy, 050474 Bucharest, Romania; vasile.potop@umfcd.ro
[2] Physio-Pathology and Immunology Department, "Carol Davila" University of Medicine and Pharmacy, 050474 Bucharest, Romania; sorin.arama@umfcd.ro
* Correspondence: irina-cristina.barca@drd.umfcd.ro; Tel.: +40-745910290

Abstract: Background: With the development of artificial intelligence (A.I.), the optical coherence tomography angiography (OCTA) analysis of progression in hypertensive retinopathy could be improved. Our purpose was to use the OCTA to study the effect of uncontrolled dyslipidemia and hypertensive retinopathy on the retinal microvasculature and to identify a potential software update of the A.I. secondary to the OCTA analysis. By using our most relevant data, the A.I. software can be upgraded by introducing new mathematic formulas between the OCTA parameters and the lipid level. **Methods**: We performed a prospective cohort study on 154 eyes of participants from Eastern Europe. We used a standardized protocol to collect data on past medical history of dyslipidemia and hypertension and OCTA to measure retinal vascular parameters. **Results**: The average age of the participants was 56.9 ± 9.1, with a minimum of 34 and a maximum of 82 and with a higher percentage of males: 55.8%. Statistically significant correlations were found for total cholesterol and skeleton total (r = −0.249; p = 0.029), foveal avascular zone (FAZ), circularity and low-density lipoprotein (LDL) (r = 0.313; p = 0.006), non-flow area (NFA) and LDL (r = 0.233; p = 0.042), and vascular flow area (VFA) and LDL (r = −0.354; p = 0.002). **Conclusions**: Subjects with dyslipidemia and progressive hypertensive retinopathy had a reduction in microvascular density and vascular flow, a focal capillary non-perfusion, and an increased FAZ. Thus, by improving the A.I. system, our research aims to provide better OCTA monitoring, which could help in the early-stage detection of progression and development of A.I. screening programs, leading to increased efficiency in diagnosing patients.

Keywords: optical coherence tomography angiography; A.I.; capillary non-perfusion; dyslipidemia; hypertensive retinopathy

1. Introduction

According to reports from the World Health Organization (WHO), in March 2023, an estimated 1.28 billion adults aged 30–79 years worldwide had hypertension, with most (two-thirds) living in low- and middle-income countries. An estimated 46% of adults with hypertension are unaware that they have the condition, less than half of adults (42%) with hypertension are diagnosed and treated, and approximately one in five adults (21%) have it under control. Hypertension is a major cause of premature death worldwide and is a major risk factor for stroke, myocardial infarction, vascular disease, and chronic kidney disease. Dyslipidemia results in abnormal levels of lipids in the blood that can also increase the risk of cardiovascular diseases [1]. Dyslipidemia is classified into two types: primary and secondary. Primary dyslipidemia is inherited and caused by genetic mutations that affect lipid metabolism, while secondary dyslipidemia is acquired and caused by lifestyle factors or other medical conditions that alter lipid levels. Both types

are associated with hypertension. The management of dyslipidemias is an essential and integral part of cardiovascular disease prevention [2].

Chronic high blood pressure or previous episodes of high blood pressure can affect retinal microcirculatory structure and function [3].

There have been many previous studies on the association between hypertensive retinopathy and hypertension [4–10].

The prevalence rate of retinopathy in patients with mild hypertension is approximately 25%, with the moderate form of the disease is 34%, and with the severe form is 84.6%. Of the patients with hypertensive retinopathy, 42.3% have a grade I, 20% a grade II, and 2.35% a grade III ocular pathology [5].

Hypertensive retinopathy is related to blood pressure variability (BPV); the degree of BPV and retinopathy in hypertensive patients can be used to track the progression of hypertension-mediated organ damage (HOMD), and the correction of BPV may help treat or postpone the progression of HOMD. Hypertension is associated with reduced retinal vessel density and an increased foveal avascular zone (FAZ), especially in the deep venous plexus, as seen on optical coherence tomography angiography (OCTA) [11].

Patients with hypertensive retinopathy show a reduction in the retinal vascular density, superficial and deep vascular plexus (SVP and DVP), ganglion cell–internal boundary membrane complex density (GCL-ILM), and retinal nerve fiber layer (RNFL) thickness [12].

The classic retinal changes in chronic hypertensive retinopathy include arteriolar narrowing and tortuosity, arteriovenous crossing changes, retinal hemorrhages, cotton-wool spots, hard exudates, papilledema. The malignant stage of hypertensive retinopathy represents a medical emergency, often associated with markedly elevated blood pressure and with the risk of organ damage [8,13].

Compared to invasive fundus fluorescein angiography (FFA), OCTA is a new, rapid, easily performed, and non-invasive 3D imaging modality, allowing healthcare professionals to study the microvasculature of the retinal and choroidal layers in large populations of patients with dyslipidemia and hypertensive retinopathy [14]. The retina and choroid are two distinct vascular beds. Using 3D OCTA, distinct visualizations of these two circulations can be obtained. In healthy eyes, retinal circulation is located between the internal limiting membrane (ILM) and the outer plexiform layer (OPL). The choroidal circulation is located beneath Bruch's membrane (BM). The vitreous (anterior to the ILM) and outer retina (between OPL and Bruch's membrane) are avascular structures [14]. OCTA allows for quantitative assessment of the retinal microvasculature, providing metrics such as vessel density and vessel caliber. These measurements help in tracking hypertensive retinopathy progression in patients with dyslipidemia and assessing the efficacy of treatment interventions over time [9,15].

OCTA changes in hypertensive retinopathy include microvascular density reduction, focal capillary non-perfusion, increased FAZ, vascular tortuosity and irregularities, and choroidal thickness alterations [3,9,15–17].

The aim of our current study was to measure FAZ, non-flow area (NFA), vascular flow area (VFA), retinal vascular density, and perfusion by using OCTA in patients with chronic hypertension and secondary dyslipidemia, identifying the potential OCTA parameters to be analyzed using A.I. We hypothesize that OCTA will show subclinical changes in retinal microvasculature in the superficial and deep plexi, i.e., changes related to secondary dyslipidemia and progression of hypertensive retinopathy.

2. Materials and Methods

2.1. Study Participants

The Ethics Committee approved the study in December 2022. Patients were selected for inclusion in this study by the ophthalmologist and referred to the ophthalmologist by a cardiologist, general practitioner, or internist.

We performed a prospective cohort study including 77 participants with dyslipidemia and systemic hypertension from Eastern Europe, Romania. From 2022 to 2024, 154 eyes

were eligible for the final analysis presented in this study. All participants underwent ophthalmological examinations, blood testing, blood pressure monitoring, and body mass index (BMI) calculation according to protocol.

For the purpose of this study, we selected subjects diagnosed with chronic hypertension and secondary dyslipidemia (based on medical history and records) as well as undiagnosed cases by determining blood pressure, performing blood tests and fundoscopy, and by referring patients to a cardiologist and internist for an accurate diagnosis.

We selected hypertensive patients with consistently high values in all measurements, including patients using and not using antihypertensive treatments at the time of their first visit. Subjects with normalized blood pressure during follow-ups were not excluded from the study.

Patients with chronic hypertensive retinopathy were included in the study based on fundoscopic examination. We selected grade I–II of hypertensive retinopathy (mild) according to the Keith–Wagener–Barker or Mitchell–Wong classification system and no other types of retinopathy based on fundoscopy [5,18]. Patients with the following pathologies were excluded by using fundoscopy and by scanning the optic nerve and retina in both eyes using OCTA: hypertensive retinopathy grade III (moderate) and IV (severe) according to the Keith–Wagener–Barker or Mitchell–Wong classification system, glaucoma, diabetic retinopathy, maculopathy, central serous retinopathy, pigment epithelial detachments, epiretinal membranes, cystoid macula edema, vitreomacular traction, and other significant co-morbidities that could affect retinal vasculature.

In our study, we included participants with newly diagnosed dyslipidemia based on blood testing performed at the time of the first visit and diagnosed before referral by the internist or the cardiologist.

Systolic and diastolic blood pressure was measured twice per month. Blood pressure was taken with an automatic blood pressure monitor (Hartmann Tensoval Comfort, PAUL HARTMANN AG, 89522 Heidenheim, Germany). Three readings were taken 5 min apart, before exercise or at least 30 min after exercise. Hypertension was defined as present in participants with elevated blood pressure at the time of examination (SPB $\geq$ 130 mmHg or DBP $\geq$ 80 mmHg), according to American Heart Association criteria [2,16]. Undiagnosed cases were defined to be hypertensive if their systolic blood pressure was more than or equal to 140 mmHg or diastolic blood pressure more than or equal to 90 mmHg or if there was a self-reported history of self-measured blood pressure above the same values.

Height, body weight, and BMI were calculated every 3 months. Height was measured using a portable stadiometer (Seca Model 217; Seca, Hamburg, Germany). Body weight was determined using a calibrated mechanical column scale with eye-level beam (Seca 700; Seca, Hamburg, Germany). BMI was calculated by dividing body weight (kg) by the square of height (m^2). The age and sex were also noted. Recording of the smoking habit was also performed. Smoking was assessed with the single categorized question: "Are you a current smoker?".

Glycemia and HbA1c were determined in order to rule out subjects with diabetes and/or undiagnosed diabetic retinopathy. Glycemia, glycated hemoglobin (HbA1c), total cholesterol, HDL, LDL, and triglycerides were measured every 3 months using the same laboratory equipment. Non-fasting blood samples were used to assess plasma lipid profiles. Measurements were performed on all patients with or without statin use at the time of the enrollment. Patients with normalized plasma lipid values after receiving statin treatment during follow-ups were not excluded from the study.

Each participant underwent an ophthalmic examination including refraction measured with Auto Kerato-Refractometer (Topcon, Model KR-800, Tokyo, Japan) and determination of intraocular pressure with Goldmann Tonometer, fundoscopy, and OCTA. All refractive errors were included. All subjects had a normal intraocular pressure ranging from 10 to 21 mmHg. Examination of the retina using a Volk 90D and a 78D power lens (Volk, Mentor, OH, USA) was performed quarterly in order to stage hypertensive retinopathy and to include or exclude participants in our study.

OCTA was performed on both eyes, on all patients, every 3 months, with the assessment of FAZ, NFA, VFA, retinal vascular density, and perfusion on the OCTA. Eyes with poor-quality images or a poor signal strength were excluded.

2.2. Imaging Protocol

OCTA of the macula was performed using Angio-OCT Optopol Revo NX 130 (OPTOPOL Technology, Zawiercie, Poland) (Angiography module including a Retina Angiography 6 × 6 mm scan program and OCTA mosaic program to merge together 3 × 3 or 6 × 6 mm scans in order to create high-resolution mosaics of a larger area, iTracking™ technology, Motion Correction Technology™ and Projection Artefact Removal algorithm to minimize artifacts). The available vascular layers identified by the OCTA can be selected from the following list: retina, vitreous, superficial capillary plexus, superficial vascular plexus (SVC), radial peripapillary capillary plexus (RPCP), deep capillary plexus, deep vascular plexus (DVC), intermediate capillary plexus (ICP), outer retina layers, choriocapillaris, and choroidal vessels, with depth coded.

For the purpose of this study, 6.0 × 6.0 mm scans were performed, centered on the fovea of the right eye and left eye from each participant. The software (Version 11.0.4 REVO NX Device SN: 1560658/16) was used to analyze the superficial layer (with the offset of inner limiting membrane (ILM) 0 μm, inner plexiform layer/ inner nuclear layer (IPL/INL) 15 μm), the deep layer (with the offset of IPL/INL—15 μm, IPL/INL—70 μm), and the choriocapillaris (with the offset of Bruch's membrane (BM) top BM 30 μm, bottom BM 45 μm).

The software was also used to assess the FAZ, NFA, VFA, and the retinal vascular density and skeleton. These parameters were derived from an En face angiogram. The FAZ, NFA, and VFA areas were detected and measured using the semi-auto area tool of the software. As the semi-automated measurement was noted to be inaccurate in a number of scans, manual measurements were performed. The calculated areas of the following parameters are provided: FAZ area in mm^2, FAZ perimeter in mm, circularity (ratio between the measured perimeter and the perimeter of a circular area of the same size), NFA in mm^2, VFA area in mm^2, and VFA flow area in mm^2. Quantification provided the quantification of vasculature in specific sectors and a heat map corresponding to the vasculature.

Density display is defined as the total area of perfused vasculature per unit area in a region of measurement. This metric is calculated by summing up the number of pixels that contain perfused vasculature and dividing the sum by the total number of pixels in the considered region. The result is an unitless number ranging from 0 (no perfusion) to 1 (fully perfused) (mm^2/mm^2).

Skeleton display is defined as the total area of skeletonized vasculature per unit area in a region of measurement. Skeletonization performs thinning of all vessels down to 1 pixel width and thus makes analysis more sensitive to small vasculature (as the large vessels lose more area than the thin ones in the skeletonization process). This metric is calculated by summing up the number of pixels that represent the skeleton of the vasculature and dividing the sum by the total number of pixels in the considered region. The result is an unitless number ranging from 0 (no perfusion) to 1 (mm^2/mm^2).

Both density and skeleton display can detect abnormal vasculature and provide repeatable quantitative results equally in normal and diseased eyes. Measurement zones for the available 6 mm width retina scans are total, superior, inferior, center, inner, superior inner, inferior inner, outer, superior outer, inferior outer, and the Early Treatment of Diabetic Retinopathy Study (ETDRS) grid.

Comparison and progression view screens were used. The comparison view screen shows the analysis results, comparing two examinations of one eye on the same side in the same scan mode but from different dates. The comparison view is used to observe follow-up changes in the eye structure. The software automatically selects outermost examinations (the oldest and the newest) in order to compare them. The progression view screen shows the analysis results comparing four examinations, all performed on the same

side in the same scan mode and on the same size of scanning area, arranged in a time sequence. Eyes with a significant artifact that obscured the vascular area of interest, with segmentation failure that could not be manually corrected, and with poor-quality images or a poor signal strength were excluded (Figures 1–3).

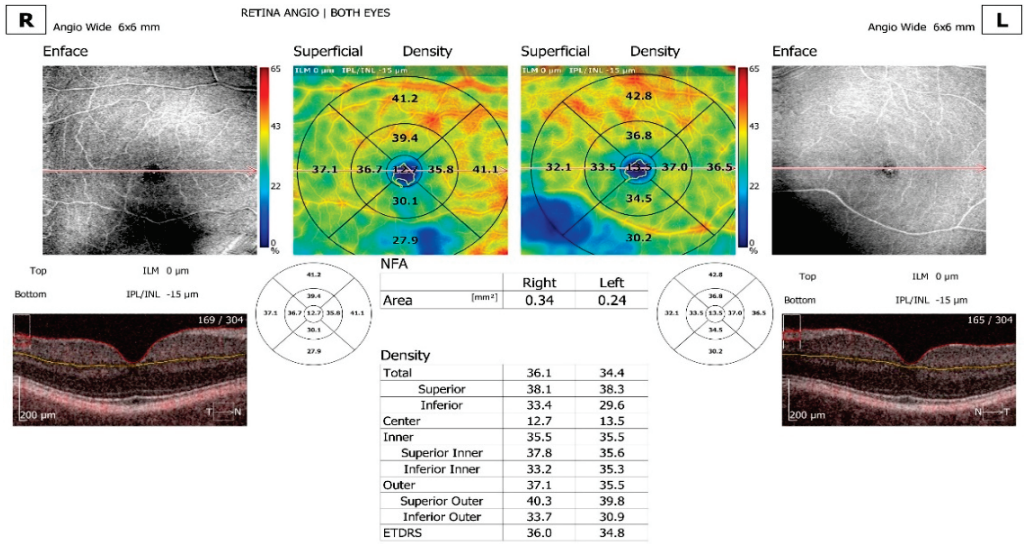


Figure 1. En face OCTA of the superficial vascular plexus with calculated NFA area, comparing right eye and left eye and calculated total, center, inner and outer density. OCT showing that the superficial plexus extends from the inner limiting membrane (top) (0 µm) to the inner plexiform layer/inner nuclear layer (bottom) (IPL/INL—15 µm). ETDRS grid is also available, comparing right eye and left eye.

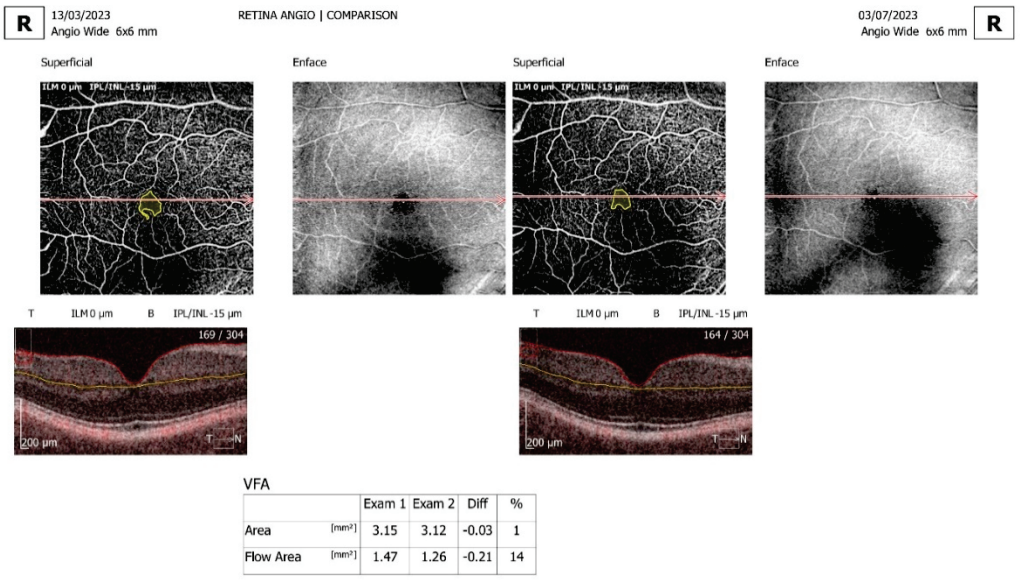
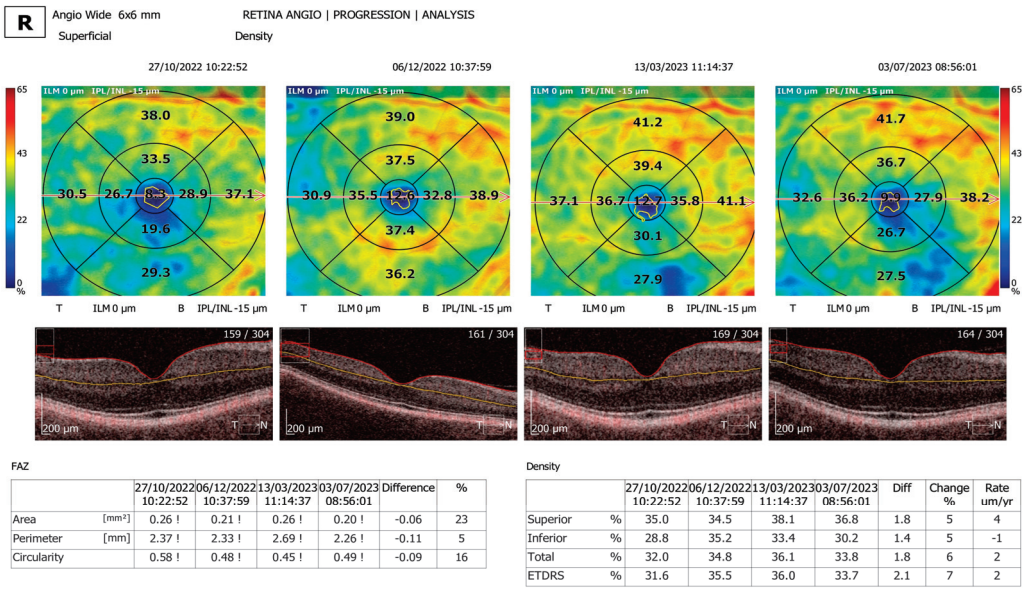


Figure 2. En face OCTA of the superficial vascular plexus with calculated VFA area and flow area, comparing right eye examination number 1 and examination number 2, with the difference between them measured in mm^2 and percentage.



FAZ

		27/10/2022 10:22:52	06/12/2022 10:37:59	13/03/2023 11:14:37	03/07/2023 08:56:01	Difference	%
Area	[mm²]	0.26 !	0.21 !	0.26 !	0.20 !	-0.06	23
Perimeter	[mm]	2.37 !	2.33 !	2.69 !	2.26 !	-0.11	5
Circularity		0.58 !	0.48 !	0.45 !	0.49 !	-0.09	16

Density

		27/10/2022 10:22:52	06/12/2022 10:37:59	13/03/2023 11:14:37	03/07/2023 08:56:01	Diff	Change %	Rate um/yr
Superior	%	35.0	34.5	38.1	36.8	1.8	5	4
Inferior	%	28.8	35.2	33.4	30.2	1.4	5	-1
Total	%	32.0	34.8	36.1	33.8	1.8	6	2
ETDRS	%	31.6	35.5	36.0	33.7	2.1	7	2

Figure 3. Retina Angiography progression analysis program comparing four scans. FAZ area, perimeter, and circularity were measured and compared. Superior, inferior, total density, and ETDRS grid are shown in a table in the right bottom corner of the OCTA, with calculated difference and change rate in µm/year between the four scans.

The current A.I. software was uploaded to the OCT machine.

The newest A.I. DeNoise™ technology improves tomogram quality powered by A.I. Advanced A.I. algorithms enhance the quality of a single tomogram to the level of an averaged tomogram obtained through multiple scans. The A.I. DeNoise™ algorithm filters out noise from the tomogram for the highest and smoothest image quality. The function is available on all tomograms. The moment a tomogram is loaded for review, the software starts denoising it. After a short moment, the original "undenoised" tomogram is replaced with a noise-free image obtaining a better visualization.

The A.I. Retina is a new layer segmentation for the posterior segment that is based on artificial intelligence, resulting in more accurate recognition of retinal layer boundaries.

The maximum intensity projection (MIP) algorithm provides better visualization of OCTA data for analysis. This MIP tool is useful for visualizing OCTA data, as it enables easier identification and tracking of high-intensity structures such as blood vessels.

The A.I. system has a direct impact on the accuracy of the clinical assessment and the assessment of the status of areas of pathology in the retina. This level of detection is more accurate and results in more detailed screening. Overall, it is a more effective way of running a pathology evaluation. A.I. segmentation is important for follow-up examinations, bringing a more accurate diagnosis when analyzing pathology over time.

2.3. Statistical Analysis

IBM Statistical Package for the Social Sciences (SPSS) V23, IBM SPSS Release 26.0.0.0, Minitab V16, JASP V 0.18.3, and Microsoft Office 2019 were used for statistical analysis.

The whole sample of 154 eyes was divided into two groups, with 77 subjects each: right eye and left eye, with each group subdivided into a hypercholesterolemia and hypertriglyceridemia group. Using descriptive statistics, the following indicators were calculated: mean, standard deviation (SD), mode, median, and 95% confidence intervals (CI). We analyzed age, sex, rural and urban areas, smoking habit, blood pressure (systolic and diastolic), cholesterol levels including total and HDL and LDL, triglycerides, BMI, glycemia, and HbA1c. For each group, the following parameters of the OCTA were determined for both eyes: FAZ area, FAZ perimeter, FAZ circularity, NFA area, VFA area, VFA flow area, density total, density ETDRS, skeleton total, and skeleton ETDRS.

The correlation between the blood pressure and the OCTA parameters was calculated using Pearson Correlation and Sig. (2-tailed) *p*-value test. Correlations between blood pressure, total cholesterol, HDL, LDL, triglycerides, and OCTA parameters were determined using Pearson correlation coefficient (r) and two-tailed *p*-value. We divided subjects with dyslipidemia into two groups, namely a hypercholesterolemia group (we analyzed total cholesterol, HDL, and LDL) and hypertriglyceridemia group (we analyzed triglycerides), and we determined correlation with OCTA parameters for the right eye and left eye.

A dependent *t*-test was used to determine if there is a significant difference between OCTA parameters in the right eye and the left-eye group, analyzing t-value (t), *p*-value (*p*), and degrees of freedom (df).

A multivariate analysis exploring how these factors interact collectively was performed. Factor analysis was performed using Principal component analysis of IBM SPSS Release 26.0.0.0. Kaiser–Meyer–Olkin test, Bartlett's test, rotated component matrix, total variance explained, and scree plot were used.

3. Results

3.1. Descriptive Analysis

A total of 154 eyes were eligible for the final analysis presented in this study.

Glycemia and HbA1c were determined in order to rule out subjects with diabetes and/or undiagnosed diabetic retinopathy. Age, sex, smoking habit, and rural and urban areas were analyzed and are shown in Table 1.

Table 1. Characteristics of the patient sample.

Parameters		Frequency	Valid Percent (%)	Age (Years)	
				Mean ± SD	Median (Min to Max)
Sex	Females	34	44.2	58.8 ± 8.8	58.0 (35 to 79)
	Males	43	55.8	55.4 ± 9.1	55.0 (34 to 82)
Areas	Rural	24	31.2	56.0 ± 8.5	55.5 (34 to 71)
	Urban	53	68.8	57.3 ± 9.4	56.0 (35 to 82)
Smokers	Yes	21	27.3	56.0 ± 8.5	55.5 (34 to 71)
	No	56	72.7	57.3 ± 9.4	56.0 (35 to 82)
Total		77	100.0	56.9 ± 9.1	56.0 (34 to 82)

The average age of the participants was 56.9 ± 9.1 years, with a minimum of 34 years and a maximum age of 82 years, and with a higher percentage of males: 55.8%. Overall, 31.2% subjects were recruited from rural areas and 68.8% subjects from urban areas; 27.3% were smokers, and 72.7% non-smokers.

The main clinical characteristics of the whole sample are shown in Table 1.

Table 2 shows the mean, standard deviation (SD), mode, median, and 95% confidence intervals (CI) mean calculated for blood pressure (SBP and DBP) and clinical variables in subjects with hypertension and dyslipidemia (Table S1).

Hypertensive participants with a higher mean SBP and DBP (129.5 ± 11.1, respectively 78.5 ± 9.6) had a significantly higher mean total cholesterol (197.7 ± 38.3) and mean LDL (124.0 ± 32.7) and a significantly higher mean BMI (28.8 ± 4.7). The mean HDL was 52.6 ± 13.3. Other clinical variables of the whole population study, which were not significant among subjects with hypertension and dyslipidemia, such as triglycerides, glycemia, and HbA1c, are shown in the Table 2 and Figure 4. Subjects with diabetes were excluded from this study.

Table 2. Statistical indicators calculated for clinical variables.

Clinical Variables	Statistical Indicators			
	Mean ± SD	Median (Min to Max)	Mode	95% CI Mean (Lower to Upper)
SBP (mmHg)	129.5 ± 11.1	131.0 (90.0 to 153.0)	138.0	(127.0 to 132.0)
DBP (mmHg)	78.5 ± 9.6	78.0 (43.0 to 108.0)	74.0	(76.4 to 80.7)
Total Cholesterol (mg/dL)	197.7 ± 38.3	197.1 (109.3 to 275.6)	234.0	(189.0 to 206.4)
HDL (mg/dL)	52.6 ± 13.3	50.4 (29.3 to 103.4)	60.8	(49.5 to 55.6)
LDL (mg/dL)	124.0 ± 32.7	121.4 (53.4 to 194.6)	53.4	(116.6 to 131.4)
Triglycerides (mg/dL)	122.7 ± 48.8	112.4 (65.2 to 427.4)	95.8	(111.6 to 133.8)
BMI (kg/m^2)	28.8 ± 4.7	28.4 (19.2 to 44.5)	25.1	(27.7 to 29.9)
Glycemia (mg/dL)	92.7 ± 10.4	92.2 (69.7 to 128.3)	95.5	(90.4 to 95.1)
HbA1c (%)	5.5 ± 0.4	5.5 (4.2 to 6.2)	5.9	(5.4 to 5.6)

SBP: systolic blood pressure; DBP diastolic blood pressure; HDL: high-density cholesterol, LDL: low-density cholesterol; BMI: body mass index; HbA1c: glycated hemoglobin. All data are shown as mean ± standard deviation (SD), median (min to max) for continuous variables, mode, and 95% confidence interval mean (lower to upper).

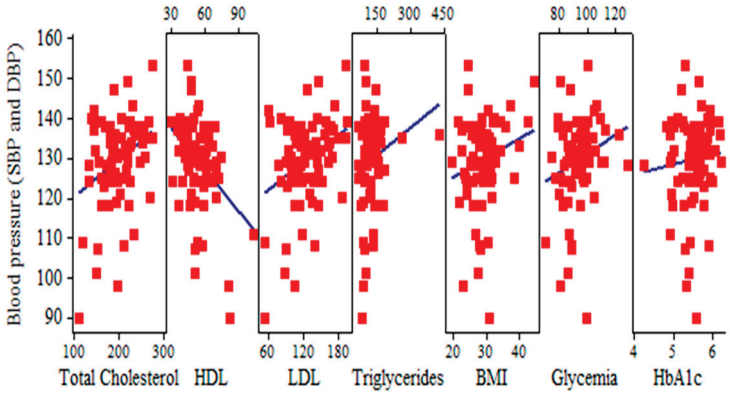


Figure 4. Statistical indicators of clinical variables.

3.2. Analysis of Ophthalmological Parameters

The analysis of ophthalmological parameters is shown in Tables 3–13.

Table 3. The mean, median, mode, 95% CI mean, standard deviation (SD), and minimum and maximum for ophthalmological parameters in the right-eye group.

OCTA Right Eye	Statistical Indicators			
	Mean ± SD	Median (Min to Max)	Mode	95% CI Mean (Lower to Upper)
NFA Area (mm^2)	0.4 ± 0.2	0.3 (0.1 to 1.5)	0.4	(0.3 to 0.4)
FAZ Area (mm^2)	0.5 ± 0.6	0.4 (0.1 to 2.8)	0.5	(0.4 to 0.7)
FAZ Perimeter (mm)	3.7 ± 2.1	2.9 (1.3 to 11.4)	3.7	(3.2 to 4.2)
FAZ Circularity	0.4 ± 0.1	0.4 (0.2 to 0.8)	0.4	(0.4 to 0.5)
VFA Area (mm^2)	3.2 ± 0.0	3.2 (3.1 to 3.2)	3.2	(3.1 to 3.2)
VFA Flow (mm^2)	1.3 ± 0.3	1.3 (0.4 to 2.1)	1.3	(1.3 to 1.4)
Density Total	34.5 ± 4.0	35.5 (18.3 to 39.0)	34.5	(33.6 to 35.4)
Density ETDRS	34.4 ± 4.0	35.3 (18.9 to 39.1)	34.4	(33.5 to 35.3)

Table 3. *Cont.*

OCTA Right Eye	Statistical Indicators			
	Mean ± SD	Median (Min to Max)	Mode	95% CI Mean (Lower to Upper)
Skeleton Total	20.8 ± 1.1	20.8 (18.1 to 22.7)	20.8	(20.5 to 21.0)
Skeleton ETDRS	20.8 ± 1.2	20.9 (18.2 to 24.4)	20.8	(20.5 to 21.0)

FAZ: foveal avascular zone; NFA: non-flow area; VFA: vascular flow area; ETDRS: Early Treatment of Diabetic Retinopathy Study. All data are shown as mean ± standard deviation (SD), median (min to max) for continuous variables, mode, and 95% confidence interval mean (lower to upper).

Table 4. Statistical indicators for the left-eye group.

OCTA Left Eye	Statistical Indicators			
	Mean ± SD	Median (Min to Max)	Mode	95% CI Mean (Lower to Upper)
NFA Area (mm^2)	0.4 ± 0.2	0.4 (0.1 to 0.8)	0.2	(0.3 to 0.4)
FAZ Area (mm^2)	0.4 ± 0.5	0.3 (0.1 to 3.5)	0.2	(0.3 to 0.5)
FAZ Perimeter (mm)	3.5 ± 1.7	3.1 (1.4 to 10.3)	2.3	(3.1 to 3.9)
FAZ Circularity	0.5 ± 0.1	0.5 (0.2 to 0.7)	0.5	(0.4 to 0.5)
VFA Area (mm^2)	3.2 ± 0.0	3.2 (3.1 to 3.2)	3.2	(3.1 to 3.2)
VFA Flow (mm^2)	1.3 ± 0.3	1.4 (0.6 to 2.0)	1.4	(1.2 to 1.4)
Density Total	34.6 ± 4.0	35.6 (20.3 to 40.1)	34.5	(33.7 to 35.5)
Density ETDRS	33.9 ± 4.4	34.7 (19.8 to 40.0)	33.4	(32.9 to 34.9)
Skeleton Total	20.9 ± 2.0	20.9 (16.9 to 35.7)	21.0	(20.5 to 21.4)
Skeleton ETDRS	20.7 ± 1.9	20.9 (16.3 to 30.7)	20.4	(20.3 to 21.1)

FAZ: foveal avascular zone; NFA: non-flow area; VFA: vascular flow area; ETDRS: Early Treatment of Diabetic Retinopathy Study. All data are shown as mean ± standard deviation (SD), median (min to max) for continuous variables, mode, and 95% confidence interval mean (lower to upper).

Table 5. Right-eye group with hypercholesterolemia and correlations with OCTA parameters.

	Correlations			
OCTA Parameters: Right Eye	Pearson Correlation	Total Cholesterol	HDL	LDL
FAZ Area	Pearson Correlation	0.020	−0.091	−0.108
	Sig. (2-tailed)	0.864	0.431	0.350
FAZ Perimeter	Pearson Correlation	0.056	−0.079	−0.058
	Sig. (2-tailed)	0.630	0.495	0.615
FAZ Circularity	Pearson Correlation	0.198	−0.053	0.216
	Sig. (2-tailed)	0.085	0.645	0.059
NFA Area	Pearson Correlation	0.126	−0.103	0.016
	Sig. (2-tailed)	0.275	0.371	0.893
VFA Area	Pearson Correlation	0.016	0.274 *	0.023
	Sig. (2-tailed)	0.891	0.016	0.841
VFA Flow	Pearson Correlation	−0.160	0.050	−0.056
	Sig. (2-tailed)	0.163	0.667	0.628

Table 5. *Cont.*

Correlations				
OCTA Parameters: Right Eye	**Pearson Correlation**	**Total Cholesterol**	**HDL**	**LDL**
Density Total	Pearson Correlation	−0.124	0.148	−0.145
	Sig. (2-tailed)	0.282	0.198	0.209
Density ETDRS	Pearson Correlation	−0.136	0.149	−0.149
	Sig. (2-tailed)	0.238	0.194	0.197
Skeleton Total	Pearson Correlation	−0.125	0.198	−0.108
	Sig. (2-tailed)	0.278	0.084	0.348
Skeleton ETDRS	Pearson Correlation	0.037	0.143	0.076
	Sig. (2-tailed)	0.750	0.214	0.511

FAZ: foveal avascular zone; NFA: non-flow area; VFA: vascular flow area; ETDRS: Early Treatment of Diabetic Retinopathy Study: right eye; HDL: high-density cholesterol, LDL: low-density cholesterol. * Correlation is significant at the 0.05 level (2-tailed).

Table 6. Left-eye group with hypercholesterolemia and correlations with OCTA parameters.

Correlations				
OCTA Parameters: Left Eye	**Pearson Correlation**	**Total cholesterol**	**HDL**	**LDL**
FAZ Area	Pearson Correlation	0.167	−0.141	−0.066
	Sig. (2-tailed)	0.146	0.220	0.569
FAZ Perimeter	Pearson Correlation	0.099	−0.082	−0.033
	Sig. (2-tailed)	0.393	0.477	0.777
FAZ Circularity	Pearson Correlation	0.287 *	−0.178	0.313 **
	Sig. (2-tailed)	0.011	0.121	0.006
NFA Area	Pearson Correlation	0.245 *	−0.025	0.232 *
	Sig. (2-tailed)	0.032	0.826	0.042
VFA Area	Pearson Correlation	−0.273 *	0.018	−0.354 **
	Sig. (2-tailed)	0.016	0.878	0.002
VFA Flow	Pearson Correlation	−0.257 *	−0.021	−0.142
	Sig. (2-tailed)	0.024	0.856	0.218
Density Total	Pearson Correlation	−0.112	0.119	−0.131
	Sig. (2-tailed)	0.332	0.305	0.256
Density ETDRS	Pearson Correlation	−0.078	0.100	−0.042
	Sig. (2-tailed)	0.501	0.385	0.716
Skeleton Total	Pearson Correlation	−0.249 *	0.230 *	−0.178
	Sig. (2-tailed)	0.029	0.044	0.120
Skeleton ETDRS	Pearson Correlation	−0.254 *	0.186	−0.177
	Sig. (2-tailed)	0.026	0.105	0.124

FAZ: foveal avascular zone; NFA: non-flow area; VFA: vascular flow area; ETDRS: Early Treatment of Diabetic Retinopathy Study: left eye; HDL: high-density cholesterol, LDL: low-density cholesterol. * Correlation is significant at the 0.05 level (2-tailed). ** Correlation is significant at the 0.01 level (2-tailed).

Table 7. Right-eye group with hypertriglyceridemia. Correlations with OCTA parameters.

Correlations		
OCTA Parameters	**Pearson Correlation**	**Triglycerides**
FAZ Area	Pearson Correlation	−0.092
	Sig. (2-tailed)	0.425
FAZ Perimeter	Pearson Correlation	−0.076
	Sig. (2-tailed)	0.511
FAZ Circularity	Pearson Correlation	0.058
	Sig. (2-tailed)	0.615
NFA Area	Pearson Correlation	−0.139
	Sig. (2-tailed)	0.227
VFA Area	Pearson Correlation	−0.189
	Sig. (2-tailed)	0.099
VFA Flow	Pearson Correlation	0.090
	Sig. (2-tailed)	0.437
Density Total	Pearson Correlation	0.033
	Sig. (2-tailed)	0.773
Density ETDRS	Pearson Correlation	0.036
	Sig. (2-tailed)	0.753
Skeleton Total	Pearson Correlation	−0.113
	Sig. (2-tailed)	0.326
Skeleton ETDRS	Pearson Correlation	−0.168
	Sig. (2-tailed)	0.144

FAZ: foveal avascular zone; NFA: non-flow area; VFA: vascular flow area; ETDRS: Early Treatment of Diabetic Retinopathy Study: right eye.

Table 8. Left-eye group with hypertriglyceridemia. Correlations with OCTA parameters.

Correlations		
OCTA Parameters	**Pearson Correlation**	**Triglycerides**
FAZ Area	Pearson Correlation	−0.068
	Sig. (2-tailed)	0.557
FAZ Perimeter	Pearson Correlation	−0.073
	Sig. (2-tailed)	0.527
FAZ Circularity	Pearson Correlation	0.071
	Sig. (2-tailed)	0.542
NFA Area	Pearson Correlation	−0.072
	Sig. (2-tailed)	0.534
VFA Area	Pearson Correlation	−0.103
	Sig. (2-tailed)	0.371
VFA Flow	Pearson Correlation	−0.048
	Sig. (2-tailed)	0.679

Table 8. *Cont.*

Correlations		
OCTA Parameters	**Pearson Correlation**	**Triglycerides**
Density Total	Pearson Correlation	0.116
	Sig. (2-tailed)	0.315
Density ETDRS	Pearson Correlation	0.083
	Sig. (2-tailed)	0.473
Skeleton Total	Pearson Correlation	−0.053
	Sig. (2-tailed)	0.650
Skeleton ETDRS	Pearson Correlation	−0.001
	Sig. (2-tailed)	0.990

FAZ: foveal avascular zone; NFA: non-flow area; VFA: vascular flow area; ETDRS: Early Treatment of Diabetic Retinopathy Study: left eye.

Table 9. The ophthalmological parameters in the right-eye group with hypertensive retinopathy and the correlation with SBP and DBP, calculated using Pearson Correlation and Sig. (2-tailed) test.

RIGHT EYE OCTA		RIGHT EYE OCTA										
		SBP and DBP	**NFA Area**	**FAZ Area**	**FAZ Perimeter**	**FAZ Circularity**	**VFA Area**	**VFA Flow Area**	**Density Total**	**Density ETDRS**	**Skeleton Total**	**Skeleton ETDRS**
SBP and DBP	Pearson Correlation	1	0.276 *	0.236 *	0.253 *	0.256 *	−0.368 **	−0.364 **	−0.382 **	−0.399 **	−0.364 **	−0.351 **
	Sig. (2-tailed)		0.015	0.039	0.026	0.025	0.001	0.001	0.001	0.000	0.001	0.002
NFA Area	Pearson Correlation	**0.276 ***	1	0.512 **	0.491 **	0.112	−0.289 *	−0.242 *	−0.433 **	−0.449 **	−0.382 **	−0.103
	Sig. (2-tailed)	0.015		0.000	0.000	0.333	0.011	0.034	0.000	0.000	0.001	0.371
FAZ Area	Pearson Correlation	**0.236 ***	**0.512 ** **	1	0.927 **	−0.253 *	−0.302 **	−0.281 *	−0.268 *	−0.320 **	−0.393 **	−0.246 *
	Sig. (2-tailed)	0.039	0.000		0.000	0.026	0.008	0.013	0.018	0.005	0.000	0.031
FAZ Perimeter	Pearson Correlation	**0.253 ***	**0.491 ** **	**0.927 ** **	1	−0.374 **	−0.255 *	−0.322 **	−00.211	−0.269 *	−0.348 **	−0.235 *
	Sig. (2-tailed)	0.026	0.000	0.000		0.001	0.025	0.004	0.065	0.018	0.002	0.040
FAZ Circularity	Pearson Correlation	**0.256 ***	0.112	**−0.253 ***	**−0.374 ** **	1	−0.137	−0.085	−0.125	−0.109	−0.078	0.088
	Sig. (2-tailed)	0.025	0.333	0.026	0.001		0.234	0.461	0.277	0.347	0.499	0.445
VFA Area	Pearson Correlation	**−0.368 ** **	**−0.289 ***	**−0.302 ** **	**−0.255 ***	−0.137	1	0.039	0.230 *	0.267 *	0.397 **	0.279 *
	Sig. (2-tailed)	0.001	0.011	0.008	0.025	0.234		0.739	0.044	0.019	0.000	0.014
VFA Flow Area	Pearson Correlation	**−0.364 ** **	**−0.242 ***	**−0.281 ***	**−0.322 ** **	−0.085	0.039	1	0.145	0.142	0.221	0.253 *
	Sig. (2-tailed)	0.001	0.034	0.013	0.004	0.461	0.739		0.209	0.218	0.053	0.026
Density Total	Pearson Correlation	**−0.382 ** **	**−0.433 ** **	**−0.268 ***	−0.211	−0.125	**0.230 ***	0.145	1	0.989 **	0.775 **	0.472 **
	Sig. (2-tailed)	0.001	0.000	0.018	0.065	0.277	0.044	0.209		0.000	0.000	0.000
Density ETDRS	Pearson Correlation	**−0.399 ** **	**−0.449 ** **	**−0.320 ** **	**−0.269 ***	−0.109	**0.267 ***	0.142	**0.989 ** **	1	0.788 **	0.490 **
	Sig. (2-tailed)	0.000	0.000	0.005	0.018	0.347	0.019	0.218	0.000		0.000	0.000
Skeleton Total	Pearson Correlation	**−0.364 ** **	**−0.382 ** **	**−0.393 ** **	**−0.348 ** **	−0.078	**0.397 ** **	0.221	**0.775 ** **	**0.788 ** **	1	0.613 **
	Sig. (2-tailed)	0.001	0.001	0.000	0.002	0.499	0.000	0.053	0.000	0.000		0.000
Skeleton ETDRS	Pearson Correlation	**−0.351 ** **	−0.103	**−0.246 ***	**−0.235 ***	0.088	**0.279 ***	**0.253 ***	**0.472 ** **	**0.490 ** **	**0.613 ** **	1
	Sig. (2-tailed)	0.002	0.371	0.031	0.040	0.445	0.014	0.026	0.000	0.000	0.000	

SBP: systolic blood pressure; DBP: diastolic blood pressure; FAZ: foveal avascular zone; NFA: non-flow area; VFA: vascular flow area; ETDRS: Early Treatment of Diabetic Retinopathy Study. * Correlation is significant at the 0.05 level (2-tailed). ** Correlation is significant at the 0.01 level (2-tailed).

Table 10. OCTA parameters in the left-eye group with hypertensive retinopathy. Correlation with blood pressure and between OCTA parameters, also calculated using Pearson Correlation and Sig. (2-tailed) test.

LEFT EYE OCTA		LEFT EYE OCTA										
		SBP and DBP	**NFA Area**	**FAZ Area**	**FAZ Perimeter**	**FAZ Circularity**	**VFA Area**	**VFA Flow Area**	**Density Total**	**Density ETDRS**	**Skeleton Total**	**Skeleton ETDRS**
SBP and DBP	Pearson Correlation	1	0.301 **	0.314 **	0.322 **	0.359 **	−0.373 **	−0.379 **	−0.345 **	−0.373 **	−0.354 **	−0.363 **
	Sig. (2-tailed)		0.008	0.005	0.004	0.001	0.001	0.001	0.002	0.001	0.002	0.001
NFA Area	Pearson Correlation	**0.301 ** **	1	0.385 **	0.338 **	0.171	−0.1519	−0.19888	−0.287 *	−0.330 **	−0.306 **	−0.424 **
	Sig. (2-tailed)	0.008		0.001	0.003	0.136	0.187	0.083	0.011	0.003	0.007	0.000

Table 10. *Cont.*

LEFT EYE OCTA		LEFT EYE OCTA										
		SBP and DBP	NFA Area	FAZ Area	FAZ Perimeter	FAZ Circularity	VFA Area	VFA Flow Area	Density Total	Density ETDRS	Skeleton Total	Skeleton ETDRS
FAZ Area	Pearson Correlation	**0.314** **	**0.385** **	1	0.726 **	0.02113677	0.03426	−0.342 **	−0.240 *	−0.273 *	−0.280 *	−0.394 **
	Sig. (2-tailed)	0.005	0.001		0.000	0.855	0.767	0.002	0.036	0.016	0.014	0.000
FAZ Perimeter	Pearson Correlation	**0.322** **	**0.338** **	**0.726** **	1	−0.226 *	−0.0318	−0.293 **	−0.297 **	−0.326 **	−0.270 *	−0.355 **
	Sig. (2-tailed)	0.004	0.003	0.000		0.048	0.784	0.010	0.009	0.004	0.018	0.002
FAZ Circularity	Pearson Correlation	**0.359** **	0.171	0.021	**−0.226** *	1	−0.160	−0.263 *	−0.252 *	−0.196	−0.198	−0.082
	Sig. (2-tailed)	0.001	0.136	0.855	0.048		0.165	0.021	0.027	0.088	0.085	0.476
VFA Area	Pearson Correlation	**−0.373** **	−0.152	0.034	−0.032	−0.160	1	0.157	0.322 **	0.239 *	0.239 *	0.239 *
	Sig. (2-tailed)	0.001	0.187	0.767	0.784	0.165		0.172	0.004	0.036	0.036	0.036
VFA Flow Area	Pearson Correlation	**−0.379** **	−0.199	**−0.342** **	**−0.293** **	**−0.263** *	0.157	1	0.195	0.343 **	0.315 **	0.366 **
	Sig. (2-tailed)	0.001	0.083	0.002	0.010	0.021	0.172		0.089	0.002	0.005	0.001
Density Total	Pearson Correlation	**−0.345** **	**−0.287** *	**−0.240** *	**−0.297** **	**−0.252** *	**0.322** **	0.195	1	0.887 **	0.386 **	0.482 **
	Sig. (2-tailed)	0.002	0.011	0.036	0.009	0.027	0.004	0.089		0.000	0.001	0.000
Density ETDRS	Pearson Correlation	**−0.373** **	**−0.330** **	**−0.273** *	**−0.326** **	−0.196	**0.239** *	**0.343** **	**0.887** **	1	0.441 **	0.539 **
	Sig. (2-tailed)	0.001	0.003	0.016	0.004	0.088	0.036	0.002	0.000		0.000	0.000
Skeleton Total	Pearson Correlation	**−0.354** **	**−0.306** **	**−0.280** *	**−0.270** *	−0.198	**0.239** *	**0.315** **	**0.386** **	**0.441** **	1	0.870 **
	Sig. (2-tailed)	0.002	0.007	0.014	0.018	0.085	0.036	0.005	0.001	0.000		0.000
Skeleton ETDRS	Pearson Correlation	**−0.363** **	**−0.424** **	**−0.394** **	**−0.355** **	−0.082	**0.239** *	**0.366** **	**0.482** **	**0.539** **	**0.870** **	1
	Sig. (2-tailed)	0.001	0.000	0.000	0.002	0.476	0.036	0.001	0.000	0.000	0.000	

* Correlation is significant at the 0.05 level (2-tailed). ** Correlation is significant at the 0.01 level (2-tailed).

Table 11. The difference between right-eye and left-eye OCTA parameters. The mean, median, mode, 95% CI mean, standard deviation (SD), and minimum and maximum were calculated.

OCTA Parameters	Mean	Median	Mode	95% CI Mean		SD	Min.	Max.
				Lower	Upper			
NFA Area	0.02	0.01	0.08	−0.03	0.07	0.21	−0.53	0.68
FAZ Area	0.10	0.00	0.00	−0.01	0.21	0.48	−0.86	1.85
FAZ Perimeter	0.19	−0.03	0.00	−0.22	0.60	1.82	−3.27	6.97
FAZ Circularity	−0.02	0.00	0.00	−0.06	0.01	0.15	−0.39	0.28
VFA Area	0.00	0.00	0.00	−0.01	0.00	0.02	−0.06	0.09
VFA Flow	0.03	−0.04	0.19	−0.06	0.11	0.35	−0.85	0.98
Density Total	−0.07	−0.40	−0.40	−0.70	0.57	2.78	−4.40	10.80
Density ETDRS	0.45	0.00	−1.00	−0.31	1.22	3.37	−6.10	15.70
Skeleton Total	−0.17	−0.10	−1.10	−0.60	0.26	1.88	−13.00	4.00
Skeleton ETDRS	0.06	−0.10	−1.20	−0.40	0.50	1.98	−10.60	4.70

FAZ: foveal avascular zone; NFA: non-flow area; VFA: vascular flow area; ETDRS: Early Treatment of Diabetic Retinopathy Study; SD: standard deviation.

Table 12. The dependent 2-tailed *t*-test comparing right-eye and left-eye OCTA parameters.

OCTA Right Eye	OCTA Left Eye	t	df	*p*
NFA Area	NFA Area	0.793	76	0.430
FAZ Area	FAZ Area	1.810	76	0.074
FAZ Perimeter	FAZ Perimeter	0.913	76	0.364
FAZ Circularity	FAZ Circularity	1.322	76	0.190
VFA Area	VFA Area	0.770	76	0.444
VFA Flow	VFA Flow	0.579	76	0.564
Density Total	Density Total	0.205	76	0.838
Density ETDRS	Density ETDRS	1.181	76	0.241
Skeleton Total	Skeleton Total	0.788	76	0.433
Skeleton ETDRS	Skeleton ETDRS	0.242	76	0.810

FAZ: foveal avascular zone; NFA: non-flow area; VFA: vascular flow area; ETDRS: Early Treatment of Diabetic Retinopathy Study; t: t-value, df: degrees of freedom; *p*: *p*-value.

Table 13. Kaiser–Meyer–Olkin and Bartlett's test in the right-eye group.

Kaiser–Meyer–Olkin Measure of Sampling Adequacy.		**0.619**
Bartlett's Test of Sphericity	Approx. Chi-Square	834.421
	df	91
	Sig.	**<0.001**

Subjects were divided into two groups: right eye and left eye. For each group, the mean, median, mode, 95% CI mean, standard deviation (SD), and minimum and maximum were calculated and are shown in Tables 3 and 4. Also, for each eye group, we determined OCTA parameters and correlations with hypercholesterolemia (Tables 5 and 6) and with hypertriglyceridemia (Tables 7 and 8).

Tables 3 and 4 provide statistical indicators for the OCTA parameters for each eye group.

By analyzing the hypercholesterolemia in the right-eye group, we found the correlation between VFA area and HDL level to be statistically significant (r = 0.274; p = 0.016), showing that the higher the HDL level, the more the vascular flow is improved. Other results are also shown in Table 5 and are not of statistical significance.

In the left-eye group with hypercholesterolemia shown in Table 6, statistically significant results were determined for total cholesterol and FAZ circularity (r = 0.287; p = 0.011), NFA area (r = 0.245; p = 0.032), VFA area (r = 0.273; p = 0.016), VFA flow (r= −0.257; p = 0.024), skeleton total (r= −0.249; p = 0.029), and skeleton ETDRS (r= −0.254; p = 0.026).

The correlation between FAZ circularity and LDL (r = 0.313; p = 0.006), NFA area and LDL (r = 0.233; p = 0.042), and VFA area and LDL (r= −0.354; p = 0.002) were statistically significant. Higher values of skeleton total were correlated with higher values of HDL (r = 0.230; p = 0.044). This shows that by improving the HDL level, an improvement in the capillary circulation is obtained.

Our results show that subjects with higher values of LDL and total cholesterol lose more capillary vessels and have a higher non-flow area and poorer vascular flow; this results in the progression of hypertensive retinopathy in patients with poor dyslipidemia control.

No statistically significant correlations were found between hypertriglyceridemia and OCTA parameters in both eye groups. The results are shown in Tables 7 and 8.

By analyzing the right-eye group of hypertensive retinopathy shown in Table 9 and Figure 5, we found that four OCTA parameters were significantly higher in subjects with hypertension. The following results were obtained during our study for NFA area (r = 0.276; p = 0.015), FAZ area (r = 0.236; p = 0.039), FAZ perimeter (r = 0.253; p = 0.026), and FAZ circularity (r = 0.256; p = 0.025).

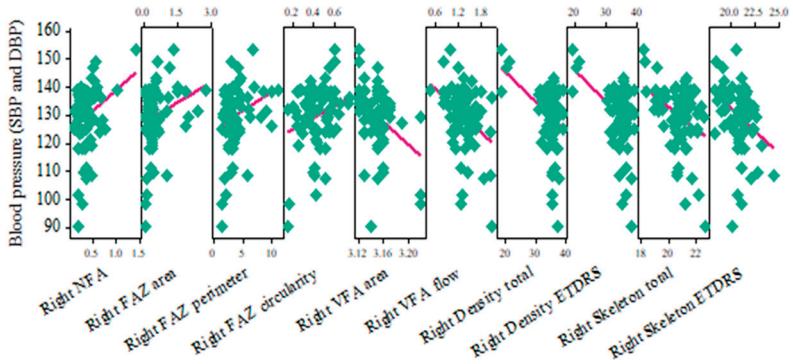


Figure 5. A matrix plot showing correlation of right-eye OCTA parameters with blood pressure.

All other OCTA parameters, namely VFA area (r = −0.368; p = 0.001), VFA flow area (r = −0.364; p = 0.001), density total (r = −0.382; p = 0.001), density ETDRS (r = −0.399; p < 0.01), skeleton total (r = −0.364; p = 0.001), and skeleton ETDRS (r = −0.351; p = 0.002), were significantly lower in subjects with hypertension.

The results show avascular and non-flow areas increase in subjects with higher blood pressure, and a capillary drop-out as well as a decreased vascular flow are present in these cases. So, progression of hypertensive retinopathy in patients with higher blood pressure can be proven based on OCTA (Table S2).

In the left-eye group of hypertensive retinopathy shown in Table 10 and Figure 6, the NFA area (r = 0.301; p = 0.008), FAZ area (r = 0.314; p = 0.005), FAZ perimeter (r = 0.322; p = 0.004), and FAZ circularity (r = 0.359; p = 0.001) values were also significantly higher in subjects with hypertension, just as with the right-eye group. All other OCTA parameters, namely VFA area (r = −0.373; p = 0.001), VFA flow area (r = −0.379; p = 0.001), density total (r = −0.345; p = 0.002), density ETDRS (r = −0.373; p = 0.001), skeleton total (r = −0.354; p = 0.002), and ETDRS (r = −0.363; p = 0.001), were lower when the blood pressure (SBP and DBP) was higher.

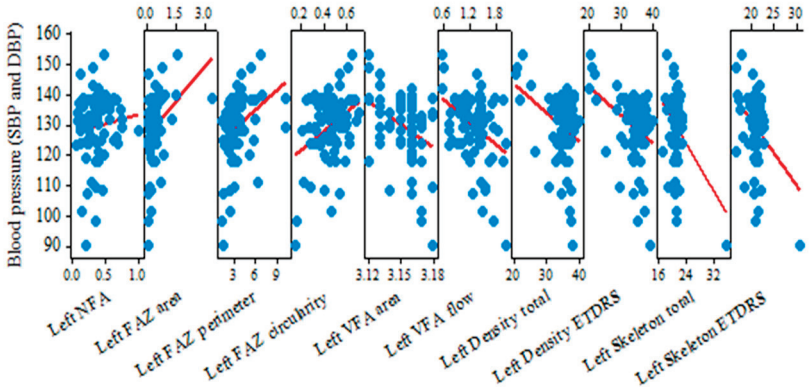


Figure 6. A matrix plot showing correlation of OCTA parameters in the left eye with blood pressure.

Patients with higher values of the blood pressure showed higher avascular zones, reduced blood flow, and a capillary drop-out. So, a progression of hypertensive retinopathy is demonstrated and measured with OCTA (Table S3).

The correlation between OCTA parameters was also calculated and shown in Tables 9 and 10, using Pearson Correlation and Sig. (2-tailed) test and showing r-value and p-value. In subjects with a higher NFA area, the values of the FAZ area and perimeter were higher, while VFA area and flow area, density (total and ETDRS), and skeleton total were significantly lower.

In both groups, participants with higher values of density (total and ETDRS parameters) showed higher skeleton (total and ETDRS) levels.

Differences between right-eye OCTA parameters and left-eye OCTA parameters were also analyzed and are shown in Tables 11 and 12.

A dependent 2-tailed t-test was used to determine if there is a significant difference between OCTA parameters in the right-eye group and OCTA parameters in the left-eye group, analyzing t-value (t), p-value (p), and degrees of freedom (df).

The t-test determined a value of $p > 0.05$, so no statistical significance was found between right-eye OCTA parameters and left-eye OCTA parameters. The highest values of p resulted for NFA area ($p > 0.430$), VFA area ($p > 0.444$), VFA flow ($p > 0.564$), density total ($p > 0.838$), and skeleton ETDRS ($p > 0.810$).

The MIP algorithm of the machine provides better visualization of OCTA data for analysis, as it enables easier identification and tracking of high-intensity structures such as blood vessels and more specific measurements of the OCTA parameters.

By associating the most relevant data obtained in our study, such as the statistically significant correlations calculated, the A.I. software can be improved by integrating new mathematic formulas to the current software to link OCTA parameters to blood pressure and lipids levels. Our most statistically significant correlations show that some OCTA variables, such as VFA area and flow, density total, and ETDRS and skeleton total, are inversely

proportional to blood pressure, total cholesterol, and LDL, while other OCTA parameters, such as NFA area and FAZ area, perimeter, and circularity, are directly proportional to blood pressure, LDL, and total cholesterol.

In order to enhance the interpretation of the results, a multivariate analysis exploring how these factors interact collectively was performed using Kaiser–Meyer–Olkin (KMO) test, Bartlett's test, rotated component matrix, total variance explained, and scree plot.

In order to demonstrate that the data used in correlations were reliable for factor analysis, a KMO and a Bartlett's test were performed. The results are shown in Tables 13 and 14. The KMO value in the right-eye group was 0.619 higher than the KMO value in the left eye, 0.531; thus, the data in right-eye group is more relevant.

Table 14. Kaiser–Meyer–Olkin and Bartlett's test in the left-eye group.

Kaiser–Meyer–Olkin Measure of Sampling Adequacy.		**0.531**
Bartlett's Test of Sphericity	Approx. Chi-Square	631.589
	df	91
	Sig.	**<0.001**

Bartlett's test resulted in a Sig. value < 0.01 in both eyes, showing that factor analysis was relevant in both groups.

The rotated component matrix is presented in Tables 15 and 16. The rotation method used was Varimax with Kaiser normalization. For the right-eye group, five components were found to be relevant, and for the left-eye group, four relevant components were found.

Table 15. Rotated component matrix in the right-eye group.

	Component				
	1	**2**	**3**	**4**	**5**
Total Density	0.928	−0.171			−0.164
ETDRS density	0.925	−0.217			−0.139
Total Skeleton	0.847	−0.262		0.212	
Skeleton ETDRS	0.699			0.247	0.321
FAZ Perimeter	−0.124	0.873			−0.390
FAZ Area	−0.173	0.870			−0.287
NFA Area	−0.264	0.724			0.338
VFA Flow	0.106	−0.478	−0.125		
Total Cholesterol		0.125	0.966		
LDL			0.931		0.127
Triglycerides		−0.240	0.366	−0.750	−0.198
HDL	0.104		0.258	0.727	−0.119
VFA Area	0.214	−0.327		0.601	−0.188
FAZ Circularity		−0.117	0.176	−0.126	0.858

Extraction method: principal component analysis. Rotation method: Varimax with Kaiser normalization. Rotation converged in five iterations.

The total variance explained is presented in Table 17 for both eyes. Values above 1 are statistically significant. In the right-eye group, five components with a cumulative % of 76.743% were found, while in the left-eye group, four components with a cumulative % value of 65.866% were obtained.

Table 16. Rotated component matrix in the left-eye group.

	Component			
	1	2	3	4
Total Density	0.876	−0.177		
Density ETDRS	0.859	−0.270		
Skeleton ETDRS	0.588	−0.542	−0.1424	0.211
Skeleton Total	0.558	−0.418	−0.169	0.322
VFA Area	0.435		−0.415	
FAZ Circularity	−0.402	−0.195	0.397	−0.244
FAZ Area		0.870		
FAZ Perimeter		0.869	−0.114	
NFA Area	−0.284	0.508	0.233	
VFA Flow	0.247	−0.441	−0.254	
Total Cholesterol		0.232	0.930	
LDL			0.927	
HDL	0.189		0.283	0.811
Triglycerides	0.191		0.289	−0.734

Extraction method: principal component analysis. Rotation method: Varimax with Kaiser normalization. Rotation converged in five iterations.

Table 17. Total variance explained—right eye and left eye.

Component	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.211	30.081	30.081	3.116	22.256	22.256
2	2.262	16.160	46.241	2.614	18.669	40.925
3	1.764	12.597	58.838	2.071	14.794	55.719
4	1.381	9.866	68.704	1.595	11.391	67.110
5	1.125	8.039	76.743	1.349	9.633	76.743
1	4.156	29.683	29.683	2.737	19.552	19.552
2	2.213	15.807	45.490	2.640	18.858	38.410
3	1.505	10.747	56.236	2.405	17.178	55.588
4	1.348	9.629	65.866	1.439	10.278	65.866

Extraction method: principal component analysis.

The scree plot calculated for both eyes is shown in Figures 7 and 8. Values above 1 are statistically significant.

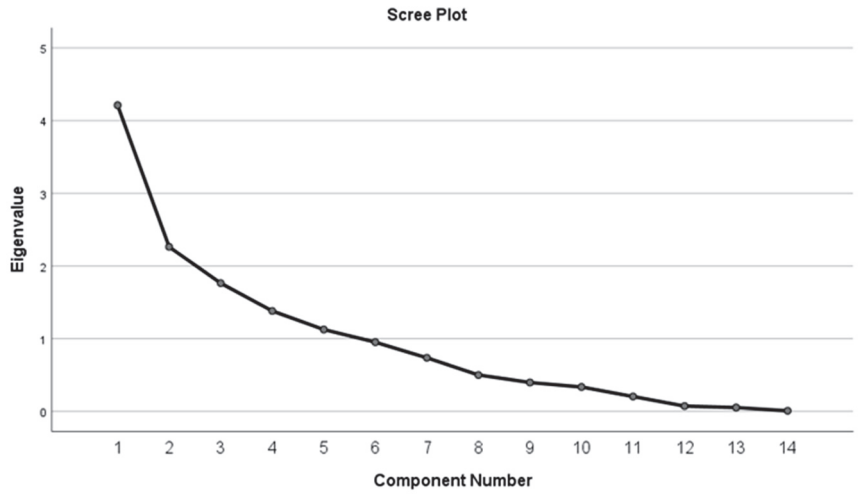


Figure 7. Scree plot—right eye.

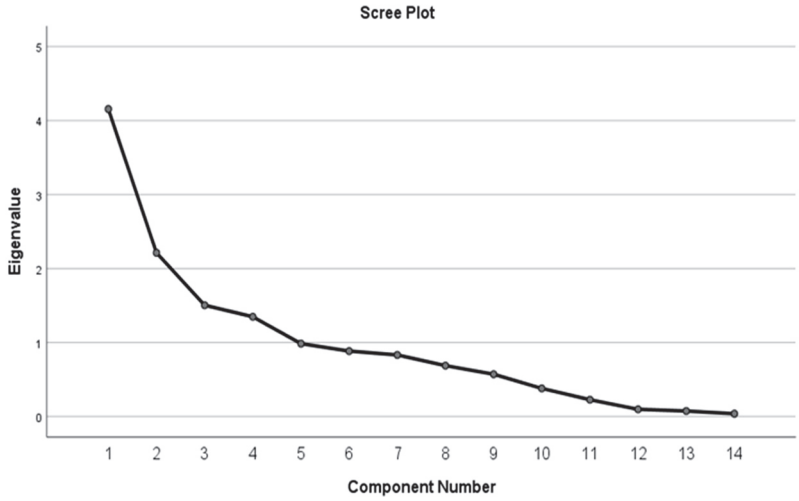


Figure 8. Scree plot—left eye.

4. Discussion

Hypertension is a very common condition worldwide and an important risk factor for cardiovascular and cerebrovascular events, renal diseases, and microvascular eye diseases [6]. Established hypertension leads to the chronic vasoconstriction of small arteries, impairment of the organ flow reserve, and the progressive worsening of hypertensive disease.

One of the latest studies about hypertensive retinopathy was published by Lei Jiao, Chaoran Lv, and Hongling Zhang [9]. This study concluded that hypertensive retinopathy is related to blood pressure variability (BPV), that the degree of BPV and retinopathy in hypertensive patients can be used to track the progression of hypertension-mediated organ damage (HOMD), and that the correction of BPV may help treat or postpone the progression of HOMD. Another study by Christopher Sun et al. showed that hypertension is associated with reduced retinal vessel density and an increased FAZ, especially in the deep venous plexus, as seen on OCTA [11].

In 2022, Niro, A et al. demonstrated in their study that patients with hypertensive retinopathy show a reduction in the retinal vascular density, superficial and deep vascular plexi (SVP and DVP), ganglion cell–internal boundary membrane complex density (GCL-ILM), and retinal nerve fiber layer (RNFL) thickness [12].

In the present study conducted on subjects with hypertension and dyslipidemia, the correlation between blood pressure (systolic and diastolic) and clinical variables showed that higher blood pressure is associated with higher total cholesterol and LDL and with lower HDL values. Triglycerides were not demonstrated to be statistically related to changes in OCTA parameters.

In our sample, participants with hypertension and a higher BMI and waist circumference had a higher blood pressure, suggesting a relationship between obesity and blood hypertension [19].

Our subjects had no diabetes and no undiagnosed diabetic retinopathy, and no relationship between hypertension and glycemia or HbA1c was found.

The relationship between blood pressure control, retinal blood vessels, choriocapillaris, and systemic vascular risk factors in hypertension was examined in our study.

We used the OCTA to confirm the association of blood pressure variability and dyslipidemia with OCTA parameters in hypertensive patients.

The choriocapillaris is composed of a dense network of capillaries, so it may be susceptible to damage as a result of uncontrolled systemic hypertension [6].

In the present study, we calculated the VFA (area and flow area parameters) at the level of the choriocapillaris. We reported a decrease in the capillary network in hypertensive subjects, which may contribute to flow reduction in and the impairment of oxygen delivery to different organs. The following values resulted for VFA area ($r = -0.368$; $p = 0.001$)

and VFA flow area (r = −0.364; $p = 0.001$) in the right-eye group. In the left-eye group, VFA area (r = −0.373; $p = 0.001$) and VFA flow area (r = −0.379; $p = 0.001$) were obtained. These microvascular alterations may represent an early step in hypertension-mediated organ damage, and thus, they can be markers in monitoring the beneficial effects of antihypertensive treatment, according to Rizzoni D, Agabiti-Rosei C, Boari G.E.M., Muiesan ML, and De Ciuceis C. [10].

In our study conducted in a hypertensive retinopathy-based sample with OCTA performed in both eyes, higher NFA area and FAZ parameters (area, perimeter, and circularity) and lower VFA (area and flow area), density, and skeleton (total and ETDRS) were observed in hypertensive subjects.

In participants with higher FAZ area and perimeter, a reduction in VFA flow area, density, and skeleton was also observed. This shows that an increase in areas with no perfusion is associated with a decrease in blood flow and a decrease in perfused blood vessels.

Subjects with higher VFA (area and flow area) showed higher density and skeleton values, meaning that subjects with an increased area of blood vessels and an increased blood flow have a larger area of perfused vasculature.

Using the Pearson Correlation and Sig. (2-tailed) test, we discovered that higher density (total and ETDRS parameters) was closely connected with higher skeleton (total and ETDRS) in hypertensive individuals.

Analyzing both eyes groups with hypercholesterolemia, in the right-eye group, we found correlation between VFA area and HDL only to be statistically significant, while in the left-eye group, statistically significant correlations were determined for total cholesterol, HLD, LDL, and multiple OCTA parameters. Our results showed that higher values of total cholesterol and LDL were associated with increasing avascular areas, reduced vascular flow, and reduced vascular density.

In our study, no statistically significant correlations between hypertriglyceridemia and OCTA parameters in the right-eye and left-eye groups were found as well as for BP and hypertriglyceridemia.

We found no statistical significance between right-eye OCTA parameters and left-eye OCTA parameters in our study. The dependent *t*-test determined a value of $p > 0.05$.

OCTA is a more suitable technique for imaging the retinal microvasculature than the conventional invasive injectable dye angiography because OCTA allows for more efficient, rapid, non-invasive, and widely available imaging. Fluorescein angiography is unsuitable for use in large cohort studies and for consecutive follow-up visits. Compared to OCTA of the retina, no other widely available imaging modality has yet been able to measure microvasculature with such good resolution, reproducibility, and short examination time [20,21].

Furthermore, studying changes in the microvasculature on OCTA in relation to systemic vascular disease may have the potential to detect early disease, to monitor disease progression, to potentially provide a better staging of hypertensive retinopathy, and to detect improvements in vasculature in patients with a better response to treatment.

Patients with progressive retinopathy and dyslipidemia are susceptible to coronary artery disease, peripheral vascular disease, and cerebrovascular events due to similar damage in other end-organ systems [22].

Ong et al. noted in the Atherosclerosis Risk in Communities (ARIC) Study, which included 15,792 participants, that hypertensive retinopathy was associated with an increased risk of stroke, independent of other risk factors. They recommended the use of fundus examination to help assess the risk of stroke in these patients [23].

Chen and co-workers studied a number of 9793 hypertensive patients within the China Stroke Primary Prevention Trial. Of these, 592 patients suffered from a stroke at the end of a 4.4-year follow-up, and a significant association was found between the presence of hypertensive retinopathy and the risk of a cerebrovascular event [24]. Hua D, Xu Y, Zeng X, et al. noted that the microvascular structure of the retina and the cerebral and cardiac tissue have multiple similarities, so fundus examination may have an important role in the

detection of end-organ damage. They suggested that the use of OCTA could increase the prediction of end-organ damage [4].

Additional damage to the eye is also possible. The persistence of arteriolar sclerosis can potentially lead to an increased incidence of arteriolar vein occlusions within the eye, as observed by Kotruchin P, Tangpaisarn T, Mitsungnern T, et al. [24].

Our findings were similar to the study performed by Chua et al. and by Sun et al. in that all our analyses confirmed a decrease in the deep capillary density in patients with systemic hypertension [6,11]. However, the difference with the study carried out by Sun et al. is that in our study did not determine intraocular pressure, visual acuity, and spherical equivalent, and our study did not compare the characteristics between groups, i.e., control and hypertensive. In addition, our study did not correlate choriocapillaris microvasculature with BP and renal parameters, which was analyzed by Chua et al. in their study. We had no control group, unlike other studies. Similar to the study performed by C Cuspidi et al. on non-diabetic hypertensive patients, retinal lesions were correlated with blood pressure, age, body mass index, smoking habit, and total serum cholesterol [7]. The difference is that we used OCTA as a tool for monitoring retinopathy progression, and we performed correlations between OCTA parameters and dyslipidemia.

In 2022, Rebecca Zeng et al. performed an observational study on retinal microvasculature and vasoreactivity changes in hypertension using OCTA [25]. Their study concluded that OCTA showed significant differences in the retinal vasculature of hypertensive patients, and it was also the first to demonstrate the potential of OCTA to investigate retinal vascular reactivity in patients with hypertension. Vessel density, vessel skeletonized density, and fractal dimension were calculated. A total of 23 eyes with hypertension (17 patients) and 17 control eyes (15 patients) were included, and a 6 mm $\times$ 6 mm angiogram centered on the fovea was performed. The superficial capillary plexus, vascular density, and the choriocapillaris were significantly decreased in hypertensive patients compared to control eyes. Our study had similar finding on microvascular changes, but we calculated the correlation between progressive retinopathy and microvascular changes in patients with dyslipidemia, with a larger number of subjects and without a control group. We did not study cerebral and cardiac tissue, like Hua D, Xu Y, Zeng X, et al. [26], and over a period of 18 months of monitoring our patients, we did not record any subjects with stroke, unlike Chen and co-workers [4].

A.I. computer systems are used extensively in medical sciences for diagnosing patients, discovering new drugs, improving communication with patients, and remotely treating patients [27]. In 2016, Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. developed and trained A.I. using 128,175 retinal fundus images to classify images as diabetic retinopathy and macular edema for adults with diabetes [28].

The new version of the Optopol OCT provides an A.I. algorithm for posterior-layer segmentations. By improving the current A.I. system and using OCTA findings, our research could provide new OCTA parameters for automated measurements, could help in early-stage detection of progressive hypertensive retinopathy, and could help develop screening programs, leading to increased efficiency in diagnosing patients in shorter time.

Therefore, based on our statistically significant results, we propose a few OCTA correlations to be analyzed using A.I., thus potentially resulting in new OCTA parameters: correlation between a lower VFA area and a lower HDL, a higher FAZ circularity with a higher LDL and total cholesterol, a higher LDL with a higher NFA area, and a lower skeleton total with a lower HDL.

However, there are a few limitations to this study. Firstly, we did not determine the values of apolipoprotein A1 and apolipoprotein B in correlation with OCTA parameters. Secondly, our changes in retinal vasculature were not correlated with renal function, heart function, other peripheral microvasculature parameters, or other end-organ damage measurements. Our current OCTA imaging (Angio-OCT Optopol Revo NX 130) depends on motion contrast from blood flow to visualize the retinal vasculature; hence, vessels with flow below the detectable threshold were not displayed. We acknowledge that our manual

measurement performed in order to delimit FAZ, NFA, and VFA may not be as exact as the automated measurements performed by the OCTA, but this is unlikely to significantly affect our results. Also, our cohort only included 77 participants. Consecutive follow-up visits are required after our 18-month observational study.

We used the OCTA as a tool for monitoring retinopathy associated with dyslipidemia, but other complementary methods, such as fluorescein angiography, suitable for use in small cohort studies, and indocyanine green angiography (ICGA), which is another diagnostic procedure used to examine choroidal blood flow and associated pathology, should be required in order to obtain more precise results.

In our study, our primary objective was to use OCTA as a tool for analyzing the progression of hypertensive retinopathy in patients with dyslipidemia. Our secondary objective was to suggest that an improvement of the current A.I. system of the OCTA can be made.

The current A.I. system has a direct impact on the accuracy of the clinical assessment and the assessment of the status of areas of pathology in the retina. This level of detection is based on the A.I. Retina layer segmentation and MIP algorithm, thus resulting in more accurate results and more detailed screening. A.I. segmentation is important for follow-up examinations, bringing a more accurate diagnosis when analyzing pathology over time. The MIP tool enables easier identification and tracking of blood vessels. We suggest updating the current A.I system by introducing mathematic formulas in the system to show that some OCTA parameters, lipid levels, and blood pressure are directly proportional to each other, while some OCTA variables are inversely proportional to blood pressure and lipid levels, thus resulting in a more accurate evaluation and monitoring of retinal pathology in patients with dyslipidemia and hypertension.

5. Conclusions

In conclusion, the present findings confirmed the correlation between hypertensive retinopathy, dyslipidemia, and OCTA parameters, and apart from blood pressure reduction, targeting the microcirculation itself could be beneficial in preventing or reducing end-organ damage and thus reducing mortality and morbidity.

OCTA is a useful tool in detecting in detail the various alterations in retinal microvasculature [14,17,21,25,29]. Earlier detection upon OCTA of the small blood vessel changes, prior to the appearance of changes on the fundus examination, can provide a more accurate staging of the hypertensive retinopathy and can help us to better understand the pathophysiology involved.

Further studies, particularly with a lager cohort, are needed to test the effectiveness of retinal capillary density as a novel biomarker in predicting the incidence and progression of hypertension-related microvascular complications.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/jcm13247584/s1, Table S1. Matrix Plot. Correlation between blood pressure (systolic and diastolic) and clinical variables; Table S2. Matrix Plot. Correlation between blood pressure (systolic and diastolic) and Right-Eye OCTA parameters; Table S3. Matrix Plot. Correlation between blood pressure (systolic and diastolic) and Left-Eye OCTA parameters.

Author Contributions: Conceptualization, I.C.B. and S.S.A.; methodology, I.C.B., V.P. and S.S.A.; software, I.C.B.; validation, V.P. and S.S.A.; formal analysis, I.C.B.; investigation, I.C.B.; resources, I.C.B.; writing, I.C.B.; writing—review and editing, I.C.B., V.P. and S.S.A.; supervision, V.P. and S.S.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: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the "Regina Maria: The Healthcare Network". Approval Date: 14 December 2022.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data that support the findings of the present study are available from the corresponding author (G.S.) upon reasonable request.

Acknowledgments: We thank Nicu Duna for the technical support.

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

References

1. Pappan, N.; Awosika, A.O.; Rehman, A. Dyslipidemia. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2024.
2. Williams, B.; Mancia, G.; Spiering, W.; Rosei, E.A.; Azizi, M.; Burnier, M.; Clement, D.L.; Coca, A.; de Simone, G.; Dominiczak, A.; et al. 2018 ESC/ESH Guidelines for the management of arterial hypertension. The Task Force for the management of arterial hypertension of the European Society of Cardiology (ESC) and the European Society of Hypertension (ESH). *G. Ital. Cardiol.* **2018**, *19* (Suppl. S1), 3S–73S.
3. Lee, W.H.; Park, J.H.; Won, Y.; Lee, M.W.; Shin, Y.I.; Jo, Y.J.; Kim, J.Y. Retinal Microvascular Change in Hypertension as measured by Optical Coherence Tomography Angiography. *Sci. Rep.* **2019**, *9*, 156.
4. Chen, X.; Liu, L.; Liu, M.; Huang, X.; Meng, Y.; She, H.; Zhao, L.; Zhang, J.; Zhang, Y.; Gu, X.; et al. Hypertensive Retinopathy and the Risk of Stroke Among Hypertensive Adults in China. *Investig. Ophthalmol. Vis. Sci.* **2021**, *62*, 28. [CrossRef] [PubMed]
5. Cheung, C.Y.; Biousse, V.; Keane, P.A.; Schiffrin, E.L.; Wong, T.Y. Hypertensive eye disease. *Nat. Rev. Dis. Primers* **2022**, *8*, 14. [PubMed]
6. Chua, J.; Le, T.T.; Tan, B.; Ke, M.; Li, C.; Wong, D.W.K.; Tan, A.C.S.; Lamoureux, E.; Wong, T.Y.; Chin, C.W.L.; et al. Choriocapillaris microvasculature dysfunction in systemic hypertension. *Sci. Rep.* **2021**, *11*, 4603. [CrossRef] [PubMed]
7. Cuspidi, C.; Meani, S.; Valerio, C.; Fusi, V.; Catini, E.; Sala, C.; Magrini, F.; Zanchetti, A. Prevalence and correlates of advanced retinopathy in a large selected hypertensive population. The Evaluation of Target Organ Damage in Hypertension (ETODH) study. *Blood Press.* **2005**, *14*, 25–31.
8. Dziedziak, J.; Zaleska-Zmijewska, A.; Szaflik, J.P.; Cudnoch-Jedrzejewska, A. Impact of Arterial Hypertension on the Eye: A Review of the Pathogenesis, Diagnostic Methods, and Treatment of Hypertensive Retinopathy. *Med. Sci. Monit.* **2022**, *28*, e935135.
9. Jiao, L.; Lv, C.; Zhang, H. Effect of blood pressure variability on hypertensive retinopathy. *Clin. Exp. Hypertens* **2023**, *45*, 2205050. [PubMed]
10. Rizzoni, D.; Agabiti-Rosei, C.; Boari, G.E.M.; Muiesan, M.L.; De Ciuceis, C. Microcirculation in Hypertension: A Therapeutic Target to Prevent Cardiovascular Disease? *J. Clin. Med.* **2023**, *12*, 4892. [CrossRef]
11. Sun, C.; Ladores, C.; Hong, J.; Nguyen, D.Q.; Chua, J.; Ting, D.; Schmetterer, L.; Wong, T.Y.; Cheng, C.-Y.; Tan, A.C.S. Systemic hypertension associated retinal microvascular changes can be detected with optical coherence tomography angiography. *Sci. Rep.* **2020**, *10*, 9580.
12. Niro, A.; Sborgia, G.; Lampignano, L.; Giuliani, G.; Castellana, F.; Zupo, R.; Bortone, I.; Puzo, P.; Pascale, A.; Pastore, V.; et al. Association of Neuroretinal Thinning and Microvascular Changes with Hypertension in an Older Population in Southern Italy. *J. Clin. Med.* **2022**, *11*, 1098. [CrossRef]
13. Wong, W.; Gopal, L.; Yip, C.C. Hypertensive retinopathy and choroidopathy. *Can. Med Assoc. J.* **2020**, *192*, E371. [CrossRef] [PubMed]
14. Jia, Y.; Bailey, S.T.; Hwang, T.S.; McClintic, S.M.; Gao, S.S.; Pennesi, M.E.; Flaxel, C.J.; Lauer, A.K.; Wilson, D.J.; Hornegger, J.; et al. Quantitative optical coherence tomography angiography of vascular abnormalities in the living human eye. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E2395–E2402. [CrossRef] [PubMed]
15. Rotsos, T.; Andreanos, K.; Blounas, S.; Brouzas, D.; Ladas, D.S.; Ladas, I.D. Multimodal imaging of hypertensive chorioretinopathy by swept-source optical coherence tomography and optical coherence tomography angiography: Case report. *Medicine* **2017**, *96*, e8110. [CrossRef]
16. Takayama, K.; Kaneko, H.; Ito, Y.; Kataoka, K.; Iwase, T.; Yasuma, T.; Matsuura, T.; Tsunekawa, T.; Shimizu, H.; Suzumura, A.; et al. Novel Classification of Early-stage Systemic Hypertensive Changes in Human Retina Based on OCTA Measurement of Choriocapillaris. *Sci. Rep.* **2018**, *8*, 15163. [CrossRef] [PubMed]
17. Mirshahi, A.; Karkhaneh, R.; Roohipour, R.; Rajabi, M.; Vahedian, Z.; Bazvand, F. Optical Coherence Tomography Angiography Findings in Malignant Hypertensive Retinopathy. *J. Ophthalmic Vis. Res.* **2022**, *17*, 432–436.
18. Downie, L.E.; Hodgson, L.A.; Dsylva, C.; McIntosh, R.L.; Rogers, S.L.; Connell, P.; Wong, T.Y. Hypertensive retinopathy: Comparing the Keith-Wagener-Barker to a simplified classification. *J. Hypertens* **2013**, *31*, 960–965. [CrossRef]
19. Zhang, W.; He, K.; Zhao, H.; Hu, X.; Yin, C.; Zhao, X.; Shi, S. Association of body mass index and waist circumference with high blood pressure in older adults. *BMC Geriatr.* **2021**, *21*, 260. [CrossRef] [PubMed]
20. Levy, B.I.; Ambrosio, G.; Pries, A.R.; Struijker-Boudier, H.A. Microcirculation in hypertension: A new target for treatment? *Circulation* **2001**, *104*, 735–740. [CrossRef] [PubMed]
21. Tan, A.C.S.; Tan, G.S.; Denniston, A.K.; Keane, P.A.; Ang, M.; Milea, D.; Chakravarthy, U.; Cheung, C.M.G. An overview of the clinical applications of optical coherence tomography angiography. *Eye* **2018**, *32*, 262–286. [CrossRef]
22. Mehta, S. Optical Coherence Tomography Angiography Findings Associated with Accelerated Hypertension. *Cureus* **2024**, *16*, e59290. [CrossRef]

23. Ong, Y.T.; Wong, T.Y.; Klein, R.; Klein, B.E.; Mitchell, P.; Sharrett, A.R.; Couper, D.J.; Ikram, M.K. Hypertensive retinopathy and risk of stroke. *Hypertension* **2013**, *62*, 706–711. [CrossRef]
24. Kotruchin, P.; Tangpaisarn, T.; Mitsungnern, T.; Sukonthasarn, A.; Hoshide, S.; Turana, Y.; Siddique, S.; Buranakitjaroen, P.; Van Huynh, M.; Chia, Y.C.; et al. Hypertensive emergencies in Asia: A brief review. *J. Clin. Hypertens* **2022**, *24*, 1226–1235. [CrossRef] [PubMed]
25. Zeng, R.; Garg, I.; Bannai, D.; Kasetty, M.; Katz, R.; Park, J.Y.; Lizano, P.; Miller, J.B. Retinal microvasculature and vasoreactivity changes in hypertension using optical coherence tomography-angiography. *Graefes Arch. Clin. Exp. Ophthalmol.* **2022**, *260*, 3505–3515. [CrossRef]
26. Hua, D.; Xu, Y.; Zeng, X.; Yang, N.; Jiang, M.; Zhang, X.; Yang, J.; He, T.; Xing, Y. Use of optical coherence tomography angiography for assessment of microvascular changes in the macula and optic nerve head in hypertensive patients without hypertensive retinopathy. *Microvasc. Res.* **2020**, *129*, 103969. [CrossRef]
27. Basu, K.; Sinha, R.; Ong, A.; Basu, T. Artificial Intelligence: How is It Changing Medical Sciences and Its Future? *Indian J. Dermatol.* **2020**, *65*, 365–370. [CrossRef]
28. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **2016**, *316*, 2402–2410. [CrossRef] [PubMed]
29. Koustenis, A., Jr.; Harris, A.; Gross, J.; Januleviciene, I.; Shah, A.; Siesky, B. Optical coherence tomography angiography: An overview of the technology and an assessment of applications for clinical research. *Br. J. Ophthalmol.* **2017**, *101*, 16–20. [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.

tomography

MDPI

Article

BAE-ViT: An Efficient Multimodal Vision Transformer for Bone Age Estimation

Jinnian Zhang [1,†], Weijie Chen [1,†], Tanmayee Joshi [1], Xiaomin Zhang [2], Po-Ling Loh [3], Varun Jog [3], Richard J. Bruce [4], John W. Garrett [4] and Alan B. McMillan [1,4,5,6,*]

[1] Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI 53706, USA
[2] Department of Computer Science, University of Wisconsin-Madison, Madison, WI 53706, USA
[3] Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge CB2 1TN, UK
[4] Department of Radiology, University of Wisconsin-Madison, Madison, WI 53706, USA
[5] Department of Medical Physics, University of Wisconsin-Madison, Madison, WI 53706, USA
[6] Department of Biomedical Engineering, University of Wisconsin-Madison, Madison, WI 53706, USA
* Correspondence: abmcmillan@wisc.edu
† These authors contributed equally to this work.

Abstract: This research introduces BAE-ViT, a specialized vision transformer model developed for bone age estimation (BAE). This model is designed to efficiently merge image and sex data, a capability not present in traditional convolutional neural networks (CNNs). BAE-ViT employs a novel data fusion method to facilitate detailed interactions between visual and non-visual data by tokenizing non-visual information and concatenating all tokens (visual or non-visual) as the input to the model. The model underwent training on a large-scale dataset from the 2017 RSNA Pediatric Bone Age Machine Learning Challenge, where it exhibited commendable performance, particularly excelling in handling image distortions compared to existing models. The effectiveness of BAE-ViT was further affirmed through statistical analysis, demonstrating a strong correlation with the actual ground-truth labels. This study contributes to the field by showcasing the potential of vision transformers as a viable option for integrating multimodal data in medical imaging applications, specifically emphasizing their capacity to incorporate non-visual elements like sex information into the framework. This tokenization method not only demonstrates superior performance in this specific task but also offers a versatile framework for integrating multimodal data in medical imaging applications.

Keywords: bone age regression; machine learning; vision transformer; gender embedding; multimodal data

1. Introduction

Bone age assessment is one of the most common radiological studies, and is essential for experts to determine growth disorders from differences between bone and chronological age. Historically, skeletal development was assessed manually from metacarpals, phalanges, and carpals in X-rays using techniques such as Greulich–Pyle [1] and Tanner–Whitehouse [2]. However, these methods rely heavily on radiologists' experience, leading to massive time and labor costs.

Recently, automatic bone age estimation (BAE) using deep learning has demonstrated superior efficiency and accuracy. The existing literature may be categorized into two types: methods that focus on specific anatomical Regions of Interest (ROI-based) and those that analyze entire images (image-based methods). ROI methods first extract image features of the palm or local bone parts using experts [3] or deep networks [4], which are then used for BAE by convolutional neural networks (CNNs) [5], residual attention networks [6], or

regression models [7]. Although ROI methods are accurate, feature extraction requires much manual labor. The multi-stage pipeline also has low prediction efficiency, limiting clinical applicability. Image-based methods directly feed images into CNNs [8–11], which is efficient for both training and testing. The winning solution for the RSNA Pediatric Bone Age Challenge 2017 [12] used an InceptionV3-based model with a mean absolute error (MAE) of 4.2 months.

Sex information is important for BAE [13], and is often concatenated to image features either directly [14] or after being converted to an embedding vector by linear layers [15]. The combined features are then fed into a multilayer perceptron (MLP) for prediction. Notably, this method does not permit interactions between sex and image information.

Machine learning techniques have also been used to enhance traditional methods of bone age assessment by automating feature analysis. Subtle skeletal features, such as cortical decalcification and microfractures in CT scans, have been analyzed using texture analysis and deep learning methods to refine age and sex detection [16]. Texture-based local binary patterns and spatial attention mechanisms have been integrated into deep residual networks, addressing challenges specific to ethnicity in datasets [17]. Multiscale imaging approaches have been employed to study skeletal aging by combining traditional radiography with advanced microscopy techniques [18], while MRI-based models have been developed as non-invasive alternatives for age estimation through textural feature extraction [19].

Recently, vision transformers (ViTs) [20–22] have shown great potential in computer vision. Unlike CNNs, characterized by locality, the self-attention module [23] in ViTs allows each pixel in the input image to interact with all other pixels. ViTs can further handle multi-source inputs by converting them into sequences of tokens, making them a popular choice for multimodal data [24–26].

To the best of our knowledge, there is no existing work exploring multimodal ViTs in BAE. In this paper, we design a multimodal vision transformer, BAE-ViT, that integrates both image and sex information more efficiently than CNNs, and demonstrate its improved performance on normal and challenging samples.

2. Materials and Methods

2.1. Dataset

We consider two datasets of left-hand radiographs. The first dataset is publicly available from the 2017 RSNA Pediatric Bone Age Challenge. There are 12,611 training images (6833 male, 5778 female), 1425 validation images (773 male, 652 female), and 200 test images (100 male, 100 female). Based on the study in [27], this dataset is distinguished as one of the largest and most comprehensive, extensively published and validated, which enhances the reproducibility and applicability of our research. For training and validation, images were sourced from the Children's Hospital Colorado (Aurora, Colo) and Lucile Packard Children's Hospital at Stanford, while the test set images were exclusively obtained from the Lucile Packard Children's Hospital [12], making this a multi-site dataset. The second dataset is an external validation dataset consisting of data from our own institution from 100 different patients (61 male, 39 female). This retrospective study was approved by our Institutional Review Board. The need for informed consent was waived for this Health Insurance Portability and Accountability Act-compliant study.

2.2. Models

Two traditional model architectures for BAE are shown in Figure 1a,b. The regression model only uses CNNs or ViTs to obtain image features and a linear layer for regression. Sex information is ignored. The ensemble model has two branches to encode image and sex information, which are concatenated as inputs to three linear layers for prediction. However, extracting features from different data types using independent branches may be suboptimal. Therefore, we convert image and sex information into tokens of the same dimension, which interact in the transformer blocks and are aggregated and fed into the

linear regressor in Figure 1c. The hyperparameters used were those recommended by the original model designers, leveraging their extensive expertise and experimentation, while ensuring consistency and fairness in comparisons by selecting configurations with a similar number of parameters.

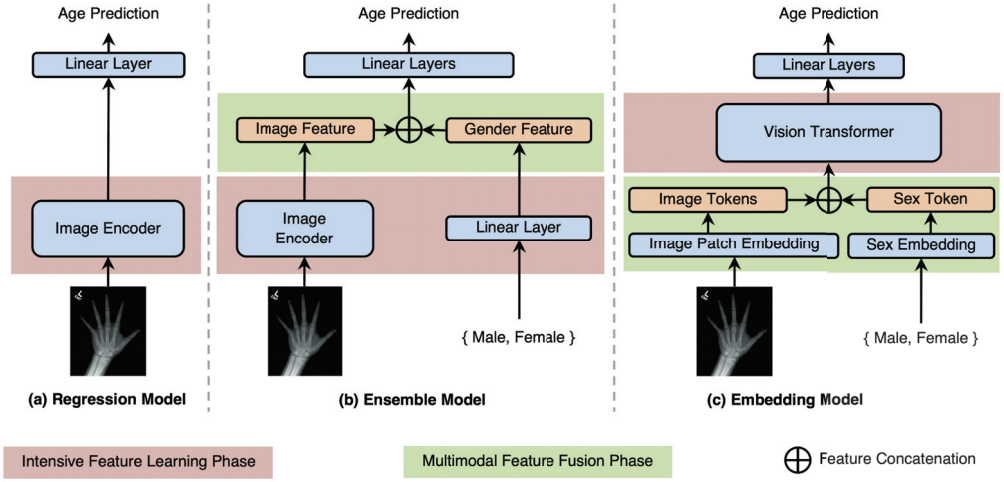


Figure 1. A comparison of architectures between the regression, ensemble, and embedding models. The red area indicates the intensive feature learning phase (red), which requires significant computational resources, while the green area indicates the multimodal feature fusion phase (green), where image features are integrated with non-visual features such as sex. (**a**) Regression model: This model only takes images as inputs, without using biological sex information. (**b**) Ensemble model: In this model, two branches encode the image and sex information into feature vectors separately. These features are concatenated and fed into linear layers for bone age estimation. The biological sex information is integrated after the image encoder and processed only by the linear layer. (**c**) Embedding model (proposed): Our proposed multimodal vision transformer converts both the image and sex information into tokens. These tokens interact in the transformer blocks through attention mechanisms and are finally projected through a linear layer for predictions.

In our methodology, we integrate visual and non-visual data, leveraging the strengths of ViTs for their comprehensive data processing capabilities, a task for which the Swin Transformer and DeiT model's distillation strategy are particularly suited. The Swin Transformer, customized for the nuanced processing of image data, and the DeiT's efficiency-focused distillation approach, pave the way for TinyViT, a model that embodies the essence of rapid learning through progressive distillation from large, pretrained models. This model harmonizes performance with computational efficiency, a testament to its sophisticated design. Additionally, we employ an architectural innovation by utilizing inverted residual blocks, which adopt a narrow -> wide -> narrow channel structure, contrasting with traditional residual blocks. This choice is complemented by the incorporation of lightweight and efficient MBConv layers in the initial stages of our model, optimizing the learning of low-level representations through their strong inductive biases, thereby enhancing both the efficiency and effectiveness of our integrated approach.

From a methodological perspective, tokenization is a process that transforms multimodal data into uniform vectors, facilitating cross-validation among features from diverse modalities. Typically, features from images are extracted through convolutional blocks, while sex information is derived using linear layer projection. The specific technical details are elaborated upon in the following sections.

Our proposed multimodal ViT architecture is illustrated in Figure 2. We follow the design of TinyViT-21M [28], which is more efficient than DeiT-S [21] and Swin-T [20], and has even better performance than Swin-L on ImageNet-1k [29]. It consists of a patch-embedding layer, patch merging layers, and transformer blocks with shifted windows.

Additionally, we generate a novel sex embedding using a linear layer with 2C hidden nodes (the dimension of the image tokens).

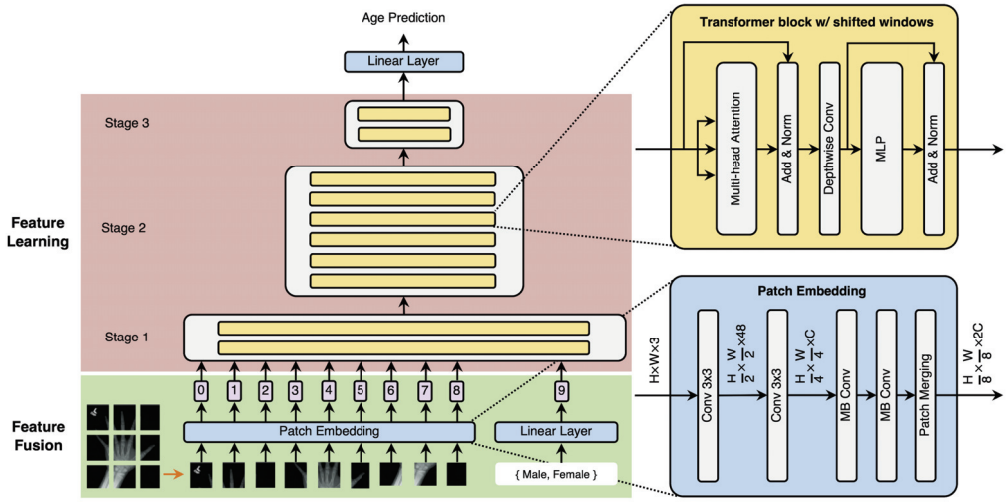


Figure 2. BAE-ViT architecture with biological sex embedding. This diagram illustrates the three-stage design of the BAE-ViT architecture. In the feature fusion phase (green), the model uses patch-embedding, convolutions, and MBConv blocks, followed by patch merging to create token sequences. The biological sex information is tokenized through a linear layer and processed by the transformer alongside other visual patch tokens. In the feature learning phase (red), stages 1, 2, and 3 involve transformer blocks with shifted window attention, processing features at varying scales. The architecture is designed for efficient feature extraction and integrates sex information, facilitating enhanced classification performance.

Considering the default input dimension for a visual transformer, the image is initially resized or randomly cropped to dimensions of 224×224 pixels. We obtain the image tokens through the patch-embedding layer, involving two 3×3 convolutional layers with stride 3 to reduce the input resolution from (H, W) to $(H/4, W/4)$ and increase the channel size from 3 to C. After two inverted residual blocks (also known as MBConv) [30], we use the patch-merging layer to decrease the resolution of feature maps to $(H/8, W/8)$, in which there are two pointwise convolutional layers and one depthwise convolutional layer in the middle with kernel size 3 and stride 2. Finally, the feature maps are rearranged to $HW/64$ image tokens with dimension 2C. Combined with the sex token, all $HW/64 + 1$ tokens are fed into the transformer blocks.

Traditional transformer blocks contain multi-head attention [23], layer normalization [31], and MLP. However, the global computational complexity is quadratic in the number of tokens N. To improve efficiency, TinyViT introduces shifted windows from Swin-T [20] to separate image tokens into groups. Attention is performed within each group:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T/\sqrt{d} + B)V, \tag{1}$$

where $Q, K, V \in \mathbb{R}^{M^2 \times d}$ are the query, key, and value matrices; d is the hidden dimension; and M^2 is the group size. B is the relative position bias matrix taken from a learnable matrix $\hat{B} \in \mathbb{R}^{(2M+1)\times(2M+1)}$ according to the relative position of each token pair in each axis, lying in $[-M+1, M-1]$. This reduces the computational complexity to $\mathbb{O}(NM^2)$. Notably, the group partition is shifted by half the group size in adjacent blocks to avoid performance degradation due to lack of interaction across groups. Moreover, a depthwise convolution with kernel size 3 and stride 1 is applied to capture local features to supplement global attention. In BAE-ViT, the sex token is included in each group for interaction. Therefore, Q,K,V have dimensions $(M^2 + 1) \times d$ in the attention module of transformer blocks. We fix the last row and column in $B \in \mathbb{R}^{(M^2+1)\times d}$ to 0, indicating no relative positional bias in

the sex token. We did not observe performance improvements after introducing relative positional bias.

2.3. Evaluation

We compare BAE-ViT to regression and ensemble models using three different CNNs (InceptionV3 [32], ResNet50 [33], and EfficientNet-B5 [34]) and TinyViT as the image encoder, after removing their original classification head. In the ensemble model, we fix the size of the sex feature vector as 8 and the hidden dimension of the first two linear layers as 1000. The last linear layer is a single neuron.

For model evaluation, we calculated the MAE between model predictions and the ground truth. Each test image was resized to make the shorter edge equal to the input resolution, and we obtained a square crop of the resized image at the center along the longer edge. We also adopted a multi-crop test method from [35]: for each test image, we obtained 10 random square crops with the given resolution and fed them into the model. The median of the outputs was used as the final prediction. The number of random crops needs to be sufficiently large to yield stable predictions. The adoption of a multi-crop testing methodology, specifically the utilization of 10 random square crops, represents a standard and widely recognized approach in the evaluation of deep learning models, particularly in the domain of image classification. This methodological choice is predicated on the objective of enhancing model robustness and generalization by simulating a diverse array of real-world imaging conditions. By incorporating crops from the four corners and center of the image, along with their horizontal flips, this approach systematically evaluates the model's performance across varied spatial presentations, thereby mimicking the multifaceted nature of real-world visual perception. Such a strategy not only aims to mitigate potential overfitting by challenging the model with less predictable image segments but also leverages ensemble learning principles, wherein the aggregation of predictions across these crops serves to bolster prediction accuracy and stability. This comprehensive evaluation framework thus ensures that the model's high performance is indicative of genuine learning rather than mere memorization of training data biases. Despite the increased computational demands associated with processing multiple crops per image, the adoption of this popular and standardized method is justified by its significant contributions to improving the depth and reliability of model evaluations, thereby facilitating the development of more adaptable and dependable image classification models.

To evaluate the influence of demographic attributes on the model evaluation, we designed a label perturbation experiment during the testing stage. In this experiment, biological sex labels in the RSNA dataset were intentionally perturbed by assigning incorrect values. This altered demographic information was then used to assess model performance on X-ray scans, allowing us to observe the impact of such inconsistencies on prediction accuracy.

In our study, we employ ScoreCAM [36] for generating class activation maps due to its distinct advantages over gradient-based methods like Grad-CAM [37]. ScoreCAM's reliance on forward-pass activations, rather than gradients, offers clearer and more interpretable visual explanations of model decisions by directly assessing the impact of different image regions on the model's output. This approach mitigates issues associated with gradient-based visualizations, such as noisy gradients, providing a more accurate reflection of the model's focus areas. Although ScoreCAM demands higher computational resources due to its need for multiple forward passes, its capacity for producing more faithful and interpretable activation maps justifies its selection in our methodology, enhancing the transparency and accuracy of our model evaluations.

In our analysis, the Wilcoxon signed-rank test was employed to identify statistically significant differences between model predictions and ground-truth labels, with a predefined significance level set at 0.05. This choice reflects standard practice in statistical analysis, balancing the detection of genuine effects against the risk of false discoveries. To address the concern of increased type I errors due to multiple comparisons, the Bonferroni

correction method was rigorously applied, thereby adjusting the significance thresholds to a more stringent criterion. This step is crucial in preserving the integrity of our findings amid multiple statistical tests. It is pertinent to note, however, that while the Wilcoxon signed-rank test is robust to the non-normality of data, it operates under the assumption that the differences between matched pairs are symmetrically distributed around the median. This assumption is a potential limitation of the method, suggesting cautious interpretation of results when the distribution of differences is expected to deviate from symmetry. Our adherence to these statistical protocols underscores our commitment to methodological rigor and the reliability of our conclusions, within the acknowledged constraints of the chosen statistical tests.

3. Results

All models were trained on the RSNA training dataset; the RSNA validation dataset was used to optimize the hyperparameters. The trained models were evaluated on both the RSNA and external datasets. All experiments were conducted on two NVIDIA A100 GPUs (NVIDIA Corporation, Santa Clara, CA, USA) with 80 GB memory using PyTorch (software version 1.7.1, created by Meta Platforms, Inc., Menlo Park, CA, USA). For image preprocessing, we obtained the mean and standard deviation for z-score normalization based on the RSNA dataset. Data augmentation techniques included RandAugment [38] and random erasing [39]. For training, we inherited the hyperparameters from Swin-T: when using CNNs as image encoders, we set the learning rate to 2.5×10^{-4}, batch size to 256, and number of epochs to 500. All models were trained from scratch. For the ensemble model that uses TinyViT, we used the same settings. The dropout rate for all ensemble models was 0.5. For the regression model with TinyViT as the image encoder, we used a learning rate of 3.1×10^{-5}, batch size of 32, and number of epochs of 300. Weights pretrained on ImageNet-1k were used for initialization. The training settings for BAE-ViT were the same as the regression model with TinyViT.

3.1. Comparison of Performance of CNNs and ViTs

Table 1 summarizes the performance of our proposed multimodal BAE-ViT and the regression and ensemble models with CNNs and ViTs as the image encoder. We observe that sex information is critical for BAE: the ensemble model outperforms the regression model by more than 1.0 month in MAE. This conforms to the expectation that models are prone to fitting the average age of images with similar skeletal maturity, but since females generally achieve maturity nearly two years earlier than males, models that disregard sex information have larger errors.

Table 1. Performance comparison between our proposed BAE-ViT and the regression and ensemble models that use CNNs and ViTs as the image encoder. For regression models, we consider not only a single model for all images (Regression), but also two models for male images and female images separately (Regression-S). For Regression-S, we show the average MAE on both the whole test dataset and the male and female test data in months. The best performance is shown in bold.

Model		Param. No. (M)	Input Res. (px)	Sex (Y/N)	RSNA MAE (↓)		External MAE (↓)	
					Center	Multi-Crop	Center	Multi-Crop
RSNA Challenge winner		>24	500^2	Y	-	4.2	-	-
Ensemble-VGG 16		>138	600^2	Y	-	-	8.8	-
Inception-V3	Regression	25	500^2	N	6.1	5.7	8.1	8.0
	Regression-S	25	500^2	-	4.6 M: 4.4 F: 4.7	4.2 M: 4.0 F: 5.0	7.2 M: 7.4 F: 6.9	7.5 M: 7.5 F: 7.4
	Ensemble	25	500^2	Y	4.8	4.4	7.1	7.0

Table 1. *Cont.*

Model		Param. No. (M)	Input Res. (px)	Sex (Y/N)	RSNA MAE (↓)		External MAE (↓)	
					Center	Multi-Crop	Center	Multi-Crop
ResNet50	Regression	24	500^2	N	6.8	6.7	8.6	8.4
	Regression-S	24	500^2	-	5.0 M: 4.6 F: 5.3	4.5 M: 4.3 F: 4.7	7.6 M: 7.7 F: 7.4	7.7 M: 7.7 F: 7.6
	Ensemble	24	500^2	Y	**4.3**	4.2	7.1	7.2
EfficientNet-B5	Regression	28	456^2	N	6.6	5.7	8.1	8.0
	Regression-S	30	456^2	-	5.4 M: 4.9 F: 5.8	4.9 M: 4.5 F: 5.3	7.2 M: 7.5 F: 6.7	7.1 M: 7.1 F: 6.9
	Ensemble	30	456^2	Y	5.5	4.9	7.4	7.2
TinyViT	Regression	21	500^2	N	6.0	5.6	8.4	7.8
	Regression-S	21	512^2	-	4.6 M: 4.5 F: 4.7	4.4 M: 4.5 F: 4.3	7.0 M: 7.2 F: 6.7	7.1 M: 7.2 F: 6.8
	Ensemble	21	512^2	Y	4.9	4.7	6.9	7.0
BAE-ViT		21	512^2	Y	4.4	**4.1**	**6.7**	**6.9**

We found that using separate sex-specific models led to better performance than using a single model. The regression model for males always had a lower MAE than that of females on the RSNA test dataset, but the opposite was true on the external test dataset. This suggests that distribution shifts exist between the training and test datasets. Compared to the ensemble model, the average performance of separate models was slightly better when using Inception-V3, EfficientNet-B5, or TinyViT as the image encoder. A possible reason is that simply concatenating image and sex features is inefficient, so performance depends on the architecture of the image encoder. For example, when using ResNet50, the ensemble model achieved the best MAE on the RSNA dataset.

Our proposed BAE-ViT achieved the best performance on both the RSNA and external datasets with the multi-crop test method. When using only the center crop of test images, BAE-ViT was only 0.1 worse in MAE than the ensemble model with ResNet50 on the RSNA dataset. However, it performs much better on the external dataset, showing better generalization. This demonstrates that the interaction between image and sex information by using attention in transformer blocks is more efficient than the ensemble model.

3.2. Visualization of Heatmaps

We generate heatmaps for different age groups using ScoreCAM in Figure 3. Compared to CNNs, that have coarser receptive fields, ViTs can focus on key local regions due to the global attention of transformer blocks. Note that some models fail to attend to the correct region for images with a bone age less than 100 months, e.g., the ensemble model with Inception-V3 as the image encoder also attends to the border of images with a bone age of 34.2 (first column), 87.5 (sixth column), and 89.9 (second column). This may be due to a lack of training samples in this age group. Moreover, since sex information was used in all models, there is no obvious difference between the heatmaps for males and females. Notably, our proposed BAE-ViT additionally attends to the radius and ulna, which will be the focus of future research.

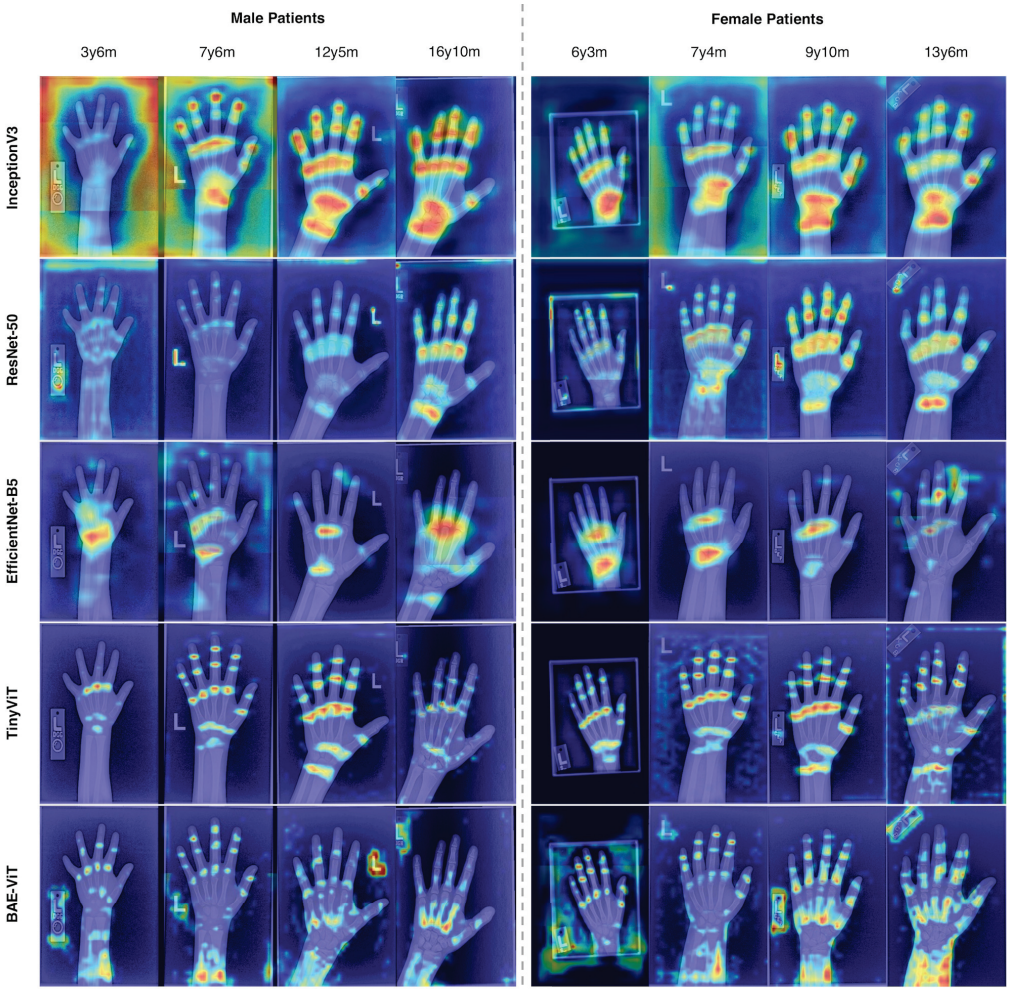


Figure 3. Heatmaps of our proposed BAE-ViT and ensemble models using CNNs or TinyViT as the image encoder by ScoreCAM. The left four columns are male with bone ages of 34.2, 89.9, 149.1, and 202.3 months, respectively. The right four columns are female with bone ages of 75.0, 87.5, 118.2, and 162.0 months, respectively. Heatmaps tend to highlight joints within the fingers and hand.

3.3. Performance of Models in Different Age Groups

Our BAE-ViT achieves the highest performance on the RSNA test dataset, with an average MAE of 4.1 s using the multi-crop test. The male MAE is 4.4 months and the female MAE is 3.9. Its predictions show a high agreement with the labels Figure 4a, with a mean bias of -0.66 and a standard deviation of 5.40 months (Figure 4b), slightly better than the ensemble model with ResNet50 (mean bias of -0.7 and standard deviation of 5.50 months). Based on a Wilcoxon signed-rank test, the statistical results showed no significant differences between model predictions and labels (p-values of ResNet50 and BAE-ViT are 0.2 and 0.4, respectively). Overall, BAE-ViT is accurate and reliable.

3.4. Training Techniques for Ensemble Models

We can train the ensemble model either from scratch or using a two-phase method, where we first pretrain the image encoder, and then train the model with the weights of the image encoder fixed or variable. This should intuitively lead to better performance. However, the results in Table 2 show that training from scratch yields a lower MAE than the two-phase method for different models as the image encoder. We investigate the curves of both training and test error, as shown in Figure 5. Although pretrained weights can accelerate the convergence rate, they lead to overfitting. Hence, the model is more sensitive to training hyperparameters.

Table 2. MAE scores in months for ensemble models with different image encoders, comparing end-to-end and two-phase training with fixed and non-fixed weights.

Models	End-to-End	Pretrained on ImageNet-1k		Pretrained on RSNA data	
		Fixed	Non-Fixed	Fixed	Non-Fixed
Inception-V3	4.8	9.5	5.2	5.3	4.9
ResNet50	4.5	9.7	5.1	5.8	5.2
TinyViT	4.9	9.4	5.8	5.2	5.8

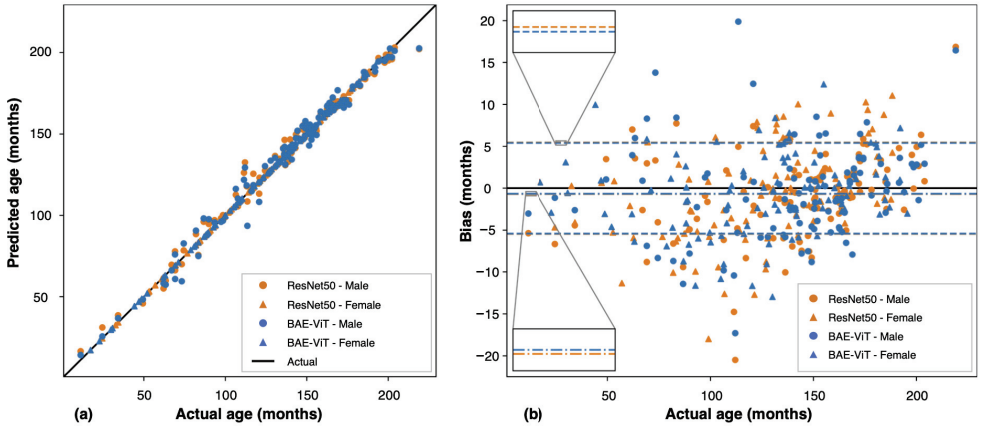


Figure 4. Performance comparison between our proposed BAE-ViT model and the ResNet50 ensemble model. (**a**) Correlation between actual and predicted bone ages, demonstrating high correlation for both models. (**b**) Mean bias and standard deviation of bias for the models, with dot-dashed lines representing mean bias and dashed lines indicating the standard deviation. The mean bias, defined as the signed average error, indicates whether the model's predictions are consistently higher or lower than the true values. Both models exhibit mean bias values close to zero (-0.66 for BAE-ViT and -0.70 for ResNet50), suggesting no significant overestimation or underestimation. BAE-ViT shows a slightly lower mean bias and standard deviation (5.40) compared to the ResNet50 ensemble model (5.50 standard deviation), indicating its superior accuracy and consistency in predictions.

For two-phase methods with a fixed image encoder, it is noteworthy that pretraining the image encoder on ImageNet-1k results in an MAE exceeding 4.0 months compared to pretraining on RSNA data. This discrepancy is likely due to the domain gap between ImageNet-1k and bone images. In contrast, for two-phase methods with a variable image encoder, the final MAE is closer between pretraining on the ImageNet-1k and RSNA datasets, indicating a reduced impact of the pretraining domain on the model's performance given the variable image encoder weights. Moreover, fixing the pretrained weights of the image encoder is generally less efficient. Therefore, even after pretraining, the image encoder needs to adjust the weights for better interaction with the sex branch to improve accuracy.

3.5. Robustness of Models over Bad Examples

We evaluate the robustness of BAE-ViT and different ensemble models over 44 selected images from different patients deemed to be of poor quality, e.g., containing multiple hands, the right hand, or bad positioning. Radiographers often need to retake poor-quality images, expending more resources. However, our proposed BAE-ViT shows better robustness on these bad examples than ensemble models, with only a 9.9 month MAE. Therefore, one can improve efficiency by avoiding image reacquisition even in the presence of low-quality data.

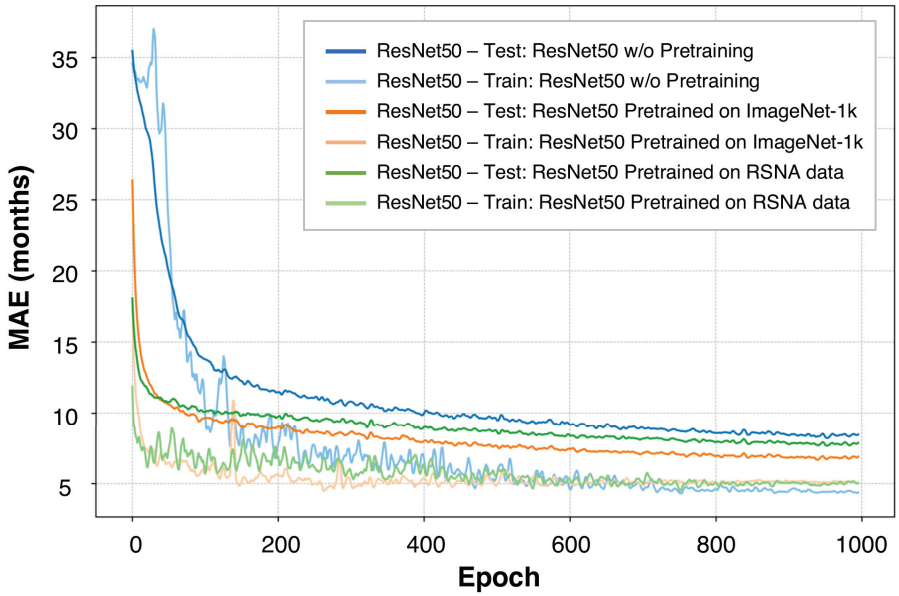


Figure 5. ResNet50 training and testing loss comparison. This figure presents the loss curves for a ResNet50 model under different training conditions. Lighter solid lines show the training loss, and darker dashed lines represent the testing loss. The graph compares end-to-end training with pretraining on ImageNet-1k and RSNA datasets. The lower testing loss across all methods could imply a limited diversity in the testing dataset compared to the training dataset.

3.6. Sensitivity to Demographic Label Perturbation

The label perturbation experiment revealed a substantial performance degradation when incorrect biological sex labels were provided. Specifically, the mean absolute error (MAE) increased dramatically from 4.1 months to 20.9 months with center cropping and from 4.2 months to 21.5 months with multi-cropping. These results underscore the sensitivity of the model to accurate demographic information.

4. Discussion

We have designed a new architecture to combine different types of data more efficiently in bone age regression. Due to their inherent architectural limitations, CNNs struggle to process multimodal data directly. Traditional methods include converting the sex information to an additional channel of the input image, which is highly inefficient, or using an ensemble model to concatenate image features from CNNs and sex features from another branch, which does not allow non-visual features to interact with vision features. However, by leveraging properties of vision transformers, our multimodal BAE-ViT facilitates the encoding of non-visual attributes within its transformer blocks, thereby enabling intricate interactions between image-based and non-visual information. This capability paves the way for the integration of additional data modalities, such as demographic distributions, into a unified network architecture. Such an approach holds promise for enhancing both the performance and interpretability of the model.

4.1. Model Comparison

Several recent studies have presented advancements in bone age estimation techniques. For instance, a study [40] utilized CNNs combined with the Tanner–Whitehouse method to extract phalangeal and carpal bone features for predicting bone age. This study employed an Extreme Learning Machine (ELM) algorithm [41], achieving an MAE of 6.1 months on the RSNA dataset, but did not incorporate sex information. Another study [42] implemented a CNN model to estimate bone age from metacarpal and carpal bones, reporting MAEs of 5.6 and 6.0 months on a private dataset of 3871 children. A further study [43] leveraged the YOLOv5 network to extract region-of-interest (ROI) patches, incorporating biological

sex information by passing it through linear and softmax layers to align with the patch dimensions. This information was then used to modulate ROI patches through a Swin Transformer network, achieving an MAE of 4.6 months on the RSNA dataset.

In contrast, our proposed method utilizes entire X-ray scans and introduces a novel approach by converting non-visual information (e.g., biological sex) into a token. This token undergoes attention computation with visual tokens, allowing for a generalized application to other data formats (e.g., demographic, audio, image, text) through tokenization. Our approach achieved an MAE of 4.1 months on the RSNA dataset. Furthermore, we generate heatmaps by using ScoreCAM to better understand the behavior of CNNs and ViTs in Figure 3. We also investigate different training techniques for ensemble models, and observe that pretraining the image encoder leads to overfitting, and hence worse performance. Additionally, we evaluate BAE-ViT and ensemble models on challenging examples and find that our proposed model is more robust to distortions in images in Table 3.

Table 3. Evaluation of BAE-ViT against various ensemble models using different image encoders on challenging examples with image distortions, measured by MAE. The table compares performance under center-crop and multi-crop test conditions. The lowest MAE, indicating the best performance, is in bold, underscoring the robustness of the proposed BAE-ViT model to image distortions.

Test Method	Center-Crop	Multi-Crop
Inception-V3	10.7	10.5
ResNet50	10.6	10.8
EfficientNet-B5	11.3	11.2
TinyViT	10.3	10.5
BAE-ViT	**9.9**	**10.1**

4.2. Limitations

This work is not without limitations. First, we only evaluated models for one task, using a publicly available dataset. It would be interesting to see if our proposed BAE-ViT can also outperform the other models on similar datasets that contain both image and sex information, such as the UTKFace dataset [44], that is popular in age regression tasks. However, the RSNA dataset is one of the largest bone age regression datasets and has been widely studied in the literature. Our work is therefore still useful and can serve as a strong baseline for comparison in the future. Second, we only evaluate the robustness of models on a few challenging examples. There are more types of distortions in bone images in practice, and it is important to investigate if our proposed model is still resistant to them. Third, the ensemble models we studied primarily focus on combining different types of input data. However, ensemble learning also encompasses integrating multiple models with varying architectures that process the same inputs. Models with heterogeneous architectures may possess distinct strengths in handling different aspects of the input data, potentially leading to improved overall performance. This approach has been shown to be effective, as demonstrated in the RSNA Challenge [45], and is widely adopted in medical imaging. Therefore, it is crucial to investigate whether our proposed BAE-ViT can maintain its superior performance when integrated into an ensemble framework with heterogeneous models. It is notable that the ensemble method can also be applied to our proposed model.

4.3. Clinical Implementation

Potential challenges when implementing this model in a real clinical setting include the requirement for high computational power and integration with existing hospital systems. Although TinyViT is a knowledge distillation model with high performance and fewer parameters, it still requires a GPU for inference. There currently exist hundreds of FDA-approved AI software tools in radiology. Some tools utilize local inference, others utilize cloud providers. The allocation and availability of GPU resources does not appear to be preventing the utilization of AI in radiology. BAE-ViT does not have excessively large computational needs. Integration into hospital systems: Additional work would be

needed to fully integrate any AI solution into radiology practice. This was not a major focus of our work. There exist several solutions that provide AI orchestration across various vendor platforms, for example, NVIDIA Clara, but many other competitors already exist in the field.

In the present study, our primary target has been on the regression analysis of bone age, utilizing a limited set of two distinct types of input data. Expanding the diversity of input data sources is anticipated to more comprehensively elucidate the advantages of ViT architectures over CNNs in the field of medical imaging. In conclusion, leveraging self-supervised pretraining techniques on large-scale public datasets has been shown to potentially improve the performance of ViT models [22]. From a clinical perspective, enhancing the interpretability of these machine learning models by providing detailed insights in the intermediate computational stages and robustness analysis could bolster radiologists' confidence in the automated system.

4.4. Impact of Demographic Label Integrity

The label perturbation experiment underscores the critical role of accurate demographic data in our model's performance. The significant performance drop observed with incorrect labels is likely attributable to the early fusion approach employed in our model. In this design, the biological sex attribute is first converted into a unified token, which later contributes to the feature learning phase (Shown in Figure 1c). This integration allows the model to observe specific correlations between visual features and sex attributes, enabling it to learn both shared and distinct features. However, incorrect labels disrupt this process, impairing feature extraction and leading to erroneous predictions.

In real-world healthcare scenarios, such mismatches in demographic labels are rare, as robust data validation protocols and regulatory standards typically minimize such errors. Nevertheless, our findings underscore the importance of maintaining data integrity in automated systems. Addressing these vulnerabilities further could involve incorporating mechanisms to detect and mitigate the impact of label inconsistencies, enhancing both robustness and reliability in clinical applications.

5. Conclusions

In this study, we designed a vision transformer model for bone age estimation, named BAE-ViT, which enables interactions between image and sex information via a tokenization method within transformer blocks. BAE-ViT demonstrated superior performance by achieving a lower mean absolute error (MAE) compared to traditional models and exhibited greater robustness when handling degraded or poor-quality images. This highlights the model's potential in improving bone age estimation accuracy and its applicability in challenging imaging conditions.

Author Contributions: Conceptualization: J.Z., W.C., T.J., X.Z., P.-L.L., V.J., R.J.B., J.W.G. and A.B.M.; data curation: J.Z. and W.C.; formal analysis: J.Z., W.C., P.-L.L., V.J., R.J.B., J.W.G. and A.B.M.; funding acquisition: P.-L.L., V.J., R.J.B., J.W.G. and A.B.M.; investigation: J.Z.; methodology: J.Z., W.C., T.J., P.-L.L., V.J., R.J.B., J.W.G. and A.B.M.; project administration: P.-L.L., V.J., R.J.B., J.W.G. and A.B.M.; resources: J.Z., W.C., P.-L.L., V.J., R.J.B., J.W.G. and A.B.M.; software: J.Z. and W.C.; supervision: P.-L.L., V.J., R.J.B., J.W.G. and A.B.M.; validation: J.Z., W.C., P.-L.L., V.J., R.J.B., J.W.G. and A.B.M.; visualization: J.Z. and W.C.; writing—original draft: J.Z., W.C., P.-L.L., V.J., R.J.B., J.W.G. and A.B.M.; writing—review and editing: J.Z., W.C., P.-L.L., V.J., R.J.B., J.W.G. and A.B.M. All authors have read and agreed to the published version of the manuscript.

Funding: The research reported in this publication was supported by National Library of Medicine under award number R01LM013151, as well as the National Science Foundation under award number DMS-1749857. This work was partially performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk (accessed on 5 November 2024)), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council. Moreover,

the Department of Radiology at the University of Wisconsin-Madison is financially supported by GE Healthcare.

Institutional Review Board Statement: This retrospective study was approved by the Institutional Review Board (IRB) of the University of Wisconsin-Madison (protocol code: 2016-0418, and approval date: 9 August 2016). All research was conducted in accordance with the principles of the Declaration of Helsinki.

Informed Consent Statement: The need for informed consent was waived for this Health Insurance Portability and Accountability Act-compliant study.

Data Availability Statement: The primary dataset supporting this study, consisting of 14,236 cases, is accessible through the 2017 RSNA Pediatric Bone Age Challenge, available at https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/rsna-pediatric-bone-age-challenge-2017 (accessed on 5 November 2023). A supplementary dataset comprising 100 cases, obtained in compliance with institutional protocols, is not available for public access due to privacy and ethical restrictions. The code is available at https://github.com/mimrtl/BAE_ViT (accessed on 5 November 2023).

Conflicts of Interest: Professor McMillan maintains a consulting position with Weinberg Medical Physics, LLC; however, no resources from this company were utilized in our study, and no financial ties exist between any author and the firm.

References

1. Greulich, W.W.; Pyle, S.I. *Radiographic Atlas of Skeletal Development of the Hand and Wrist*; Stanford University Press: Redwood City, CA, USA, 1959. Available online: http://www.sup.org/books/title/?id=2696 (accessed on 25 October 2022).
2. Poznanski, A.K. Assessment of Skeletal Maturity and Prediction of Adult Height (TW2 Method). *Am. J. Dis. Child.* **1977**, *131*, 1041–1042. [CrossRef]
3. Lee, J.H.; Kim, Y.J.; Kim, K.G. Bone age estimation using deep learning and hand X-ray images. *Biomed. Eng. Lett.* **2020**, *10*, 323–331. [CrossRef] [PubMed] [PubMed Central]
4. Lee, H. Fully Automated Deep Learning System for Bone Age Assessment. *J. Digit. Imaging* **2017**, *30*, 427–441. [CrossRef] [PubMed]
5. Bui, T.D.; Lee, J.J.; Shin, J. Incorporated region detection and classification using deep convolutional networks for bone age assessment. *Artif. Intell. Med.* **2019**, *97*, 1–8. [CrossRef] [PubMed]
6. Wu, E.; Kong, B.; Wang, X.; Bai, J.; Lu, Y.; Gao, F.; Zhang, S.; Cao, K.; Song, Q.; Lyu, S.; et al. Residual Attention Based Network for Hand Bone Age Assessment. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 1158–1161. [CrossRef]
7. Han, J.; Jia, Y.; Zhao, C.; Gou, F. Automatic Bone Age Assessment Combined with Transfer Learning and Support Vector Regression. In Proceedings of the 2018 9th International Conference on Information Technology in Medicine and Education (ITME), Hangzhou, China, 19–21 October 2018; pp. 61–66. [CrossRef]
8. Liu, Y.; Zhang, C.; Cheng, J.; Chen, X.; Wang, Z.J. A multi-scale data fusion framework for bone age assessment with convolutional neural networks. *Comput. Biol. Med.* **2019**, *108*, 161–173. [CrossRef] [PubMed]
9. Umer, M.; Eshmawi, A.A.; Alnowaiser, K.; Mohamed, A.; Alrashidi, H.; Ashraf, I. Skeletal age evaluation using hand X-rays to determine growth problems. *Peerj Comput. Sci.* **2023**, *9*, e1512. [CrossRef] [PubMed] [PubMed Central]
10. Liu, Z.Q.; Hu, Z.J.; Wu, T.Q.; Ye, G.X.; Tang, Y.L.; Zeng, Z.H.; Ouyang, Z.M.; Li, Y.Z. Bone age recognition based on mask R-CNN using xception regression model. *Front. Physiol.* **2023**, *14*, 1062034. [CrossRef]
11. Pan, X.; Zhao, Y.; Chen, H.; Wei, D.; Zhao, C.; Wei, Z. Fully Automated Bone Age Assessment on Large-Scale Hand X-Ray Dataset. *Int. J. Biomed. Imaging* **2020**, *2020*, 8460493. [CrossRef]
12. Halabi, S.S.; Prevedello, L.M.; Kalpathy-Cramer, J.; Mamonov, A.B.; Bilbily, A.; Cicero, M.; Pan, I.; Pereira, L.A.; Sousa, R.T.; Abdala, N.; et al. The RSNA Pediatric Bone Age Machine Learning Challenge. *Radiology* **2019**, *290*, 498–503. [CrossRef]
13. González, C.; Escobar, M.; Daza, L.; Torres, F.; Triana, G.; Arbeláez, P. SIMBA: Specific Identity Markers for Bone Age Assessment. In *Medical Image Computing and Computer Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12266, pp. 753–763. [CrossRef]
14. Ren, X.; Li, T.; Yang, X.; Wang, S.; Ahmad, S.; Xiang, L.; Stone, S.R.; Li, L.; Zhan, Y.; Shen, D.; et al. Regression Convolutional Neural Network for Automated Pediatric Bone Age Assessment From Hand Radiograph. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 2030–2038. [CrossRef]
15. Wang, C.; Wu, Y.; Wang, C.; Zhou, X.; Niu, Y.; Zhu, Y.; Gao, X.; Wang, C.; Yu, Y. Attention-based multiple-instance learning for Pediatric bone age assessment with efficient and interpretable. *Biomed. Signal Process. Control* **2023**, *79*, 104028. [CrossRef]
16. Nurzynska, K.; Piórkowski, A.; Strzelecki, M.; Kociołek, M.; Banyś, R.P.; Obuchowicz, R. Differentiating age and sex in vertebral body CT scans—Texture analysis versus deep learning approach. *Biocybern. Biomed. Eng.* **2024**, *44*, 20–30. [CrossRef]

17. Guo, Z.; Wang, X.; Yang, L.; Yang, X.; Qi, Y.; Zhao, Z. An intelligent bone age assessment model incorporating multilayer superimposed texture enhancement and the China-05 attention mechanism. *Biomed. Signal Process. Control* **2025**, *99*, 106852. [CrossRef]
18. Hering, R.N.; von Kroge, S.; Delsmann, J.; Simon, A.; Ondruschka, B.; Püschel, K.; Schmidt, F.N.; Rolvien, T. Pronounced cortical porosity and sex-specific patterns of increased bone and osteocyte lacunar mineralization characterize the human distal fibula with aging. *Bone* **2024**, *182*, 117068. [CrossRef]
19. Obuchowicz, R.; Nurzynska, K.; Pierzchala, M.; Piorkowski, A.; Strzelecki, M. Texture Analysis for the Bone Age Assessment from MRI Images of Adolescent Wrists in Boys. *J. Clin. Med.* **2023**, *12*, 2762. [CrossRef] [PubMed]
20. 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 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002. [CrossRef]
21. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jegou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the Machine Learning Research, Boulder, CO, USA, 18–24 July 2021; pp. 10347–10357.
22. 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. *arXiv* **2022**. [CrossRef]
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
24. Hu, R.; Singh, A. UniT: Multimodal Multitask Learning with a Unified Transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 1419–1429. [CrossRef]
25. Wang, Y.; Chen, X.; Cao, L.; Huang, W.; Sun, F.; Wang, Y. Multimodal Token Fusion for Vision Transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 12176–12185. [CrossRef]
26. Xu, Y.; Xu, Y.; Lv, T.; Cui, L.; Wei, F.; Wang, G.; Lu, Y.; Florencio, D.; Zhang, C.; Che, W.; et al. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; pp. 2579–2591. [CrossRef]
27. Prokop-Piotrkowska, M.; Marszałek-Dziuba, K.; Moszczyńska, E.; Szalecki, M.; Jurkiewicz, E. Traditional and New Methods of Bone Age Assessment-An Overview. *J. Clin. Res. Pediatric Endocrinol.* **2021**, *13*, 251–262. . [CrossRef] [PubMed] [PubMed Central]
28. Wu, K.; Zhang, J.; Peng, H.; Liu, M.; Xiao, B.; Fu, J.; Yuan, L. TinyViT: Fast Pretraining Distillation for Small Vision Transformers. *arXiv* **2022**. [CrossRef]
29. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
30. Howard, A.; Sandler, M.; Chen, B.; Wang, W.; Chen, L.C.; Tan, M.; Chu, G.; Vasudevan, V.; Zhu, Y.; Pang, R.; et al. Searching for MobileNetV3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324. [CrossRef]
31. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**. [CrossRef]
32. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. *arXiv* **2015**. [CrossRef]
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**. [CrossRef]
34. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6105–6114.
35. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]
36. Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; Hu, X. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 111–119. [CrossRef]
37. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. Conf. Comput. Vis.* **2017**, *10*, 618–626. [CrossRef]
38. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18613–18624.
39. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random Erasing Data Augmentation. *arXiv* **2017**. [CrossRef]
40. Guo, L.; Wang, J.; Teng, J.; Chen, Y. Bone Age Assessment Based on Deep Convolutional Features and Fast Extreme Learning Machine Algorithm. *Front. Energy Res.* **2022**, *9*, 813650. [CrossRef]
41. Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme learning machine: Theory and applications. *Neurocomputing* **2006**, *70*, 489–501. [CrossRef]
42. Wu, J.; Mi, Q.; Zhang, Y.; Wu, T. SVTNet: Automatic bone age assessment network based on TW3 method and vision transformer. *Int. J. Imaging Syst. Technol.* **2024**, *34*, e22990. [CrossRef]

43. Mao, X.; Hui, Q.; Zhu, S.; Du, W.; Qiu, C.; Ouyang, X.; Kong, D. Automated Skeletal Bone Age Assessment with Two-Stage Convolutional Transformer Network Based on X-ray Images. *Diagnostics* **2023**, *13*, 1837. [CrossRef]
44. Zhang, Z.; Song, Y.; Qi, H. Age Progression/Regression by Conditional Adversarial Autoencoder. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
45. Pan, I.; Thodberg, H.H.; Halabi, S.S.; Kalpathy-Cramer, J.; Larson, D.B. Improving Automated Pediatric Bone Age Estimation Using Ensembles of Models from the 2017 RSNA Machine Learning Challenge. *Radiol. Artif. Intell.* **2019**, *1*, 6. [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.

Journal of
Clinical Medicine

MDPI

Article

Enhancing Surgical Wound Monitoring: A Paired Cohort Study Evaluating a New AI-Based Application for Automatic Detection of Potential Infections

Andrea Craus-Miguel [1,2,*,†], Marc Munar [3,4,5], Gabriel Moyà-Alcover [4,5,6,7], Ana María Contreras-Nogales [8], Manuel González-Hidalgo [3,4,5,7] and Juan José Segura-Sampedro [9,10,11,†]

1 General and Digestive Surgery Department, Clínica Universidad de Navarra, 31008 Madrid, Spain
2 PhD Program in Translational Research in Public Health and High Prevalence Diseases, Health Sciences, University of the Balearic Islands (UIB), 07122 Palma, Spain
3 Soft Computing, Image Processing and Aggregation (SCOPIA) Research Group, University of the Balearic Islands (UIB), 07122 Palma, Spain
4 Artificial Intelligence Research Institute of the UIB (IAIB), University of the Balearic Islands (UIB), 07122 Palma, Spain
5 Health Research Institute of the Balearic Islands (IdISBa), 07120 Palma, Spain
6 Computer Graphics and Vision and AI (UGiVIA) Research Group, University of the Balearic Islands (UIB), 07122 Palma, Spain
7 Laboratory of Artificial Intelligence Applications (LAIA@UIB), University of the Balearic Islands (UIB), 07122 Palma, Spain
8 Nursing Team, Wound Care Unit, Outpatient Department, Son Espases University Hospital, 07120 Palma, Spain
9 General & Digestive Surgery Service, Hospital Universitario La Paz, 28046 Madrid, Spain; segusamjj@hotmail.com
10 Hospital La Paz Institute for Health Research (IdiPAZ), 28046 Madrid, Spain
11 School of Medicine, Universidad CEU San Pablo, 28003 Madrid, Spain
* Correspondence: acrausm@gmail.com
† Previous affiliation during study: General & Digestive Surgery, Son Espases University Hospital, 07120 Palma, Spain.

Abstract: Background/Objectives: This study assessed the feasibility and security of remote surgical wound monitoring using the RedScar© smartphone app, which employs automated diagnosis for early visual detection of infections without direct healthcare personnel involvement. Additionally, patient satisfaction with telematic care was evaluated as a secondary aim. Surgical site infection (SSI) is the second leading cause of healthcare-associated infections (HAIs), leading to prolonged hospital stays, heightened patient distress, and increased healthcare costs. **Methods:** The study employed a prospective paired-cohort and single-blinded design, with a sample size of 47 adult patients undergoing abdominal surgery. RedScar© was used for remote telematic monitoring, evaluating the feasibility and security of this approach. A satisfaction questionnaire assessed patient experience. The study protocol was registered at ClinicalTrials.gov under the identifier NCT05485233. **Results:** Out of 47 patients, 41 successfully completed both remote and in-person follow-ups. RedScar© demonstrated a sensitivity of 100% in detecting SSIs, with a specificity of 83.13%. The kappa coefficient of 0.8171 indicated substantial agreement between the application's results and human observers. Patient satisfaction with telemonitoring was high: 97.6% believed telemonitoring reduces costs, 90.47% perceived it prevents work/school absenteeism, and 80.9% found telemonitoring comfortable. **Conclusions:** This is the first study to evaluate an automatic smartphone application on real patients for diagnosing postoperative wound infections. It establishes the safety and feasibility of telematic follow-up using the RedScar© application for surgical wound assessment. The high sensitivity suggests its utility in identifying true cases of infection, highlighting its potential role in clinical practice. Future studies are needed to address limitations and validate the efficacy of RedScar© in diverse patient populations.

Keywords: surgical wound monitoring; smartphone application; artificial intelligence; telemedicine; surgical site infection (SSI); remote monitoring; digital health; postoperative care; telematic follow-up; patient satisfaction; active patient

1. Introduction

Surgical site infection (SSI) is the second cause of healthcare-associated infections (HAIs), with a prevalence of 19.6% of all HAIs in Europe [1]. SSIs are associated with significant morbidity, including prolonged postoperative hospital stays, additional surgical procedures, increased mortality rates, and an impact on patient health-related quality of life (HRQoL) [2,3]. The incidence of SSIs varies from 0.5% to 10.1%, depending on the type of surgical procedure. Delayed diagnosis of SSIs, which often occurs after patient discharge [4], exacerbates patient distress and increases healthcare costs. In Europe, the economic burden of SSIs is substantial, with direct costs estimated to range from EUR 10,443 to EUR 19,638 per infection [5]. These costs include extended hospital stays, additional treatments, and follow-up visits. Indirect costs, such as lost productivity and long-term disability, further amplify the financial impact on healthcare systems and society [6].

In recent years, the rise of new devices and smartphone apps has advanced e-health, aligning it with patients' real-life routines for early disease detection and proactive healthcare [7–11]. This deepens our understanding of disease progression, resulting in reduced mortality and improved well-being. With mobile technology accessible to around 85% of the population [12], telemedicine emerges as a game-changer, enhancing routine care while concurrently diminishing wound infection occurrences and medical center visits without compromising patient satisfaction [13–15]. It is important to use user-friendly smartphone apps with navigation, instant assessments, and standardization, alongside predictive models powered by machine learning and artificial intelligence for better infection detection [16–20]

We created RedScar©, a smartphone app for remote post-discharge assessment of surgical wounds. It uses automated diagnosis to detect and evaluate surgical site infections without human involvement, enabling early identification and intervention.

Innovative Approach: This study represents a pioneering effort, as it is the first to utilize a smartphone application in real patients, relying on photographs for diagnosing wound infections without the direct intervention of healthcare personnel. This innovative approach holds great promise for enhancing the early detection of surgical wound infections and improving patient care.

The main aim of the study was to establish the feasibility and security of telematic monitoring of surgical wounds using the RedScar© application. As a secondary aim, the satisfaction of patients under telematic care with the RedScar© app was evaluated.

2. Methods

2.1. Study Design

This study utilized a prospective paired cohort study design with a non-randomized, single-blinded approach, where each participant served as their own control. The investigator remained unaware of the application's recommendations during wound assessments. The study included adult patients undergoing abdominal surgery at a tertiary care hospital in 2021. Patient follow-up care involved a combination of remote telematic monitoring and in-person revisions. The study was conducted following the guidelines of the 2015 Standards for Reporting Diagnostic Accuracy Studies (STARD) Initiative [21], thus ensuring transparency and quality in the presentation of diagnostic accuracy results. The study's design and protocol underwent thorough review and obtained approval from the Research Ethics Committee of the Balearic Islands (CEI-IB) under reference number IB: 4100/20. Informed consent was obtained from all participants, ensuring patient privacy and data

security throughout the study. The study protocol was registered at ClinicalTrials.gov under the identifier NCT05485233.

2.2. Sample Size and Recruitment

The sample size was estimated considering that this was a pilot study and the frequency of surgical site infections (SSI). We used a 95% confidence level and an estimated SSI frequency of 5%. To achieve an initial sample size of 30 participants, including a 10% loss rate, we adjusted the desired precision to 0.075 [22].

2.3. Inclusion Criteria

We exclusively included adult inpatients who had undergone abdominal surgery with staple closure and who possessed an Android smartphone. Due to the technical limitations of the Android system, only patients whose smartphones had a minimum version of Android 10, with a camera capable of capturing images of a minimum size of 500×500 pixels, were accepted.

2.4. Exclusion Criteria

Patients lacking access to a compatible smartphone or those unable to use the app correctly were excluded from the study. Additionally, patients unable to undergo personal revisions were excluded.

2.5. Identification and Categorization

Potential eligible patients were identified based on our inclusion criteria. We meticulously categorized patients as one of the following: "included", "missed", "rejected", or "excluded". Detailed records were maintained to track their respective statuses throughout the recruitment process.

2.6. Variables

The collected variables included gender, age, ASA risk score, surgery date and type (open/laparoscopic), type of laparotomy (medial incision, subcostal, Pfannenstiel, McBurney, other), discharge date, and length of hospitalization.

2.7. Visiting Schedule

Hospitalization and Informed Consent

Patients were discharged post-surgery at the discretion of the responsible physician. Before discharge, participants in the study underwent the same hospitalization protocol as other patients treated at our institution. The study's protocol was fully explained by the researcher, and written consent was obtained from the patients during this discussion. Subsequently, patients downloaded the RedScar© application (Balearic Island University, Palma, Spain), and the researcher provided a detailed explanation of its functionality, including how to respond to all wound-related questions outlined in Table 1. Patients received a brief training session on how to use the application.

Table 1. Questionnaire designed to obtain information on the health status of each patient for the RedScar© app.

Question	Response
Visual Analog Scale	0–10 (Scale of pain severity)
Redness around the wound	No (0) Yes (1)
Burning sensation or heat at the wound	No (0) Yes (1)
Wound opening	No (0) Yes (1)

Table 1. *Cont.*

Question	Response
Swelling	No (0) Yes (1)
Liquid secretion	No (0) Yes (1)
Fever	No (0) Yes (1) > 37.7 °C

2.8. Post-Surgery Follow-Up

On the 3rd and 10th days following surgery, each patient used the RedScar© application to upload an image of their surgical wound (Figure 1). The application evaluated the risk of wound complications and categorized patients into two groups: those requiring a new consultation due to complications and those with satisfactory wound evolution, eligible for discharge.

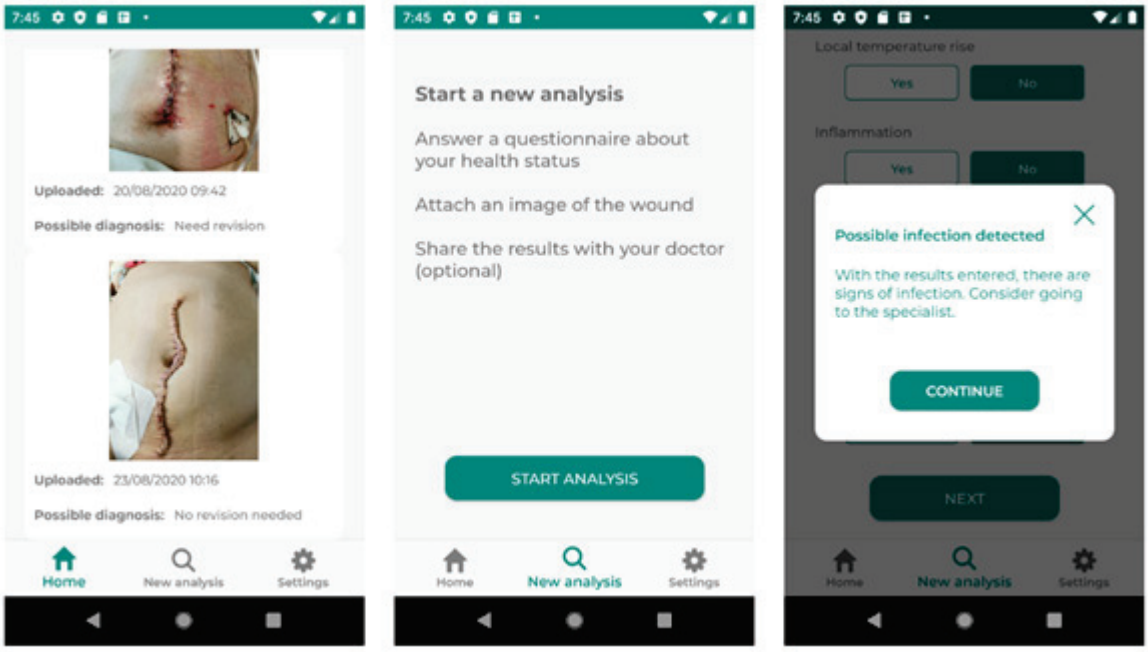


Figure 1. Screenshots of some functionalities of the RedScar© app in the patient interface. From left to right: the history of the images uploaded by the patient, the area to start a new automatic analysis, and finally the result given by the application when a sign of infection has been detected in the analyzed image.

2.8.1. Clinical Assessments

On the 3rd and 10th days post-surgery, patients underwent in-person examinations of their surgical wounds by a healthcare professional, either a physician or a nurse. These professionals confirmed the presence or absence of complications and, if deemed necessary, collected microbiological samples to enhance the diagnostic process. The healthcare professionals also completed the questionnaire outlined in Table 1 to compare responses with those provided through the application.

2.8.2. Patient Satisfaction Assessment

Upon concluding the study, a patient satisfaction survey based on the questionnaire developed by Yip et al. [23] and used in similar studies [13] was administered (Table 2). This questionnaire provided valuable insights into patient experiences with the telemedicine-based wound assessment process.

Table 2. Satisfaction questionnaire (17.).

Satisfaction Questionnaire	Answer
1. The information received before participating in the study was adequate.	Yes/No
2. I do not have difficulties with the questionnaire used within the application.	Yes/No
3. I do not have difficulties when taking the photo and uploading it to the mobile application.	Yes/No
4. I have spent less time with telematic follow-up than with personal revision.	Yes/No
5. I think that telematic follow-up is cheaper than the personal revision.	Yes/No
6. I think that telematic review could prevent absenteeism from work/school.	Yes/No
7. Telematic follow-up seems more comfortable to me than follow-up in consultations.	Yes/No
8. I think that telematic monitoring could detect a bad evolution earlier.	Yes/No
9. I think the healthcare provided via telemedicine is consistent.	Yes/No
10. The care received seems adequate to me.	Yes/No
11. Overall, I am satisfied with the quality of service being provided via telemedicine.	Yes/No
12. I would prefer telematic follow-up to face-to-face consultations.	Yes/No
13. I will use telemedicine services again.	Yes/No
14. Following the instructions of the application, have you checked your wound at your health center?	Yes/No, Why?
15. Although the application had ruled out the infection of the wound, have you preferred to check it in person?	Yes/No, Why?
16. Despite using the application, have you consulted any other source (Google-type web search, opinion of close relatives)?	Yes/No, Why?
17. How do you think we can improve the RedScar© application? Leave your comment.	

This visiting schedule outlines the comprehensive approach taken to monitor patients' surgical wounds, combining remote monitoring via the RedScar© application with in-person clinical assessments to ensure the accuracy and reliability of the study's findings.

2.9. Data Analysis Plan

The analysis was conducted on an intention-to-treat basis. Patients removed from the study were not replaced or re-included and were excluded from the analysis. Descriptive analysis of quantitative variables was performed using means and standard deviations or medians and percentiles, with a 95% bilateral confidence interval and range (minimum and maximum) or P50 [P25–P75]. Qualitative variables were summarized in a table that included absolute and relative frequencies. Statistical significance was considered at $p < 0.05$. The IBM® SPSS® Statistics 19 package was used for data analysis.

To measure the agreement between the evaluations provided by two human observers and RedScar©, we employed the kappa statistic as a quantitative measure. To assess concordance, we utilized all the photographs gathered throughout the study, not just those specifically from days 3 and 10. For the intervention, the sensitivity and specificity of the RedScar© application for the detection of wound infection were calculated, and the correlation of clinical diagnoses with the application's automatic response was compared. The level of redness in infected wound photographs was recorded to improve the infection cut-off point.

Data on patient satisfaction were also collected via a follow-up questionnaire and analyzed separately.

2.10. Development of the RedScar© Application

The RedScar© app employs a multifaceted approach to assess surgical wound infections, combining patient-reported data and automatic computer image analysis. The user interface includes a straightforward questionnaire addressing various wound-related aspects, such as pain severity (measured on a visual analog scale, VAS), redness, burning sensation, wound leakage, swelling, liquid secretion, and fever (as outlined in Table 1). To proceed to the next step, which involves uploading a photograph of the wound, patients must respond to all questions in the questionnaire. As shown in González-Hidalgo et al. [24], the application initiates a multistep computerized image analysis process using fuzzy sets, based on gray-scale mathematical morphology (see [20] for more information) and its extension to color images proposed by Bibiloni et al. [25]. Then, by combining this image analysis with the result of the questionnaire, the application autonomously generates a diagnosis without requiring the direct supervision of a healthcare professional (see Image 1). The application has been implemented using Java language and the Android Studio development environment. For the analysis of the image and according to the method proposed in [24], the following repositories implemented by the authors in Python language, which are freely accessible, were used:

- Staple detection [26], which implements the automatic localization of the staples and the wound in the image.
- InPYinting [27], which implements the image reconstruction previously proposed [25].

These repositories have been integrated into the application for image analysis.

All image processing and machine learning methods have been applied and trained, respectively, on the same data set divided into training and testing, randomly prefixed and accessible on redscar.uib.es [28].

2.11. Data Protection and Management

Each patient was assigned an identification number to maintain anonymity. A data file was established in which all patient data were anonymized. When a patient was included in the study, a random code was automatically generated by the application, and only the investigator knew the correspondence between the patient and the code.

The application was supported by two servers. The first, a computing server, only received the images, applied algorithms for detecting infection, and returned the value to the device. This service was located at the University of the Balearic Islands (UIB). The second server, which hosted the entire database, was located on Google's Firebase service. All information was encrypted, and only the researchers involved in the project had access to it. The connection with the RedScar© application was secure, and access was only allowed from the application. Patient information was saved in JSON format (key-value system). Researchers had accounts with the same system, allowing them to be linked with the corresponding patient–physician pairs.

3. Results

A total of 56 patients met the inclusion criteria. Nine patients were excluded. The exclusion criteria were the absence of a compatible device (three patients), refusal to participate in the study (four patients), and failure to download the mobile application (two patients). Out of the 47 included patients, 41 successfully completed both in-person follow-up and application-based follow-up. The median age within the study cohort was 59 years; 46.8% were females, and 53.2% were males. However, six patients did not complete the study because they either failed to upload the required photo or did not attend the second in-person visit. Nevertheless, the photograph taken on the third postoperative day was utilized in these six cases, resulting in a total of 88 images being analyzed. See Figure 2 for a graphical scheme of the study.

Five (12.19%) patients developed surgical site infections, all of which were detected by the application, resulting in a sensitivity of 100%. The application produced 14 false posi-

tives, for which it was indicated that a presential review was needed, but no complications were identified by the clinician. The specificity was 83.13%.

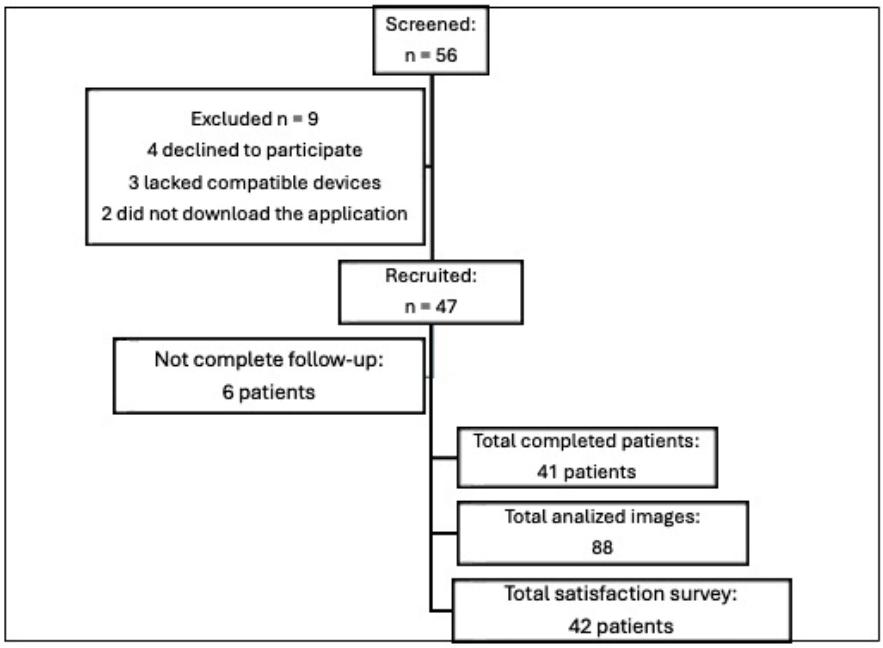


Figure 2. Flow chart of the patients included in the study.

The detection of infection using automatic image processing was approached in this work from the point of view of the redness of the wound, i.e., the proportion of pixels that had a color close to red. Therefore, determining the optimal threshold for the proportion of red pixels in the image allowed us to detect redness in the wound and to have a criterion for classifying according to the presence of infection. After the analysis of the images, the optimal cutoff point was calculated using the ROC curve, establishing an optimal red proportion value of 0.63. In Figure 3, we display the ROC curve for the automatic red proportion detection.

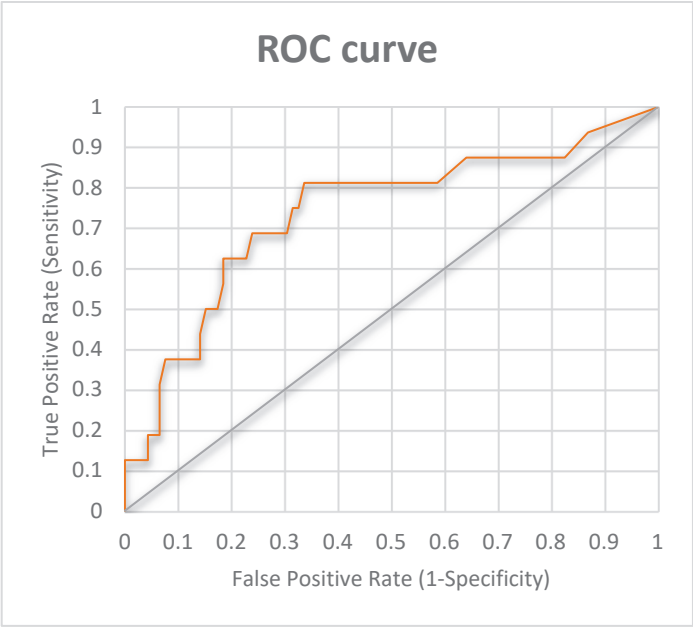


Figure 3. ROC curve for the automatic red proportion detection in the RedScar© app. A cutoff of 0.63 was set.

3.1. Concordance

Throughout the study, a total of 363 photographs were captured. These images were used to assess the wound's progression and evaluate the app's performance.

The statistical analysis using kappa was conducted on the full set of images (363), which were independently categorized by two researchers. The calculated kappa coefficient of concordance between the application's results and the observations yielded a value of 0.8171. This value signifies a substantial level of agreement between the human observers and the RedScar© application.

3.2. Satisfaction and Patient-Reported Outcomes

A total of 42 patients answered the questionnaire.

Concerning photograph submission and analysis, 54.76% of patients experienced no issues, whereas 42.85% encountered app freezes during upload and analysis, and 2.38% did not answer (Figure 4).

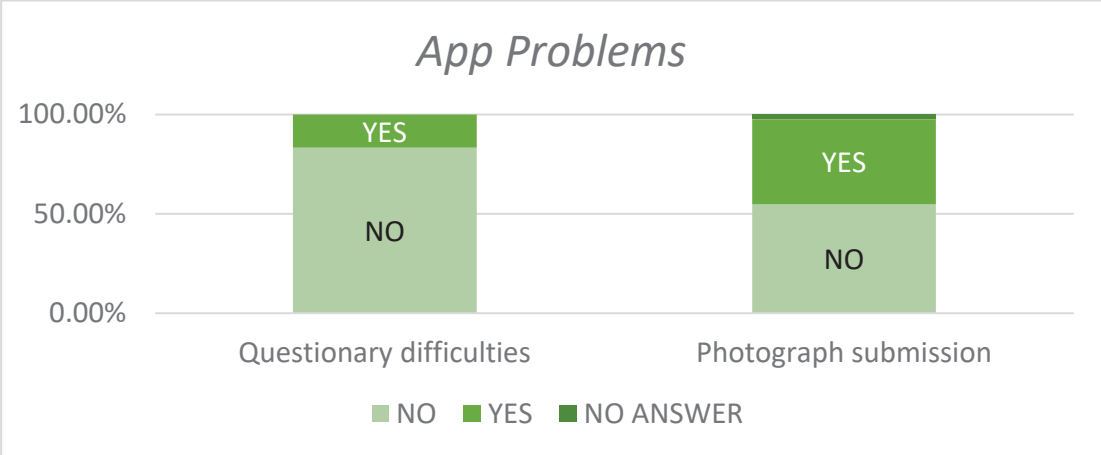


Figure 4. Percentage of patients included in the study who, when using the RedScar© app, encountered some type of problem when completing the questionnaire or attaching an image for automatic analysis.

The average app time of use was 3 min and 37 s, with a minimum of 1 min of use and a maximum of 15 min, caused by network issues.

Almost all (97.6%) of the respondents believed that telemonitoring would reduce costs, and 90.47% perceived that telemonitoring prevented work/school absenteeism. Additionally, 80.9% found telemonitoring more comfortable (Figure 5).

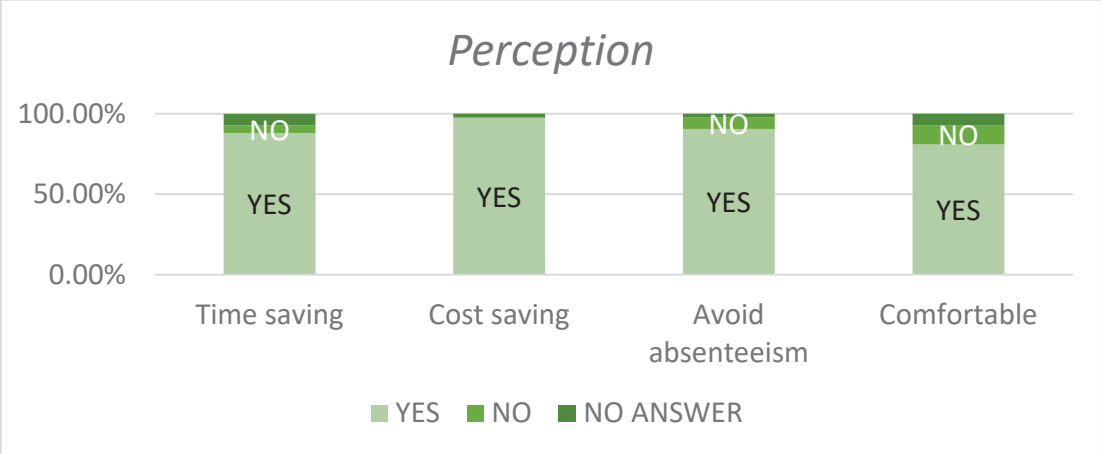


Figure 5. Percentage of patients' perception of different items on the RedScar application.

Regarding reliability, 73.80% of patients found the telemonitoring system trustworthy. Furthermore, 92.85% expressed satisfaction with the provided care, and 80.95% were content with the app's service quality. Among the 38 patients who responded to the preference question, only 20 patients expressed a preference for telemonitoring. However, the majority (78.57%) stated they would use a similar app for future health monitoring. In terms of app result consultations, 30.95% consulted healthcare professionals based on app indications and 35.71% consulted alternative sources (Figure 6).

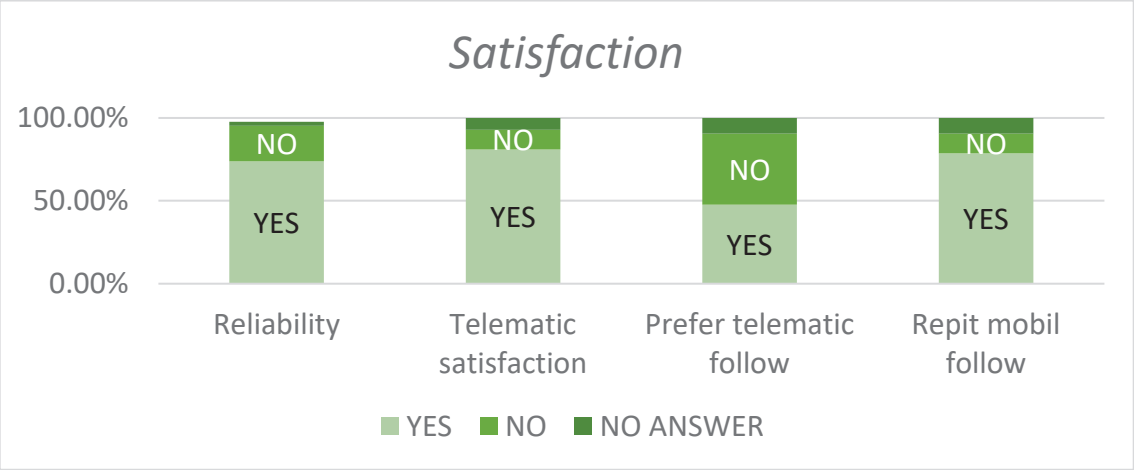


Figure 6. Percentage of patient satisfaction on different items of the RedScar© application.

4. Discussion

The RedScar© application proved to be safe, as no complications derived from telematic follow-up were detected. Telematic follow-up with the RedScar© application was feasible, as most of the patients (85.7%) were able to finish the study. This differs from other studies in which rates of incomplete telemonitoring were high (ranging from 26% to 77%) [19,29,30]. The fact that most of our patients successfully completed the study could be attributed to the user-friendly app and the relatively short 10-day monitoring period [14,17,18].

The kappa value of 0.8171, which measured the agreement between two human blinded observers and RedScar©, indicated a substantial agreement between human observers and the RedScar© application.

To address the absence of professional intervention in both the image acquisition and their subsequent diagnosis, our design aimed for heightened sensitivity, resulting in a 100% final sensitivity outcome. While ensuring the non-underdiagnosis of patients with infection, this comes at the expense of reduced specificity (83.13%), with 16 false-positive detections. Nevertheless, both the sensitivity and specificity are higher than those achieved in other studies [31]. This trade-off means that while a few patients might undergo additional checks unnecessarily, the high sensitivity ensures that most true cases of surgical site infections (SSIs) are detected. In a clinical setting, this is crucial for patient safety and timely intervention. Early detection and treatment of SSIs can significantly improve patient outcomes and reduce complications. This approach is taken to mitigate the risk of false negatives, ensuring that cases of complicated surgical wounds do not go without proper assessment. This helps us avoid the scenario reported by Jia et al. [32], in which out of 81 surgical wound infections, only 66 were diagnosed, and 14 cases escaped detection.

The key innovation of the RedScar© application lies in the automatic analysis of surgical wounds without human intervention. Current telemedicine solutions for wound monitoring typically involve human intervention for diagnosis [19,33]. While various mobile applications exist for diagnosing complicated wounds, most focus on chronic wounds (e.g., diabetic foot ulcers) rather than surgical wounds [34,35]. With RedScar©, patient-uploaded images receive an independent presumptive diagnosis, suggesting the

need for clinic visits. This app utilizes only the patient's smartphone, in contrast to other applications that use costly technologies such as infrared cameras and thermal imaging [17,18,34,36]. The automated analysis eliminates the need for individual clinician assessments, facilitating integration into routine clinical practice, a feature not found in these studies. RedScar© is unique in its fully automated diagnostic process, which does not require direct human supervision. This innovation reduces the burden on healthcare professionals and allows for more scalable and efficient monitoring.

Patient engagement stands as a baseline metric delineated in the proper design and implementation of digital surgical care navigation tools [37]. Our study demonstrates a commendable adherence to this criterion, as evidenced by participants extending their app usage beyond the stipulated duration, utilizing it for more days than required. This behavior may be attributed to its ease of use, immediacy, and user-friendly interface. This circumstance has not been described in other similar papers in the literature, which makes us optimistic about the development of the app. This resulted in a significantly higher number of images than expected, empowering all patients to play an active role in their treatment and recovery. Thanks to the multiple images provided by the patients, a virtual repository of surgical wound photographs was created and made available to other researchers. This repository can be accessed through the website redscar.uib.es (accessed on 10 December 2024) [28].

The great innovation in this app is the diagnostic methodology. There is no app or article that describes the diagnosis of surgical wound images without human intervention. The closest counterparts to our app are those designed for telemedicine-based diagnosis of surgical wound infections, which involve human intervention [33,38–40]. In fact, only a study by Scheper et al. [19] comes close to a similar methodology. They employed a mobile phone app with a predetermined score, assessing the likelihood of infection and advising contact with a clinician. However, it is worth noting that the diagnosis relied on the researcher's evaluation of the patient's provided photo, rather than the app itself generating the diagnosis.

Similar to the studies analyzed [19,33], our results also reported a prominent level of confidence in telemonitoring (73.80%), although only 47.61% preferred this type of monitoring over in-person follow-up. The observations expressed by some patients in uploading photographs, along with initial malfunctions in the application, could explain this phenomenon. However, 78.57% of the participants would still use a mobile device for monitoring their health status. One area for improvement in future studies is the image upload time, as well as the response time of the application. Additionally, based on the results obtained, enhancing the specificity of the application would help increase patient confidence in RedScar©.

The main limitation of our study is the low wound infection rate, with only five patients (12.19%) experiencing infections. While this aligns with the average infection rate in abdominal surgery [41], further research is currently underway through a multicenter clinical trial with a larger patient population, including those at higher risk of wound infection and in other surgical contexts, such as orthopedic or gynecologic surgeries. This aims to determine the broader applicability of the RedScar© app and better validate these results.

Our findings demonstrate the potential of the RedScar© app to enhance the detection of surgical site infections (SSIs). However, it is crucial to consider potential ethical issues, such as over-reliance on AI in clinical decision-making. Recent studies [42,43] have emphasized that AI can improve prediction accuracy and support surrogate decision-making, but it cannot replicate the nuanced ethical deliberation inherent to human clinicians, and maintaining human oversight is essential to prevent biases and ensure fairness in AI applications in healthcare. Ensuring that AI tools complement human judgment is essential to maintaining high standards of patient care.

An acknowledged constraint in m-health pertains to the age discrepancy in technology adoption [12]. Notably, three out of the four patients who declined participation were aged over 70, while the primary age within the study cohort was 59 years. This observation underscores a propensity for this technology among younger individuals. Addressing this age-related challenge emerges as a promising avenue for bolstering future studies, underscoring the imperative to scrutinize age restrictions and streamline application usage for elderly patients. Studies have shown that older adults often face barriers such as perceived complexity and lack of confidence in using new technologies [44,45]. Ensuring inclusivity in future research will help validate the broader applicability and effectiveness of the RedScar© app across diverse populations.

Currently, the app is only available for Android, resulting in three recruited patients being unable to participate in the study. However, this limitation is in the process of being solved, as the iOS application is under development.

During the study, 42.85% of patients faced challenges when uploading photographs. This may have contributed to not achieving a 100% completion rate. These technical issues have been resolved in the next versions of the app but should be further validated in a subsequent study with a larger patient sample.

Regarding app usage, despite its user-friendliness, some patients struggled to take proper photographs. They failed to center the images correctly, exposed dressings, or used disinfectants with colored substances that hindered the accurate reading of redness. The same difficulties have been reported in other studies [15,46,47], in which the task of taking photographs of the wound was reported as difficult by patients. In future studies, clearer instructions should be provided to patients to ensure optimal results, and the app should be improved regarding these limitations.

5. Conclusions

Our study successfully demonstrated the safety and feasibility of telematic follow-up with the RedScar© application. Its high sensitivity suggests its utility in identifying true cases of infection, while the specificity could benefit from further refinement. There was a substantial agreement between the application's results and human observations, underscoring its promising role in clinical practice. The user-friendly design and the automated diagnosis empowered patients, resulting in a high completion rate and highlighting the application's key innovation—automatic analysis without human intervention with simple technology. Despite certain limitations, such as a low wound infection rate and technical challenges, our findings emphasize the potential of RedScar© and AI-based apps in routine clinical practice, calling for future studies to validate its efficacy with larger and more diverse patient samples.

Author Contributions: Conceptualization: J.J.S.-S., A.C.-M., M.M., G.M.-A. and M.G.-H.; Data curation: A.C.-M. and J.J.S.-S.; Formal Analysis: A.C.-M., M.G.-H., M.M. and G.M.-A.; Funding acquisition: J.J.S.-S., G.M.-A. and M.G.-H.; Investigation: A.C.-M., A.M.C.-N. and J.J.S.-S.; Methodology: A.C.-M., J.J.S.-S. and M.G.-H.; Project administration: J.J.S.-S. and M.G.-H.; Resources: J.J.S.-S., A.C.-M., M.M., G.M.-A. and M.G.-H.; Software: M.M., G.M.-A. and M.G.-H.; Supervision: J.J.S.-S., G.M.-A. and M.G.-H.; Validation: J.J.S.-S., A.C.-M., M.M., G.M.-A. and M.G.-H.; Writing—original draft: J.J.S.-S. and A.C.-M.; Writing—review and editing: J.J.S.-S., A.C.-M., G.M.-A. and M.G.-H. All authors have read and agreed to the published version of the manuscript.

Funding: This study is part of the R+D+i Project PID2020-113870GB-I00- "Desarrollo de herramientas de Soft Computing para la Ayuda al Diagnóstico Clínico y a la Gestión de Emergencias (HESOCODICE)" funded by MCIN/AEI/10.13039/501100011033. This study received methodological and statistical support from the Clinical Research Unit and Clinical Trials Unit (UICEC) of the Balearic Islands Health Research Institute (IdISBa).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Research Ethics Committee of the Balearic Islands (CEI-IB) under reference number IB: 4100/20 in May 2020.

Informed Consent Statement: Informed consent was obtained from all participants, ensuring patient privacy and data security throughout the study.

Data Availability Statement: Data are available through a virtual repository of surgical wound photographs was created. This repository can be accessed through the website redscar.uib.es (accessed on 10 December 2024) [28].

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

References

1. Suetens, C.; Hopkins, S.; Kolman, J.; Diaz Högberg, L. *Point Prevalence Survey of Healthcare-Associated Infections and Antimicrobial Use in European Acute Care Hospitals*; Publications Office of the European Union: Luxembourg, 2011. [CrossRef]
2. Perencevich, E.N.; Sands, K.E.; Cosgrove, S.E.; Guadagnoli, E.; Meara, E.; Platt, R. Health and Economic Impact of Surgical Site Infections Diagnosed after Hospital Discharge. *Emerg. Infect. Dis.* **2003**, *9*, 196. [CrossRef] [PubMed]
3. Lorente-Herce, J.M.; Parra-Membrives, P.; Martínez-Baena, D.; Cañete-Gómez, J.; Segura-Sampedro, J.J. Influencia de La Infección Del Sitio Quirúrgico En El Pronóstico Oncológico Tras Resección Con Intención Curativa Del Cáncer Colorrectal. *Cir. Cir.* **2021**, *89*, 574–582. [CrossRef] [PubMed]
4. Woelber, E.; Schrick, E.J.; Gessner, B.D.; Evans, H.L. Proportion of Surgical Site Infections Occurring after Hospital Discharge: A Systematic Review. *Surg. Infect.* **2016**, *17*, 510–519. [CrossRef] [PubMed]
5. Badia, J.M.; Casey, A.L.; Petrosillo, N.; Hudson, P.M.; Mitchell, S.A.; Crosby, C. Impact of Surgical Site Infection on Healthcare Costs and Patient Outcomes: A Systematic Review in Six European Countries. *J. Hosp. Infect.* **2017**, *96*, 1–15. [CrossRef]
6. Costabella, F.; Patel, K.B.; Adepoju, A.V.; Singh, P.; Attia Hussein Mahmoud, H.; Zafar, A.; Patel, T.; Watekar, N.A.; Mallesh, N.; Fawad, M.; et al. Healthcare Cost and Outcomes Associated With Surgical Site Infection and Patient Outcomes in Low- and Middle-Income Countries. *Cureus* **2023**, *15*, e42493. [CrossRef]
7. Biancuzzi, H.; Dal Mas, F.; Bidoli, C.; Pegoraro, V.; Zantedeschi, M.; Negro, P.A.; Campostrini, S.; Cobianchi, L. Economic and Performance Evaluation of E-Health before and after the Pandemic Era: A Literature Review and Future Perspectives. *Int. J. Env. Res. Public. Health* **2023**, *20*, 4038. [CrossRef]
8. Sanger, P.C.; Hartzler, A.; Han, S.M.; Armstrong, C.A.L.; Stewart, M.R.; Lordon, R.J.; Lober, W.B.; Evans, H.L. Patient Perspectives on Post-Discharge Surgical Site Infections: Towards a Patient-Centered Mobile Health Solution. *PLoS ONE* **2014**, *9*, e114016. [CrossRef]
9. Karim, M.Z.A.; Thamrin, N.M.; Shauri, R.L.A.; Jailani, R.; Manaf, M.H.A.; Mustapa, N.A. Evaluating Telemedicine Diabetes Mellitus: A Mobile Health App for Type-2 Diabetes. *Int. J. Adv. Appl. Sci.* **2024**, *13*, 787–795. [CrossRef]
10. Xu, W.; Huang, X.; Lin, Q.; Wu, T.; Guan, C.; Lv, M.; Hu, W.; Dai, H.; Chen, P.; Li, M.; et al. Application of Alfalfa App in the Management of Oral Anticoagulation in Patients with Atrial Fibrillation: A Multicenter Randomized Controlled Trial. *BMC Med. Inf. Decis. Mak.* **2024**, *24*, 294. [CrossRef]
11. Ruggiero, A.; Martora, F.; Fabbrocini, G.; Villani, A.; Marasca, C.; Megna, M.; Fornaro, L.; Comune, R.; Potestio, L. The Role of Teledermatology During the COVID-19 Pandemic: A Narrative Review. *Clin. Cosmet. Investig. Dermatol.* **2022**, *15*, 2785–2793. [CrossRef]
12. Demographics of Mobile Device Ownership and Adoption in the United States | Pew Research Center. Available online: https://www.pewresearch.org/internet/fact-sheet/mobile/ (accessed on 22 June 2023).
13. Segura-Sampedro, J.J.; Rivero-Belenchón, I.; Pino-Díaz, V.; Rodríguez Sánchez, M.C.; Pareja-Ciuró, F.; Padillo-Ruiz, J.; Jimenez-Rodriguez, R.M. Feasibility and Safety of Surgical Wound Remote Follow-up by Smart Phone in Appendectomy: A Pilot Study. *Ann. Med. Surg.* **2017**, *21*, 58–62. [CrossRef] [PubMed]
14. Armstrong, K.A.; Coyte, P.C.; Brown, M.; Beber, B.; Semple, J.L. Effect of Home Monitoring via Mobile App on the Number of In-Person Visits Following Ambulatory Surgery a Randomized Clinical Trial. *JAMA Surg.* **2017**, *152*, 622–627. [CrossRef] [PubMed]
15. Kufel, J.; Bargieł, K.; Koźlik, M.; Czogalik, Ł.; Dudek, P.; Jaworski, A.; Magiera, M.; Bartnikowska, W.; Cebula, M.; Nawrat, Z.; et al. Usability of Mobile Solutions Intended for Diagnostic Images—A Systematic Review. *Healthcare* **2022**, *10*, 2040. [CrossRef] [PubMed]
16. Wang, S.C.; Au, Y.; Ramirez-Garcialuna, J.L.; Lee, L.; Berry, G.K. The Promise of Smartphone Applications in the Remote Monitoring of Postsurgical Wounds: A Literature Review. *Adv. Ski. Wound Care* **2020**, *33*, 489–496. [CrossRef]
17. Prey, B.J.; Colburn, Z.T.; Williams, J.M.; Francis, A.D.; Vu, M.; Lammers, D.; McClellan, J.; Bingham, J.R. The Use of Mobile Thermal Imaging and Machine Learning Technology for the Detection of Early Surgical Site Infections. *Am. J. Surg.* **2023**, *231*, 60–64. [CrossRef]
18. Wang, S.C.; Anderson, J.A.E.; Evans, R.; Woo, K.; Beland, B.; Sasseville, D.; Moreau, L. Point-of-Care Wound Visioning Technology: Reproducibility and Accuracy of a Wound Measurement App. *PloS ONE* **2017**, *12*, e0183139. [CrossRef]
19. Scheper, H.; Derogee, R.; Mahdad, R.; van der Wal, R.J.P.; Nelissen, R.G.H.H.; Visser, L.G.; de Boer, M.G.J. A Mobile App for Postoperative Wound Care after Arthroplasty: Ease of Use and Perceived Usefulness. *Int. J. Med. Inf.* **2019**, *129*, 75–80. [CrossRef]

20. Strzelecki, M.; Kociołek, M.; Strąkowska, M.; Kozłowski, M.; Grzybowski, A.; Szczypiński, P.M. Artificial Intelligence in the Detection of Skin Cancer: State of the Art. *Clin. Dermatol.* **2024**, *42*, 280–295. [CrossRef]
21. Bossuyt, P.M.; Reitsma, J.B.; Bruns, D.E.; Gatsonis, C.A.; Glasziou, P.P.; Irwig, L.; Lijmer, J.G.; Moher, D.; Rennie, D.; De Vet, H.C.W.; et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. *BMJ* **2015**, *351*, h5527. [CrossRef]
22. Viechtbauer, W.; Smits, L.; Kotz, D.; Budé, L.; Spigt, M.; Serroyen, J.; Crutzen, R. A Simple Formula for the Calculation of Sample Size in Pilot Studies. *J. Clin. Epidemiol.* **2015**, *68*, 1375–1379. [CrossRef]
23. Yip, M.P.; Chang, A.M.; Chan, J.; Mackenzie, A.E. Development of the Telemedicine Satisfaction Questionnaire to Evaluate Patient Satisfaction with Telemedicine: A Preliminary Study. *J. Telemed. Telecare* **2003**, *9*, 46–50. [CrossRef] [PubMed]
24. González-Hidalgo, M.; Moyà-Alcover, G.; Munar, M.; Bibiloni, P.; Craus-Miguel, A.; González-Argenté, X.; Segura-Sampedro, J.J. Detection and Automatic Deletion of Staples in Images of Wound of Abdominal Surgery for M-Health Applications. In *Lecture Notes in Computational Vision and Biomechanics*; Springer: Dordrecht, The Netherlands, 2019; Volume 34, pp. 219–229.
25. Bibiloni, P.; González-Hidalgo, M.; Massanet, S. Soft Color Morphology: A Fuzzy Approach for Multivariate Images. *J. Math. Imaging Vis.* **2019**, *61*, 394–410. [CrossRef]
26. GitHub—Mmunar97/Staples-Detection. Available online: https://github.com/mmunar97/staples-detection (accessed on 10 December 2024).
27. GitHub—Mmunar97/InPYinting. Available online: https://github.com/mmunar97/inPYinting (accessed on 10 December 2024).
28. Redscar | UIB. Available online: http://redscar.uib.es/ (accessed on 10 December 2024).
29. Semple, J.L.; Sharpe, S.; Murnaghan, M.L.; Theodoropoulos, J.; Metcalfe, K.A. Using a Mobile App for Monitoring Post-Operative Quality of Recovery of Patients at Home: A Feasibility Study. *JMIR Mhealth Uhealth.* **2015**, *3*, e3929. [CrossRef] [PubMed]
30. McLean, K.A.; Mountain, K.E.; Shaw, C.A.; Drake, T.M.; Pius, R.; Knight, S.R.; Fairfield, C.J.; Sgrò, A.; Bouamrane, M.; Cambridge, W.A.; et al. Remote Diagnosis of Surgical-Site Infection Using a Mobile Digital Intervention: A Randomised Controlled Trial in Emergency Surgery Patients. *NPJ Digit. Med.* **2021**, *4*, 160. [CrossRef] [PubMed]
31. Nkurunziza, T.; Williams, W.; Kateera, F.; Riviello, R.; Niyigena, A.; Miranda, E.; Bikorimana, L.; Nkurunziza, J.; Velin, L.; Goodman, A.S.; et al. MHealth-Community Health Worker Telemedicine Intervention for Surgical Site Infection Diagnosis: A Prospective Study among Women Delivering via Caesarean Section in Rural Rwanda. *BMJ Glob. Health* **2022**, *7*, 9365. [CrossRef]
32. Jia, H.; Ng, H.; Huang, D.; Rajaratnam, V. Diagnosing Surgical Site Infections Using Telemedicine: A Systematic Review. *Surgeon* **2021**, *20*, e78–e85. [CrossRef]
33. Gunter, R.L.; Fernandes-Taylor, S.; Rahman, S.; Awoyinka, L.; Bennett, K.M.; Weber, S.M.; Greenberg, C.C.; Kent, K.C. Feasibility of an Image-Based Mobile Health Protocol for Postoperative Wound Monitoring. *J. Am. Coll. Surg.* **2018**, *226*, 277–286. [CrossRef]
34. Lin, D.S.H.; Lee, J.K. Mobile Health-Based Thermometer for Monitoring Wound Healing after Endovascular Therapy in Patients with Chronic Foot Ulcer: Prospective Cohort Study. *JMIR Mhealth Uhealth* **2021**, *9*, e26468. [CrossRef]
35. Keegan, A.C.; Bose, S.; McDermott, K.M.; Starks White, M.P.; Stonko, D.P.; Jeddah, D.; Lev-Ari, E.; Rutkowski, J.; Sherman, R.; Abularrage, C.J.; et al. Implementation of a Patient-Centered Remote Wound Monitoring System for Management of Diabetic Foot Ulcers. *Front. Endocrinol.* **2023**, *14*, 1157518. [CrossRef]
36. Li, F.; Wang, M.; Wang, T.; Wang, X.; Ma, X.; He, H.; Ma, G.; Zhao, D.; Yue, Q.; Wang, P.; et al. Smartphone-Based Infrared Thermography to Assess Progress in Thoracic Surgical Incision Healing: A Preliminary Study. *Int. Wound J.* **2022**, *20*, 2000–2009. [CrossRef]
37. Panda, N.; Perez, N.; Tsangaris, E.; Edelen, M.; Pusic, A.; Zheng, F.; Haynes, A.B. Enhancing Patient-Centered Surgical Care With Mobile Health Technology. *J. Surg. Res.* **2022**, *274*, 178–184. [CrossRef] [PubMed]
38. McLean, K.A.; Sgrò, A.; Brown, L.R.; Buijs, L.F.; Daines, L.; Potter, M.A.; Bouamrane, M.M.; Harrison, E.M. Evaluation of Remote Digital Postoperative Wound Monitoring in Routine Surgical Practice. *NPJ Digit. Med.* **2023**, *6*, 85. [CrossRef] [PubMed]
39. Salati, S.A. Smartphone Photography for Smart Assessment of Post-Surgical Wounds—An Experience during the COVID-19 Pandemic. *Pol. J. Surg.* **2021**, *94*, 70–74. [CrossRef] [PubMed]
40. Symer, M.M.; Abelson, J.S.; Milsom, J.; McClure, B.; Yeo, H.L. A Mobile Health Application to Track Patients After Gastrointestinal Surgery: Results from a Pilot Study. *J. Gastrointest. Surg.* **2017**, *21*, 1500–1505. [CrossRef]
41. Eckhauser, F.; Azoury, S.; Farrow, N.; Hu, Q.; Soares, K.; Hicks, C.; Azar, F.; Rodriguez-Unda, N.; Poruk, K.; Cornell, P.; et al. Postoperative Abdominal Wound Infection & Ndash; Epidemiology, Risk Factors, Identification, and Management. *Chronic Wound Care Manag. Res.* **2015**, *137*, 137–148. [CrossRef]
42. Karimian, G.; Petelos, E.; Evers, S.M.A.A. The Ethical Issues of the Application of Artificial Intelligence in Healthcare: A Systematic Scoping Review. *AI Ethics* **2022**, *2*, 539–551. [CrossRef]
43. Benzinger, L.; Ursin, F.; Balke, W.T.; Kacprowski, T.; Salloch, S. Should Artificial Intelligence Be Used to Support Clinical Ethical Decision-Making? A Systematic Review of Reasons. *BMC Med. Ethics* **2023**, *24*, 48. [CrossRef]
44. Mitzner, T.L.; Boron, J.B.; Fausset, C.B.; Adams, A.E.; Charness, N.; Czaja, S.J.; Dijkstra, K.; Fisk, A.D.; Rogers, W.A.; Sharit, J. Older Adults Talk Technology: Technology Usage and Attitudes. *Comput. Hum. Behav.* **2010**, *26*, 1710–1721. [CrossRef]
45. Vaportzis, E.; Clausen, M.G.; Gow, A.J. Older Adults Perceptions of Technology and Barriers to Interacting with Tablet Computers: A Focus Group Study. *Front. Psychol.* **2017**, *8*, 1687. [CrossRef]

46. Gunter, R.; Fernandes-Taylor, S.; Mahnke, A.; Awoyinka, L.; Schroeder, C.; Wiseman, J.; Sullivan, S.; Bennett, K.; Greenberg, C.; Kent, K.C. Evaluating Patient Usability of an Image-Based Mobile Health Platform for Postoperative Wound Monitoring. *JMIR Mhealth Uhealth* **2016**, *4*, e6023. [CrossRef]
47. Wiseman, J.T.; Fernandes-Taylor, S.; Barnes, M.L.; Tomsejova, A.; Saunders, R.S.; Kent, K.C. Conceptualizing Smartphone Use in Outpatient Wound Assessment: Patients' and Caregivers' Willingness to Use Technology. *J. Surg. Res.* **2015**, *198*, 245–251. [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.

tomography

MDPI

Article

Automated Measurement of Effective Radiation Dose by ^{18}F-Fluorodeoxyglucose Positron Emission Tomography/Computed Tomography

Yujin Eom [1], Yong-Jin Park [2], Sumin Lee [2], Su-Jin Lee [2], Young-Sil An [2], Bok-Nam Park [2] and Joon-Kee Yoon [2,*]

[1] Department of AI Mobility Engineering, Ajou University, Suwon 16499, Republic of Korea; eomyj2@ajou.ac.kr
[2] Department of Nuclear Medicine and Molecular Imaging, Ajou University School of Medicine, Suwon 16499, Republic of Korea; yjpark@aumc.ac.kr (Y.-J.P.); suesj202@aumc.ac.kr (S.-J.L.); aysays@aumc.ac.kr (Y.-S.A.); curies@aumc.ac.kr (B.-N.P.)
* Correspondence: jkyoon3@ajou.ac.kr; Tel.: +82-31-219-7839

Abstract: Background/Objectives: Calculating the radiation dose from CT in ^{18}F-PET/CT examinations poses a significant challenge. The objective of this study is to develop a deep learning-based automated program that standardizes the measurement of radiation doses. Methods: The torso CT was segmented into six distinct regions using TotalSegmentator. An automated program was employed to extract the necessary information and calculate the effective dose (ED) of PET/CT. The accuracy of our automated program was verified by comparing the EDs calculated by the program with those determined by a nuclear medicine physician (n = 30). Additionally, we compared the EDs obtained from an older PET/CT scanner with those from a newer PET/CT scanner (n = 42). Results: The CT ED calculated by the automated program was not significantly different from that calculated by the nuclear medicine physician (3.67 ± 0.61 mSv and 3.62 ± 0.60 mSv, respectively, $p = 0.7623$). Similarly, the total ED showed no significant difference between the two calculation methods (8.10 ± 1.40 mSv and 8.05 ± 1.39 mSv, respectively, $p = 0.8957$). A very strong correlation was observed in both the CT ED and total ED between the two measurements ($r^2 = 0.9981$ and 0.9996, respectively). The automated program showed excellent repeatability and reproducibility. When comparing the older and newer PET/CT scanners, the PET ED was significantly lower in the newer scanner than in the older scanner (4.39 ± 0.91 mSv and 6.00 ± 1.17 mSv, respectively, $p < 0.0001$). Consequently, the total ED was significantly lower in the newer scanner than in the older scanner (8.22 ± 1.53 mSv and 9.65 ± 1.34 mSv, respectively, $p < 0.0001$). Conclusions: We successfully developed an automated program for calculating the ED of torso ^{18}F-PET/CT. By integrating a deep learning model, the program effectively eliminated inter-operator variability.

Keywords: ^{18}F-FDG; positron emission tomography; computed tomography; effective dose; deep learning

1. Introduction

Medical radiation is the second largest source of radiation exposure, surpassed only by natural background radiation [1]. The repeated use of diagnostic medical imaging procedures that emit ionizing radiation increases the risk of cancers, thereby necessitating proper management [2,3]. In Korea, it has been reported that the estimated population attributable fraction for all cancers due to diagnostic medical procedures is 0.9%, corresponding to 1915 cancer incidences and 637 cancer-related deaths in 2013 [4].

Positron emission tomography (PET) with ^{18}F-fluorodeoxyglucose (FDG) is a non-invasive diagnostic tool valuable for the staging, management, and prognostication of various malignant tumors [5]. When combined with computed tomography (CT), which provides an anatomical localization of lesions, hybrid PET/CT offers superior diagnostic accuracy compared to PET alone. Additionally, ^{18}F-FDG PET/CT has been utilized for

cancer screening in healthy volunteers due to its advantage of whole-body coverage and high specificity [6].

PET/CT is a significant source of medical radiation, with low-dose CT contributing substantially to the overall radiation dose of PET/CT [7–9]. The radioactivity of radiopharmaceuticals injected for PET imaging primarily depends on the patient's weight. In contrast, radiation exposure from CT varies more widely due to multiple factors, including the patient's weight, tube current, tube voltage, scan range and scan time, ranging from sub-mSv to tens of mSv [10–12]. Calculating the radiation dose from PET/CT is complex, time-consuming and burdensome for nuclear medicine physicians, particularly for CT, which requires whole-body image segmentation. The results may vary depending on the method and the operator.

Several CT dosimetry programs are currently available to calculate organ-absorbed doses, effective doses (ED) or both. However, these programs typically use standard human phantoms instead of individualized segmentation data, potentially leading to inaccurate estimates of whole-body CT radiation doses [13]. Furthermore, variations among dosimetry programs arise from differences in the phantoms used [13,14]. Many existing methods rely on separate calculation tools for CT and PET, where radiation doses are assessed independently. Some CT dosimetry programs also require manual data input by users [15–18], which increases the risk of human error and adds significant time and labor demands. Consequently, hospitals often do not provide information on PET/CT radiation exposure, leaving patients unaware of the amount of radiation they are being exposed to [19,20]. To date, no automated program that can simultaneously calculate radiation doses from both CT and PET has been developed.

Recently, Jacob Wasserthal developed a publicly available CT segmentation program, TotalSegmentator, which leverages deep learning technology [21]. Based on the nnU-Net framework, it is designed to segment 117 anatomical structures, including organs, bones, muscles, and vessels.

This study aimed to develop an automated program that measures radiation doses from ^{18}F-FDG PET/CT while minimizing labor and eliminating operator variability. By integrating a deep-learning based segmentation model (TotalSegmentator) into our program, we aimed to streamline the calculation process and enhance the accuracy in radiation dose estimation.

2. Materials and Method

2.1. FDG PET/CT Protocol

In this study, ^{18}F-FDG PET/CT imaging was performed using either the Discovery STE PET/CT scanner or the Discovery MI PET/CT scanner, both of which are products of GE Healthcare (Milwaukee, WI, USA). Prior to the PET/CT examination, all participants were required to fast for a minimum of 6 h. The blood glucose levels of the participants at the time of ^{18}F-FDG injection were maintained below 150 mg/dL.

The imaging protocol began with the acquisition of unenhanced CT images, which spanned from the top of the skull to the upper thigh, encompassing the torso. Immediately after the CT scan, emission torso PET data were acquired. The protocol for the Discovery STE PET/CT scanner was as follows: an ^{18}F-FDG dosage of 5.1 MBq/kg, a frame duration of 180 s in three-dimensional mode, and a matrix size of 128 $\times$ 128 for PET. The CT parameters were 120 kV, a rotation of 17.5 mm, a tube rotation time of 0.5 s/rev, an acquisition time of 58.6 s, and a matrix size of 512 $\times$ 512. In contrast, the protocol for the Discovery MI PET/CT scanner was slightly different: an ^{18}F-FDG dosage of 4.0 MBq/kg and a frame duration of 90 s in three-dimensional mode for PET. The CT parameters were 120 kV, a rotation of 39.4 mm, and an acquisition time of 50.0 s. CT images from both scanners were acquired using automated tube current modulation mode. The number of CT slices was 263, 299 or 335 slices for the older scanner and 341, 395 or 449 slices for the newer scanner. The PET images were reconstructed using an ordered-subset expectation-maximization

algorithm with 20 subsets and two iterations, and the attenuation correction was based on the CT data.

2.2. Collection and Preprocessing of PET/CT Data

CT and PET images of each patient's torso were retrieved in the Digital Imaging and Communication in Medicine (DICOM) format from the Picture Archiving and Communication System (PACS, INFINITT Healthcare, Seoul, Republic of Korea) (Figure 1). To calculate the CT ED via an automated program, the CT images were converted from DICOM to Neuroimaging Informatics Technology Initiative (NIfTI) format. Subsequently, the NIfTI files were used as input for the TotalSegmentator (ver. 2.2.1, available at https://github.com/wasserth/TotalSegmentator, accessed on 25 March 2024) [21]. We have incorporated this into our automated program for the segmentation of whole-body CTs. In addition, CT dose reports were collected to acquire necessary information for the calculation of CT ED, including the Volume Computed Tomography Dose Index (CTDIvol) and Dose Length Product (DLP) of the torso (Figure 1). The reliability of the segmentation by TotalSegmentator was visually checked in all CT data.

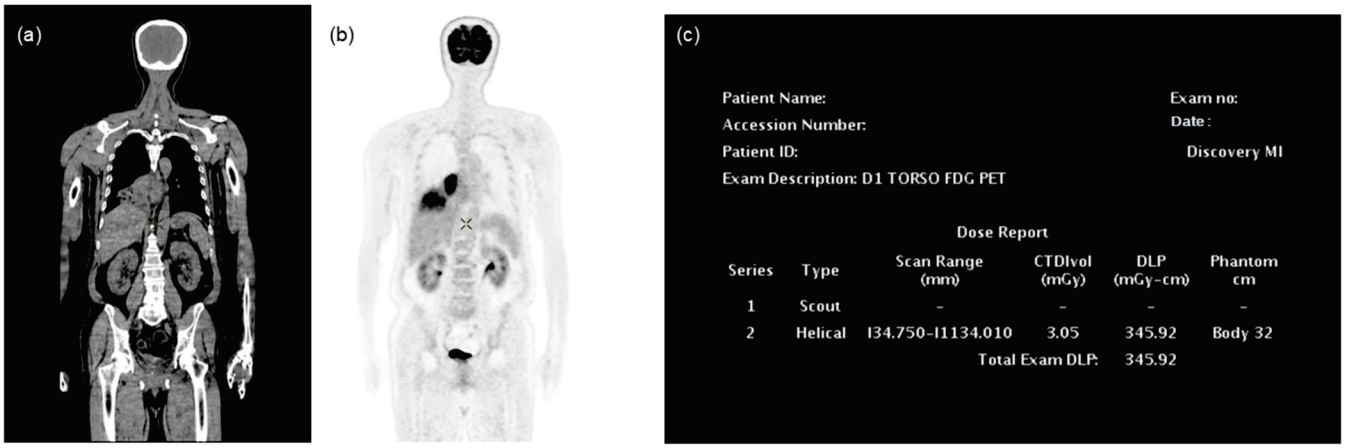


Figure 1. Representative images of PET/CT and a CT dose report from a patient with lung cancer. (**a**) coronal CT (**b**) coronal PET (**c**) CT dose report.

2.3. Effective Dose Calculation

Given that torso CT encompasses several organs, each with distinct tissue conversion factors, we initially segmented it into six body parts: the head (from the top of the skull to C1), the neck (from C1 to T1), the chest (from T1 to the hepatic dome), the abdomen (from the hepatic dome to the iliac crest), the pelvis (from the iliac crest to the bottom of the pelvic bones), and the femur (from the bottom of the pelvic bones to the last slice) (Figure 2). The top of the skull was identified as the fourth slice above from the first slice of the brain, rather than the first slice of the head, because the TotalSegmentator occasionally recognized the radius as the skull when a PET/CT was performed with the arm raised. The iliac crest was defined as the first slice of the iliac bone.

The CT ED was defined as the sum of the EDs of each body part (1), calculated by multiplying the DLP of each body part by its respective conversion factor (2). The DLP for each body part was determined by multiplying the irradiated length of the body part by CTDIvol (3). The irradiated length was calculated by dividing the torso DLP by torso CTDIvol (4). The CT ED (in mSv) can be calculated using the following formulas:

$$\text{CT ED(mSv)} = \sum \text{Part's ED(mSv)} \tag{1}$$

$$\text{Part's ED(mSv)} = \text{Part's DLP(mGy} \times \text{cm)} \times \text{Conversion factor}\left(\frac{\text{mSv}}{\text{mGy} \times \text{cm}}\right) \tag{2}$$

$$\text{Part's DLP}(\text{mGy} \times \text{cm}) = \frac{\text{Number of Part's slice}}{\text{Number of Total slice}} \times \text{Irradiated length}(\text{cm}) \times \text{CTDIvol}(\text{mGy}) \quad (3)$$

$$\text{Irradiated length}(\text{cm}) = \frac{\text{DLP } (\text{mGy} \times \text{cm})}{\text{CTDIvol } (\text{mGy})} \quad (4)$$

In this study, we opted to use the irradiated length instead of the scan range length. The radiation dose, when calculated using the scan range length, could be underestimated compared to the actual values. This is because the irradiated length tends to be longer than the scan range length when CT images are acquired in a helical rotation [22]. The conversion factors, as adopted from ICRP 102, are as follows: 0.0021 mSv/mGy·cm for the head, 0.0059 mSv/mGy·cm for the neck, 0.014 mSv/mGy·cm for the chest, 0.015 mSv/mGy·cm for the abdomen, 0.015 mSv/mGy·cm for the pelvis, and 0.015 mSv/mGy·cm for the femur [23].

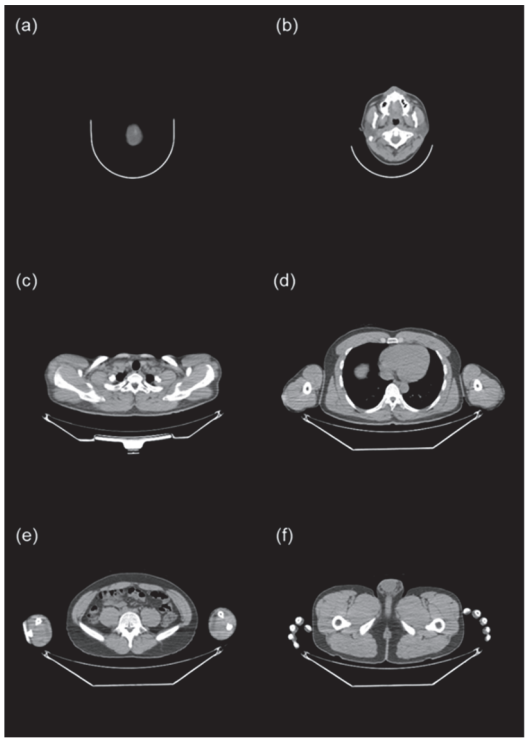


Figure 2. Reference images used for segmenting the torso into six body parts. (**a**) Top of the skull (**b**) C1 (**c**) T1 (**d**) hepatic dome (**e**) iliac crest (**f**) bottom of the pelvic bones.

PET ED was defined as the product of the total radionuclide dose and age-specific weighting factor (5). The PET ED (in mSv) can be calculated using the following formula:

$$\text{PET ED}(\text{mSv}) = \text{Total radionuclide dose}(\text{MBq}) \times \text{Weighting factor} \left(\frac{\text{mSv}}{\text{MBq}}\right) \quad (5)$$

The weighting factors for ^{18}F-FDG, as adopted from ICRP 128, are as follows: 0.095 mSv/MBq for ages 0 to 1 year, 0.056 mSv/MBq for ages 2 to 5 years, 0.037 mSv/MBq for ages 6 to 10 years, 0.024 mSv/MBq for ages 11 to 15 years, and 0.019 mSv/MBq for ages 16 and older [24]. The ED of PET/CT was calculated by adding the EDs of CT and PET scans.

2.4. Automated Effective Dose Calculation Program

In this study, we have developed a program that automatically calculates the ED of ^{18}F-FDG torso PET/CT scans using a Python script (ver. 3.6). The inputs for the ED

calculation program include CT images (in NIfTI format), PET images (in DICOM format), a dose report (in DICOM format), and anatomical part information from TotalSegmentator (in NIfTI format) for each patient. These inputs were uploaded into a single folder. The program is designed to locate the path and extract the necessary information for ED calculation.

The program utilized the *z*-axis information from CT images (in NIfTI format) to calculate the total number of slices. It also identified the maximum and minimum values for the brain, C1, T1, liver, and hip and subsequently calculated the lengths of each body part. To determine the range of the abdomen and the pelvis, the right and left hips were unified as a single entity, referred to as the hip. Information on DLP, CTDIvol, the total dose of radionuclide, and age were obtained from the DICOM tags of dose reports and PET images (Figure 3).

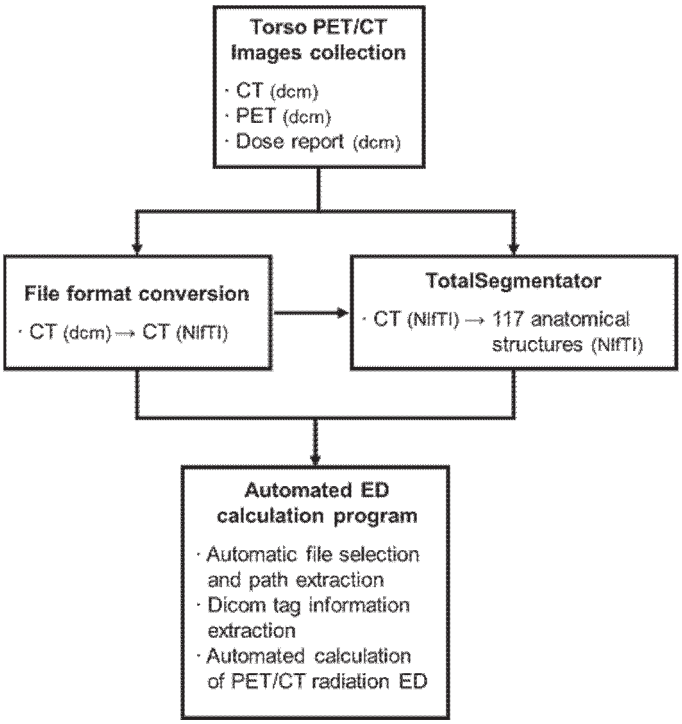


Figure 3. The diagram of the automated ED calculation program for ^{18}F-FDG PET/CT.

2.5. Verification of the Automated ED Calculation Program

We prepared two datasets to verify the automated ED calculation program. The first dataset comprises 30 consecutive ^{18}F-FDG PET/CT scans performed in March 2024 for preoperative staging of cancers (Group 1). All scans in Group 1 were acquired using a newer scanner model (GE Discovery MI, GE Healthcare, Milwaukee, WI, USA). Using this dataset, we conducted three evaluations: (1) a comparison of the CT ED and total ED calculated by the automated program with those calculated by a nuclear medicine physician (J-K.Y.), the 'CT-Expo' program (ver. 2.5, SASCRAD), and whole-body conversion factors (WB CF, 0.0082 and 0.0015) [25–27]; (2) an assessment of the correlation between the EDs derived from the automated program and those determined by the nuclear medicine physician; and (3) an evaluation of the program's repeatability and reproducibility using ED calculations from ten randomly selected PET/CT scans.

The second dataset includes 84 paired ^{18}F-FDG PET/CT scans from 42 patients, conducted from April 2023 to April 2024 for either staging or monitoring the therapeutic response of cancers (Group 2). Each pair of PET/CT scans were acquired using different equipment (either older or newer PET/CT scanners) within a one-year period. With this dataset, we evaluated whether our automated program worked properly with both scanners, as the newer scanner incorporates significantly different technologies. In addition,

we assessed whether there was a significant difference in radiation dose between these scanners and investigated the causes of this difference.

The clinical characteristics of the patients in Group 1 (n = 30) and Group 2 (n = 42) are shown in Table 1. Group 1 consisted of 12 male and 18 female patients, with an average age of 60 ± 11 years. Group 2 comprised 24 male and 18 female patients, with an average age of 60 ± 12 years. The mean body weight was 60.4 ± 12.5 kg for Group 1 and 59.4 ± 11.3 kg for Group 2. In Group 1, lung cancer was the most common disease, accounting for 8 out of 30 patients (26.7%), followed by endometrial cancer, which affected 7 out of 30 patients (23.3%). In Group 2, lymphoma was the most common disease, accounting for 7 out of 42 patients (16.7%), with lung cancer being the second most common, affecting 6 out of 42 patients (14.3%).

Table 1. Clinical characteristics of patients.

		Number of Patients (%)	**Mean ± SD**
Group 1 (n = 30)			
Age (years)			60 ± 11
Sex	Male/female	12:18 (40.0:60.0)	
Weight (kg)			60.4 ± 12.5
Type of Cancers	Lung	8 (26.7)	
	Endometrial	7 (23.3)	
	Esophageal	3 (10.0)	
	Ovary	3 (10.0)	
	Renal cell	2 (6.7)	
	Breast	1 (3.3)	
	Cervix	1 (3.3)	
	Colon	1 (3.3)	
	Lymphoma	1 (3.3)	
	Nasal	1 (3.3)	
	Rectal	1 (3.3)	
	Vocal cord	1 (3.3)	
Group 2 (n = 42)			
Age (years)			60 ± 12
Sex	Male/female	24:18 (57.1:42.9)	
Weight (kg)			59.4 ± 11.3
Type of Cancers	Lymphoma	7 (16.7)	
	Lung	6 (14.3)	
	Stomach	5 (11.9)	
	Breast	4 (9.5)	
	Rectal	3 (7.1)	
	Renal cell	3 (7.1)	
	Cervix	2 (4.8)	
	Colon	2 (4.8)	
	Pancreatic	2 (4.8)	
	Prostate	2 (4.8)	
	AoV	1 (2.4)	
	Esophageal	1 (2.4)	
	GB	1 (2.4)	
	Head and Neck	1 (2.4)	
	Klatskin	1 (2.4)	
	Ovary	1 (2.4)	

Abbreviations: SD, standard deviation; AoV, Ampulla of Vater; GB, gallbladder.

In Group 2, 21 patients underwent ^{18}F-FDG PET/CT scans using an older scanner for the initial imaging, followed by a newer scanner for follow-up imaging. The remaining 21 patients underwent the process in reverse order. The mean interval between the initial imaging and follow-up imaging for Group 2 was 165 ± 76 days.

2.6. External Validation of the Automated ED Calculation Program

For the external validation of the automated program, we collected 34 ^{18}F-FDG PET/CT scans acquired from 25 other hospitals between September and November 2024, which were brought by patients upon transfer (Group 3). All 34 scans were performed to evaluate patients with cancers. Sixteen scans were acquired using PET/CT scanners from Siemens Healthineers (four with Biograph 128, six with Biograph 64, three with Biograph 40, and three with Biograph 16, Forchheim, Germany). Twelve scans were acquired using PET/CT scanners from GE Healthcare (seven with Discovery MI, four with Discovery 710, and one with Discovery 690, Milwaukee, WI, USA). The remaining six scans were acquired using PET/CT scanners from Philips Healthcare (four with Gemini TF and two with Ingenuity TF, Eindhoven, The Netherlands). We aimed to determine whether our program could measure the radiation dose from these scans and to investigate the causes of any measurement failures.

2.7. Statistical Analysis

Data are represented as the mean ± standard deviation. The Student's *t*-test was employed to compare the CT ED, PET ED and total ED between the nuclear medicine physician and the automated program, and between the older and newer PET/CT scanners. A *p*-value of less than 0.05 was considered statistically significant. The correlation coefficient (r^2) was also determined. Bland–Altman plots were used to display the differences (with a 95% confidence interval) in EDs as calculated by the nuclear medicine physician and the automated program. The maximum allowed differences were calculated as described [28]. Statistical analyses were performed using Medcalc software (ver. 22.030, Medcalc Software Ltd., Ostend, Belgium) or RStudio (AGPL ver. 3). To evaluate the repeatability of the automated program, intraclass correlation coefficients were analyzed.

3. Results

3.1. Comparison of PET/CT EDs: Automated Program vs. Conventional Methods

Table 2 displays the EDs of ^{18}F-FDG PET/CT calculated using various methods. Figure 4a compares the EDs for CT and PET/CT as determined by the automated program, a nuclear medicine physician, the 'CT-Expo' program, and WB CFs.

The CT ED calculated by the automated program (3.67 ± 0.61 mSv) showed no significant difference compared to that calculated by the nuclear medicine physician (3.62 ± 0.60 mSv, $p = 0.7623$). However, the CT ED from the automated program was significantly lower than those obtained using the 'CT-Expo' program (by 30.2% (4.78 ± 0.74 mSv, $p < 0.0001$)) or WB CF (0.015) (by 26.1% (4.63 ± 0.75 mSv, $p < 0.0001$)) and significantly higher than that calculated using WB CF (0.0082) (by 31.1% (2.53 ± 0.41 mSv, $p < 0.0001$)). The PET ED was identical for both the automated program and the nuclear medicine physician (4.43 ± 0.86 mSv), as it depends on the injected dose of ^{18}F-FDG and the patient's age. Similarly, there was no significant difference in total EDs between the automated program and the physician (8.10 ± 1.40 mSv and 8.05 ± 1.39 mSv, respectively, $p = 0.8957$). In contrast, the total ED calculated by the automated program was significantly lower than those obtained using the 'CT-Expo' program (by 13.7% (9.21 ± 1.52 mSv, $p = 0.0047$)) or WB CF (0.015) (by 11.8% (9.06 ± 1.53 mSv, $p = 0.0143$)) and significantly higher than that calculated using WB CF (0.0082) (by 14.1% (6.96 ± 1.22 mSv, $p = 0.0014$)).

The CT and total EDs calculated by the automated program demonstrated a very strong correlation with those calculated by the physician ($r^2 = 0.9981$ for CT and $r^2 = 0.9996$ for PET/CT, Figure 4b). The regression analysis revealed slopes of 1.010 and 1.004, with intercepts of 0.008 and 0.005 for the CT ED and total ED, respectively.

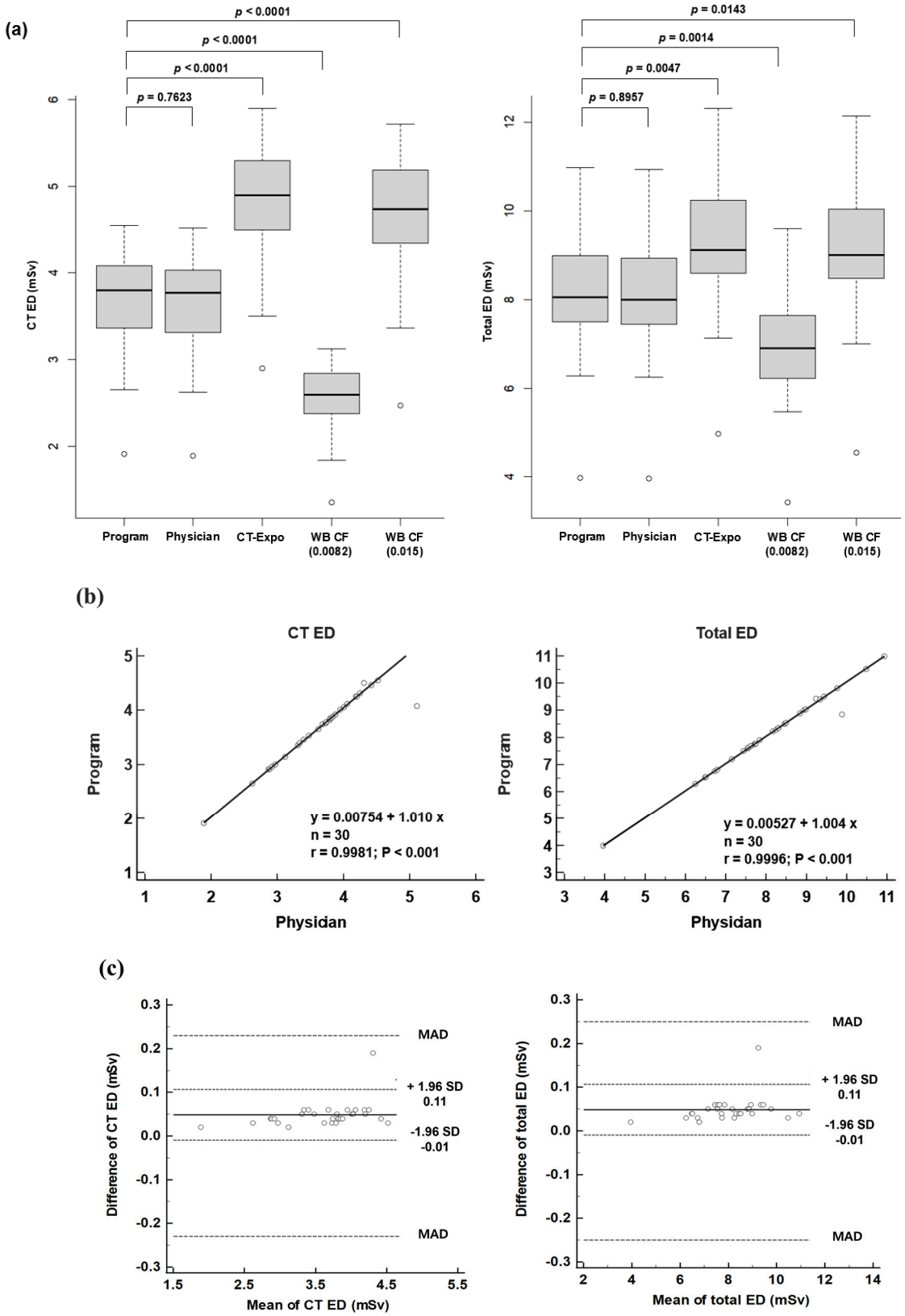


Figure 4. (**a**) Box and whisker plots of effective doses for CT and PET/CT (total) calculated by various methods; automated program, nuclear medicine physician, CT-Expo' program and whole-body conversion factors (0.0082 and 0.0015). (**b**) Correlation between effective doses calculated by the automated program and the nuclear medicine physician. (**c**) Bland–Altman plots for the differences in effective doses of CT and PET/CT (total) calculated by the automated program and a nuclear medicine physician; MAD, maximum allowed difference.

Figure 4c illustrates the differences in EDs between the automated program and the nuclear medicine physician. For CT ED, the difference was 0.05, with a limit of agreement (LoA) ranging from −0.01 to 0.11. Similarly, for the total ED, the difference was 0.05, with a LoA ranging from −0.01 to 0.11. The maximum allowed differences, calculated from the coefficients of variation of both methods, were 0.23 for the CT ED and 0.25 for the total ED. One patient exhibited a higher CT ED (0.19) and total ED (0.19) than the upper LoA, but both values were within the range of the maximum allowed difference. The other values were within the LoAs for both the CT ED and total ED.

To assess repeatability and reproducibility, ED measurements were repeated three times for 10 randomly selected PET/CT scans at a different time point from the initial measurement. The results showed that all repeated ED calculations were identical across the 10 scans and consistent with the initial measurements (Supplementary File S1). Interclass correlation coefficients for the CT ED, PET ED and total ED were 1.0, 1.0 and 1.0, respectively. These findings demonstrate that the automated program exhibits excellent repeatability and reproducibility in measuring PET/CT ED.

Table 2. Comparison of the CT EDs and total EDs among various methods.

Patient	PET ED	Automated Program		Physician		CT-Expo		WB CF (0.0082 *)		WB CF (0.015 *)	
		CT	Total	CT	Total	CT	Total	CT	Total	CT	Total
1	6.52	4.46	10.98	4.42	10.94	5.80	12.32	3.08	9.60	5.63	12.15
2	5.76	4.06	9.82	4.01	9.77	5.30	11.06	2.84	8.60	5.19	10.95
3	5.96	4.55	10.51	4.52	10.48	5.90	11.86	3.12	9.08	5.72	11.68
4	5.38	4.12	9.50	4.06	9.44	5.30	10.68	2.81	8.19	5.14	10.52
5	5.14	3.89	9.03	3.85	8.99	5.10	10.24	2.68	7.82	4.91	10.05
6	5.15	4.25	9.40	4.19	9.34	5.60	10.75	2.94	8.09	5.38	10.53
7	4.93	4.50	9.43	4.31	9.24	5.60	10.53	2.99	7.92	5.47	10.40
8	4.69	4.31	9.00	4.25	8.94	5.60	10.29	2.95	7.64	5.39	10.08
9	4.46	3.82	8.28	3.79	8.25	4.90	9.36	2.59	7.05	4.74	9.20
10	4.64	4.25	8.89	4.20	8.84	5.60	10.24	2.96	7.60	5.41	10.05
11	4.57	3.78	8.35	3.74	8.31	4.90	9.47	2.59	7.16	4.73	9.30
12	4.77	4.08	8.85	4.03	8.80	5.40	10.17	2.85	7.62	5.22	9.99
13	4.58	3.92	8.50	3.88	8.46	5.10	9.68	2.69	7.27	4.92	9.50
14	4.68	3.86	8.54	3.82	8.50	5.00	9.68	2.68	7.36	4.90	9.58
15	4.16	3.74	7.90	3.68	7.84	4.80	8.96	2.57	6.73	4.70	8.86
16	4.37	3.85	8.22	3.80	8.17	4.90	9.27	2.62	6.99	4.79	9.16
17	4.23	3.40	7.63	3.34	7.57	4.50	8.73	2.39	6.62	4.37	8.60
18	4.23	3.46	7.69	3.40	7.63	4.50	8.73	2.40	6.63	4.40	8.63
19	3.49	4.01	7.50	3.95	7.44	5.10	8.59	2.73	6.22	4.99	8.48
20	4.11	3.65	7.76	3.62	7.73	4.70	8.81	2.50	6.61	4.58	8.69
21	3.67	3.53	7.20	3.48	7.15	4.60	8.27	2.44	6.11	4.47	8.14
22	4.01	3.76	7.77	3.73	7.74	4.90	8.91	2.59	6.60	4.73	8.74
23	4.23	3.36	7.59	3.31	7.54	4.50	8.73	2.38	6.61	4.35	8.58
24	3.67	3.14	6.81	3.12	6.79	4.10	7.77	2.17	5.84	3.97	7.64
25	3.76	3.00	6.76	2.97	6.73	3.90	7.66	2.08	5.84	3.80	7.56
26	3.63	2.65	6.28	2.62	6.25	3.50	7.13	1.84	5.47	3.37	7.00
27	4.8	2.96	7.76	2.92	7.72	3.80	8.60	2.03	6.83	3.71	8.51
28	3.62	2.92	6.54	2.88	6.50	3.80	7.42	2.00	5.62	3.66	7.28
29	3.62	2.91	6.53	2.87	6.49	3.80	7.42	2.02	5.64	3.70	7.32
30	2.07	1.91	3.98	1.89	3.96	2.90	4.97	1.35	3.42	2.47	4.54

Abbreviations: PET, positron emission tomography; CT, computed tomography; ED, effective dose; WB CF, whole-body conversion factor (* mSv/mGy·cm).

3.2. Comparison of PET/CT EDs: Older Scanner vs. Newer Scanner

The average tube currents of CT scanners were 61.1 ± 9.8 mA (ranging from 35.4 to 77.0 mA) for the newer scanner and 66.3 ± 0.7 mA (ranging from 64.1 to 68.2 mA) for the older scanner. The scan lengths were 1099 ± 35 mm for the newer scanner and

1052 ± 42 mm for the older scanner. The injected doses of ^{18}F-FDG were 228.3 ± 45.1 MBq for the newer scanner and 318.6 ± 59.2 MBq for the older scanner.

Figure 5 shows the EDs of CT, PET and PET/CT calculated from both newer and older scanners. The ED for CT was similar between the newer and older scanners (3.65 ± 0.26 mSv and 3.83 ± 0.75 mSv, respectively, p = 0.1667). Using 1 mSv as a cut-off, the CT ED was higher with the newer scanner in four patients (5.30 vs. 2.93 mSv; 5.63 vs. 3.89 mSv; 4.50 vs. 3.33 mSv; and 5.65 vs. 3.92 mSv). In contrast, the CT ED was higher with the older scanner in three patients (2.14 vs. 3.37 mSv; 2.67 vs. 3.89 mSv; and 2.28 vs. 3.40 mSv).

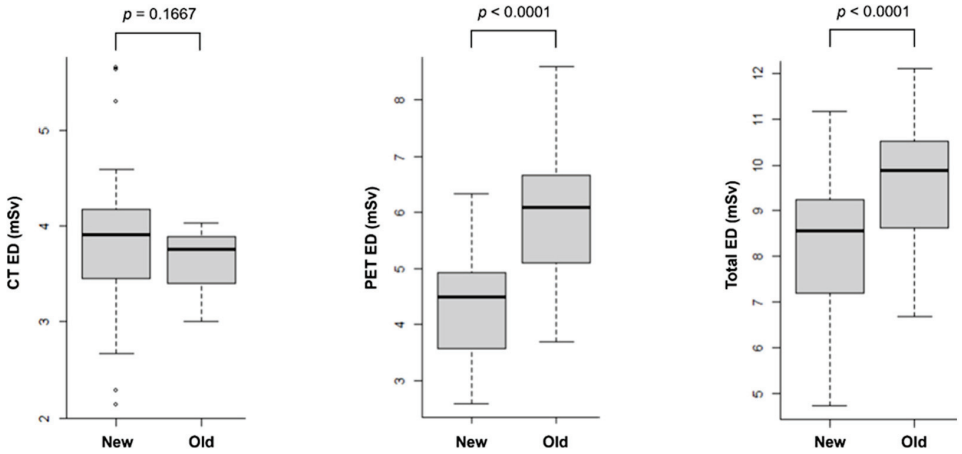


Figure 5. Box and whisker plots of effective doses for CT, PET, and PET/CT (total) from newer and older scanners.

Meanwhile, the PET ED was significantly lower with the newer scanner (4.39 ± 0.91 mSv) compared to the older scanner (6.00 ± 1.17 mSv, p < 0.0001). Consequently, the total ED for PET/CT was also significantly lower with the newer scanner (8.22 ± 1.53 mSv) than with the older scanner (9.65 ± 1.34 mSv, p < 0.0001).

For the 21 patients who underwent PET/CT examinations using the newer scanner followed by the older scanner, the average weight change was −2.4 ± 4.2 kg. The CT ED did not differ significantly between the two examinations (3.96 ± 0.70 mSv for the newer scanner vs. 3.70 ± 0.27 mSv for the older scanner, p = 0.141). However, the PET ED was significantly lower with the newer scanner (4.61 ± 0.95 mSv) compared to the older scanner (6.26 ± 1.29 mSv, p < 0.001). Consequently, the total ED was significantly reduced with the newer scanner (8.58 ± 1.46 mSv) compared to the older scanner (9.96 ± 1.44 mSv, p < 0.001).

In a separate group of 21 patients who initially underwent PET/CT examinations using the older scanner and subsequently the newer scanners, the average weight change was −0.4 ± 3.1 kg. Consistent with the first group, there was no significant difference in the CT ED between the two scanners (3.60 ± 0.25 mSv for the older scanner vs. 3.69 ± 0.79 mSv for the newer scanner, p = 0.638). Similarly, the PET ED showed a significant reduction with the newer scanner (4.18 ± 0.84 mSv) compared to the older scanner (5.73 ± 0.99 mSv, p < 0.001). Consequently, the total ED was significantly lower with the newer scanner (7.87 ± 1.56 mSv) compared to the older scanner (9.34 ± 1.17 mSv, p < 0.001).

3.3. External Validation of the Automated ED Calculation Program

The automated program successfully calculated EDs for 18 out of 34 scans (52.9%). For these scans, the EDs were 3.67 ± 1.89 mSv (range: 1.83–10.31) for CT, 4.85 ± 1.12 mSv (range: 3.10–6.68) for PET, and 8.52 ± 2.47 mSv (range: 5.99–16.99) for PET/CT. Measurement failures occurred in 16 scans (47.1%) due to the absence of a dose report form in 12 cases (75%) and the lack of ^{18}F-FDG dose information in the DICOM headers in four cases (25%). These issues may have arisen during the process of archiving data from PACS for transfer.

4. Discussion

In this study, we successfully developed a program to automate the calculation of the effective radiation dose for torso ^{18}F-FDG PET/CT. While several studies have addressed the calculation of EDs for PET/CT, to the best of our knowledge, this is the first study to automate calculation of ED by using a deep learning model. Our results showed that the EDs of the PET/CT obtained by the automated program showed a strong correlation with those calculated by a nuclear medicine physician. Additionally, the CT segmentation in this program was standardized using TotalSegmentator-derived anatomic information, thereby eliminating the inter-operator variability.

PET/CT is a hybrid imaging modality that combines both PET and CT, both of which involve the emission of ionizing radiation. Consistent with our findings, the ED from ^{18}F-FDG PET generally remains below 10 mSv per examination [29,30]. Advancements in technology have further reduced radiation exposure by enhancing PET sensitivity without compromising image quality [31]. In contrast, the CT ED varies significantly, with low-dose protocols typically ranging from 1 to 10 mSv, while diagnostic protocols often generally exceed 10 mSv [29,30]. As a result, the total ED can surpass 20 mSv, potentially leading clinicians or patients to reconsider undergoing PET/CT examinations. Unlike PET, the CT component of PET/CT lacks such strict standardization, making the calculation of radiation exposure from torso or whole-body CT scans more complex and challenging.

Radiation exposure from PET/CT can be calculated in various ways. As demonstrated in our study, one common approach involves segmenting the torso into five to six regions and calculating the ED using CTDIvol and DLP values measured by CT scanners in conjunction with tissue conversion factors (DLP method) [32]. However, the DLP method has notable limitations. Its accuracy depends on the size of the phantom used, often leading to an underestimation of the ED, particularly in pediatric patients [33]. Beyond the use of phantoms, how DLP values are determined by CT scanners remains largely unknown. Moreover, the DLP method is less ideal than organ-absorbed dose calculations, as it does not assess radiation risks to individual organs [13]. Despite these drawbacks, the DLP method remains widely used due to its simplicity and practicality. It provides a single, standardized metric, making it convenient for cancer risk assessment related to radiation exposure. Additionally, it is valuable for comparing radiation doses across imaging modalities that emit ionizing radiation, thereby supporting protocol optimization [33,34].

Radiation exposure can alternatively be estimated by inputting CT parameters into dosimetry programs, which calculate both organ-specific and total radiation doses [13]. These dosimetry programs are broadly categorized into three types: stand-alone programs that operate with (e.g., CT-Expo, ImPACT) [8,16,17] or without Microsoft Excel (e.g., ImpactDose, NCICT, WinDose) [35]; PACS-integrated tools (e.g., DoseWatch, Radimetrics) [36]; and web-based platforms (e.g., VirtualDose). Summing the organ-absorbed doses calculated by these tools aligns well with the ICRP definition of ED. Furthermore, these programs allow users to select patient-specific phantoms such as age, gender, and body size [13]. However, limitations persist, particularly regarding the phantoms used in these programs [13]. Variability in computational human phantoms among programs leads to significant differences in ED estimates [14,37,38]. For organs within the scan range, discrepancies as high as 67% have been reported. Additionally, these tools are constrained by limited scanner model options and mismatches between computational human phantoms and actual patient anatomy, which can result in inaccurate dose estimations [13]. Using individualized anatomical models derived from patients instead of standard computational phantoms remains a labor-intensive process [33].

In this study, we compared EDs calculated by an automated program to those obtained using the "CT-Expo" software. As reported in the results, the automated program yielded significantly lower CT and total EDs compared to "CT-Expo", by 30.2% and 13.7%, respectively. These findings align with previous studies indicating that EDs calculated using the DLP method were generally lower—ranging from 4% to 74%—compared to dosimetry programs, except for head scans [22].

The simplest method to estimate the ED of whole-body CT is to multiply the DLP by whole-body tissue conversion factor [25–27]. In this study, applying a conversion factor of 0.0082 mSv/mGy·cm to Group 1 resulted in estimated CT ED and total ED values that were 31.1% and 14.1% lower, respectively, than those calculated by the automated program. Conversely, using a conversion factor of 0.015 mSv/mGy·cm produced estimated CT ED and total ED values that were 26.1% and 11.8% higher, respectively, than those calculated by the automated program [39]. Therefore, using a whole-body tissue conversion factor to estimate CT ED can lead to either underestimation or overestimation of the ED for ^{18}F-FDG PET/CT.

When comparing older and newer PET/CT scanners, the CT ED was not significantly different. The newer CT scanner had a longer scan length and a higher number of the CT slices, while the average tube current was lower than that of the older scanner. In four patients, the CT ED was more than 1 mSv higher with the newer scanner compared to the older scanner, resulting from the higher tube voltage (140 keV vs. 120 keV) [10]. PET/CT technicians increased the tube voltage to accommodate the body shape of these patients. The tube currents for these CT scans were similar between the two scanners. Conversely, in three patients, the CT ED was lower with the newer scanner compared to the older scanner. In these cases, reduced tube currents (35.4, 38.1 and 44.5 mA) contributed to the lower CT ED, as the device automatically adjusted the tube current settings.

The PET ED was significantly lower in the newer scanner compared to the older scanner. The newer PET scanner has higher sensitivity for gamma photons, leading to a significant reduction in the injected dose of ^{18}F-FDG [31]. Additionally, weight changes due to surgery or anti-cancer chemotherapy between the initial and follow-up imaging influenced the difference in PET ED. For the 21 patients who underwent PET/CT scans in the order of older to newer scanners, the PET ED decreased by 3.8% due to a 2.3 kg reduction in weight. In contrast, for the remaining 21 patients who underwent scans in the reverse order, the PET ED decreased by 0.8% due to a 0.4 kg reduction in weight. However, the influence of weight change on the PET ED was relatively small compared to the change in the injected dose. A comparison of PET/CT EDs between an older scanner and a newer scanner confirmed that the automated program functions accurately, regardless of the scanner model and acquisition-reconstruction parameters.

One limitation of this automated program is that when PET/CT is performed with the arms raised, TotalSegmentator may incorrectly identify certain anatomical structures. For example, it may mistakenly recognize the radius as the skull and the carpal bones as the C1 vertebra. Consequently, when dividing the torso into body parts, subtle discrepancies can arise between the actual CT ED and the calculated CT ED due to the incorrect segmentation of the head and neck regions. To address this issue, this study set the top of the skull at four slices above the brain. However, attention must be paid to patients' arm positions, and the segmentation standards should be adjusted in accordance with the protocols of each institution.

Another limitation is that this study was conducted at a single center. To address this, the automated program was tested on 34 PET/CT scans from other hospitals. We found that the measurement failures were primarily due to the absence of dose report forms and missing dose information in the DICOM files. While these findings provide valuable insights, future multi-center studies are necessary to improve the program's generalizability and robustness.

5. Conclusions

In summary, we developed an automated program to measure radiation doses for ^{18}F-FDG PET/CT, demonstrating strong correlation with manually calculated ED. The program exhibited excellent repeatability and reproducibility while effectively reducing inter-operator variability through the integration of a deep learning-based program. Furthermore, it performed reliably regardless of the scanner model or acquisition-reconstruction parameters. By simplifying the workflow to require only file conversion and placement in a designated folder, this automated program is expected to significantly reduce a physician's workload.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/tomography10120151/s1, File S1: Repeated measurement of the CT, PET and total EDs using the automated program.

Author Contributions: Conceptualization, J.-K.Y.; methodology, J.-K.Y., Y.-J.P. and Y.E.; software, Y.E.; validation, S.-J.L., S.L., B.-N.P. and Y.-S.A.; formal analysis, Y.E.; investigation, J.-K.Y. and Y.E.; data curation, Y.E.; writing—original draft preparation, Y.E.; writing—review and editing, J.-K.Y.; funding acquisition, J.-K.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1F1A1076500) and 'Korea Government Grant Program for Education and Research in Medical AI' through the Korea Health Industry Development Institute (KHIDI), funded by the Korea government (MOE, MOHW).

Institutional Review Board Statement: This study was approved by the institutional review board (approval no. AJOUIRB-DB-2024-345; approval date: 17 July 2024) and was conducted in accordance with the ethical standards of the 1964 Declaration of Helsinki and its later amendments.

Informed Consent Statement: The institutional review board waived the requirement for informed consent.

Data Availability Statement: A part of the data is available in a Supplementary File. The rest of the data is available upon reasonable request, by contacting the corresponding author.

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

References

1. UNEP. Radiation Effects and Sources. In *What Is Radiation? What Does Radiation Do to Us? Where Does Radiation Come from?* UNED, ed.; UNSCEAR: Vienna, Austria, 2016; pp. 1–10, 13–17, 27–37.
2. Richardson, D.B.; Leuraud, K.; Laurier, D.; Gillies, M.; Haylock, R.; Kelly-Reif, K.; Bertke, S.; Daniels, R.D.; Thierry-Chef, I.; Moissonnier, M.; et al. Cancer mortality after low dose exposure to ionising radiation in workers in France, the United Kingdom, and the United States (INWORKS): Cohort study. *BMJ* **2023**, *382*, e074520. [CrossRef] [PubMed]
3. Martin, C.J.; Barnard, M. How much should we be concerned about cumulative effective doses in medical imaging? *J. Radiol. Prot.* **2022**, *42*, 011514. [CrossRef] [PubMed]
4. Lim, H.; Choi, J.; Kim, J.H.; Cheong, H.K.; Ha, M. Estimation of Cancer Incidence and Mortality Risks Attributed to Diagnostic Medical Radiation Exposure in Korea, 2013. *J. Korean Med. Sci.* **2018**, *33*, e211. [CrossRef] [PubMed]
5. Hustinx, R.; Benard, F.; Alavi, A. Whole-body FDG-PET imaging in the management of patients with cancer. *Semin. Nucl. Med.* **2002**, *32*, 35–46. [CrossRef]
6. Kojima, S.; Zhou, B.; Teramukai, S.; Hara, A.; Kosaka, N.; Matsuo, Y.; Suzuki, H.; Torigoe, S.; Suzuki, T.; Uno, K.; et al. Cancer screening of healthy volunteers using whole-body ^{18}F-FDG-PET scans: The Nishidai clinic study. *Eur. J. Cancer* **2007**, *43*, 1842–1848. [CrossRef]
7. Hosono, M.; Takenaka, M.; Monzen, H.; Tamura, M.; Kudo, M.; Nishimura, Y. Cumulative radiation doses from recurrent PET-CT examinations. *Br. J. Radiol.* **2021**, *94*, 20210388. [CrossRef]
8. Kwon, H.W.; Kim, J.P.; Lee, H.J.; Paeng, J.C.; Lee, J.S.; Cheon, G.J.; Lee, D.S.; Chung, J.K.; Kang, K.W. Radiation Dose from Whole-Body F-18 Fluorodeoxyglucose Positron Emission Tomography/Computed Tomography: Nationwide Survey in Korea. *J. Korean Med. Sci.* **2016**, *31* (Suppl. S1), S69–S74. [CrossRef]
9. Huang, B.; Law, M.W.; Khong, P.L. Whole-body PET/CT scanning: Estimation of radiation dose and cancer risk. *Radiology* **2009**, *251*, 166–174. [CrossRef]
10. Elojeimy, S.; Tipnis, S.; Huda, W. Relationship between radiographic techniques (kilovolt and milliampere-second) and CTDI(VOL). *Radiat. Prot. Dosim.* **2010**, *141*, 43–49. [CrossRef]
11. Son, H.K.; Lee, S.H.; Nam, S.; Kim, H.J. Radiation dose during CT scan with PET/CT clinical protocols. In Proceedings of the 2006 IEEE Nuclear Science Symposium Conference Record, San Diego, CA, USA, 29 October–1 November 2006; pp. 2210–2214.
12. Adeleye, B.; Chetty, N. Radiation dose from ^{18}F-FDG PET/CT procedures: Influence of specific CT model and protocols. *Radioprotection* **2018**, *53*, 107–113. [CrossRef]
13. Lee, C. A review of organ dose calculation tools for patients undergoing compuated tomography scans. *J. Radiat. Prot. Res.* **2021**, *46*, 151–159. [CrossRef]
14. De Mattia, C.; Campanaro, F.; Rottoli, F.; Colombo, P.E.; Pola, A.; Vanzulli, A.; Torresin, A. Patient organ and effective dose estimation in CT: Comparison of four software applications. *Eur. Radiol. Exp.* **2020**, *4*, 14. [CrossRef] [PubMed]
15. Inoue, Y.; Nagahara, K.; Inoki, Y.; Hara, T.; Miyatake, H. Clinical evaluation of CT radiation dose in whole-body ^{18}F-FDG PET/CT in relation to scout imaging direction and arm position. *Ann. Nucl. Med.* **2019**, *33*, 169–176. [CrossRef] [PubMed]
16. Inoue, Y.; Nagahara, K.; Tanaka, Y.; Miyatake, H.; Hata, H.; Hara, T. Methods of CT Dose Estimation in Whole-Body ^{18}F-FDG PET/CT. *J. Nucl. Med.* **2015**, *56*, 695–700. [CrossRef]

17. Quinn, B.; Dauer, Z.; Pandit-Taskar, N.; Schoder, H.; Dauer, L.T. Radiation dosimetry of ^{18}F-FDG PET/CT: Incorporating exam-specific parameters in dose estimates. *BMC Med. Imaging* **2016**, *16*, 41. [CrossRef]
18. Quinn, B.M.; Gao, Y.; Mahmood, U.; Pandit-Taskar, N.; Behr, G.; Zanzonico, P.; Dauer, L.T. Patient-adapted organ absorbed dose and effective dose estimates in pediatric ^{18}F-FDG positron emission tomography/computed tomography studies. *BMC Med. Imaging* **2020**, *20*, 9. [CrossRef]
19. Ribeiro, A.S.F.; Husson, O.; Drey, N.; Murray, I.; May, K.; Thurston, J.; Oyen, W.J.G. Radiation exposure awareness from patients undergoing nuclear medicine diagnostic ^{99m}Tc-MDP bone scans and 2-deoxy-2-(^{18}F) fluoro-D-glucose PET/computed tomography scans. *Nucl. Med. Commun.* **2020**, *41*, 582–588. [CrossRef]
20. Ukkola, L.; Oikarinen, H.; Henner, A.; Haapea, M.; Tervonen, O. Patient information regarding medical radiation exposure is inadequate: Patients' experience in a university hospital. *Radiography* **2017**, *23*, e114–e119. [CrossRef]
21. Wasserthal, J.; Breit, H.C.; Meyer, M.T.; Pradella, M.; Hinck, D.; Sauter, A.W.; Heye, T.; Boll, D.T.; Cyriac, J.; Yang, S.; et al. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiol. Artif. Intell.* **2023**, *5*, e230024. [CrossRef]
22. Christner, J.A.; Kofler, J.M.; McCollough, C.H. Estimating effective dose for CT using dose-length product compared with using organ doses: Consequences of adopting International Commission on Radiological Protection publication 103 or dual-energy scanning. *AJR Am. J. Roentgenol.* **2010**, *194*, 881–889. [CrossRef]
23. Valentin, J. *Managing Patient Dose in Multi-Detector Computed Tomography(MDCT)*; ICRP Publication 102; The International Commission on Radiological Protection: Ottawa, ON, Canada, 2007; Volume 37, pp. 1–79. [CrossRef]
24. Mattsson, S.; Johansson, L.; Leide Svegborn, S.; Liniecki, J.; Nosske, D.; Riklund, K.A.; Stabin, M.; Taylor, D.; Bolch, W.; Carlsson, S.; et al. Radiation Dose to Patients from Radiopharmaceuticals: A Compendium of Current Information Related to Frequently Used Substances. *Ann. ICRP* **2015**, *44*, 7–321. [CrossRef] [PubMed]
25. Baldi, D.; Tramontano, L.; Alfano, V.; Punzo, B.; Cavaliere, C.; Salvatore, M. Whole Body Low Dose Computed Tomography Using Third-Generation Dual-Source Multidetector with Spectral Shaping: Protocol Optimization and Literature Review. *Dose Response* **2020**, *18*, 1559325820973131. [CrossRef] [PubMed]
26. Sabri, A.S.A.; Wong, J.H.D. Estimation of effective dose for whole body ^{18}F-FDG PET/CT examination. *J. Phys. Conf. Ser.* **2019**, *1248*, 012006. [CrossRef]
27. Tonkopi, E.; Ross, A.A.; MacDonald, A. JOURNAL CLUB: CT dose optimization for whole-body PET/CT examinations. *AJR Am. J. Roentgenol.* **2013**, *201*, 257–263. [CrossRef]
28. Jensen, A.; Kjelgaard-Hensen, M. Diagnostic test validation. In *Schalm's Veterinary Hematology*; Wiley-Blackwell: Hoboken, NJ, USA, 2011.
29. Ferrara, D.; Shiyam Sundar, L.K.; Chalampalakis, Z.; Geist, B.K.; Gompelmann, D.; Gutschmayer, S.; Hacker, M.; Kertész, H.; Kluge, K.; Idzko, M.; et al. Low-dose and standard-dose whole-body [^{18}F]FDG-PET/CT imaging: Implications for healthy controls and lung cancer patients. *Front. Phys.* **2024**, *12*, 1378521. [CrossRef]
30. Brix, G.; Lechel, U.; Glatting, G.; Ziegler, S.I.; Munzing, W.; Muller, S.P.; Beyer, T. Radiation exposure of patients undergoing whole-body dual-modality ^{18}F-FDG PET/CT examinations. *J. Nucl. Med.* **2005**, *46*, 608–613.
31. Alberts, I.; Sachpekidis, C.; Prenosil, G.; Viscione, M.; Bohn, K.P.; Mingels, C.; Shi, K.; Ashar-Oromieh, A.; Rominger, A. Digital PET/CT allows for shorter acquisition protocols or reduced radiopharmaceutical dose in [^{18}F]-FDG PET/CT. *Ann. Nucl. Med.* **2021**, *35*, 485–492. [CrossRef]
32. Gordic, S.; Alkadhi, H.; Hodel, S.; Simmen, H.P.; Brueesch, M.; Frauenfelder, T.; Wanner, G.; Sprengel, K. Whole-body CT-based imaging algorithm for multiple trauma patients: Radiation dose and time to diagnosis. *Br. J. Radiol.* **2015**, *88*, 20140616. [CrossRef]
33. Newman, B.; Ganguly, A.; Kim, J.E.; Robinson, T. Comparison of different methods of calculating CT radiation effective dose in children. *AJR Am. J. Roentgenol.* **2012**, *199*, W232–W239. [CrossRef]
34. Lee, C. How to estimate effective dose for CT patients. *Eur. Radiol.* **2020**, *30*, 1825–1827. [CrossRef]
35. Abuzaid, M.M.; Tekin, H.O.; Elshami, W.; Almisned, G.; Ene, A.; Zakaly, H.M.H. Radiation dose assessment in multiple injured patients using whole-body computed tomography. *J. Radiat. Res. Appl. Sci.* **2022**, *15*, 100465. [CrossRef]
36. Hemke, R.; Yang, K.; Husseini, J.; Bredella, M.A.; Simeone, F.J. Organ dose and total effective dose of whole-body CT in multiple myeloma patients. *Skelet. Radiol.* **2020**, *49*, 549–554. [CrossRef] [PubMed]
37. Saeed, M.K. A comparison of the CT-dosimetry software packages based on stylized and boundary representation phantoms. *Radiography* **2020**, *26*, e214–e222. [CrossRef] [PubMed]
38. Gao, Y.; Quinn, B.; Mahmood, U.; Long, D.; Erdi, Y.; St Germain, J.; Pandit-Taskar, N.; Xu, X.G.; Bolch, W.E.; Dauer, L.T. A comparison of pediatric and adult CT organ dose estimation methods. *BMC Med. Imaging* **2017**, *17*, 28. [CrossRef]
39. Kopp, M.; Loewe, T.; Wuest, W.; Brand, M.; Wetzl, M.; Nitsch, W.; Schmidt, D.; Beck, M.; Schmidt, B.; Uder, M.; et al. Individual Calculation of Effective Dose and Risk of Malignancy Based on Monte Carlo Simulations after Whole Body Computed Tomography. *Sci. Rep.* **2020**, *10*, 9475. [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.

cancers

MDPI

Article

The Three-Class Annotation Method Improves the AI Detection of Early-Stage Osteosarcoma on Plain Radiographs: A Novel Approach for Rare Cancer Diagnosis

Joe Hasei [1,*], Ryuichi Nakahara [2], Yujiro Otsuka [3,4,5], Yusuke Nakamura [5], Kunihiro Ikuta [6], Shuhei Osaki [7], Tamiya Hironari [8], Shinji Miwa [9], Shusa Ohshika [10], Shunji Nishimura [11], Naoaki Kahara [12], Aki Yoshida [2], Tomohiro Fujiwara [2], Eiji Nakata [2], Toshiyuki Kunisada [2] and Toshifumi Ozaki [2]

1 Department of Medical Information and Assistive Technology Development, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Okayama 700-8558, Japan
2 Science of Functional Recovery and Reconstruction, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Okayama 700-8558, Japan
3 Department of Radiology, Juntendo University School of Medicine, Tokyo 113-8431, Japan
4 Milliman, Inc., Tokyo 102-0083, Japan
5 Plusman LCC, Tokyo 103-0023, Japan
6 Department of Orthopedic Surgery, Graduate School of Medicine, Nagoya University, Nagoya 464-0083, Japan
7 Department of Musculoskeletal Oncology and Rehabilitation, National Cancer Center Hospital, Tokyo 104-0045, Japan
8 Department of Musculoskeletal Oncology Service, Osaka International Cancer Institute, Osaka 541-8567, Japan
9 Department of Orthopedic Surgery, Kanazawa University Graduate School of Medical Sciences, Ishikawa 920-8641, Japan
10 Department of Orthopaedic Surgery, Hirosaki University Graduate School of Medicine, Aomori 036-8563, Japan
11 Department of Orthopaedic Surgery, Kindai University Hospital, Osaka 589-8511, Japan
12 Department of Orthopedic Surgery, Mizushima Central Hospital, Kurashiki 712-8064, Japan
* Correspondence: py3g9rcw@s.okayama-u.ac.jp

Simple Summary: Developing effective artificial intelligence (AI) systems for rare diseases such as osteosarcoma is challenging owing to the limited available data. This study introduces a novel approach for preparing training data for AI systems that detect osteosarcoma using X-rays. Traditional methods label tumor areas as a single entity; however, our new approach divides tumor regions into three distinct classes: intramedullary, cortical, and extramedullary. This three-class annotation method enables AI systems to learn more effectively from limited datasets by incorporating detailed anatomical knowledge. This new approach to data preparation resulted in more robust AI models that could detect subtle tumor changes at lower threshold values, demonstrating how strategic data annotation methods can enhance AI performance even with limited training samples. This methodological innovation in data preparation offers a new paradigm for developing AI systems for rare diseases, for which traditional data-driven approaches often fall short.

Abstract: Background/Objectives: Developing high-performance artificial intelligence (AI) models for rare diseases is challenging owing to limited data availability. This study aimed to evaluate whether a novel three-class annotation method for preparing training data could enhance AI model performance in detecting osteosarcoma on plain radiographs compared to conventional single-class annotation. **Methods**: We developed two annotation methods for the same dataset of 468 osteosarcoma X-rays and 378 normal radiographs: a conventional single-class annotation (1C model) and a novel three-class annotation method (3C model) that separately labeled intramedullary, cortical, and extramedullary tumor components. Both models used identical U-Net-based architectures, differing only in their annotation approaches. Performance was evaluated using an independent validation

dataset. **Results**: Although both models achieved high diagnostic accuracy (AUC: 0.99 vs. 0.98), the 3C model demonstrated superior operational characteristics. At a standardized cutoff value of 0.2, the 3C model maintained balanced performance (sensitivity: 93.28%, specificity: 92.21%), whereas the 1C model showed compromised specificity (83.58%) despite high sensitivity (98.88%). Notably, at the 25th percentile threshold, both models showed identical false-negative rates despite significantly different cutoff values (3C: 0.661 vs. 1C: 0.985), indicating the ability of the 3C model to maintain diagnostic accuracy at substantially lower thresholds. **Conclusions**: This study demonstrated that anatomically informed three-class annotation can enhance AI model performance for rare disease detection without requiring additional training data. The improved stability at lower thresholds suggests that thoughtful annotation strategies can optimize the AI model training, particularly in contexts where training data are limited.

Keywords: osteosarcoma; medical image annotation; anatomical annotation method; rare cancer

1. Introduction

Primary malignant bone tumors, despite their extremely low incidence [1–3], predominantly affect children, adolescents, and young adults and require comprehensive support measures, including psychosocial care [4–6]. Diagnostic delays, owing to the rarity of these tumors, can significantly impact patient prognosis. Early detection is crucial for improving patient outcomes; however, their rarity often leads to diagnostic delays, particularly among adolescents and young adults, where these conditions are frequently misdiagnosed as growing pain or exercise-related disorders [7]. The nonspecific nature of the initial symptoms, combined with the challenges that general orthopedic surgeons face in making accurate initial diagnoses, compounds this problem.

Primary screening using radiographic imaging plays an important role in the diagnosis of primary malignant bone tumors. However, detecting subtle changes in bone structure requires considerable expertise, and even experienced specialists face the risk of oversight. Furthermore, disparities in access to specialist care across regions present additional barriers to early diagnosis [8–10].

Therefore, the development of artificial intelligence (AI)-assisted diagnostic imaging systems has attracted considerable attention. Recent advances in AI for medical imaging have demonstrated transformative potential across multiple applications, ranging from image segmentation and structure identification to automated abnormality detection and predictive analytics. In particular, deep learning-based methods have shown improved performance in oncologic imaging tasks, including tumor characterization and treatment response prediction, thereby contributing to better patient outcomes [11]. Currently, two main AI-based approaches are widely used in medical imaging: traditional machine learning (ML) algorithms that learn from predefined features such as texture and morphology and deep learning methods utilizing complex neural networks that can automatically extract and optimize meaningful parameters directly from raw image data [11]. These approaches offer enhanced sensitivity, specificity, and reliability in radiological evaluations and are of special interest in the context of rare diseases, where data scarcity has historically limited the effectiveness of conventional diagnostic methods. Advances in ML technology have enabled the construction of AI models capable of learning features from large volumes of medical imaging data and detecting abnormalities with high precision [12,13]. Nevertheless, for rare conditions such as primary malignant bone tumors, the limited availability of

training data poses significant challenges. Moreover, conventional data-driven approaches have limitations in development under such data-scarce conditions [14].

Previous studies on AI development for X-ray image interpretation of primary malignant bone tumors have solely utilized models trained on unannotated data, achieving notably limited performance, with sensitivity rates of approximately 70%, showing minimal improvement in recent years [15–17]. Thus, in the interpretation of X-ray images of primary malignant bone tumors, the handling of training data may be more crucial than the selection of baseline models.

Unlike in MRI, abnormal shadows on X-ray images of primary malignant bone tumors are often extremely subtle and difficult to detect, making this a particularly challenging task. Although various reports exist on MRI image interpretation models [18–20], studies on X-ray image interpretation remain scarce. Additionally, while MRI and CT scans can yield multiple images per case, X-ray imaging provides fewer images, making it particularly challenging to accumulate substantial image data for rare conditions, such as primary malignant bone tumors. Consequently, creating high-quality datasets is crucial for improving model performance.

Therefore, we focused on developing high-quality annotated data as a novel approach to maximize the utility of limited data. Previously, through detailed tumor region segmentation performed by bone and soft tissue tumor specialists while referencing MRI images, we successfully developed high-performance AI models by creating annotated data that capture subtle changes typically difficult to discern in X-ray images alone [21].

In this study, we report the development of a novel methodology called the "three-class annotation method" for annotating primary malignant bone tumor image data, which enabled us to further enhance the performance of X-ray image interpretation AI models using the same dataset. Herein, we present the procedures and effectiveness of the proposed approach. This study addressed the critical challenges in rare disease diagnosis through AI, focusing on enhancing model performance with limited datasets through anatomically informed annotation strategies. Our findings demonstrated that strategic annotation methods can significantly improve diagnostic capabilities, even with limited data availability, offering a new paradigm for AI development in rare cancer diagnosis.

2. Materials and Methods

2.1. Study Population and Dataset

This study utilized 468 X-ray images from 31 patients with osteosarcoma of the distal femur who were treated at Okayama University Hospital. All patients were pathologically confirmed to have osteosarcoma. Of these images, 328 were used for training. For the normal knee images, 265 of the 378 images from 54 patients without degenerative changes were used for training. The remaining images were used for validation. Given the rarity of osteosarcoma, developing models with limited image datasets necessitated the adoption of strategies to maximize the image diversity in each case. Considering the constraints of the available cases, we implemented a unique approach to partition the dataset into training and internal validation subsets. To optimize the model generalization capability and diagnostic accuracy by maximizing the image diversity in both subsets, we pooled all available cases and randomly selected 70% of the images for training, with the remaining 30% designated for internal validation. Although this methodology resulted in images from the same patients appearing in both the training and validation sets, potentially leading to overfitting, it was chosen as a strategic tradeoff to mitigate the limitations imposed by the small number of available cases. To address potential overfitting concerns, we employed a completely independent dataset for the final performance evaluation, enabling

the assessment of the model's generalization capabilities using unpublished data from different institutions.

2.2. Image Annotation Methods

Two distinct annotation methodologies were applied to the same dataset. The first approach annotated the tumor region as a single class (1C method), whereas the second approach divided the tumor into three classes (3C method). Primary malignant bone tumors are malignant neoplasms that originate within the bone and exhibit extraosseous extension. These tumors can spread through the Haversian canals and may manifest notable periosteal reactions such as Codman's triangle [22]. This results in distinct clinical radiographic findings in the intraosseous and extraosseous regions. Therefore, we aimed to enhance the model performance by separately annotating tumors in three distinct regions: intraosseous, cortical, and extraosseous zones. The 3C annotation method does not divide the dataset into separate categories but rather annotates each tumor image into three distinct anatomical regions: intramedullary, cortical, and extramedullary. This approach enriches the training data by providing detailed anatomical information within each image, allowing the model to learn region-specific features more effectively. The segmentation was performed manually by a single author with over 15 years of experience in the diagnosis and treatment of musculoskeletal tumors. This expert leveraged both MRI and X-ray images to ensure accuracy and consistency in identifying anatomical boundaries. MRI images provided superior contrast for delineating soft tissue and intramedullary structures, while X-ray images offered complementary anatomical information. The segmentation regions were then validated by another board-certified orthopedic surgeon. The intramedullary tumor component is annotated in yellow, cortical bone involvement in orange, and extramedullary tumor extension in blue (Figure 1).

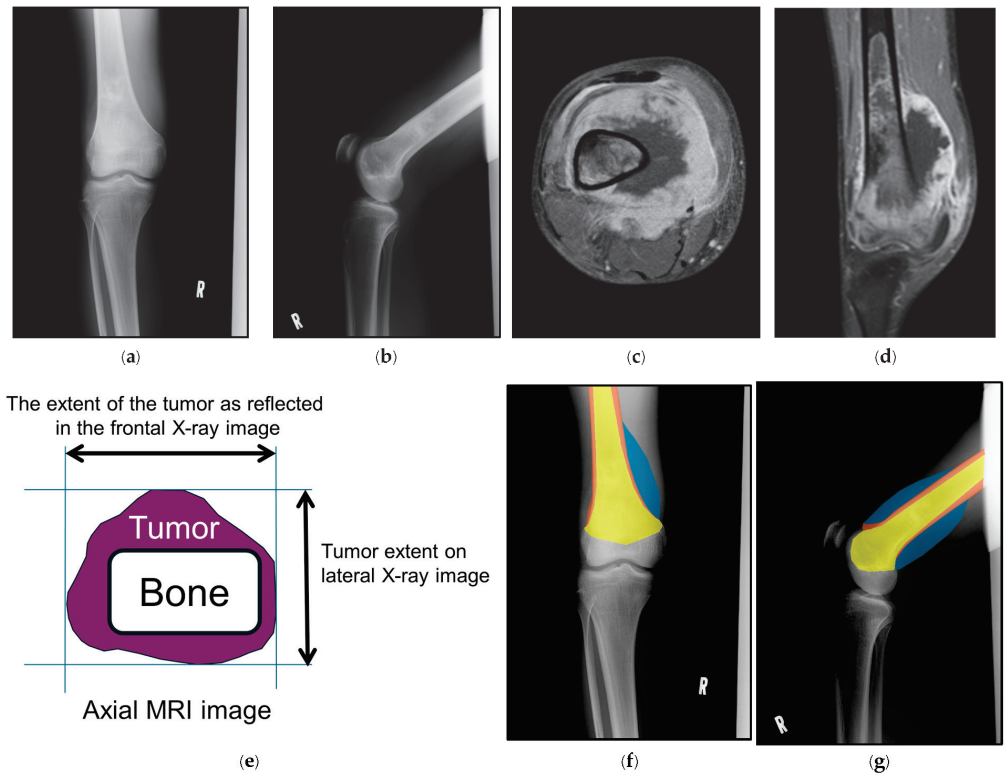


Figure 1. Radiographic images showing the process of tumor segmentation and annotation using both X-ray and MRI guidance. This figure demonstrates our systematic approach to tumor segmentation,

in which MRI images provide crucial guidance for the accurate annotation of tumor boundaries on X-ray images, which alone may not clearly reveal the full extent of the lesion: (**a**,**b**) anteroposterior (AP) and lateral X-ray views of the distal femur showing subtle bone changes, demonstrating the challenge of determining tumor boundaries on plain radiographs; (**c**,**d**) axial and sagittal T1-weighted fat-suppressed contrast-enhanced MRI images clearly delineating the tumor extent; (**e**) schematic diagram illustrating how the tumor extent visible on axial MRI correlates with the measurements on AP and lateral X-ray projections; (**f**,**g**) annotated AP and lateral X-ray images utilizing the MRI-guided three-class annotation method, where yellow indicates the intramedullary tumor component, orange represents cortical bone involvement, and blue indicates extramedullary tumor extension.

2.3. Network Architecture and Model Training Process

In this study, we employed the standard U-Net architecture as originally described by Ronneberger et al. (2015) [23] without incorporating additional modifications or enhancements. Specifically, we did not implement variations such as attention mechanisms, residual connections, or other advanced modules. The primary reason for this choice was to evaluate the effectiveness of the proposed three-class annotation method in isolation, without the confounding effects of architectural complexities. Furthermore, transfer learning was not utilized. Instead, the model was trained from scratch using the annotated dataset prepared for this study. This decision was made to ensure that the results reflect the direct impact of our annotation methodology rather than pre-trained features from unrelated tasks or datasets. Pixel deshuffling was applied before the first convolution to reduce the total number of network parameters, and pixel shuffling was performed after the final convolution. Convolutional operations were conducted for each unit, referred to in this study as a "Res block". The Res block features an architecture similar to that of a residual block [24] with a skip connection. The Res block is designed to enhance feature propagation and mitigate the vanishing gradient problem, thereby improving the stability of deeper networks. The network output utilizes multiresolution maps in the same manner as the U-Net. Each X-ray image was first renormalized using a window level within the 1^{st}–99^{th} percentile of the pixel values to exclude outliers. This step ensures consistent pixel intensity across images, which is critical for effective feature learning. The images were then resampled into 1 mm squares in the imager pixel spacing space. During training, 128 mm square patches were cropped to include a lesion area with a 0.5 ratio and augmented through random intensity shifts (probability = 0.5, offsets = 0.05) and random affine transformations (probability = 0.5) using Project MONAI (version 1.3.1) [25]. These augmentation techniques were chosen to enhance model generalization and robustness against limited data availability. The model was trained in an end-to-end manner using a stochastic gradient descent optimizer with a cross-entropy loss. We employed the One Cycle LR [26] with a maximum learning rate of 1×10^{-1} and completed a total of 100 steps over 40 cycles. At the end of each epoch, the DICE coefficient [27] for the entire annotated tumor region was calculated to identify the best metric epoch. The DICE coefficient serves as an evaluation metric that assesses the accuracy of the AI-segmented region compared to the ground-truth-annotated tumor region. To evaluate the performance of the diagnostic tool in detecting lesions, we calculated the sensitivity, specificity, and DICE coefficient and determined the optimal cutoff value for lesion detection probability. The entire program was developed using Python 3.11 with Pytorch 2.0.1 (Python Software Foundation). All analyses were conducted using Pytorch, sklearn, and scipy packages. All patient data utilized in this study were obtained retrospectively, and patient consent was obtained via the opt-out method. The participants were allowed to withdraw from the study at any point. This study was approved by the institutional ethics committee and conducted following all ethical guidelines to ensure patient privacy and transparent data usage.

2.4. Model Evaluation Process

For the performance evaluation, we analyzed 268 X-ray images from 32 cases of distal femoral osteosarcoma from the Osaka International Cancer Center and 554 normal knee X-ray images from 148 cases from Mizushima Central Hospital as evaluation data. The evaluation dataset contained no images used during training. In this study, we employed receiver operating characteristic (ROC) curve analysis to determine the optimal cutoff value for evaluating the performance of the imaging equipment. Initially, we calculated sensitivity and specificity scores using an image dataset labeled as abnormal (label "1") and normal (label "0"). These sensitivity and specificity values were then used to generate the ROC curves. The ROC curve plots sensitivity (true-positive rate) on the vertical axis against one-specificity (false-positive rate) on the horizontal axis, and the cutoff value was determined using the Youden index (J) [28]. We calculated the Youden index for cutoff values ranging from 0 to 1 in increments of 0.001 for each equipment's scores and determined the optimal cutoff as the value that maximized the Youden index. This approach represents an effective methodology for maximizing the classifier performance while maintaining a balance between sensitivity and specificity [29]. Additionally, we calculated the area under the curve (AUC) to evaluate the overall classifier performance. For the calculation of the "score", which represents the AI's determination, we employed the softmax function to compute probabilities for each class. The softmax function implemented in the final output layer of the neural network converts the class scores into probability distributions. Specifically, we calculated the probability of belonging to each class using the following equation, where zi represents the output from the final layer:

$$\text{Score} = \text{softmax}(\text{zi}) = \frac{\boldsymbol{exp}(\boldsymbol{zi})}{\sum_{j=1}^{K} \boldsymbol{exp}(\boldsymbol{zj})}$$

Here, zi represents the i-th output from the final layer, and K denotes the total number of classes. The scores calculated using this softmax function were interpreted as the probabilities of belonging to each class and were utilized as the model's predictive output. This methodology enabled the comparison of scores between different classes and allowed the model to perform classifications with confidence levels for each class. Furthermore, this approach facilitated the appropriate evaluation of class scores in the calculation of ROC curves and AUC values.

2.5. Statistical Analysis

For a comparative evaluation of the diagnostic performances between the 1C and 3C models, we conducted statistical analyses focusing on sensitivity and specificity. Sensitivity (true-positive rate) was defined as the proportion of actual positives correctly identified by the model and specificity (true-negative rate) as the proportion of actual negatives correctly identified. To determine whether there were statistically significant differences in the sensitivity and specificity between the two models, we employed McNemar's test. This test is appropriate for paired nominal data, allowing the assessment of differences in classification outcomes for the same subjects between the two models. McNemar's test was selected based on its suitability for evaluating paired dichotomous outcomes, particularly for diagnostic test comparisons [30]. Additionally, we calculated the 95% confidence intervals (Cis) for sensitivity and specificity using the Wilson score method [31], providing a measure of the precision of our estimates. Statistical significance was set at $p < 0.05$. This study utilized the same patient cohort as in our previous study [21], which demonstrated the effectiveness of expert annotations in AI-based osteosarcoma detection. Building on these findings, we developed a novel three-class annotation method to further enhance the model performance. For detailed patient demographics and clinical characteristics, the

readers can refer to our previous publication. This test is appropriate for paired nominal data, specifically in our case where two different annotation methods (1C and 3C) were applied to the same dataset, allowing for the assessment of differences in classification outcomes for the same subjects between the two models.

3. Results

3.1. Comparison of Overall Model Performance

In this study, we compared the performance of 1C and 3C models. The analysis revealed several distinct characteristics between the two models. Regarding the diagnostic accuracy (scores), the 1C model yielded a mean score of 0.954 (standard deviation 0.137) versus 0.771 (standard deviation 0.270) of the 3C model. In the discrimination performance evaluation using ROC curves and AUC, the 1C model achieved an AUC of 0.99, and the 3C model achieved 0.98, confirming the superior discriminative ability of both models (Figure 2a). The optimal cutoff values balancing sensitivity and specificity were as follows: the 1C model's optimal cutoff value was 0.769, whereas the 3C model showed a significantly lower value of 0.131. At these points, the 1C model demonstrated a sensitivity of 95.52% and specificity of 96.21%, while the 3C model showed a sensitivity of 96.64% and specificity of 90.11%. While there was no statistically significant difference in sensitivity, the specificity showed a statistically significant difference ($p < 0.05$). However, the 3C model maintained a satisfactory specificity of >90% (Figure 2b,c).

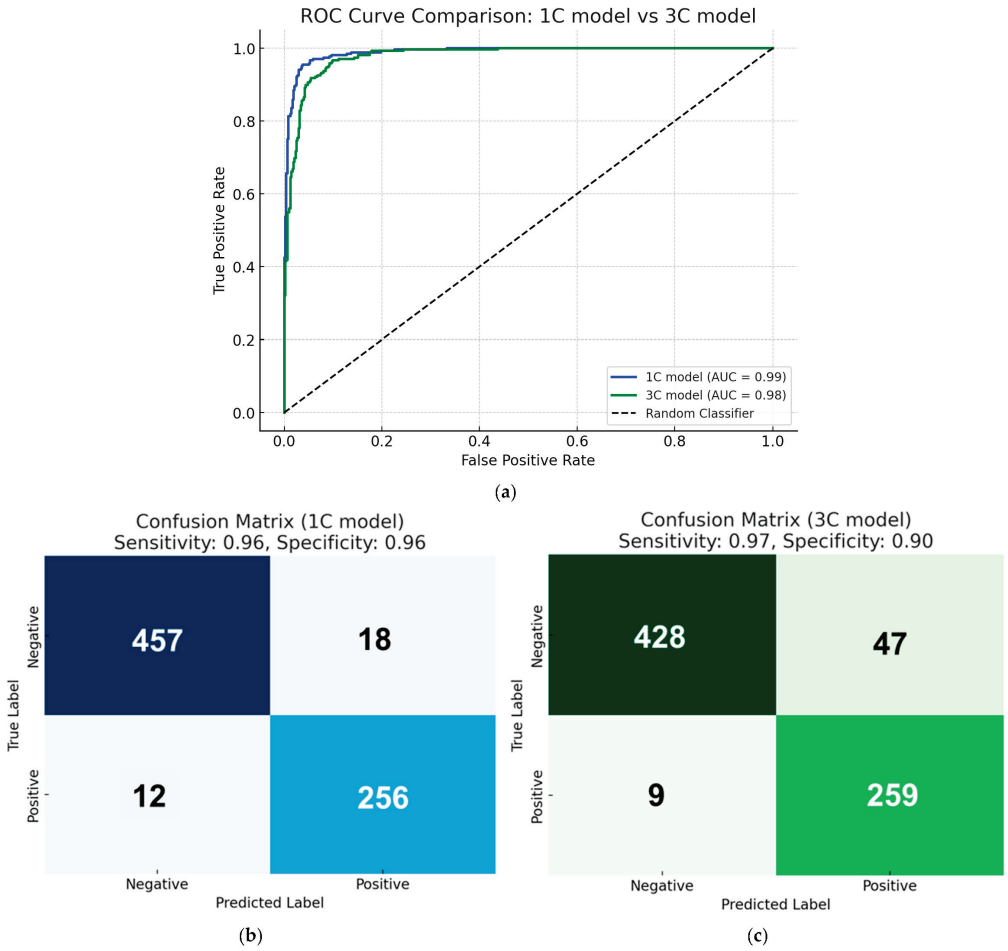


Figure 2. Comparison of the performance of the 1C model and the 3C model for image classification tasks: (**a**) ROC curve comparison between the 1C model (blue line) and the 3C model (green line). The

area under the curve (AUC) for the 1C model is 0.99, while the AUC for the 3C model is 0.98. The dashed diagonal line represents the performance of a random classifier (AUC = 0.50); (**b**) the confusion matrix for the 1C model. The 1C model achieved a sensitivity of 96% and a specificity of 96%. The matrix shows the number of true negatives (457), false positives (18), false negatives (12), and true positives (256); (**c**) the confusion matrix for the 3C model. The 3C model achieved a sensitivity of 97% and a specificity of 90%. The matrix shows the number of true negatives (428), false positives (47), false negatives (9), and true positives (259).

3.2. Analysis of Model Sensitivity and Specificity Patterns

Further analysis of sensitivity and specificity changes across continuous cutoff variations revealed characteristic differences between the models (Figure 3). The 3C model consistently maintained higher specificity than the 1C model. While the 1C model required higher cutoff values to achieve high specificity, it showed a sharp decline in sensitivity beyond its optimal cutoff of 0.769. In contrast, the 3C model demonstrated a stable sensitivity–specificity balance at lower cutoff values, confirming its ability to reliably detect abnormalities even at a low cutoff. When setting the score cutoff value at 0.2, the 1C model showed a specificity of 83.58% and sensitivity of 98.88%, while the 3C model demonstrated more balanced results, with a specificity of 92.21% and sensitivity of 93.28%. At the 25th percentile cutoff values (1C model: 0.985, 3C model: 0.661), both models showed identical false-negative counts of 67, indicating that the 3C model could detect abnormalities at lower scores while maintaining a sensitivity comparable to that of the 1C model.

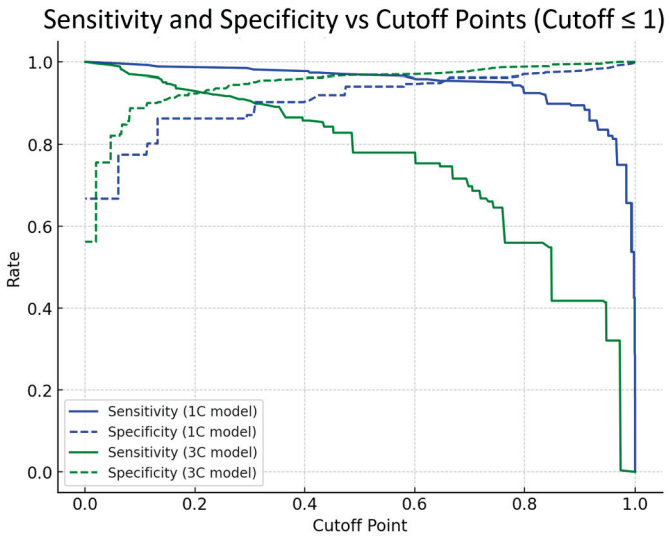


Figure 3. Sensitivity and specificity curves plotted against cutoff values for the 1C and 3C models. The graph displays sensitivity (solid lines) and specificity (dashed lines) values on the *y*-axis (rate) against cutoff values (0.0 to 1.0) on the *x*-axis. The blue lines represent the 1C model's performance metrics (sensitivity: solid blue line; specificity: dashed blue line), while the green lines show the 3C model's performance metrics (sensitivity: solid green line; specificity: dashed green line). The values range from 0.0 to 1.0 on both axes, with grid lines marking intervals of 0.2.

3.3. Evaluation of Abnormal Region Detection Capability

In evaluating the abnormal region capture capability using DICE coefficients, the 1C model averaged 0.663 (standard deviation 0.243) and the 3C model averaged 0.644 (standard deviation 0.247), showing no statistically significant difference ($p = 0.380$). This confirms that the 3C model can capture abnormal regions with an accuracy equivalent to that of the 1C model.

3.4. Correlation Analysis

The correlation analysis between the scores and DICE coefficients showed moderate correlations for both models, with correlation coefficients of 0.682 for the 1C model and 0.594 for the 3C model (Figure 4). These results indicate that while both models maintain equivalent sensitivity and abnormal region capture capabilities with very high specificity, the 3C model demonstrates the ability to detect abnormalities at lower scores than those of the 1C model.

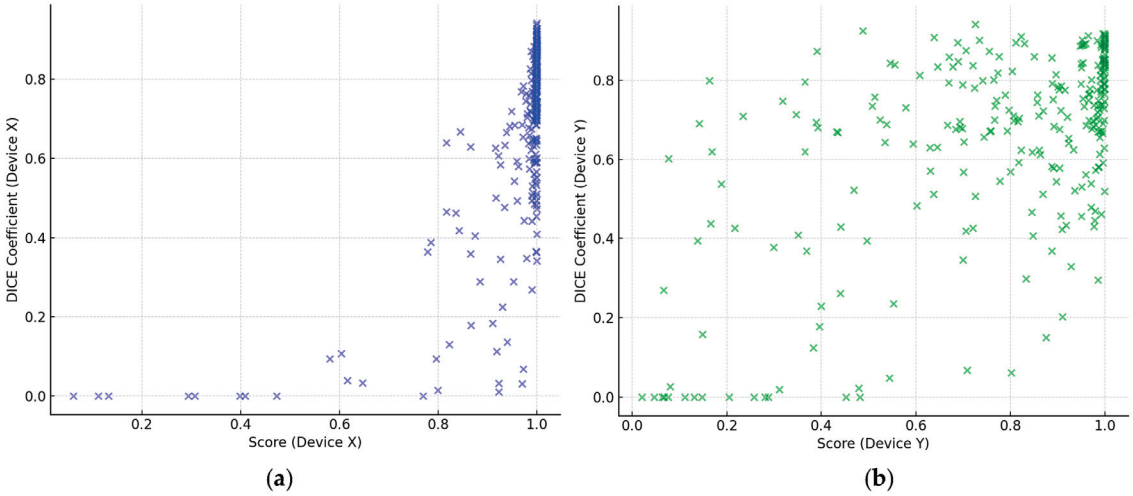


Figure 4. Scatter plots depicting the relationship between the score and DICE coefficient for the 1C and 3C models: (**a**) the relationship between the score (on the *x*-axis) and the DICE coefficient (on the *y*-axis) for the 1C model. The correlation coefficient between the score and the DICE coefficient is 0.682; (**b**) the relationship between the score (on the *x*-axis) and the DICE coefficient (on the *y*-axis) for the 3C model. The correlation coefficient between the score and the DICE coefficient is 0.594.

4. Discussion

4.1. Comparative Analysis of Model Performance

The three-class annotation method AI model (3C model) for X-ray image interpretation developed in this study demonstrated distinctive characteristics compared with the conventional single-class annotation method (1C model), introducing a novel perspective in image data preparation for primary malignant bone tumor X-ray interpretation models. This is the first study to demonstrate how AI model characteristics can be altered by annotating a single tumor into multiple classes based on clinically distinct findings. In the conventional single-class annotation method, treating the entire tumor as a single class results in mixed information from regions showing different X-ray characteristics, such as extraosseous tumors, intraosseous tumors, and cortical regions. When the cutoff was set based on the 25th percentile value, this approach showed very high specificity (99.16%) but required an extremely high cutoff value (0.985), suggesting the potential oversight of subtle abnormalities or early-stage tumors with lower scores. Conversely, the three-class annotation method introduced in this study enabled the AI model to learn the detailed characteristics of each region by separating the tumor annotations into three distinct areas. Consequently, when the cutoff was set based on the 25th percentile value, the model could detect abnormalities at a lower cutoff value (0.661), proving it to be more suitable for screening purposes, which is the most critical application of this model. Comparing the performance of both models, while the 1C model (mean 0.954) outperformed the 3C model (mean 0.771) in terms of diagnostic accuracy scores, there was no statistically significant difference in abnormal region capture capability, evaluated using DICE coefficients (1C model: 0.663, 3C model: 0.644, $p = 0.380$). This suggests that the three-class annotation method can provide more

detailed information without compromising the detection accuracy. Indeed, in cases where intraosseous tumor shadows were particularly subtle, the 3C model demonstrated a more accurate tumor range capture (Figure 5).

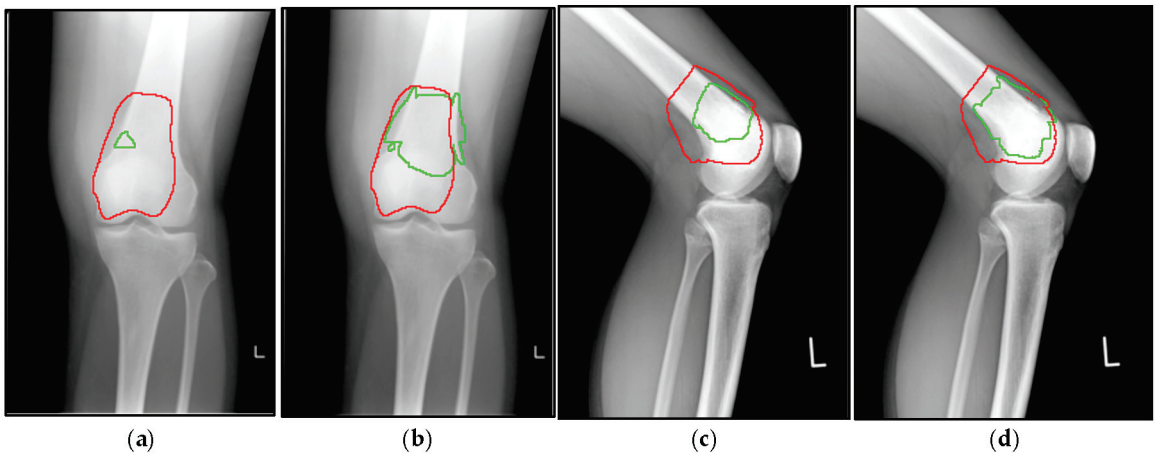


Figure 5. Comparison of lesion detection performance between 1C and 3C AI models in identifying osteosarcoma on X-ray images: (**a**,**b**) anteroposterior radiographs showing osteosarcoma detection results of (**a**) 1C model and (**b**) 3C model outcomes, where the red line represents the osteosarcoma area (as determined by expert human annotation), and the green line indicates the lesion area identified by the AI model; (**c**,**d**) lateral radiographs showing the same comparison, namely (**c**) 1C model and (**d**) 3C model outcomes, with red lines representing expert human annotation and green lines showing AI model detection. "L" indicates the left side of the patient.

4.2. Theoretical Framework and Mechanisms

The impact of this three-class annotation method on the AI model learning processes can be considered from multiple theoretical perspectives. First, from the viewpoint of feature space partitioning and representation learning, although the 1C model recognized the entire tumor region as a single feature space with mixed image characteristics, the 3C model enabled clearer learning of image features specific to each anatomically distinct region (intraosseous, cortical, and extraosseous). This allowed for a more detailed recognition of the differences between normal structures and pathological changes in each region. Second, the 3C model could effectively perform multi-task learning, enhancing the model generalization capability through parallel learning of recognition in each region. Learning inter-regional relationships enables the acquisition of more robust feature representations, which is particularly advantageous when learning from limited datasets, such as in this study. Furthermore, the 3C model formed independent decision boundaries for each region, enabling finer discrimination. The presence of multiple decision boundaries enables more flexible abnormality detection, providing theoretical grounds for the high detection capability at low cutoff values observed in this study.

4.3. Clinical Implications and Applications

From a clinical application perspective, the characteristic differences between the models have significant implications. The 1C model represents a "certainty-focused" approach that detects only clear abnormalities with high confidence, while the 3C model represents a "flexibility-focused" approach suitable for initial screening and follow-up monitoring; thus, it is capable of evaluating a broader range of abnormalities incrementally. Particularly noteworthy is the potential contribution of the 3C model's lower cutoff value to the early detection of primary malignant bone tumors, which is a rare cancer type. This implies an increased possibility of detecting subtle changes in the early stages, potentially leading to more opportunities for early diagnosis and appropriate therapeutic intervention. Setting

the 1C model's cutoff at 0.985 poses clinical challenges owing to its extremely high value, potentially flagging normal physiological variations or minor image noise as abnormal. This could lead to unnecessary examinations and treatments and increased patient burden, potentially overconsuming medical resources. For instance, when setting the cutoff of both models at 0.2, the 1C model showed 98.88% sensitivity and 83.58% specificity, whereas the 3C model demonstrated 93.28% sensitivity and 92.21% specificity, confirming the 3C model's capability for balanced diagnosis. While the 1C model could be used with a low cutoff value when prioritizing sensitivity, considering the need to reduce false positives, the 3C model appeared to be more appropriate. At a 0.2 cutoff, the false-positive rates were 16.42% for the 1C model and 7.79% for the 3C model, indicating the 3C model's superior suitability for clinical use. These theoretical characteristics provide a foundation for explaining the experimental results demonstrated by the 3C model, particularly its high detection capability at lower cutoff and more stable specificity maintenance. This theoretical framework also suggests potential applications in other medical image diagnostic tasks.

4.4. Future Directions and Challenges

The limitations of this study include the scale and diversity of the datasets used. Verification using more diverse cases is necessary, and considering the characteristics of rare cancers, data collection through multi-institutional collaborative research is desirable. To enhance data diversity and mitigate overfitting, we employed augmentation techniques such as intensity shifts, affine transformations, and patch-based cropping during model training. These methods were specifically designed to preserve the subtle radiographic shadows characteristic of bone tumors in X-ray images, which are critical for accurate detection. The effectiveness of these techniques was validated through the model's performance on an independent external dataset, demonstrating their ability to improve robustness and generalization. While advanced techniques such as generative adversarial network (GAN)-based synthetic data generation hold promise for further expanding dataset diversity, their application to high-resolution X-ray images with subtle shadow details remains challenging, particularly in the context of rare diseases like malignant bone tumors. This is primarily due to the limited availability of training data for such diseases, which makes it difficult to develop high-performance GAN models capable of accurately capturing and replicating the fine-grained features necessary for clinical applications [32]. Our proposed 3C method, with its anatomically informed structured annotations, could potentially serve as a foundational approach for developing more sophisticated data augmentation techniques, including GANs specifically tailored for rare diseases. The anatomical separation of tumor regions in our method provides a framework that could guide the generation of more realistic and clinically relevant synthetic data in future research. As GAN architectures and domain-specific adaptations continue to evolve, further investigation will explore how our 3C annotation method can contribute to the development of more effective data augmentation strategies for rare diseases. Additionally, further investigation is needed regarding the optimal methods for three-class annotation and the weighting of each class. Additionally, the ethical challenges associated with AI applications in medicine must be carefully considered. It is crucial for healthcare professionals, researchers, and ethics experts to collaborate to develop guidelines and appropriate educational programs that address various issues, including privacy protection, algorithmic bias, and the interpretation and responsibility of AI diagnoses [33]. The primary focus of this study was to evaluate the effectiveness of the proposed three-class annotation method for improving AI detection of early-stage osteosarcoma. To achieve this, we intentionally employed the standard U-Net architecture without incorporating variations or additional enhancements. This approach allowed us to

isolate and assess the impact of the annotation strategy independently of other architectural factors. However, we recognize that a thorough comparison with different U-Net variations and alternative deep learning models would provide a more comprehensive understanding of the proposed method's robustness and generalizability. Examples of such variations include attention mechanisms, residual connections, or transfer learning approaches, all of which have demonstrated potential in enhancing segmentation and classification tasks in medical imaging. Evaluating the three-class annotation method using these architectures would allow us to better determine its adaptability and potential benefits in diverse clinical scenarios. As part of future work, we plan to conduct additional experiments to compare the performance of the three-class annotation approach across multiple U-Net variations and other architectures. These comparisons will not only validate our findings but also help optimize the annotation strategy for broader applications. This limitation has been added to acknowledge the scope of the current study and to provide transparency about its boundaries. Despite these limitations, we believe that the results of this study provide a significant contribution to the field by demonstrating the potential of anatomically informed annotation strategies in rare cancer diagnosis.

4.5. Ethical Considerations and Implementation

The AI model using the three-class annotation method developed in this study introduces new possibilities as a diagnostic approach for primary malignant bone tumors. Particularly, for early-stage abnormality detection and detailed lesion evaluation, it demonstrated characteristics distinct from conventional single-class models, potentially contributing to improved diagnostic accuracy and early detection rates depending on its clinical implementation. Future expectations include verification using larger and more diverse datasets, application to other modalities, and addressing the ethical challenges to further enhance the effectiveness and safety of this new approach.

4.6. Global Impact and Future Perspectives

Furthermore, the distinct characteristics revealed between the single- and three-class models demonstrate the importance of matching model characteristics with clinical needs beyond basic performance metrics in AI model development. The successful development of an AI model for primary malignant bone tumors indicates the potential for AI technology applications in other rare diseases, contributing to the advancement of personalized medicine. Furthermore, the three-class annotation method presents an approach for integrating diverse imaging findings within a single disease, a concept that is applicable to other imaging modalities and disease areas. This study emphasized the significance of annotation design for future medical AI development by demonstrating how annotation methods can significantly influence AI model performance and characteristics. Future challenges include verification using larger and more diverse datasets, application to other imaging modalities, and optimization of the three-class annotation method. Moreover, developing explainable AI to visualize AI decision-making and address the ethical and legal issues associated with the clinical implementation of AI diagnostic support tools is crucial. Building on these research outcomes, establishing international research collaboration networks and sharing insights into rare cancer diagnosis and treatment are expected to contribute to improving healthcare quality on a global scale. This study presents a novel approach for improving the diagnostic accuracy and early detection rates of rare cancers by combining AI technology and medical insights. This innovative method is expected to serve as a guideline for future medical imaging diagnostic AI development and contribute to improving the quality of life of patients worldwide. Collaboration among healthcare professionals, AI researchers, and policymakers in promoting further development and

appropriate clinical application of this new technology represents a significant step toward realizing next-generation precision medicine.

5. Conclusions

The AI model for X-ray image interpretation using the three-class annotation method developed in this study presents an innovative approach for diagnosing primary malignant bone tumors and has significant implications for future medical image diagnostic AI development. By overcoming the limitations of conventional single-class annotation methods and enabling detailed classification of different regions within tumors, this approach has demonstrated the potential to contribute to the early detection of rare cancers and appropriate treatment selection through enhanced AI model learning capabilities and diagnostic characteristics. In screening tests for rare diseases with low prevalence, high specificity is particularly crucial because false positives can lead to excessively detailed examinations. The 3C model demonstrated significant clinical value through its flexibility in detecting subtle abnormalities at lower cutoff values (0.661) while maintaining abnormal region capture capability (DICE coefficient 0.644). This three-class annotation method presents a novel modeling approach for complex diseases with regions of different characteristics within a single lesion, suggesting potential applications in AI diagnostic model development for other rare cancers and diseases with complex lesions.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cancers17010029/s1, Table S1: Score values and DICE coefficients data for 1C and 3C models shown in Figure 4.

Author Contributions: Conceptualization, J.H., R.N., T.H., N.K., S.M., S.O. (Shuhei Osaki), S.N., K.I., S.O. (Shusa Ohshika), A.Y., T.F., E.N., T.K. and T.O.; methodology, J.H., R.N., Y.O. and Y.N.; software, Y.O. and Y.N.; validation, J.H., Y.O., Y.N., T.F., E.N., T.K. and T.O.; formal analysis, J.H., Y.O. and Y.N.; investigation, J.H., Y.O., Y.N., T.F. and E.N.; resources, R.N., T.H., N.K., S.M., S.O. (Shuhei Osaki), S.N., K.I., S.O. (Shusa Ohshika), A.Y., T.F. and E.N.; data curation, J.H. and R.N.; writing—original draft preparation, J.H.; writing—review and editing, J.H., T.H., T.F., E.N., T.K. and T.O.; visualization, J.H.; supervision, J.H. and T.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Japan Society for the Promotion of Science (JSPS) as part of the Grants-in-Aid for Scientific Research (KAKENHI), grant number 21K09228 [to J.H.], and the Japan Agency for Medical Research and Development (AMED), grant number 22579674 [to T.O.]. The APC was funded by the Japan Agency for Medical Research and Development (AMED), grant number 22579674.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Research Ethics Committee of Okayama University (protocol code 2304-004, approved on 26 January 2023).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study using an opt-out method, in which patients were informed about the study and given the opportunity to decline participation.

Data Availability Statement: The numeric data presented in Figure 4 are available in Table S1 in the Supplementary Materials. Additional data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: J.H. is affiliated with an endowed chair funded by Plusman LLC (Tokyo, Japan). This affiliation represents institutional support only, and J.H. has no other financial relationships or interests that are directly relevant to the content of this study. Author Y.O. was employed by the company Milliman, Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of

interest. he 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. Savage, S.A.; Mirabello, L. Using Epidemiology and Genomics to Understand Osteosarcoma Etiology. *Sarcoma* **2011**, *2011*, 548151. [CrossRef] [PubMed]
2. Stiller, C.A.; Bielack, S.S.; Jundt, G.; Steliarova-Foucher, E. Bone Tumours in European Children and Adolescents, 1978–1997. Report from the Automated Childhood Cancer Information System Project. *Eur. J. Cancer* **2006**, *42*, 2124–2135. [CrossRef] [PubMed]
3. Mirabello, L.; Troisi, R.J.; Savage, S.A. International Osteosarcoma Incidence Patterns in Children and Adolescents, Middle Ages and Elderly Persons. *Int. J. Cancer* **2009**, *125*, 229–234. [CrossRef] [PubMed]
4. Gardner, M.H.; Barnes, M.J.; Bopanna, S.; Davis, C.S.; Cotton, P.B.; Heron, B.L.; Henninger, A.; Alva, E.; Gleason, M.W.; Whelan, K.F.; et al. Barriers to the Use of Psychosocial Support Services Among Adolescent and Young Adult Survivors of Pediatric Cancer. *J. Adolesc. Young Adult Oncol.* **2014**, *3*, 112–116. [CrossRef] [PubMed]
5. Bleyer, A.; Barr, R.; Ries, L.; Whelan, J.; Ferrari, A. *Cancer in Adolescents and Young Adults*; Springer: Berlin/Heidelberg, Germany, 2016; ISBN 9783319336794.
6. Zebrack, B.J.; Block, R.; Hayes-Lattin, B.; Embry, L.; Aguilar, C.; Meeske, K.A.; Li, Y.; Butler, M.; Cole, S. Psychosocial Service Use and Unmet Need among Recently Diagnosed Adolescent and Young Adult Cancer Patients. *Cancer* **2013**, *119*, 201–214. [CrossRef]
7. Longhi, A.; Errani, C.; De Paolis, M.; Mercuri, M.; Bacci, G. Primary Bone Osteosarcoma in the Pediatric Age: State of the Art. *Cancer Treat. Rev.* **2006**, *32*, 423–436. [CrossRef]
8. Johnston, K.J.; Wen, H.; Joynt Maddox, K.E. Lack of Access to Specialists Associated with Mortality and Preventable Hospitalizations of Rural Medicare Beneficiaries. *Health Aff.* **2019**, *38*, 1993–2002. [CrossRef]
9. Resnick, M.J. Re: Lack of Access to Specialists Associated with Mortality and Preventable Hospitalizations of Rural Medicare Beneficiaries. *J. Urol.* **2020**, *203*, 1052. [CrossRef] [PubMed]
10. Grineski, S.; Morales, D.X.; Collins, T.; Wilkes, J.; Bonkowsky, J.L. Geographic and Specialty Access Disparities in US Pediatric Leukodystrophy Diagnosis. *J. Pediatr.* **2020**, *220*, 193–199. [CrossRef]
11. Obuchowicz, R.; Strzelecki, M.; Piórkowski, A. Clinical Applications of Artificial Intelligence in Medical Imaging and Image Processing-A Review. *Cancers* **2024**, *16*, 1870. [CrossRef] [PubMed]
12. Li, M.; Jiang, Y.; Zhang, Y.; Zhu, H. Medical Image Analysis Using Deep Learning Algorithms. *Front. Public Health* **2023**, *11*, 1273253. [CrossRef] [PubMed]
13. Sistaninejhad, B.; Rasi, H.; Nayeri, P. A Review Paper about Deep Learning for Medical Image Analysis. *Comput. Math. Methods Med.* **2023**, *2023*, 7091301. [CrossRef]
14. 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]
15. He, Y.; Pan, I.; Bao, B.; Halsey, K.; Chang, M.; Liu, H.; Peng, S.; Sebro, R.A.; Guan, J.; Yi, T.; et al. Deep Learning-Based Classification of Primary Bone Tumors on Radiographs: A Preliminary Study. *EBioMedicine* **2020**, *62*, 103121. [CrossRef] [PubMed]
16. von Schacky, C.E.; Wilhelm, N.J.; Schäfer, V.S.; Leonhardt, Y.; Gassert, F.G.; Foreman, S.C.; Gassert, F.T.; Jung, M.; Jungmann, P.M.; Russe, M.F.; et al. Multitask Deep Learning for Segmentation and Classification of Primary Bone Tumors on Radiographs. *Radiology* **2021**, *301*, 398–406. [CrossRef]
17. Li, J.; Li, S.; Li, X.; Miao, S.; Dong, C.; Gao, C.; Liu, X.; Hao, D.; Xu, W.; Huang, M.; et al. Primary Bone Tumor Detection and Classification in Full-Field Bone Radiographs via YOLO Deep Learning Model. *Eur. Radiol.* **2023**, *33*, 4237–4248. [CrossRef] [PubMed]
18. Ouyang, T.; Yang, S.; Gou, F.; Dai, Z.; Wu, J. Rethinking U-Net from an Attention Perspective with Transformers for Osteosarcoma MRI Image Segmentation. *Comput. Intell. Neurosci.* **2022**, *2022*, 7973404. [CrossRef] [PubMed]
19. Wu, J.; Liu, Z.; Gou, F.; Zhu, J.; Tang, H.; Zhou, X.; Xiong, W. BA-GCA Net: Boundary-Aware Grid Contextual Attention Net in Osteosarcoma MRI Image Segmentation. *Comput. Intell. Neurosci.* **2022**, *2022*, 3881833. [CrossRef] [PubMed]
20. Wu, J.; Zhou, L.; Gou, F.; Tan, Y. A Residual Fusion Network for Osteosarcoma MRI Image Segmentation in Developing Countries. *Comput. Intell. Neurosci.* **2022**, *2022*, 7285600. [CrossRef] [PubMed]
21. Hasei, J.; Nakahara, R.; Otsuka, Y.; Nakamura, Y.; Hironari, T.; Kahara, N.; Miwa, S.; Ohshika, S.; Nishimura, S.; Ikuta, K.; et al. High-Quality Expert Annotations Enhance Artificial Intelligence Model Accuracy for Osteosarcoma X-Ray Diagnosis. *Cancer Sci.* **2024**, *115*, 3695–3704. [CrossRef]
22. McDonald, J.; DenOtter, T.D. Codman Triangle. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2023.
23. 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, Munich, Germany, 5–9 October 2015; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241.

24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. Chlap, P.; Finnegan, R.N. PlatiPy: Processing Library and Analysis Toolkit for Medical Imaging in Python. *J. Open Source Softw.* **2023**, *8*, 5374. [CrossRef]
26. Smith, L.N.; Topin, N. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. In Proceedings of the Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, Baltimore, MD, USA, 14–18 April 2019; SPIE: Bellingham, WA, USA, 2019; Volume 11006, pp. 369–386.
27. Carass, A.; Roy, S.; Gherman, A.; Reinhold, J.C.; Jesson, A.; Arbel, T.; Maier, O.; Handels, H.; Ghafoorian, M.; Platel, B.; et al. Evaluating White Matter Lesion Segmentations with Refined Sørensen-Dice Analysis. *Sci. Rep.* **2020**, *10*, 8242. [CrossRef] [PubMed]
28. Fluss, R.; Faraggi, D.; Reiser, B. Estimation of the Youden Index and Its Associated Cutoff Point. *Biom. J.* **2005**, *47*, 458–472. [CrossRef] [PubMed]
29. Dhamnetiya, D.; Jha, R.P.; Shalini, S.; Bhattacharyya, K. How to Analyze the Diagnostic Performance of a New Test? Explained with Illustrations. *J. Lab. Physicians* **2022**, *14*, 90–98. [CrossRef] [PubMed]
30. Mcnemar, Q. Note on the Sampling Error of the Difference between Correlated Proportions or Percentages. *Psychometrika* **1947**, *12*, 153–157. [CrossRef] [PubMed]
31. Wilson, E.B. Probable Inference, the Law of Succession, and Statistical Inference. *J. Am. Stat. Assoc.* **1927**, *22*, 209. [CrossRef]
32. Islam, S.; Aziz, M.T.; Nabil, H.R.; Jim, J.R.; Mridha, M.F.; Kabir, M.M.; Asai, N.; Shin, J. Generative Adversarial Networks (GANs) in Medical Imaging: Advancements, Applications, and Challenges. *IEEE Access* **2024**, *12*, 35728–35753. [CrossRef]
33. Topol, E.J. High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nat. Med.* **2019**, *25*, 44–56. [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.

tomography

MDPI

Article

Effective Dose Estimation in Computed Tomography by Machine Learning

Matteo Ferrante [1,†], Paolo De Marco [1], Osvaldo Rampado [2], Laura Gianusso [2] and Daniela Origgi [1,*]

[1] Medical Physics Unit, IEO, European Institute of Oncology IRCCS, 20141 Milan, Italy; matteo-ferrante@hotmail.it (M.F.); paolo.demarco@ieo.it (P.D.M.)
[2] Medical Physics Unit, A.O.U. Città della Salute e della Scienza di Torino, 10126 Turin, Italy; orampado@cittadellasalute.to.it (O.R.); lgianusso@cittadellasalute.to.it (L.G.)
* Correspondence: daniela.origgi@ieo.it
† Affiliation at the time of the study.

Abstract: Background: Computed tomography scans are widely used in everyday medical practice due to speed, image reliability, and detectability of a wide range of pathologies. Each scan exposes the patient to a radiation dose, and performing a fast estimation of the effective dose (E) is an important step for radiological safety. The aim of this work is to estimate E from patient and CT acquisition parameters in the absence of a dose-tracking software exploiting machine learning. Methods: In total, 69,037 CT acquisitions were collected with the dose-tracking software (DTS) available at our institution. E calculated by DTS was chosen as the target value for prediction. Different machine learning algorithms were selected, optimizing parameters to achieve the best performance for each algorithm. Effective dose was also estimated using DLP and k-factors, and with multiple linear regression. Mean absolute error (MAE, mean absolute percentage error (MAPE), and R^2 were used to evaluate predictions in the test set and in an external dataset of 3800 acquisitions. Results: The random forest regressor (MAE: 0.416 mSv; MAPE: 7%; and R^2: 0.98) showed best performances over the neural network and the support vector machine. However, all three machine learning algorithms outperformed effective dose estimation using k-factors (MAE: 2.06; MAPE: 26%) or multiple linear regression (MAE: 0.98; MAPE: 44.4%). The random forest regressor on the external dataset showed an MAE of 0.215 mSv and an MAPE of 7.1%. Conclusions: Our work demonstrated that machine learning models trained with data calculated by a dose-tracking software can provide good estimates of the effective dose just from patient and scanner parameters, without the need for a Monte Carlo approach.

Keywords: artificial intelligence (AI); patient radiation protection; dose tracking

1. Introduction

Computed tomography is one of the most widely used diagnostic imaging techniques, due to its speed and ability to detect anatomical details that are fundamental to the diagnosis of numerous pathologies. However, it is based on the use of ionizing radiation, a possible source of risk for the patient. The estimation of the effective dose (E) administered during radiological investigations of computed tomography in the medical field is of particular interest for radiation protection. According to ICRP 147 [1], the effective dose is used in medicine for comparing doses from different medical procedures, informing judgements on justification, and establishing constraints for carers and volunteers in medical research.

There are several approaches to the estimation of this parameter, ranging from the simplest, in which it assumes a linear relationship between Dose Length Product (DLP)

and effective dose, to more complicated approaches based on Monte Carlo simulations on mathematical phantoms. In particular, there is dose-tracking software that automatically queries the hospital's PACS and retrieves information from the structured dose report and the image itself. A great deal of this software includes Monte Carlo simulations on a set of mathematical anthropomorphic phantoms. They include, for example, representative phantoms of men and women of different sizes, ages, and situations, such as ones dedicated to pregnant women. The scout image of the patient is selected and the most similar phantom is associated with the patient, and then the values of the simulation are corrected through the use of look-up tables, which provide the corrective effective dose values to the different organs according to the acquisition parameters.

The range of possible approaches is very wide and different software implements multiple solutions, with different levels of complexity, ranging from estimates based on coefficients to the use of the same images acquired.

The primary aim of our work is to explore the power of machine learning in effective dose estimation, starting from values obtained using a commercial dose-tracking software (Radimetrics, Bayer Healthcare [2]). Such software provides effective doses using some patient information like age, gender, height, and weight, the acquisition parameters, and the dose indices provided by the acquisition devices.

The secondary aim of our work is to predict the right dose class with high accuracy, starting from effective dose estimates. The new Italian regulation (D.Lgs. 101/20), a transposition of the Euratom 59/2013, establishes that every radiological procedure must indicate in the radiological report information relating to patient exposure in terms of classes of effective dose (E). There are four possible dose classes: class I if $E < 1$ mSv, class II if $1 < E < 5$ mSv, class III if $5 < E < 1$ 0 mSv, and class IV if $E > 10$ mSv.

2. Materials and Methods

We collected and anonymized dose-tracking information and developed a pipeline in Python to preprocess them and feed the data to machine learning models, to estimate the effective dose.

The pipeline explained hereafter is summarized in Figure 1.

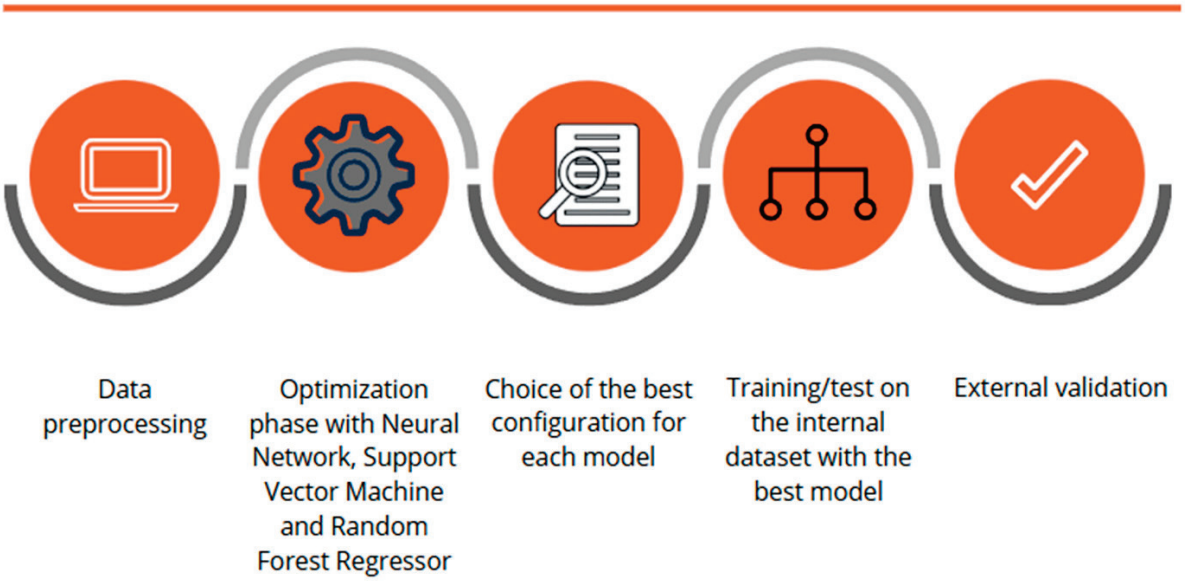


Figure 1. Pipeline for effective dose estimation.

Most of this work was carried out using the scikit-learn (https://scikit-learn.org/stable/ accessed on 20 March 2021) [3] and tensorflow (https://www.tensorflow.org/ accessed on the 20 March 2021) [4] Python libraries.

2.1. Data

2.1.1. Internal Dataset and Variable Description

We collected 69,037 CT acquisitions of two different devices of the same vendor from January 2019 to February 2021 and sent them to the commercial dose-tracking software (Radimetrics, Bayer Healthcare). This software is based on mathematical phantoms coupled with precomputed Monte Carlo and produces an estimate of the effective dose for each acquisition.

The output of this software was used to explore the performances of several machine learning models and whether they are able to predict the effective dose just by using patient and scanner information, to produce a model able to estimate the dose and the class dose just by using the information available in the header of the DICOM image.

We collected a great deal of information from the dose-tracking software, including descriptions, parameters, and more. We chose to feed our model with a subset of the parameters that were exportable from Radimetrics^{TM}. In particular, we included in our model three different types of information:

Patient-related information: gender, weight, height, and age.

Scanner-related parameters: kVp, minimum, maximum, and mean value of mAs, nominal mA, rotation time, pitch and collimation.

Dose estimate indices: CTDI_vol and DLP for both head and body phantoms.

Patient-related information is useful to include corrections about patient size.

Scanner-related parameters are useful to include corrections about modulation to follow patient shape and attenuation.

Dose estimate indices provide a baseline value because of the well-known linear relationship between DLP and E. These indices are computed based on the values computed on two types of uniform water phantom, head and body, with, respectively, 16 and 32 cm of diameter.

2.1.2. External Dataset

To ensure the generalization capability of our best model, we also employed a dataset from a different institution with a different dose-tracking software (Physico, Emme Esse, Italy) [5], collecting 3800 acquisitions from another hospital (Azienda Ospedaliero Universitaria (A.O.U.) Città della Salute e della Scienza di Torino).

2.2. Preprocessing

We preprocessed the data exported from Radimetrics^{TM} developing a pipeline able to reduce both the features and the outliers present in this dataset.

We removed all samples that included missing values and we excluded features like series description and hospital-related information, to anonymize data and reduce hospital custom name dependence. This process simplified the dataset for an eventual generalization procedure on external datasets. At the end of the preprocessing procedure, the result was a cleaned dataset of over 69,000 samples, that was split into training (70%) and test (30%).

The dataset included 16 features describing the acquisition procedure and the target variable, i.e., the effective dose, estimated by Radimetrics^{TM} using the mathematical phantoms and look-up tables with the definition of effective dose as in the ICRP 103 [6].

2.3. Exploratory Data Analysis

To be sure that machine learning models could learn the right correlation between data and outputs, we explored our data to gain insights of the most important features.

As a first step, we computed a clustermap of the features in the training set using the Python seaborn [7] library. We also computed a correlation map (see Figure 2) between features and the target variable, effective dose, defined as "ICRP_103_mSv" in Radimetrics™.

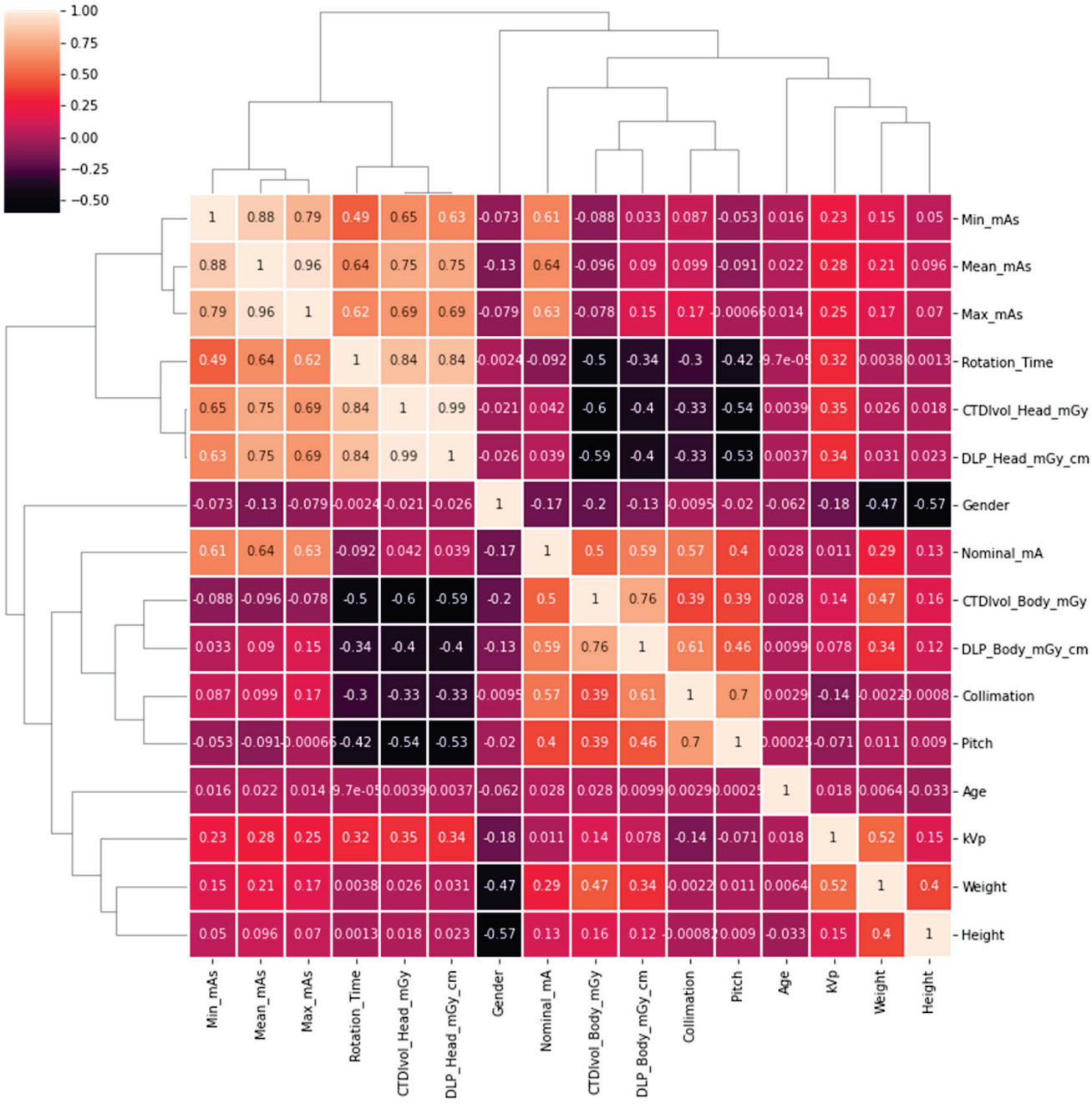


Figure 2. Cluster map computed from the correlation between features in the training dataset. We can observe that features can be clustered into patient information, scanner parameters, and dose descriptors.

We can observe that is possible to cluster together scanner features and dosimetric indices because they are highly correlated, containing information about the parameters of the scanning procedure and the expected dosimetric output. In Figure 3, we can observe the correlation between the features in the training set and the target variable. We can observe that DLP (Body) shows a high correlation value, as expected, so we can assume that a good model will start with a nearly linear estimate using DLP and adjust this baseline value including higher order dependencies from other variables.

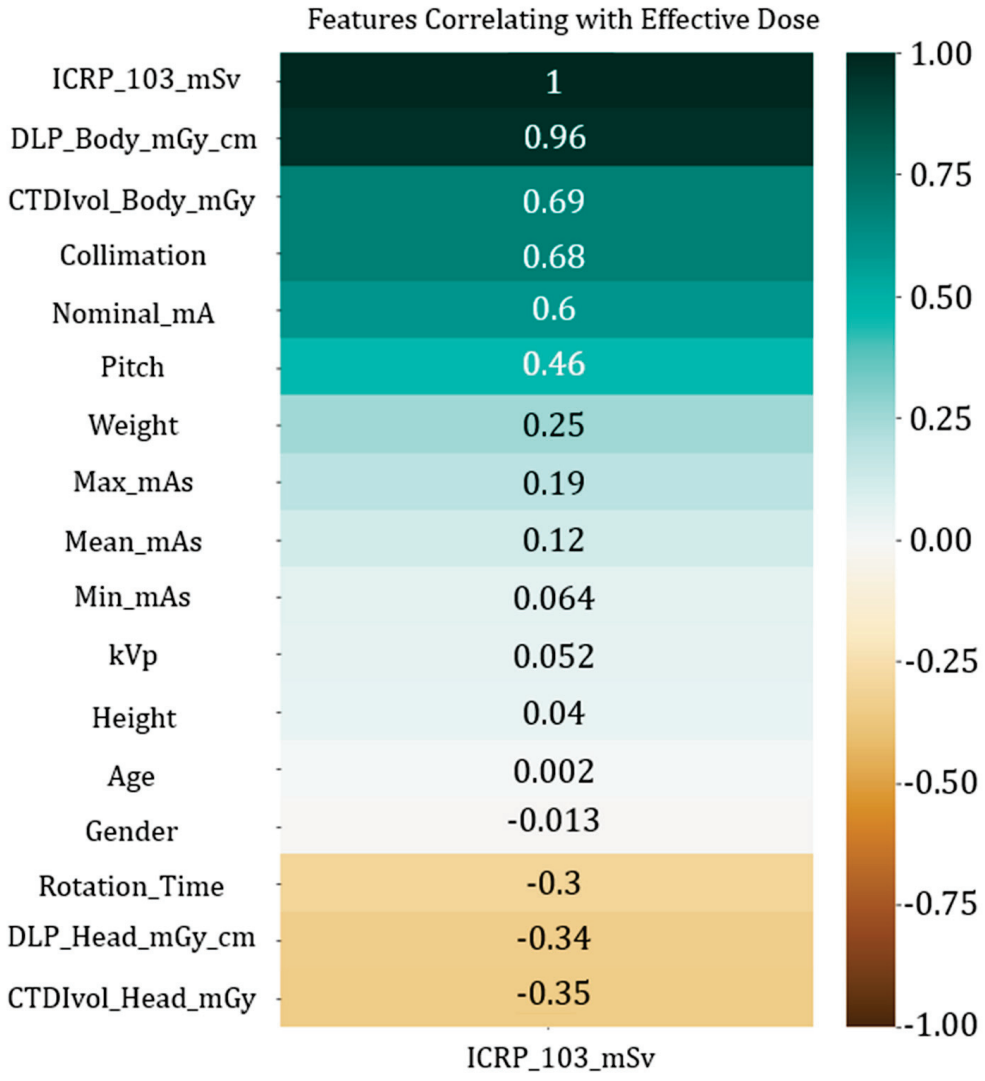


Figure 3. Correlation between features and target variable.

To complete our exploratory analysis, we also included a Principal Component Analysis (PCA) to check which features affect mostly the variability in the output. We used the scikit-learn implementation of PCA with default parameters to obtain a ranking of the features based on their importance in the variability explanation. In Figure 4, it is possible to observe this parameter in the PCA decomposition. The most important feature is DLP (Body) that can explain 73% of the total variance in the dataset. The second most important variable is DLP (Head) that accounts for about 23% of the variance in the dataset.

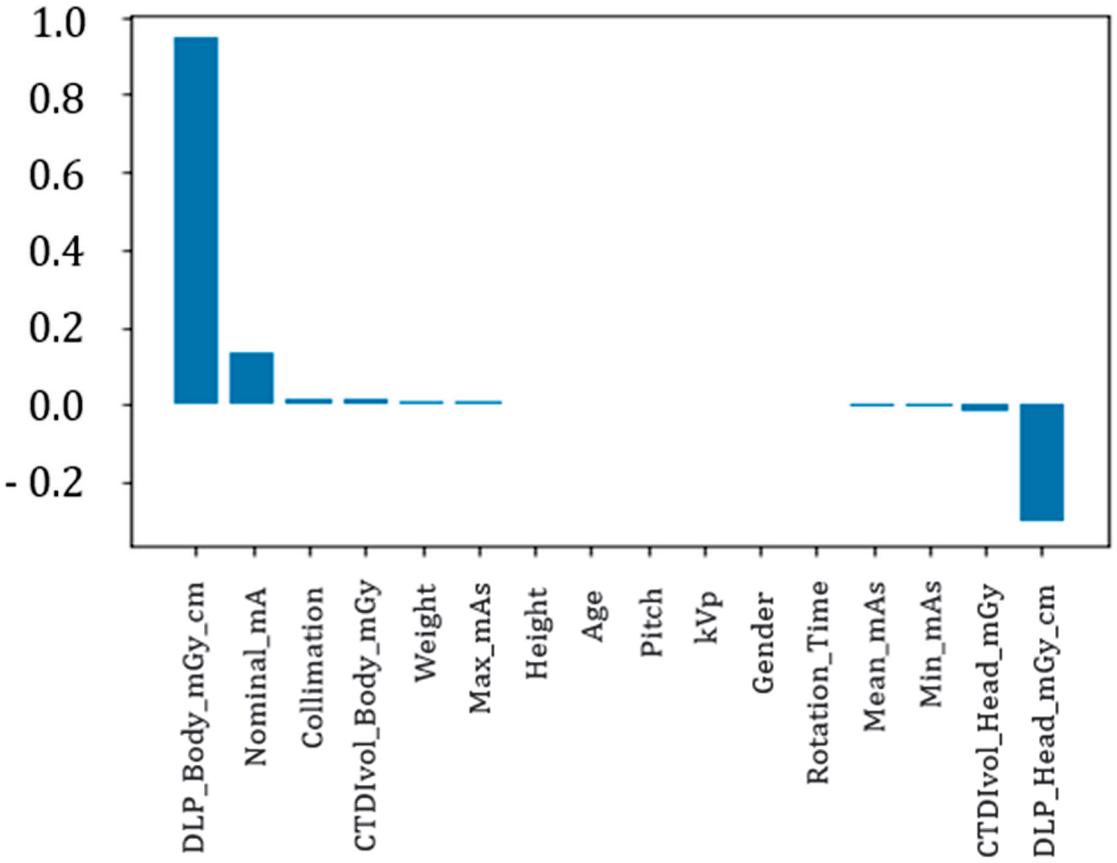


Figure 4. PCA importance (y axis) of the variable.

2.4. Model Comparison

In our study, we compared three machine learning algorithms: a neural network [8], a support vector machine [9], and a random forest [10].

For each of these models, the request was to perform a regression and predict an estimate of the effective dose starting from the features in the training dataset.

Since machine learning model performance is generally sensitive to hyperparameter values, we used the Weight & Biases (W&B) [11] library to experiment with different combinations and choose the optimal hyperparameters for this particular task.

For all optimization studies, we used the training dataset, further divided into a part for training and a part for validation with a 70–30% division.

For all the models, we explored different combinations of hyperparameters with a random search, collecting the results in a series of reports in order to study the impact of each parameter on the final results and compare the different models with each other.

2.4.1. Neural Network

A neural network is a mathematical model where a network of neurons (computational units) connected together is trained by adjusting the values of the weights in order to minimize the value of a cost function that represents a measure of the errors between the network estimates and the real values. In a regression network, the last computational unit proposes a regression that estimates the final value.

We developed a simple script that generates the architecture of a neural network using the high-level API keras [8] to build a multilayer perceptron.

In our optimization analysis, we included these possible combinations of hyperparameters, that are summarized in Table 1:

Table 1. Possible configuration of the neural network that we used in our work in the optimization phase.

Hyperparameter	Choice 1	Choice 2	Choice 3	Choice 4
hidden_layers	[128]	[256, 128, 64]	[512, 256, 128, 64]	-
do	0	0.2	0.4	0.5
activation	relu	-	-	-
solver	adam	sgd	-	-
epochs	50	-	-	-
batch_size	256	512	1024	-
Init_lr	1×10^{-2}	1×10^{-3}	1×10^{-4}	-

The "hidden_layers" parameter is a list where the number of elements means the number of hidden layers and the value of the ith item of the list is the number of computation units in the ith layer (for example, [256, 128, 64] means 3 hidden layers with, respectively, 256, 128, and 64 units).

The "do" parameter is the Drop-Out factor, namely, the percentage of connections that are cut out at each layer.

The "activation" parameter is the activation function used to introduce non-linearity in the network. We used Rectified Linear Unit, "relu".

The "solver" parameter is related to the optimizer choice.

The "epochs" parameter is the number of times that the whole training dataset is shown to the network.

The "batch_size" parameter is the number of samples that are shown at the same time to the network.

The "init_lr" parameter controls the starting learning rate of the optimizer algorithm.

For all parameters, up to four choices are provided. The W&B library randomly picks 30 different configurations to study the impact of these hyperparameters on final performances.

2.4.2. Random Forest Regressor

Random forest models average the prediction of a high number of decision trees. Each decision tree tries to split the dataset iteratively and starting from the mean value in the dataset, adjust the regression to output an estimate of the effective dose based on the values of the different features. Each tree is trained using just some of the total features and part of the dataset in a parallel way in order to reduce overfitting chance. We used the random forest regressor implementation in the scikit-learn library [10].

In our optimization analysis, we included these possible combinations of hyperparameters, that are summarized into Table 2:

Table 2. Possible configuration for random forest regression.

Hyperparameter	Choice 1	Choice 2	Choice 3	Choice 4
n_estimators	100	500	1000	1×10^{4}
max_depth	None	10	100	-
ccp_alpha	0	1×10^{-4}	1×10^{-3}	1×10^{-2}

'n_estimators' is the number of decision trees to train.

'max_depth' is the maximum number of splits allowed for each tree.

'ccp_alpha' is a regularization parameter that is involved in the pruning procedure following the Minimal Cost-Complexity Pruning procedure.

For all parameters, up to four possible choices are provided and the W&B library randomly picks up to 15 different configurations to study the impact of these hyperparameters on final performances.

2.5. Model Selection and Training

For each configuration of each model, we evaluated three parameters, namely, the r2 score, the mean absolute error (MAE), and the mean absolute percentage error (MAPE), considering the estimates from Radimetrics as our gold standard. All models were evaluated using these metrics and the best configuration of each algorithm was chosen to fit the data.

We fitted the best model with the whole training set and evaluated the performances on the test set and the external dataset, with an increasing number of features, as shown in Table 3.

Table 3. The number of features used to feed the best-performing model, for the external validation.

6 Features	8 Features	10 Features
Gender	Gender	Gender
Weight	Weight	Weight
Height	Height	Height
$CTDI_{vol}$ [mGy]	$CTDI_{vol}$ [mGy]	$CTDI_{vol}$ [mGy]
DLP [mGycm]	DLP [mGycm]	DLP [mGycm]
Scan Region	Scan Region	Scan Region
	kV_p	kV_p
	Mean mAs	Mean mAs
		Collimation
		Pitch

Then, we finally chose the best-performing model and used it to perform a class dose prediction.

2.6. Effective Dose from Other Estimates

To evaluate the added value of effective dose estimates by machine learning, effective doses were also evaluated using DLP and k- factors from the work of Romanyukha et al. [12]. Scan regions were attributed according to study description, and correspondent k-factors were used for E estimate, according to the well-known formula

$$E = k \times DLP$$

In addition, E was evaluated with multiple linear regression using the 6 features of Table 4 (gender, weight, height, CTDI, DLP, and scan region), computing the R^2 of the estimates.

MAE and MAPE were calculated accordingly, considering the estimates from the dose-tracking software as our gold standard.

Table 4. Summary of the results of different models on the internal test set.

Model	MAE [mSv]	MAPE [%]	R^2
Neural Network	0.557	8.6	0.97
Support Vector Machine	0.658	23.3	0.96
Random Forest Regressor	0.416	7	0.98
k-factors and DLP	2.06	26	\
Multiple Linear Regression	0.98	44.4	0.93

3. Results

3.1. Internal Validation

In the optimization phase, we logged a lot of information about the impact of the selected hyperparameters on the performance of the model. Weights & Biases provide a lot of useful panels to explore correlations and results of different runs (see Figure 5 as an example).

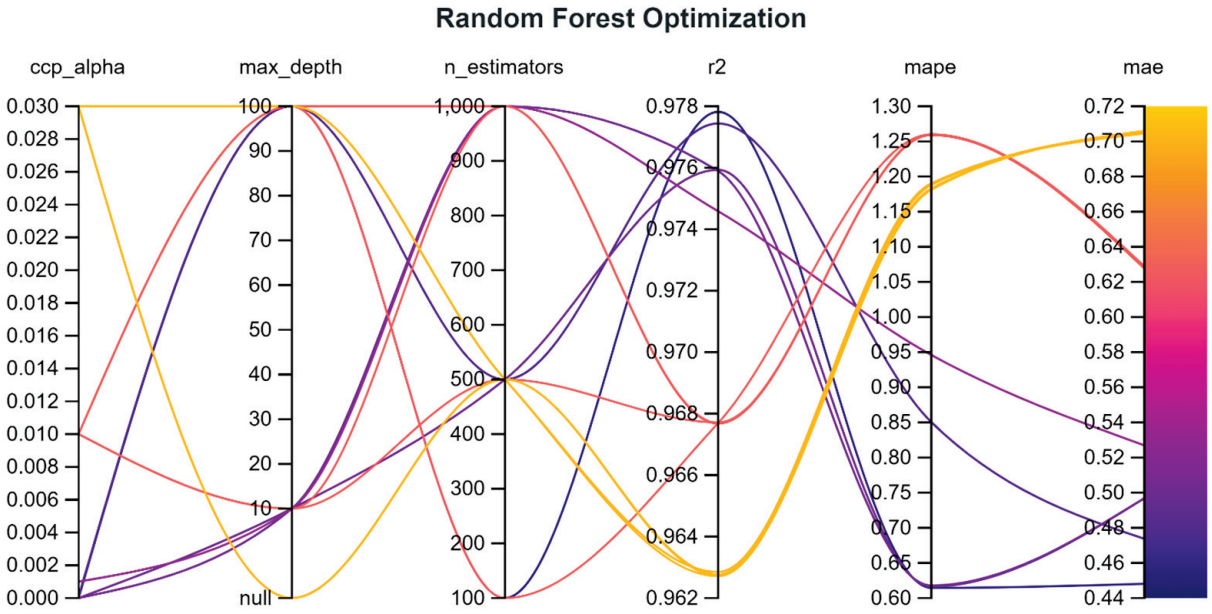


Figure 5. Example of optimization for random forest using Weights & Biases. Each configuration is a line that passes the point that represents the values of ccp_alpha, max_depth, and n_estimators chosen by the random search algorithm. R^2, MAPE, and MAE are the results.

For the neural network, in the optimization phase, MAE ranged from 0.57 mSv to 2.99 mSv, with an R2 score that ranged from 0.36 to 0.97. Analyzing the W&B report, we chose as final hyperparameters the combination of hidden_layers = [512, 256, 128, 64], do = 0, init_lr = 1×10^{-4}, solver = 'adam', and batch_size = 512.

For the support vector machine, in the optimization phase, MAE ranged from 0.68 mSv to 4 mSv, with an r2 score that ranged from 0.14 to 0.96. Analyzing the W&B report, we chose as final hyperparameters the combination of C = 1, kernel = "rbf", and gamma = "scale".

For the random forest regressor, in the optimization phase, MAE ranged from 0.44 mSv to 0.7 mSv, with an r2 score that ranged from 0.96 to 0.98.

We fitted all the selected models with the whole training dataset and then evaluated their performances using mean absolute error, mean absolute percentage error, and the coefficient of determination. Table 4 summarizes the results of each model. We found good

r2 with all models and the overall best-performing model was the random forest regressor, which achieved the lowest mean absolute error.

The visual correlation between predicted and expected doses for random forest is shown in Figure 6.

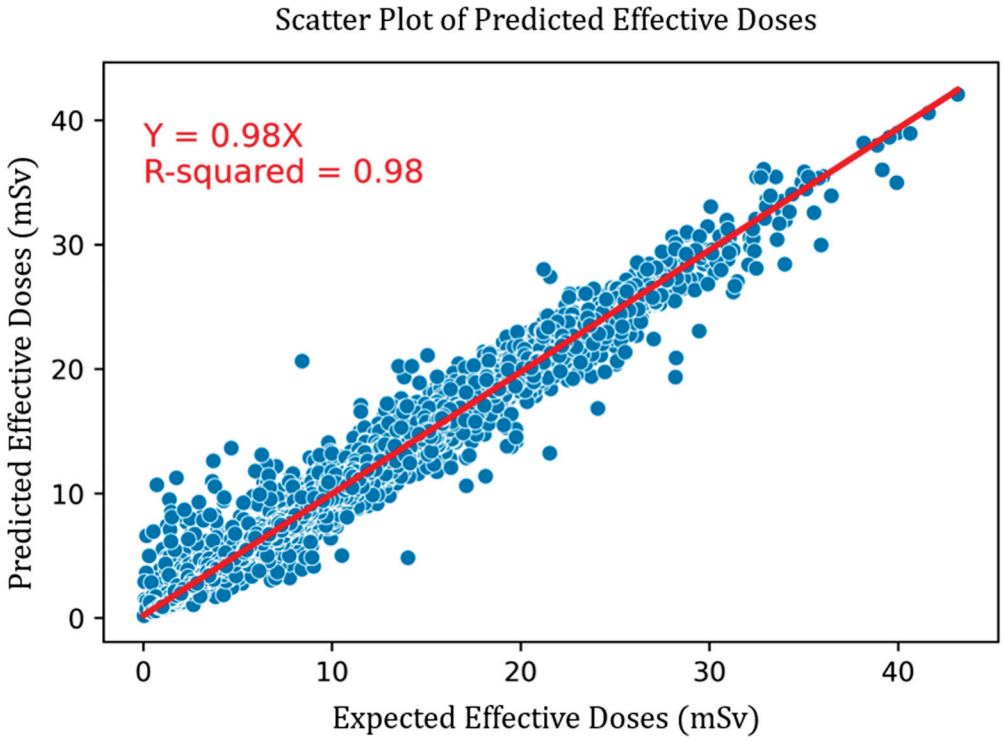


Figure 6. Predicted vs. expected doses for random forest regressor.

3.2. Dose Class Prediction

The new Italian regulation (D.Lgs. 101/20), a transposition of the Euratom 59/2013, establishes that every radiological procedure must indicate in the radiological report information relating to patient exposure in terms of classes of effective dose (E).

We chose our best model, the random forest regressor, and mapped all the Radimetrics and model estimates of the effective dose into the corresponding dose class to compute accuracy, precision, recall, and f1-score for each class to ensure that our model is able to reasonably place each acquisition in the right dose class. We achieved high results for each class, which further supports the hypothesis that random forest could be a way to estimate the effective dose and predict the right dose class (Table 5).

Table 5. Results in dose class prediction from random forest regressor effective dose estimates.

DOSE Class	Precision	Recall	F1-Score	Support
I	1	0.99	0.99	3171
II	0.96	0.96	0.96	4904
III	0.93	0.93	0.93	4871
IV	0.96	0.96	0.96	4879
accuracy	0.96			17,735

3.3. External Validation

We reached good results on test data coming from the same source, so devices, body regions, and kind of exam are really homogeneous between training and test

data (Tables 6 and 7). To further check the generalization ability of our model, we also collected data coming from different hospitals, with similar body regions (mainly body and head exams).

Table 6. Results for the external validation as a function of the number of features.

	MAE (MAPE) 6 Features	**MAE (MAPE) 8 Features**	**MAE (MAPE) 10 Features**
External Validation—Turin	0.243 (7.9%)	0.224 (7.4%)	0.215 (7.1%)

Multiple linear regression for the external dataset resulted in a MAE of 0.89 mSv and a MAPE of 32.8% with an R^2 of 0.74.

Table 7. Results in dose class prediction from the random forest regressor effective dose estimates for the external dataset.

DOSE Class	**Precision**	**Recall**	**F1-Score**	**Support**
I	0.96	0.80	0.87	205
II	0.96	0.98	0.97	2655
III	0.92	0.93	0.93	859
IV	0.93	0.85	0.89	81
Accuracy	0.95			3800

3.4. Interpretability

Assuming that the Monte Carlo approach is the gold standard for effective dose estimation and that dose-tracking software is the tool that can provide us with the best reasonably achievable estimate of this parameter, the use of simple machine learning models can also be a useful tool when trying to explore how these models use information to estimate the variable of interest, without having access to the image. Surely, it is necessary and inevitable that some biases present in the data are included and learned by the model—and this is probably the cause of the different performances if evaluated on data from other sources—and for models such as random forest, it is possible to study the weight of each feature and also the trend of the most important ones.

For this purpose, we calculated the importance of the features (see Figure 7), where we found results in agreement with what was expected from the preliminary analysis based on the correlations and the decomposition into main PCA components; specifically, we identified DLP Body as the most important feature from which an initial estimate starts, which is then adjusted using the other features to account for specific variables.

To further explore the relationships learned from the model, we developed a partial dependency analysis for three variables of particular interest: DLP, mean mAs, and kVp.

The results are shown in Figure 8: these plots are a measure of the trend of the prediction as the only variable indicated varies, perturbing the input data, and plotting how the estimate varies according to the parameter studied.

We have identified an almost linear trend for DLP and a low dependence (a practically horizontal line at the height of the average effective dose value in our dataset—6.88 mSv), in agreement with the fact that DLP (Body) is the parameter that most affects the estimate.

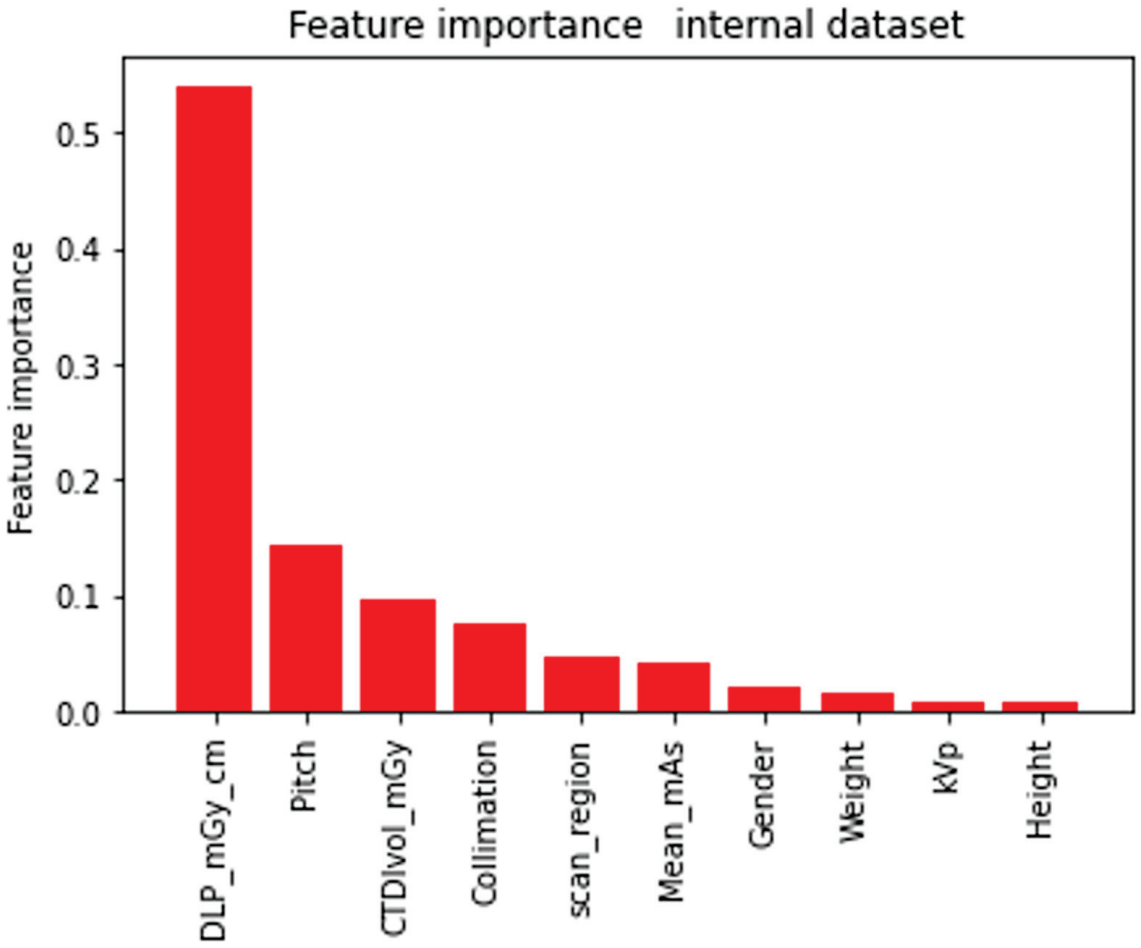


Figure 7. Feature importance for random forest regressor.

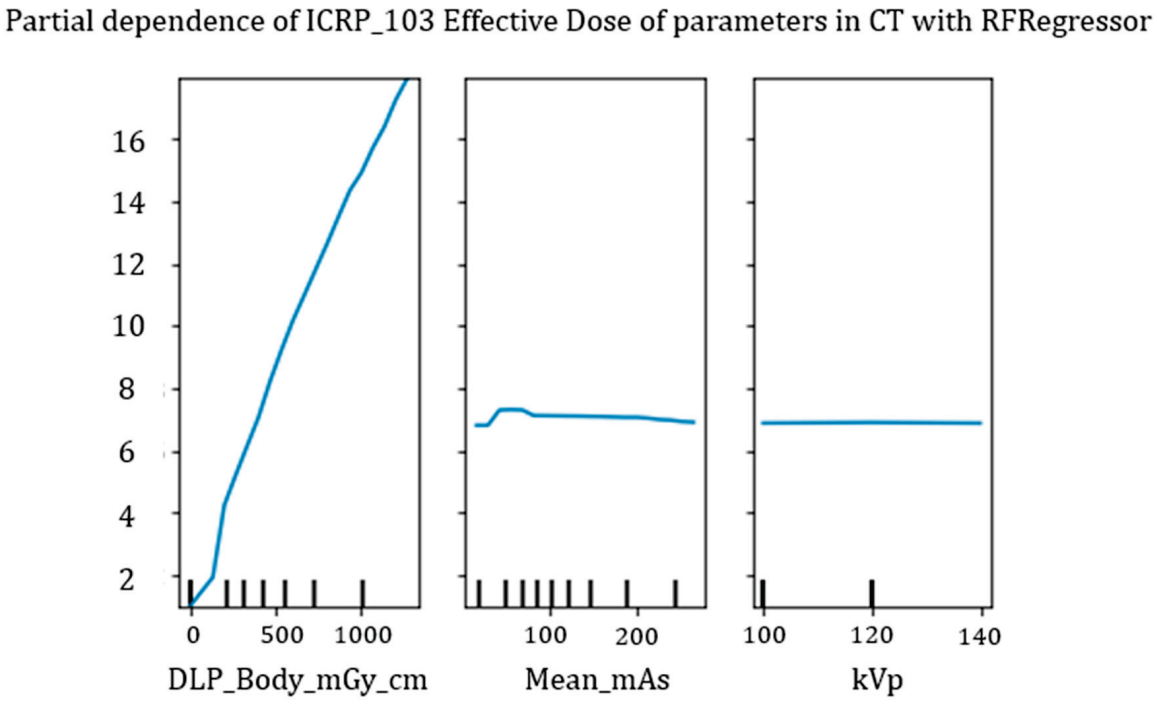


Figure 8. Dependence analysis for DLP, mean_mAs, and kVp.

4. Discussion

Estimating the effective dose is a crucial point for radiation protection and the optimization of diagnostic processes based on medical imaging with ionizing radiation. Models capable of accurately estimating the effective dose are the first step to understand which parameters have the greatest impact on the dose delivered, linked to possible risks for the patient. ICRP 147 underlines the importance of the effective dose in medical radiation exposure and the need to standardize effective dose calculations to provide reference dose coefficients for specified radiographic and CT procedures.

The definition of the effective dose includes the study of the dose to the organs, weighted by the relative coefficients that take into account the different radiosensitivity of the tissues. Different calculation methods, different software, and different investigation methodologies have an important impact on the calculation of the dose. As demonstrated by De Mattia et al. [13], there are important differences in different software, with variability in the estimate of the effective dose up to 35% and in other parameters beyond 100%. According to the authors of this work, the coefficients found in the literature and implemented in the software are not always representative of the adequate values for the different clinical protocols in different facilities.

Ultimately, despite the clear certainty of correlations with some variables such as dosimetric indices, kVp, and current modulation, differences in scanners and in patients' characteristics can lead to very important differences in effective dose estimates.

Aware of these limitations, we explored the performance of different machine learning algorithms, trained on a very large and homogeneous dataset, to obtain a fast and light model capable of estimating the effective dose with reasonable accuracy starting from the patient's anatomical data and information tied to the scanner, without the need to process the image.

The results of the optimization phase were quite expected: for example, the architecture of the neural network and the choice of the best hyperparameters are reasonable, given the type of data to be handled. For the support vector machine, 'rbf' kernel maps input data into higher-dimension space, allowing better generalization, while for random forest, a small value of 'max depth' (10) and an intermediate number of 'n_estimators' (500) are able to capture the complexity of the dataset, allowing a good generalization without overfitting.

On test data coming from the same hospital, the random forest regressor showed really good predictions, with a mean absolute error below 0.5 mSv and an average accuracy of 0.96 in the prediction of the dose class.

The same model showed the ability to incorporate all the essential correlations to produce a reasonable estimate and to reproduce the results with a low average percentage error, even on cases from different sources, with patients, scanners, and clinical protocols that certainly differ, as shown from the results of the external validation, with an absolute error well below 1 mSv and a percentage error below 10%.

The prediction of dose class for the external dataset resulted in an accuracy of 0.95.

The performance of the model proved to be sufficiently robust to account for the fact that two dose-tracking software estimated organ doses, and thus effective doses, using different phantoms.

Such results could lead to a development of a pipeline in which features from Table 4 could be extracted from each acquisition, to feed the model, obtaining the effective dose estimates. The features needed are, in fact, reported in the DICOM metadata of each CT acquisitions, thus being readily available and easily implemented in an automatic workflow.

Limitations

While the random forest regressor model performed very well on data from our center and showed promising results even on data from different centers, this type of approach includes several limitations. As already described, different scanners and protocols lead to great variability in the final estimates even on the same software, which are amplified if different software is used for the estimation. Machine learning is promising for this type of study because, as we have shown, it is able to accurately learn correlations and even expected physical trends; however, training on data from a limited number of scanners and from a single center induces the model to learn the specific biases of that center and of those scanners, which worsens the performance of generalizations on different situations. Undoubtedly, extensive training with data from multicentric studies could result in an estimate that is less affected by this type of problem and with a greater capacity for generalization.

5. Conclusions

In this work, we presented an exploration of how machine learning could help predict the effective dose in CT imaging. We proposed and evaluated different optimized models and found the best performances with random forest regressor, which could be a tool to estimate the effective dose and thus the effective dose class with high accuracy once trained on enough data.

We also checked that the model had correctly learned the physical trends that are expected from the literature and checked it against datasets coming from different centers, using different devices, protocols, and even dose-tracking software.

We found results with a mean absolute error and a mean absolute percentage error in line with those expected in the literature for that type of difference.

We can conclude that machine learning could have an increasing role in the future of effective dose estimation and could perhaps be included in dose-tracking software after extensive training on multicentric data. This could soon be a new tool for personalized, lightweight effective dose estimation to be performed online just after acquisition.

Author Contributions: Conceptualization, M.F.; methodology, M.F., P.D.M. and O.R.; software, M.F., P.D.M. and O.R.; validation, M.F., P.D.M. and O.R.; writing—original draft preparation, M.F. and P.D.M.; writing—review and editing, M.F., P.D.M., O.R., L.G. and D.O.; supervision, O.R. and D.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available upon reasonable request.

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

References

1. Harrison, J.D.; Balonov, M.; Bochud, F.; Martin, C.J.; Menzel, H.G.; Smith-Bindman, R.; Ortiz-Lopez, P.; Simmonds, J.R.; Wakeford, R. The use of dose quantities in radiological protection: ICRP publication 147 Ann ICRP 50(1) 2021. *J. Radiol. Prot.* **2021**, *41*, 410. [CrossRef] [PubMed]
2. Bayer HealthCare. *Radimetrics Enterprise Platform: Dose Management Solution*; Bayer HealthCare: Pittsburgh, PA, USA, 2014; pp. 154–196. Available online: https://myradiologysolutions.bayer.com/s/radimetrics/ (accessed on 20 March 2021).
3. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; et al. Scikit-learn: Machine Learning in Python. *arXiv* **2012**, arXiv:1201.0490. [CrossRef]

4. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems (Version 2). *arXiv* **2016**, arXiv:1603.04467. [CrossRef]
5. Physico Emme Esse Italy. Available online: https://www.emme-esse.com/physico/ (accessed on 20 March 2021).
6. ICRP. In *The 2007 Recommendations of the International Commission on Radiological Protection*; ICRP Publication 103. Ann. ICRP 37 (2-4); ICRP, 2007.
7. Waskom, M. Seaborn: Statistical data visualization. *J. Open Source Softw.* **2021**, *6*, 3021. [CrossRef]
8. Chollet, F. Keras. 2015. Available online: https://github.com/fchollet/keras (accessed on 20 March 2021).
9. Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27. [CrossRef]
10. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Routledge: New York, NY, USA, 2017. [CrossRef]
11. Biewald, L. Experiment Tracking with Weights and Biases. 2020. Available online: https://www.wandb.com (accessed on 20 March 2021).
12. Romanyukha, A.; Folio, L.; Lamart, S.; Simon, S.L.; Lee, C. Body Size-Specific Effective Dose Conversion Coefficients For CT Scans. *Radiat. Prot. Dosim.* **2016**, *172*, 428–437. [CrossRef] [PubMed]
13. De Mattia, C.; Campanaro, F.; Rottoli, F.; Colombo, P.E.; Pola, A.; Vanzulli, A.; Torresin, A. Patient organ and effective dose estimation in CT: Comparison of four software applications. *Eur. Radiol. Exp.* **2020**, *4*, 14. [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

Detection of Intracranial Hemorrhage from Computed Tomography Images: Diagnostic Role and Efficacy of ChatGPT-4o

Mustafa Koyun [1,*], Zeycan Kubra Cevval [1], Bahadir Reis [2] and Bunyamin Ece [2]

[1] Department of Radiology, Kastamonu Training and Research Hospital, Kastamonu 37150, Turkey; zeycancevval@gmail.com
[2] Department of Radiology, Kastamonu University, Kastamonu 37150, Turkey; bahadirreis@hotmail.com (B.R.); bunyaminece@kastamonu.edu.tr (B.E.)
* Correspondence: dr.mustafakoyun@gmail.com

Abstract: Background/Objectives: The role of artificial intelligence (AI) in radiological image analysis is rapidly evolving. This study evaluates the diagnostic performance of Chat Generative Pre-trained Transformer Omni (GPT-4 Omni) in detecting intracranial hemorrhages (ICHs) in non-contrast computed tomography (NCCT) images, along with its ability to classify hemorrhage type, stage, anatomical location, and associated findings. **Methods:** A retrospective study was conducted using 240 cases, comprising 120 ICH cases and 120 controls with normal findings. Five consecutive NCCT slices per case were selected by radiologists and analyzed by ChatGPT-4o using a standardized prompt with nine questions. Diagnostic accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated by comparing the model's results with radiologists' assessments (the gold standard). After a two-week interval, the same dataset was re-evaluated to assess intra-observer reliability and consistency. **Results:** ChatGPT-4o achieved 100% accuracy in identifying imaging modality type. For ICH detection, the model demonstrated a diagnostic accuracy of 68.3%, sensitivity of 79.2%, specificity of 57.5%, PPV of 65.1%, and NPV of 73.4%. It correctly classified 34.0% of hemorrhage types and 7.3% of localizations. All ICH-positive cases were identified as acute phase (100%). In the second evaluation, diagnostic accuracy improved to 73.3%, with a sensitivity of 86.7% and a specificity of 60%. The Cohen's Kappa coefficient for intra-observer agreement in ICH detection indicated moderate agreement ($\kappa = 0.469$). **Conclusions:** ChatGPT-4o shows promise in identifying imaging modalities and ICH presence but demonstrates limitations in localization and hemorrhage type classification. These findings highlight its potential for improvement through targeted training for medical applications.

Keywords: ChatGPT; artificial intelligence; intracranial hemorrhage; computed tomography; radiology

1. Introduction

Intracranial hemorrhage (ICH) is the second leading cause of stroke cases worldwide and occurs secondary to multiple factors such as trauma, infarction, aneurysm rupture, and anticoagulant therapy [1,2]. It is estimated that 37,000 to 50,000 ICH cases occur annually in the United States [3]. ICH can be classified into different types including subdural, epidural, subarachnoid, intraparenchymal, and intraventricular hemorrhages, and may lead to severe neurological damage or death if not managed appropriately [4].

A study indicates that, despite a slight decrease in ICH incidence during the ten-year period from 2000 to 2010, there was no significant reduction in mortality rates, with the 30-day mortality rate approaching 40% [5]. Half of the deaths following ICH occur within the first 24 h, and early diagnosis and treatment improve outcomes [6]. The existence of ICH constitutes a contraindication for intravenous thrombolysis therapy in patients with stroke [7]. Therefore, the early and accurate detection of ICH is crucial for patient prognosis and treatment planning.

In the detection of ICH and the identification of the cause of hemorrhage, imaging modalities such as non-contrast computed tomography (NCCT) of the brain, magnetic resonance imaging (MRI), and cerebral angiography are utilized [8]. Among these imaging modalities, NCCT is considered the gold standard for ICH detection due to its high sensitivity and is widely used in clinical practice [9]. The primary diagnostic advantage of NCCT in the early period (0–6 h) is its ability to accurately determine the presence or absence of hemorrhage [4]. Additionally, NCCT can provide valuable information in determining the hemorrhage localization, intraventricular extension, occurrence and degree of hydrocephalus and edema, presence of midline shift, and compressive effects of the hematoma on surrounding structures [9]. Currently, computed tomography (CT) images are evaluated and reported by radiologists, requiring a specific time period for accurate interpretation. The number of images that radiologists must analyze on a daily basis has grown dramatically over time; reports suggest that they are required to assess images from a single patient in as low as 3–4 s [10]. Moreover, due to the absence of round-the-clock radiologist coverage in many hospitals, CT images are evaluated by radiology residents or clinicians, particularly during night shifts [11,12]. The brief time allocated for image interpretation and the dependency of results on interpreter experience lead to both diagnostic errors and interpretation variations in ICH diagnosis [13].

Artificial intelligence (AI) technology has made rapid progress in recent years and has started to be used in many areas of our lives. The use of AI in healthcare, particularly in radiology, is steadily increasing. AI has been demonstrated to be beneficial in multiple aspects of healthcare, such as accurately identifying radiological imaging findings, improving health screening programs, and reducing medical errors [14]. An AI component integrated into radiological imaging processes is considered to have the potential to increase work efficiency, expedite diagnostic processes, and reduce human-related errors [13]. In this context, AI technologies such as deep learning algorithms have been developed to reduce workload and time spent in radiology while enhancing diagnostic power [15–17]. Deep learning is a machine learning approach that utilizes convolutional neural networks for solving both simple and complex tasks [18]. Hundreds of deep learning algorithms have been developed for ICH detection across different centers, supported by varying levels of evidence [3,19–22]. These deep learning algorithms have been reported to have numerous known limitations, including the requirement for large and diverse training datasets, biases in dataset compilation, poor generalizability, overfitting, limited clinical validation studies, and inability to interpret underlying mechanisms [22]. Moreover, it poses a challenge for radiologists to select and utilize one algorithm among hundreds of deep learning algorithms. At this point, the question arises whether Chat Generative Pre-trained Transformer (ChatGPT) (OpenAI, San Francisco, CA, USA), which has become highly popular in recent years and offers advanced image/audio processing capabilities, possesses the ability to interpret radiological images and participate in the diagnostic process.

ChatGPT, one of the AI technologies, was developed by OpenAI and introduced on 30 November 2022. ChatGPT is one of the large language models (LLMs) that has been trained unsupervised on extensive text data and can subsequently make inferences about relationships between words within text [23]. The initial versions of ChatGPT, ChatGPT-3 and

3.5, possessed multiple capabilities including text authorship, cross-language translation, and question answering. ChatGPT-4o was introduced on 13 May 2024 and offered image processing and inference capabilities in addition to its previous features. The presence of ChatGPT's image processing and inference capabilities suggests its potential to evaluate radiological images. ChatGPT is considered to have the potential to assist in evaluating radiological images, detecting and characterizing abnormalities, providing preliminary diagnoses, and even recommending appropriate follow-up examinations [24].

Several studies in the literature have employed ChatGPT to evaluate mammographic images, identify distal radius fractures from radiographs, assess COVID-19 and lung cancer through chest CT imaging, and analyze ischemic stroke utilizing diffusion-weighted imaging (DWI) [25–28]. In the literature, many AI models, particularly deep learning-based ones, have been used in ICH detection (Table 1) [3,16,20,29,30]. However, our literature review revealed no studies evaluating intracranial hemorrhage detection using ChatGPT-4o. The major contributions of our study to the literature are that it is the first research evaluating the performance of ChatGPT-4o in detecting intracranial hemorrhage in brain NCCT images, and our study demonstrates the potential of a general-purpose LLM in a specialized medical imaging field.

Table 1. Summary table showing diagnostic performance of AI models used in ICH detection.

Author and Year	AI Model Used	Study Objective	Sample Size	Main Findings
Ginat et al., 2020 [3]	Deep learning	ICH detection from NCCT images	373 cases	ICH detection Sensitivity: 88.7%, Specificity: 94.2%
Yun et al., 2023 [16]	Deep learning	ICH detection from NCCT images	49,841 cases	ICH detection Sensitivity: 95.9%, Specificity: 95.3%
Heit et al., 2024 [20]	Deep learning	ICH detection from NCCT images	308 cases	ICH detection Sensitivity: 95.6%, Specificity: 95.3%
Daiwen et al., 2024 [29]	ChatGPT-4	ICH detection from NCCT images	208 cases	72.6% ICH detection rate
Arbabshirani et al., 2018 [30]	Deep learning	ICH detection from NCCT images	46,583 cases	ICH detection Sensitivity: 78%, Specificity: 80%

ICH: intracranial hemorrhage; AI: artificial intelligence; NCCT: non-contrast computed tomography.

The aim of this study is to investigate whether ChatGPT-4o can accurately detect ICH, its type, stage, localization, and associated pathologies from brain NCCT images.

2. Materials and Methods

2.1. Study Design and Case Selection

The approval for this retrospective study was obtained from the Non-Interventional Clinical Research Ethics Committee of Erzincan Binali Yildirim University (decision no: 2024-09/05; date: 11 July 2024). Our study adhered to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) guidelines [31].

Sample size calculation was performed using G*Power software (version 3.1.9.7). Power analysis was conducted with 80% statistical power and type I error probability of $\alpha = 0.05$. To detect a 90% success rate for the tested diagnostic method compared to the 100% success rate of the reference group, it was determined that a total of 132 cases, with a minimum of 66 cases per group, should be included in the study.

The brain NCCT images of 4853 patients archived in our institution's Picture Archiving Communication Systems (PACS) database between June 2023 and June 2024 were retrospectively evaluated. Considering the inclusion and exclusion criteria detailed below, a total of 240 cases were selected, comprising 120 cases with intracranial hemorrhage and 120 cases without intracranial hemorrhage (Figure 1). Case selection was performed through consensus evaluation by two radiologists with 8 and 10 years of clinical experience, respectively. Subsequently, five brain NCCT slices were selected from each case.

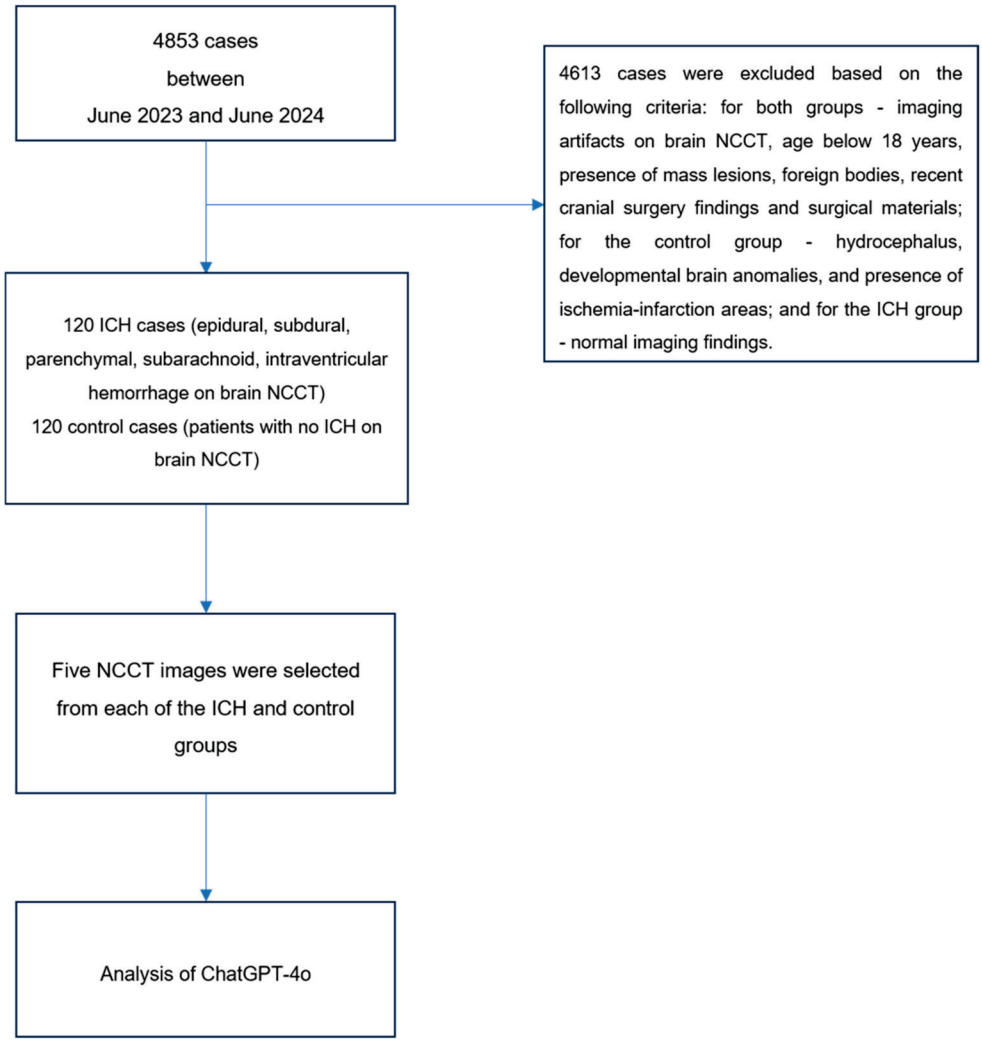
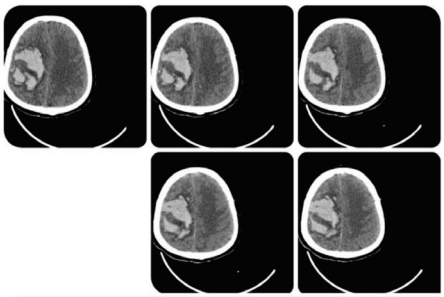


Figure 1. Flowchart of patient selection and ChatGPT-4o upload process.

Within the framework of the study protocol, a standardized prompt set consisting of nine questions was developed to be presented to ChatGPT-4o (GPT-4 Omni) (OpenAI, San Francisco, CA, USA) along with brain NCCT images. Five brain NCCT slices were selected from each case and uploaded to the model along with the designated prompt, and the resultant data were documented (Figures 2, 3, A1 and A2). The acquired data were compared with the gold standard, which was the consensus-based conclusions of radiologists. Following a two-week interval, the brain NCCT images were uploaded to ChatGPT-4o utilizing the identical prompt. This methodological approach was implemented to assess the model's response consistency across different temporal intervals. Furthermore, to minimize potential bias effects stemming from previous analyses, ICH and control group cases were uploaded to the model in a randomized sequence.



1. What is the name of this radiological imaging method?
Are there any hemorrhages in these images? Please answer "Yes" or "No."
2. What type(s) of hemorrhage(s) are present? Please answer this question with "subdural," "epidural," "subarachnoid," "intraparenchymal," or "intraventricular." If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
3. If the type of hemorrhage(s) is "subdural" or "epidural," what is the location of the hemorrhage(s)? Please specify the location with "on the patient's right" or "on the patient's left" followed by "frontal," "parietal," "temporal," "occipital," "frontoparietal," "frontotemporal," "temporoparietal," "temporooccipital," "posterior fossa," etc. If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
4. If the type of hemorrhage(s) is "intraparenchymal," what is the location of the hemorrhage(s)? P. lease specify the location with "on the patient's right" or "on the patient's left" followed by "frontal," "parietal," "temporal," "occipital," "frontoparietal," "frontotemporal," "temporoparietal," "temporooccipital," "cerebellar," "thalamus," "pons," "mesencephalon," etc. 5. If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
6. If the type of hemorrhage(s) is "subarachnoid," what is the location of the hemorrhage(s) in relation to the patient? Please specify the location with "on the patient's right" or "on the patient's left" followed by "frontal," "parietal," "temporal," "occipital," "frontoparietal," "frontotemporal," "temporoparietal," "temporooccipital," "cerebellar." If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
7. If the type of hemorrhage(s) is "intraventricular," what is the location of the hemorrhage(s) in relation to the patient? Please specify the location as "3rd ventricle," "4th ventricle," "right lateral ventricle," or "left lateral ventricle." If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
8. What is the phase of the hemorrhage(s)? Please answer this question with "acute," "subacute," or "chronic." If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
9. Are there any additional pathological findings related to the hemorrhage(s) in these images? If so, please specify the pathology/ies as "right shift," "left shift," "brain edema," "3rd ventricle compression," "right lateral ventricle compression," "left lateral ventricle compression," etc.

1. **What is the name of this radiological imaging method?**
   - This is a **non-contrast computed tomography (CT)** scan of the brain.
2. **Are there any hemorrhages in these images?**
   - **Yes**, there is a hemorrhage.
3. **What type(s) of hemorrhage(s) are present?**
   - The hemorrhage is **intraparenchymal**.
4. **If the type of hemorrhage(s) is "intraparenchymal," what is the location of the hemorrhage(s)?**
   - The hemorrhage is located **on the patient's left parietal region**.
5. **If the type of hemorrhage(s) is "subarachnoid," specify the location.**
   - Not applicable.
6. **If the type of hemorrhage(s) is "intraventricular," specify the location.**
   - Not applicable.
7. **What is the phase of the hemorrhage(s)?**
   - The hemorrhage appears **acute**, as it is hyperdense (bright white) compared to the surrounding brain tissue, which is typical for an acute bleed.
8. **Are there any additional pathological findings related to the hemorrhage(s) in these images?**
   - Yes, there is evidence of **rightward midline shift**, suggesting significant mass effect due to the hemorrhage.
   - **Brain edema** is also present around the hemorrhage.

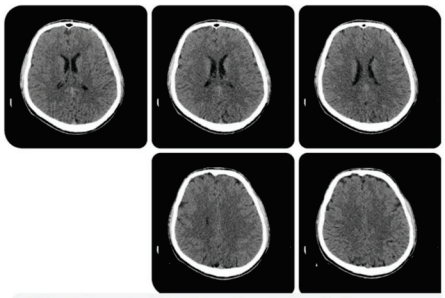
Figure 2. The evaluation of a case from the ICH group by ChatGPT-4o.

2.2. Inclusion and Exclusion Criteria

Cases with ICH (epidural, subdural, parenchymal, subarachnoid, intraventricular hemorrhage) on brain NCCT slices were included in the ICH group, while cases without ICH were included in the control group. The exclusion criteria for both ICH and control groups comprised artifacts in brain NCCT slices, patient age below 18 years, the presence of masses, foreign bodies, recent cranial surgical findings, and operative materials. Additionally, for the control group, the presence of hydrocephalus, developmental brain anomalies, and areas of ischemia–infarction were established as exclusion criteria.

2.3. Brain NCCT Images

Brain NCCT examinations were performed using a standardized protocol on a 16-slice spiral CT (Lightspeed 16, General Electric Medical Systems, Milwaukee, WI 53201, USA) with acquisition parameters as follows: a tube voltage of 120 kVp, a 128-detector configuration, and a slice thickness of 2 mm. All images were obtained in the axial plane with brain parenchymal window settings (window width: 120; window level: 40).



1. What is the name of this radiological imaging method?
Are there any hemorrhages in these images? Please answer "Yes" or "No."
2. What type(s) of hemorrhage(s) are present? Please answer this question with "subdural," "epidural," "subarachnoid," "intraparenchymal," or "intraventricular." If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
3. If the type of hemorrhage(s) is "subdural" or "epidural," what is the location of the hemorrhage(s)? Please specify the location with "on the patient's right" or "on the patient's left" followed by "frontal," "parietal," "temporal," "occipital," "frontoparietal," "frontotemporal," "temporoparietal," "temporooccipital," "posterior fossa," etc. If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
4. If the type of hemorrhage(s) is "intraparenchymal," what is the location of the hemorrhage(s)? P. lease specify the location with "on the patient's right" or "on the patient's left" followed by "frontal," "parietal," "temporal," "occipital," "frontoparietal," "frontotemporal," "temporoparietal," "temporooccipital," "cerebellar," "thalamus," "pons," "mesencephalon," etc. 5. If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
6. If the type of hemorrhage(s) is "subarachnoid," what is the location of the hemorrhage(s) in relation to the patient? Please specify the location with "on the patient's right" or "on the patient's left" followed by "frontal," "parietal," "temporal," "occipital," "frontoparietal," "frontotemporal," "temporoparietal," "temporooccipital," "cerebellar." If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
7. If the type of hemorrhage(s) is "intraventricular," what is the location of the hemorrhage(s) in relation to the patient? Please specify the location as "3rd ventricle," "4th ventricle," "right lateral ventricle," or "left lateral ventricle." If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
8. What is the phase of the hemorrhage(s)? Please answer this question with "acute," "subacute," or "chronic." If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
9. Are there any additional pathological findings related to the hemorrhage(s) in these images? If so, please specify the pathology/ies as "right shift," "left shift," "brain edema," "3rd ventricle compression," "right lateral ventricle compression," "left lateral ventricle compression," etc.

1. **What is the name of this radiological imaging method?**
   - This is a **non-contrast computed tomography (CT) scan of the brain.**
2. **Are there any hemorrhages in these images?**
   - **No**, there are no visible hemorrhages in these images.
3. **What type(s) of hemorrhage(s) are present?**
   - Not applicable.
4. **If the type of hemorrhage(s) is "subdural" or "epidural," what is the location of the hemorrhage(s)?**
   - Not applicable.
5. **If the type of hemorrhage(s) is "intraparenchymal," what is the location of the hemorrhage(s)?**
   - Not applicable.
6. **If the type of hemorrhage(s) is "subarachnoid," what is the location of the hemorrhage(s)?**
   - Not applicable.
7. **If the type of hemorrhage(s) is "intraventricular," what is the location of the hemorrhage(s)?**
   - Not applicable.
8. **What is the phase of the hemorrhage(s)?**
   - Not applicable.
9. **Are there any additional pathological findings related to the hemorrhage(s) in these images?**
   - There are **no obvious pathological findings** such as midline shift, brain edema, or ventricular compression in these images. The brain parenchyma and ventricular system appear within normal limits.

Figure 3. The evaluation of a case from the healthy control group by ChatGPT-4o.

2.4. Image Selection and Evaluation

The image upload capability of ChatGPT-4o is dependent on various factors including the number of images, image dimensions, and the volume of accompanying text, with a current size limitation of 20 megabytes per image [32]. Therefore, it was determined that five consecutive slices should be uploaded rather than all brain NCCT images of the cases to the model. From cases with ICH, 5 consecutive slices that optimally demonstrated the hemorrhagic area were selected. For the control group, sectional images were identified at anatomical planes equivalent to those selected from ICH cases.

Image formats compatible with ChatGPT-4o include JPEG, PNG, and static GIF files [29]. Therefore, the original brain NCCT images in DICOM (Digital Imaging and Communications in Medicine) format were anonymized and converted to JPEG format using AW Volumeshare 7 software (AW 4.7 version, GE Healthcare, Milwaukee, WI, USA). Irrelevant regions outside the brain were cropped and removed from the JPEG images. The JPEG images possessed variable pixel dimensions, with minimum pixel dimensions determined to be 470 $\times$ 470 pixels. Immediately prior to uploading to ChatGPT-4o for analysis, a comprehensive orientation verification was performed by radiologists. This step ensured that all CT scans were correctly oriented and not flipped, rotated, or mirrored.

2.5. Prompt Selection and Testing

Prior to finalizing the prompt for ChatGPT-4o, we conducted various prompt iterations. The objective was to consistently obtain detailed and descriptive responses from ChatGPT-4o that could be valuable for diagnostic evaluation. Ultimately, the following prompt consisting of nine comprehensive questions was observed to elicit complete responses from ChatGPT-4o.

2.6. ChatGPT Interaction and Prompting

Following the preliminary testing phase, the prompt consisting of nine questions that we established and uploaded to ChatGPT-4o was as follows:

1. What is the name of this radiological imaging method? Are there any hemorrhages in these images? Please answer "Yes" or "No".
2. What type(s) of hemorrhage(s) are present? Please answer this question with "subdural", "epidural", "subarachnoid", "intraparenchymal", or "intraventricular". If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
3. If the type of hemorrhage(s) is "subdural" or "epidural", what is the location of the hemorrhage(s)? Please specify the location with "on the patient's right" or "on the patient's left" followed by "frontal", "parietal", "temporal", "occipital", "frontoparietal", "frontotemporal", "temporoparietal", "temporooccipital", "posterior fossa", etc. If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
4. If the type of hemorrhage(s) is "intraparenchymal", what is the location of the hemorrhage(s)? Please specify the location with "on the patient's right" or "on the patient's left" followed by "frontal", "parietal", "temporal", "occipital", "frontoparietal", "frontotemporal", "temporoparietal", "temporooccipital", "cerebellar", "thalamus", "pons", "mesencephalon", etc.
5. If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
6. If the type of hemorrhage(s) is "subarachnoid", what is the location of the hemorrhage(s) in relation to the patient? Please specify the location with "on the patient's right" or "on the patient's left" followed by "frontal", "parietal", "temporal", "occipital", "frontoparietal", "frontotemporal", "temporoparietal", "temporooccipital", "cerebellar". If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
7. If the type of hemorrhage(s) is "intraventricular", what is the location of the hemorrhage(s) in relation to the patient? Please specify the location as "3rd ventricle", "4th ventricle", "right lateral ventricle", or "left lateral ventricle". If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
8. What is the phase of the hemorrhage(s)? Please answer this question with "acute", "subacute", or "chronic". If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
9. Are there any additional pathological findings related to the hemorrhage(s) in these images? If so, please specify the pathology/ies as "right shift", "left shift", "brain edema", "3rd ventricle compression", "right lateral ventricle compression", "left lateral ventricle compression", etc.

2.7. Executors and Readers

Both radiologists, with 8 and 10 years of experience, respectively, actively perform brain CT imaging reporting in routine clinical practice. To ensure standardization in the evaluation process, preliminary assessments were conducted on sample cases not included in the study, during which consensus on evaluation criteria was established between radiologists. The criteria for response categorization were established as follows: a "correct" response required complete concordance between ChatGPT-4o's answers and the radiologists' reports; a "partially correct" response indicated agreement on main findings but with some minor detail deficiencies; and an "incorrect" response was defined by discordance in main findings or the presence of diagnostically critical errors. During the evaluation process, radiologists conducted the study using medical monitors at diagnostic imaging workstations.

2.8. Statistical Analysis

The Statistical Package for the Social Sciences (SPSS) for Windows version 23 software (IBM SPSS Inc., Chicago, IL, USA) was used in the statistical analysis. The Kolmogorov–Smirnov test was utilized to assess normality distribution. The Mann–Whitney U test was employed to evaluate age differences between groups, while gender differences were assessed using the Chi-square test. Numerical variables with normal distribution were presented as mean ± standard deviation (SD), and categorical variables were expressed as numbers (n) and percentages (%). Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and diagnostic accuracy rate were calculated from ChatGPT-4o's responses. Agreement between ChatGPT-4o's responses given at two-week intervals (intra-rater correlation) was compared using Cohen's Kappa coefficient. The obtained Kappa coefficients were interpreted as follows: $\leq$0.20 indicated "very poor agreement", 0.21–0.40 "poor agreement", 0.41–0.60 "moderate agreement", 0.61–0.80 "good agreement", and 0.81–1.00 "very good agreement". A p value < 0.05 was considered statistically significant.

3. Results

A total of 240 cases were included in the study, comprising 120 cases in the ICH group and 120 cases in the healthy control group. The mean age of the 240 cases was 63.8 ± 19.6 years. No significant differences were observed between the hemorrhage group and healthy control group in terms of age and gender ($p > 0.05$) (Table 2).

Table 2. Demographic data of ICH group and healthy control group cases.

	ICH Group (n = 120)	Healthy Control Group (n = 120)	p Value
Gender *, n (%)			
Female	39 (32.5)	40 (33.3)	0.891
Male	81 (67.5)	80 (66.7)	
Age **, years, Mean ± SD	63.95 ± 19.58	63.73 ± 18.81	0.877

* Ki-kare Test; ** Mann–Whitney U test; SD: standard deviation; ICH: intracranial hemorrhage.

A total of 150 hemorrhagic areas were present in 120 cases in the ICH group. Among hemorrhage types, subdural hemorrhage was the most frequent with 78 (52%) cases, followed by intraparenchymal hemorrhage with 35 (23.3%) cases, while intraventricular hemorrhage was the least common with 7 (4.7%) cases (Table 3). In the ICH group, hemorrhage was unifocal in 92 (76.7%) cases and multifocal in 28 (23.3%) cases. While 57 (47.5%) cases in the ICH group had additional pathology associated with hemorrhage, 63 (52.5%)

cases had no hemorrhage-related additional pathology. Some cases with hemorrhage-related additional pathology presented with single pathology, while others had multiple pathologies. A total of 85 additional pathologies were identified in 120 cases in the ICH group, and these pathologies, in order of frequency, were as follows: cerebral edema in 44 (51.8%) cases, midline shift in 23 (27%) cases, left lateral ventricle compression in 10 (11.8%) cases, and right lateral ventricle compression in 8 (9.4%) cases (Table 3).

Table 3. Distribution of hemorrhage types, locations, and associated pathologies in ICH group.

	***n* (%)**
Hemorrhage type (hemorrhagic areas, *n* = 150)	
Subdural	78 (52)
Epidural	12 (8)
Intraventricular	7 (4.7)
Intraparenchymal	35 (23.3)
Subarachnoid	18 (12)
Hemorrhage location (cases, *n* = 120)	
Unifocal	92 (76.7)
Multifocal	28 (23.3)
Associated pathologies (*n* = 85)	
Cerebral edema	44 (51.8)
Midline shift	23 (27)
Right lateral ventricle compression	8 (9.4)
Left lateral ventricle compression	10 (11.8)

Of the 150 hemorrhages in the ICH group, 105 (70%) were acute, 13 (8.7%) were subacute, and 27 (18%) were chronic. Additionally, three (2%) cases presented with mixed-type hemorrhage showing both acute and chronic characteristics, while two (1.3%) cases demonstrated mixed-type hemorrhage with both acute and subacute features. All mixed-type hemorrhages were subdural in nature. When hemorrhage stages were evaluated according to hemorrhage types, acute hemorrhages constituted the majority of cases across all hemorrhage types (Table 4).

Table 4. Distribution of hemorrhage stages according to hemorrhage types in ICH group.

Hemorrhage Stage	**Subdural** ***n* (%)**	**Epidural** ***n* (%)**	**Intraventricular** ***n* (%)**	**Intraparenchymal** ***n* (%)**	**Subarachnoid** ***n* (%)**	**Total** ***n* (%)**
Acute	38 (25.3)	10 (6.7)	7 (4.7)	32 (21.3)	18 (12)	105 (70)
Subacute	11 (7.3)	1 (0.7)	0 (0)	1 (0.7)	0 (0)	13 (8.7)
Chronic	24 (16)	1 (0.7)	0 (0)	2 (1.3)	0 (0)	27 (18)
Acute–subacute	2 (1.3)	0 (0)	0 (0)	0 (0)	0 (0)	2 (1.3)
Acute–chronic	3 (2)	0 (0)	0 (0)	0 (0)	0 (0)	3 (2)
Total	78 (52)	12 (8)	7 (4.7)	35 (23.3)	18 (12)	150 (100)

In the first round, ChatGPT-4o correctly identified the imaging modality as NCCT for all cases (100%) in both ICH and control groups. Regarding the second question about the presence of hemorrhage, ChatGPT-4o correctly identified 95 (79.2%) of the 120 cases with hemorrhage, while producing false-positive results in 51 (42.5%) of the 120 cases without hemorrhage. For hemorrhage detection, ChatGPT-4o demonstrated a sensitivity of 79.2% and a specificity of 57.5%. The PPV was calculated as 65.1%, while the NPV was 73.4% (Table 5). When the same question regarding hemorrhage presence was posed to ChatGPT-4o after a two-week interval (second round), higher values were obtained for all parameters including sensitivity, specificity, PPV, NPV, and diagnostic accuracy (Table 5).

Table 5. First and second round results of ChatGPT-4o in evaluation of hemorrhage presence.

	True Positive, *n* (%)	False Positive, *n* (%)	True Negative, *n* (%)	False Negative, *n* (%)	Sensitivity, %	Specificity, %	PPV, %	NPV, %	Diagnostic Accuracy, %
ChatGPT-4o, Round 1	95 (79.2)	51 (42.5)	69 (57.5)	25 (20.8)	79.2	57.5	65.1	73.4	68.3
ChatGPT-4o, Round 2	104 (86.7)	48 (40)	72 (60)	16 (13.3)	86.7	60.0	68.4	81.8	73.3

PPV: positive predictive value; NPV: negative predictive value.

In comparing the first and second round results of ChatGPT-4o regarding hemorrhage presence, inconsistent responses were observed in 23 (19.2%) cases in the hemorrhage group and 37 (30.8%) cases in the healthy control group. There was 75% consistency between ChatGPT-4o's first and second round results across the total study population. Analysis conducted to assess intra-rater reliability revealed moderate agreement with a Cohen's Kappa value of 0.469 (Table 6).

Table 6. Comparison of first and second round ChatGPT-4o results in evaluation of hemorrhage presence.

Hemorrhage Presence Assessment							
ICH Group (*n* = 120)				Healthy Control Group (*n* = 120)			
		ChatGPT-4o, Round 1				ChatGPT-4o, Round 1	
		Negative	Positive			Negative	Positive
ChatGPT-4o, Round 2	Negative	9	7	ChatGPT-4o, Round 2	Negative	52	20
	Positive	16	88		Positive	17	31

Cohen's Kappa value: 0.469.

In identifying hemorrhage types, ChatGPT-4o correctly identified 51 (34%) of 150 hemorrhagic areas, partially correctly identified 7 (4.7%), and incorrectly identified 64 (42.6%). The partially correct category was used for cases where ChatGPT-4o responded "subdural + intraparenchymal hemorrhage" in cases with only subdural hemorrhage. The remaining 28 (18.7%) hemorrhages belonged to 25 false-negative cases where ChatGPT-4o indicated no hemorrhage was present. Among hemorrhage types, ChatGPT-4o demonstrated the highest accuracy rate for intraparenchymal hemorrhage (25/35, 71.4%). Conversely, subarachnoid hemorrhage showed the lowest accuracy rate (1/18, 5.5%) (Table 7).

Table 7. Results of ChatGPT-4o in detection of hemorrhage types.

Hemorrhage Type (*n* = 150)	Correct, *n* (%)	Partially Correct, *n* (%)	Incorrect, *n* (%)	False Negative Cases, *n* (%)	Total, *n* (%)
Subdural	21 (14)	7 (4.7)	26 (17.3)	24 (16)	78 (52)
Epidural	1 (0.7)	0 (0)	10 (6.6)	1 (0.7)	12 (8)
Intraventricular	3 (2)	0 (0)	4 (2.7)	0 (0)	7 (4.7)
Intraparenchymal	25 (16.6)	0 (0)	9 (6)	1 (0.7)	35 (23.3)
Subarachnoid	1 (0.7)	0 (0)	15 (10)	2 (1.3)	18 (12)
Total	51 (34)	7 (4.7)	64 (42.6)	28 (18.7)	150 (100)

Regarding hemorrhage stages, ChatGPT-4o responded "acute" for all 95 cases (100%) where it detected hemorrhage, regardless of whether the hemorrhage was unifocal or multifocal, and did not provide detailed specifications about which hemorrhage was designated as acute in multifocal cases. Consequently, detailed statistical data regarding hemorrhage stages according to hemorrhage locations could not be obtained.

ChatGPT-4o successfully identified only 11 (7.3%) of 150 hemorrhage locations with complete accuracy. The remaining 139 (92.7%) hemorrhage locations were not correctly identified.

Regarding hemorrhage-associated pathologies, ChatGPT-4o correctly identified 11 (47.8%) of 23 midline shifts and 29 (65.9%) of 44 cases of cerebral edema in the ICH group. However, it failed to detect both right and left lateral ventricle compressions (Table 8). Additionally, it incorrectly reported third ventricle compression in seven cases and hydrocephalus in two cases where these findings were not present.

Table 8. ChatGPT-4o results regarding pathologies associated with hemorrhage.

Associated Pathologies	True Positive, *n* (%)	False Negative, *n* (%)
Midline shift, *n* = 23	11 (47.8)	12 (52.2)
Cerebral edema, *n* = 44	29 (65.9)	15 (34.1)
Right lateral ventricle compression, *n* = 8	0 (0)	8 (100)
Left lateral ventricle compression, *n* = 10	0 (0)	10 (100)

4. Discussion

In this study, we evaluated the ability of ChatGPT-4o to detect hemorrhages, classify their types, stages, and localizations, as well as identify the presence of any pathologies accompanying the hemorrhages on brain NCCT images. To the best of our knowledge, this is the first study in the literature to assess intracranial hemorrhages using ChatGPT-4o.

For the question regarding the type of imaging modality, ChatGPT-4o provided correct answers in all cases without any errors. Since only brain NCCT images were used in this study, it was not possible to test whether ChatGPT-4o could identify other imaging modalities. The primary finding of our study reveals that ChatGPT-4o demonstrated notable diagnostic performance in detecting intracranial hemorrhage (ICH) on brain NCCT images. The model exhibited significant diagnostic efficacy in evaluating the presence of ICH, with a sensitivity of 79.2% and a specificity of 57.5%. However, the false positive rate of 42.5% indicates a considerable tendency of the model to generate incorrect diagnoses. In the study conducted by Daiwen et al. using the previous version of ChatGPT (ChatGPT-4), it was shown that the model could identify ICHs at a rate of 72.6%, which shows considerable similarity to our rate [29]. A review of the literature shows that deep learning-based AI models have achieved high diagnostic performance in the detection of ICH. For instance, one previous study reported that a deep learning model reached a sensitivity of 95.89% and a specificity of 95.33% for intracranial hemorrhage detection [16]. Similarly, another deep learning-based study reported sensitivity and specificity rates of 78% and 80%, respectively, for the detection of intracranial hemorrhage [30]. Additionally, FDA-cleared and CE-marked triage and notification software (AIDOC) has been reported to detect intracranial hemorrhage (ICH) with a sensitivity of 89–95% and a specificity of 94–99% [3,33]. In the deep learning-based study by Heit et al., a sensitivity of 95.6% and a specificity of 95.3% were obtained in ICH detection [20]. Compared to these deep learning-based studies, ChatGPT-4o achieved relatively higher sensitivity and specificity in hemorrhage detection, despite not having been specifically trained for this purpose, thus highlighting its potential if appropriately trained in the future. Indeed, an improvement in the diagnostic parameters of the model was observed in the second-round responses compared to the first-round ones. This improvement suggests that the model may demonstrate adaptability in its learning and evaluation processes. However, the model's tendency to repeat phrases from previous interactions could also be a contributing factor to this observed improvement [34].

The analysis conducted to evaluate intra-rater reliability revealed that ChatGPT-4o demonstrated a moderate level of consistency between the two rounds in assessing the presence of hemorrhage ($\kappa = 0.469$). This result is a significant indicator for questioning the model's stability, as it highlights inconsistencies in its evaluations conducted at different

time intervals. Consequently, it underscores that ChatGPT is not yet a fully reliable tool for determining the presence of hemorrhage.

In the evaluation of more specific clinical parameters such as the type, localization, and stage of hemorrhage, notable variations were observed in the performance of ChatGPT-4o. The low accuracy rate in determining hemorrhage localization is particularly striking. Hemorrhage localization is critical information, especially for surgical planning, and inaccuracies in this parameter could lead to life-threatening consequences. While high accuracy rates were achieved in assessing the stage of hemorrhage, it was observed that ChatGPT-4o consistently classified all cases as acute, thereby neglecting subacute and chronic hemorrhages. In the study by Daiwen et al., a low identification rate (37.3%) was also reported for chronic subdural hemorrhages [29]. Acute hemorrhages are relatively easier to identify as they appear hyperdense compared to the brain parenchyma. However, the isodense nature of subacute hemorrhages with the brain parenchyma and chronic hemorrhages with cerebrospinal fluid likely contributed to ChatGPT-4o's inability to recognize these stages of hemorrhage. This suggests that, while the model is more adept at understanding general concepts and processes, it still lacks the capability to accurately interpret more complex and specific medical information.

Regarding hemorrhage types, ChatGPT-4o was generally able to identify subdural, intraparenchymal, and intraventricular hemorrhages. However, it struggled to detect subarachnoid and epidural hemorrhages. The study by Daiwen et al. demonstrates that ChatGPT showed similarly high success rates in detecting intraparenchymal (86.9%) and acute subdural hemorrhages (74.4%), as in our study. However, while in our study ChatGPT's detection rates for epidural and subarachnoid hemorrhages were quite low (1/12, 8% and 1/18, 5.5%, respectively), their study found notably higher rates (89% and 71.5%, respectively) [29]. Similar to our study, in a deep learning-based study, intraparenchymal hemorrhages were identified with higher accuracy, whereas epidural hemorrhages could not be detected as effectively [3]. In another deep learning-based study, subarachnoid hemorrhages were similarly detected with lower accuracy, as observed in our study [35]. In light of these findings, we conclude that intraparenchymal hemorrhages are more easily recognized due to the contrast difference they create with brain tissue. However, the challenges in detecting subarachnoid hemorrhages are suggested to be associated with beam-hardening artifacts and partial-volume artifacts [35]. We believe that the failure of ChatGPT-4o in detecting epidural and subarachnoid hemorrhages could serve as a separate subject for further research.

Machine learning and deep learning AI applications have achieved considerable success in diagnosing from medical images. In the meta-analysis by Jung et al., the overall pooled sensitivity and specificity for bone fractures using these AI applications were found to be 91% (95% CI: 88%, 93%) and 90% (95% CI: 88%, 92%), respectively [36]. Furthermore, these AI applications can perform quantitative analysis in the diagnosis of interstitial lung diseases [37], detect brain hemorrhages with over 90% accuracy [33,35], and identify brain metastases from magnetic resonance images with up to 89% accuracy [38], and have found many clinical applications in medical imaging beyond these, demonstrating remarkable success in these areas [39,40].

However, deep learning-based AI models used in pathology detection from radiological images are purpose-specific models. They are optimized only for a particular clinical problem or imaging type. In contrast, ChatGPT-4o, as a general-purpose artificial intelligence model, has the potential to perform versatile tasks.

While developing and implementing purpose-specific AI models requires high costs and advanced technical infrastructure, general-purpose models like ChatGPT-4o offer a more accessible and cost-effective alternative.

There are dozens of purpose-specific AI models developed for ICH detection in the literature, and this diversity may cause users to experience difficulty in selecting the right model [16,19,35]. Therefore, we think that popular and general-purpose AI models like ChatGPT-4o have a higher potential to be preferred by users.

Our study shows, in light of the data we obtained, that ChatGPT-4o still has some significant limitations for clinical use. The model cannot demonstrate sufficient sensitivity and precision in detecting certain hemorrhage types like subarachnoid hemorrhage and certain hemorrhage stages like chronic phase. Another limitation of the model is its lack of ability to evaluate sequential image series while analyzing CT images. Additionally, the model requires an appropriate prompt before use. If the given prompt is not suitable, the model's diagnostic performance may be affected. Finally, the use of ChatGPT-4o in ICH detection may bring ethical and legal debates that might not be universally accepted.

This study contributes to the theoretical knowledge base in AI and radiology by demonstrating the potential of LLMs like ChatGPT in medical image analysis. ChatGPT-4o's success in detecting ICH from NCCT images shows significant potential for the future evaluation of radiological images and the reduction in radiologists' workload. Particularly in healthcare centers experiencing radiologist shortages, personnel with limited clinical experience can benefit from LLMs like ChatGPT in the initial assessment of brain NCCT images. Young radiologists in the process of gaining experience can use this technology as a supportive tool to avoid missing important findings. In this way, workflow can become more efficient and patient care quality may improve. The results of our study will guide the development of similar AI systems and their integration into clinical practice.

Our study has certain limitations. The first is that, due to ChatGPT-4o's limited capacity to process images, the model was unable to evaluate all brain NCCT images from each case. If the complete set of brain NCCT images were uploaded, the diagnostic accuracy might improve. Another limitation is that the original DICOM images had to be converted to JPEG format for analysis by ChatGPT. While this conversion was a necessary step for uploading the images to the model, we believe it might have compromised image quality, potentially affecting the model's performance and leading to misinterpretations [27]. Additionally, the selection process of brain CT images, which were determined by radiologists and subjected to specific exclusion criteria, may have introduced selection bias. Furthermore, excluding cases with conditions that could mimic hemorrhage on brain NCCT images, such as masses, metallic clips, and foreign bodies, can be considered another limitation of the study [41]. Excluding these cases from the study may have impacted the model's sensitivity and specificity values. The size of the hemorrhage is crucial for treatment planning. However, as the model lacks a tool for measuring hemorrhage size, no data could be obtained in this regard.

5. Conclusions

In conclusion, ChatGPT-4o has been evaluated as a promising auxiliary tool for detecting ICH on brain NCCT images. Considering the current limitations of the model, it should be used cautiously in clinical decision-making processes. Continuous training and development through future studies are essential to improve the model's accuracy, sensitivity, and specificity, as well as to minimize inconsistencies. In clinical practice, the results generated by ChatGPT-4o should be assessed in collaboration with radiologists, and care must be taken in making the final decision.

Author Contributions: Conceptualization, M.K.; methodology, M.K. and Z.K.C.; validation, Z.K.C. and B.R.; formal analysis, B.E.; investigation, M.K.; resources, M.K. and Z.K.C.; data curation, B.E. and B.R.; writing—original draft preparation, M.K. and B.E.; writing—review and editing, Z.K.C.

and B.R.; visualization, B.E.; supervision, B.E.; project administration, M.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 was approved by the Non-Interventional Clinical Research Ethics Committee of Erzincan Binali Yildirim University, Türkiye (decision no: 2024-09/05; date: 11 July 2024).

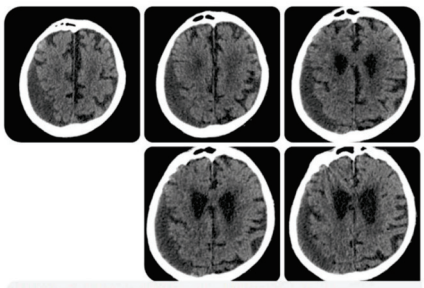
Informed Consent Statement: Patient consent was waived due to the retrospective nature of the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

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

Appendix A

Appendix A.1



1. What is the name of this radiological imaging method?
Are there any hemorrhages in these images? Please answer "Yes" or "No."
2. What type(s) of hemorrhage(s) are present? Please answer this question with "subdural," "epidural," "subarachnoid," "intraparenchymal," or "intraventricular." If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
3. If the type of hemorrhage(s) is "subdural" or "epidural," what is the location of the hemorrhage(s)? Please specify the location with "on the patient's right" or "on the patient's left" followed by "frontal," "parietal," "temporal," "occipital," "frontoparietal," "frontotemporal," "temporoparietal," "temporooccipital," "posterior fossa," etc. If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
4. If the type of hemorrhage(s) is "intraparenchymal," what is the location of the hemorrhage(s)? P. lease specify the location with "on the patient's right" or "on the patient's left" followed by "frontal," "parietal," "temporal," "occipital," "frontoparietal," "frontotemporal," "temporoparietal," "temporooccipital," "cerebellar," "thalamus," "pons," "mesencephalon," etc. 5. If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
6. If the type of hemorrhage(s) is "subarachnoid," what is the location of the hemorrhage(s) in relation to the patient? Please specify the location with "on the patient's right" or "on the patient's left" followed by "frontal," "parietal," "temporal," "occipital," "frontoparietal," "frontotemporal," "temporoparietal," "temporooccipital," "cerebellar." If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
7. If the type of hemorrhage(s) is "intraventricular," what is the location of the hemorrhage(s) in relation to the patient? Please specify the location as "3rd ventricle," "4th ventricle," "right lateral ventricle," or "left lateral ventricle." If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
8. What is the phase of the hemorrhage(s)? Please answer this question with "acute," "subacute," or "chronic." If there are multiple hemorrhages in these images, please answer this question separately for each hemorrhage.
9. Are there any additional pathological findings related to the hemorrhage(s) in these images? If so, please specify the pathology/ies as "right shift," "left shift," "brain edema," "3rd ventricle compression," "right lateral ventricle compression," "left lateral ventricle compression," etc.

What is the name of this radiological imaging method?

This radiological imaging method is a **CT scan** (Computed Tomography scan).

Are there any hemorrhages in these images?

No, there are no visible hemorrhages in these images.

What type(s) of hemorrhage(s) are present?

Since there are no visible hemorrhages in these images, this question is not applicable.

If the type of hemorrhage(s) is "subdural" or "epidural," what is the location of the hemorrhage(s)?

Not applicable.

If the type of hemorrhage(s) is "intraparenchymal," what is the location of the hemorrhage(s)?

Not applicable.

If the type of hemorrhage(s) is "subarachnoid," what is the location of the hemorrhage(s) in relation to the patient?

Not applicable.

If the type of hemorrhage(s) is "intraventricular," what is the location of the hemorrhage(s) in relation to the patient?

Not applicable.

What is the phase of the hemorrhage(s)?

Not applicable.

Are there any additional pathological findings related to the hemorrhage(s) in these images?

There are no additional pathological findings related to hemorrhage in these images.

Figure A1. The evaluation of a case with right parietal chronic subdural hematoma by ChatGPT-4o.

Appendix A.2

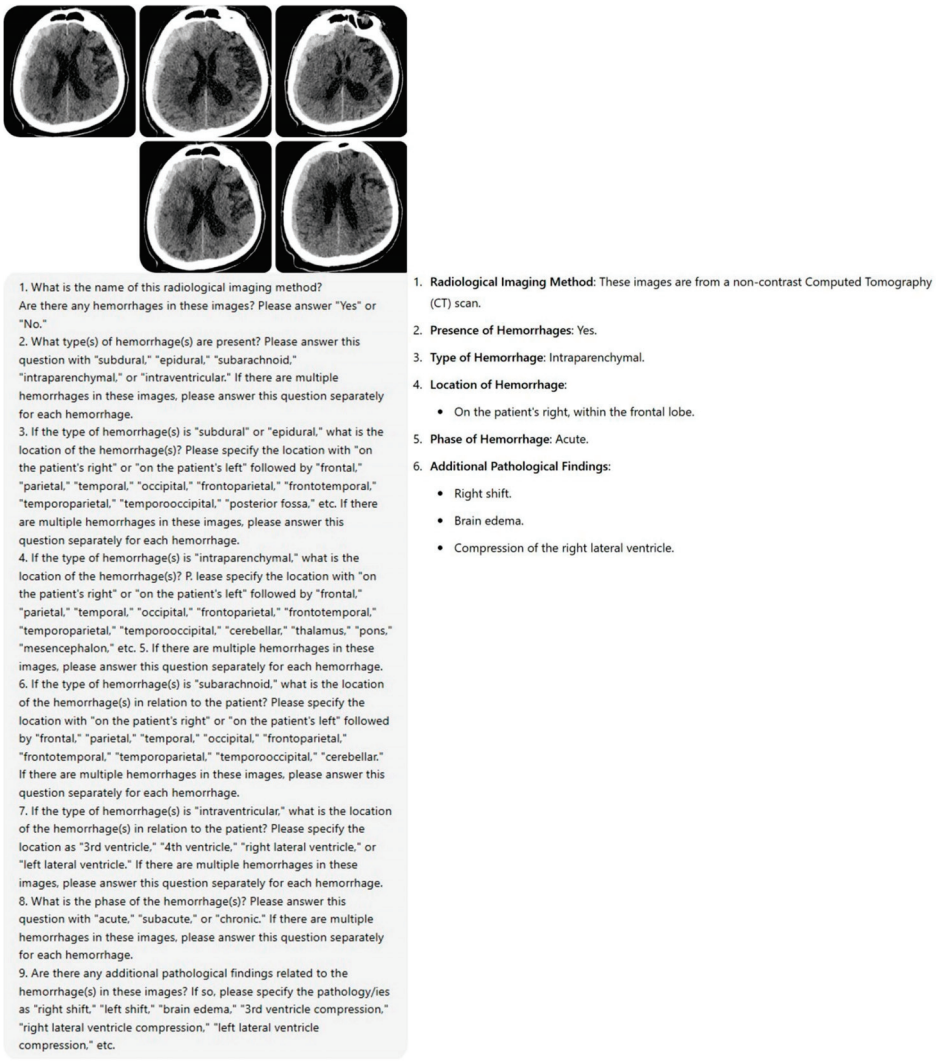


Figure A2. The evaluation of a case with right frontoparietal acute subdural hematoma by ChatGPT-4o.

References

1. Van Asch, C.J.; Luitse, M.J.; Rinkel, G.J.; van der Tweel, I.; Algra, A.; Klijn, C.J. Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time, according to age, sex, and ethnic origin: A systematic review and meta-analysis. *Lancet Neurol.* **2010**, *9*, 167–176. [CrossRef] [PubMed]
2. Heit, J.J.; Iv, M.; Wintermark, M. Imaging of intracranial hemorrhage. *J. Stroke* **2016**, *19*, 11. [CrossRef] [PubMed]
3. Ginat, D.T. Analysis of head CT scans flagged by deep learning software for acute intracranial hemorrhage. *Neuroradiology* **2020**, *62*, 335–340. [CrossRef] [PubMed]
4. Kidwell, C.S.; Chalela, J.A.; Saver, J.L.; Starkman, S.; Hill, M.D.; Demchuk, A.M.; Butman, J.A.; Patronas, N.; Alger, J.R.; Latour, L.L.; et al. Comparison of MRI and CT for detection of acute intracerebral hemorrhage. *JAMA* **2004**, *292*, 1823–1830. [CrossRef]
5. Zahuranec, D.B.; Lisabeth, L.D.; Sánchez, B.N.; Smith, M.A.; Brown, D.L.; Garcia, N.M.; Skolarus, L.E.; Meurer, W.J.; Burke, J.F.; Adelman, E.E.; et al. Intracerebral hemorrhage mortality is not changing despite declining incidence. *Neurology* **2014**, *82*, 2180–2186. [CrossRef]
6. Elliott, J.; Smith, M. The acute management of intracerebral hemorrhage: A clinical review. *Anesth. Analg.* **2010**, *110*, 1419–1427. [CrossRef]
7. Fugate, J.E.; Rabinstein, A.A. Absolute and relative contraindications to IV rt-PA for acute ischemic stroke. *Neurohospitalist* **2015**, *5*, 110–121. [CrossRef]
8. Qureshi, A.I.; Mendelow, A.D.; Hanley, D.F. Intracerebral haemorrhage. *Lancet* **2009**, *373*, 1632–1644. [CrossRef]
9. Morotti, A.; Goldstein, J.N. Diagnosis and management of acute intracerebral hemorrhage. *Emerg. Med. Clin. N. Am.* **2016**, *34*, 883–899. [CrossRef]

10. McDonald, R.J.; Schwartz, K.M.; Eckel, L.J.; Diehn, F.E.; Hunt, C.H.; Bartholmai, B.J.; Erickson, B.J.; Kallmes, D.F. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Acad. Radiol.* **2015**, *22*, 1191–1198. [CrossRef]
11. Sellers, A.; Hillman, B.J.; Wintermark, M. Survey of after-hours coverage of emergency department imaging studies by US academic radiology departments. *J. Am. Coll. Radiol.* **2014**, *11*, 725–730. [CrossRef] [PubMed]
12. Spitler, K.; Vijayasarathi, A.; Salehi, B.; Dua, S.; Azizyan, A.; Cekic, M.; Yaghmai, N.; Homer, R.; Salamonet, N. Neuroradiologist coverage improves resident perception of educational experience, referring physician satisfaction, and turnaround time. *Curr. Probl. Diagn. Radiol.* **2020**, *49*, 168–172. [CrossRef] [PubMed]
13. 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]
14. Almeida, L.C.; Farina, E.M.J.M.; Kuriki, P.E.A.; Abdala, N.; Kitamura, F.C. Performance of ChatGPT on the Brazilian radiology and diagnostic imaging and mammography board examinations. *Radiol. Artif. Intell.* **2024**, *6*, e230103. [CrossRef]
15. Lee, J.Y.; Kim, J.S.; Kim, T.Y.; Kim, Y.S. Detection and classification of intracranial haemorrhage on CT images using a novel deep-learning algorithm. *Sci. Rep.* **2020**, *10*, 20546. [CrossRef]
16. Yun, T.J.; Choi, J.W.; Han, M.; Jung, W.S.; Choi, S.H.; Yoo, R.E.; Hwang, I.P. Deep learning based automatic detection algorithm for acute intracranial haemorrhage: A pivotal randomized clinical trial. *NPJ Digit. Med.* **2023**, *6*, 61. [CrossRef]
17. Dawud, A.M.; Yurtkan, K.; Oztoprak, H. Application of deep learning in neuroradiology: Brain haemorrhage classification using transfer learning. *Comput. Intell. Neurosci.* **2019**, *2019*, 4629859. [CrossRef]
18. Kaluarachchi, T.; Reis, A.; Nanayakkara, S. A review of recent deep learning approaches in human-centered machine learning. *Sensors* **2021**, *21*, 2514. [CrossRef]
19. Lewick, T.; Kumar, M.; Hong, R.; Wu, W. Intracranial hemorrhage detection in ct scans using deep learning. In Proceedings of the 2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService), Oxford, UK, 3–6 August 2020; Volume 2020, pp. 169–172.
20. Heit, J.J.; Coelho, H.; Lima, F.O.; Granja, M.; Aghaebrahim, A.; Hanel, R.; Kwok, K.; Haerian, H.; Cereda, C.W.; Venkatasubramanian, C.; et al. Automated cerebral hemorrhage detection using RAPID. *Am. J. Neuroradiol.* **2021**, *42*, 273–278. [CrossRef]
21. Kuo, W.; Häne, C.; Mukherjee, P.; Malik, J.; Yuh, E.L. Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 22737–22745. [CrossRef]
22. Voter, A.F.; Meram, E.; Garrett, J.W.; Yu, J.P.J. Diagnostic accuracy and failure mode analysis of a deep learning algorithm for the detection of intracranial hemorrhage. *J. Am. Coll. Radiol.* **2021**, *18*, 1143–1152. [CrossRef] [PubMed]
23. Rahsepar, A.A.; Tavakoli, N.; Kim, G.H.J.; Hassani, C.; Abtin, F.; Bedayat, A. How AI responds to common lung cancer questions: ChatGPT versus Google Bard. *Radiology* **2023**, *307*, e230922. [CrossRef] [PubMed]
24. Temperley, H.C.; O'Sullivan, N.J.; Mac Curtain, B.M.; Corr, A.; Meaney, J.F.; Kelly, M.E.; Brennan, I. Current applications and future potential of ChatGPT in radiology: A systematic review. *J. Med. Imaging Radiat. Oncol.* **2024**, *68*, 257–264. [CrossRef]
25. Haver, H.L.; Bahl, M.; Doo, F.X.; Kamel, P.I.; Parekh, V.S.; Jeudy, J.; Yi, P.H. Evaluation of multimodal ChatGPT (GPT-4V) in describing mammography image features. *Can. Assoc. Radiol. J.* **2024**, *75*, 947–949. [CrossRef]
26. Mert, S.; Stoerzer, P.; Brauer, J.; Fuchs, B.; Haas-Lützenberger, E.M.; Demmer, W.; Giunta, R.E.; Nuernberger, T. Diagnostic power of ChatGPT 4 in distal radius fracture detection through wrist radiographs. *Arch. Orthop. Trauma. Surg.* **2024**, *144*, 2461–2467. [CrossRef]
27. Dehdab, R.; Brendlin, A.; Werner, S.; Almansour, H.; Gassenmaier, S.; Brendel, J.M.; Nikolaou, K.; Afat, S. Evaluating ChatGPT-4V in chest CT diagnostics: A critical image interpretation assessment. *Jpn. J. Radiol.* **2024**, *42*, 1168–1177. [CrossRef]
28. Kuzan, B.N.; Meşe, İ.; Yaşar, S.; Kuzan, T.Y. A retrospective evaluation of the potential of ChatGPT in the accurate diagnosis of acute stroke. *Diagn. Interv. Radiol.* **2024**. *online ahead of print*. [CrossRef]
29. Zhang, D.; Ma, Z.; Gong, R.; Lian, L.; Li, Y.; He, Z.; Han, Y.; Hui, J.; Huang, J.; Jiang, J.; et al. Using natural language processing (GPT-4) for computed tomography image analysis of cerebral hemorrhages in radiology: Retrospective analysis. *J. Med. Internet Res.* **2024**, *26*, e58741. [CrossRef]
30. Arbabshirani, M.R.; Fornwalt, B.K.; Mongelluzzo, G.J.; Suever, J.D.; Geise, B.D.; Patel, A.A.; Moore, G.J. Advanced machine learning in action: Identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *NPJ Digit. Med.* **2018**, *1*, 9. [CrossRef]
31. Mongan, J.; Moy, L.; Kahn, C.E. Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and Reviewers. *Radiol. Artif. Intell.* **2020**, *2*, e200029. [CrossRef]
32. ChatGPT. Available online: https://chatgpt.com/ (accessed on 22 August 2024).
33. Ojeda, P.; Zawaideh, M.; Mossa-Basha, M.; Haynor, D. The utility of deep learning: Evaluation of a convolutional neural network for detection of intracranial bleeds on non-contrast head computed tomography studies. *Proc. SPIE* **2019**, *10949*, 899–906. [CrossRef]

34. Dave, T.; Athaluri, S.A.; Singh, S. ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front. Artif. Intell.* **2023**, *6*, 1169595. [CrossRef] [PubMed]
35. Kundisch, A.; Hönning, A.; Mutze, S.; Kreissl, L.; Spohn, F.; Lemcke, J.; Sitz, M.; Sparenberg, P.; Goelz, L. Deep learning algorithm in detecting intracranial hemorrhages on emergency computed tomographies. *PLoS ONE* **2021**, *16*, e0260560. [CrossRef] [PubMed]
36. Jung, J.; Dai, J.; Liu, B.; Wu, Q. Artificial intelligence in fracture detection with different image modalities and data types: A systematic review and meta-analysis. *PLoS Digit. Health* **2024**, *3*, e0000438. [CrossRef] [PubMed]
37. Rea, G.; Sverzellati, N.; Bocchino, M.; Lieto, R.; Milanese, G.; D'Alto, M.; Bocchini, G.; Maniscalco, M.; Valente, T.; Sica, G. Beyond visual interpretation: Quantitative analysis and artificial intelligence in interstitial lung disease diagnosis "expanding horizons in radiology". *Diagnostics* **2023**, *13*, 2333. [CrossRef]
38. Ozkara, B.B.; Chen, M.M.; Federau, C.; Karabacak, M.; Briere, T.M.; Li, J.; Wintermark, M. Deep learning for detecting brain metastases on MRI: A systematic review and meta-analysis. *Cancers* **2023**, *15*, 334. [CrossRef]
39. Obuchowicz, R.; Strzelecki, M.; Piórkowski, A. Clinical applications of artificial intelligence in medical imaging and image processing—A review. *Cancers* **2024**, *16*, 1870. [CrossRef]
40. Fathi, M.; Eshraghi, R.; Behzad, S.; Tavasol, A.; Bahrami, A.; Tafazolimoghadam, A.; Bhatt, V.; Ghadimi, D.; Gholamrezanezhad, A. Potential strength and weakness of artificial intelligence integration in emergency radiology: A review of diagnostic utilizations and applications in patient care optimization. *Emerg. Radiol.* **2024**, *31*, 887–901. [CrossRef]
41. Morales, H. Pitfalls in the imaging interpretation of intracranial hemorrhage. *Semin. Ultrasound CT MR* **2018**, *39*, 457–468. [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.

cancers

MDPI

Article

XGBoost Enhances the Performance of SAFE: A Novel Microwave Imaging System for Early Detection of Malignant Breast Cancer

Ali Yurtseven [1,2,*], Aleksandar Janjic [1], Mehmet Cayoren [1,2], Onur Bugdayci [3], Mustafa Erkin Aribal [1,4] and Ibrahim Akduman [1,2]

[1] Mitos Medical Technologies, ITU Ayazaga Ari 1, Maslak, 34469 Istanbul, Turkey; aleksandarjanjic@mitosmedikal.com (A.J.); mehmet.cayoren@gmail.com (M.C.); erkin.aribal@acibadem.com (M.E.A.); akduman@itu.edu.tr (I.A.)
[2] Electrical and Electronics Engineering Faculty, Istanbul Technical University, Maslak, 34469 Istanbul, Turkey
[3] Department of Radiology, School of Medicine, Marmara University, Pendik, 34899 Istanbul, Turkey; onurbug@hotmail.com
[4] Radiology Department, Breast Health Center, Altunizade Hospital, Acibadem M.A.A. University, Atasehir, 34684 Istanbul, Turkey
* Correspondence: yurtseven23@itu.edu.tr

Simple Summary: Breast cancer is one of the most common cancers affecting women worldwide, and early detection is key to improving survival rates. However, current screening methods, like mammography, face limitations, particularly in younger women and those with dense breast tissue, where sensitivity is reduced, and there is a risk of false positives and radiation exposure. To address these challenges, we developed SAFE (Scan and Find Early), a microwave imaging device that uses non-ionizing electromagnetic waves to detect breast cancer safely and effectively. In this study, we evaluated SAFE on a large patient cohort, analyzing its ability to classify cases into malignant (cancerous) and non-cancerous (benign or healthy) categories. SAFE demonstrated high sensitivity, especially in younger patients (83%) and those with dense breasts (91%). It also successfully detected both small (<2 cm) and larger tumors, highlighting its potential for early diagnosis. Unlike traditional methods, SAFE's innovative approach offers a safer, radiation-free alternative, making it particularly suitable for frequent screenings in younger and high-risk populations. Our findings show that SAFE could complement existing tools, improving breast cancer detection and reducing diagnostic uncertainties. With further research, SAFE could play a pivotal role in transforming breast cancer screening and saving lives.

Abstract: Background/Objectives: Breast cancer is a significant global health concern, and early detection is crucial for improving patient outcomes. Mammography is widely used but has limitations, particularly for younger women with denser breasts. These include reduced sensitivity, false positives, and radiation risks. This highlights the need for alternative screening methods. In this study, we assess the performance of SAFE (Scan and Find Early), a novel microwave imaging device, in detecting breast cancer in a larger patient cohort. Unlike previous studies that predominantly relied on cross-validation, this study employs a more reliable, independent evaluation methodology to enhance generalizability. Methods: We developed an XGBoost model to classify breast cancer cases into positive (malignant) and negative (benign or healthy) groups. The model was analyzed with respect to key factors such as breast size, density, age, tumor size, and histopathological findings. This approach provides a better understanding of how these factors influence the model's performance, using an independent evaluation methodology for increased reliability. Results: Our results demonstrate that SAFE exhibits high sensitivity,

particularly in dense breasts (91%) and younger patients (83%), suggesting its potential as a supplemental screening tool. Additionally, the system shows high detection accuracy for both small (<2 cm) and larger lesions, proving effective in early cancer detection. Conclusions: This study reinforces the potential of SAFE to complement existing screening methods, particularly for patients with dense breasts, where mammography's sensitivity is reduced. The promising results warrant further research to solidify SAFE's clinical application as an alternative screening tool for breast cancer detection.

Keywords: SAFE; microwave imaging; breast cancer detection; machine learning; medical imaging; tumor classification

1. Introduction

Breast cancer is the most common form of cancer among women and remains the leading cause of cancer-related deaths in women globally [1]. For a long time, researchers have been developing diagnostic tools and promising treatments for the disease. Studies emphasize the importance of early diagnosis of tumor lesions to improve treatment success rates. Consequently, it is crucial to develop innovative diagnostic technologies that can efficiently detect tumor cells before their progression [2].

While significant advancements in breast cancer treatment strategies have contributed to a reduction in mortality rates, early detection through mammographic screening remains a key tool in improving survival outcomes [3]. However, mammography comes with several challenges, some of which limit its effectiveness, especially in certain populations [4,5].

One limitation of mammography is that it presents unique challenges for younger women, particularly those with a family history of breast cancer. Women under 50, especially those with dense breast tissue, experience reduced sensitivity and specificity in mammography, making it more difficult to detect small tumors [6–9]. The increased density of younger women's breasts makes it harder to distinguish tumors from the surrounding tissue, leading to false negatives [10–16]. Moreover, these younger women often experience a higher rate of false positives, which can lead to unnecessary follow-up tests, biopsies, and emotional distress [17,18]. One study found that over a nine-year period of annual mammograms, the likelihood of a false positive result for women aged 40–69 years was 43%, with the rate increasing to nearly 100% for women with a family history of breast cancer or other risk factors, such as benign breast disease or hormone use. This is particularly concerning, as it can lead to unnecessary treatments and anxiety, sometimes exposing women to unnecessary procedures and psychological burdens [17].

Mammography also carries risks due to the radiation exposure associated with repeated screenings [4,19]. Younger women are more susceptible to radiation-induced cancers due to their increased cell division rates, making them more vulnerable to the carcinogenic effects of ionizing radiation. This is particularly relevant for women under 50 or those with a family history of breast cancer, who may already be genetically predisposed to developing the disease [20]. Studies have shown that the cumulative radiation risk over many years of mammographic screening may outweigh the benefits, especially for those with genetic risks [21]. For women with inherited mutations such as BRCA1 or BRCA2, the risk of radiation-induced breast cancer may be higher, and early screening may not always lead to a significant reduction in mortality [21,22]. Additionally, while mammography may detect cancer at an earlier stage, no conclusive evidence exists to show that early mammography in these younger, high-risk groups leads to a decrease in breast cancer mortality [23].

Another challenge is the issue of overdiagnosis. Mammography can detect both invasive cancers and ductal carcinoma in situ (DCIS), a non-invasive form of cancer. However, some of these cancers may never grow or spread, leading to overdiagnosis and overtreatment [24]. Overdiagnosis occurs when cancers that would not have caused harm are treated unnecessarily, exposing patients to the risks and side effects of treatments like surgery, chemotherapy, or radiation. This issue is particularly relevant in younger women, where the detection of slow-growing, non-aggressive cancers could lead to unnecessary interventions that offer little benefit [25].

Furthermore, while contrast-enhanced mammography (CEM) can improve detection rates in dense breast tissue, it carries its own set of limitations and risks. CEM involves the use of contrast agents to highlight areas of abnormal blood supply, but this method increases radiation exposure compared to conventional mammography. Although it improves the sensitivity of mammograms in detecting tumors in dense breasts, the added radiation dose may be concerning, particularly for younger women and those undergoing multiple imaging studies over time. Additionally, CEM is associated with allergic reactions to the contrast agents, as well as renal toxicity in patients with impaired kidney function. The increased cost, limited accessibility, and potential adverse effects of contrast agents also make CEM less practical for widespread adoption, particularly in resource-limited settings [26–28].

Given these limitations, it is clear that mammography alone may not be sufficient for optimal breast cancer detection, especially for high-risk groups such as younger women, those with dense breast tissue, or women with a family history of breast cancer. This highlights the urgent need for alternative technologies that can overcome these challenges, improve detection accuracy, and reduce the associated risks of radiation, overdiagnosis, and overtreatment. A recent survey showed that some women in their 40s preferred to delay screening with mammography after being informed about the benefits and harms of the procedure [29]. Therefore, it is imperative to explore alternative technologies to mitigate these harms and enhance screening efficacy.

In the pursuit of improved screening technologies, one of the areas researchers have turned their attention to is microwave imaging (MWI). Unlike mammography, which uses ionizing radiation, MWI employs non-ionizing radiation, eliminating the associated risks of radiation exposure. This is especially important for younger women, who are more vulnerable to radiation-induced cancers due to their faster cell division rates. MWI works by measuring the dielectric properties of tissues, which vary significantly between healthy and cancerous tissue [30–34]. Cancerous tissues, characterized by higher water content and unique tissue decomposition and structure, exhibit elevated dielectric permittivity, while healthy tissues show lower permittivity values [35]. While edema also involves increased water content, its dielectric properties are influenced primarily by the distribution of extracellular fluid and lack the cellular irregularities and tissue decomposition associated with tumors. This distinction allows MWI to differentiate between edema and tumor tissues [36]. When microwaves interact with these tissues, the differences in water content cause distinct reflections, allowing MWI to detect even small tumors that might be missed by mammography, particularly in dense breast tissue, where overlapping densities often hinder accurate detection [34,37–39].

In addition to its superior sensitivity, MWI reduces false positives, minimizing the need for unnecessary biopsies and follow-up tests that can cause emotional distress, medical complications, and increased healthcare costs. This higher accuracy in detecting early-stage tumors and differentiating malignant from benign lesions makes MWI a reliable diagnostic tool, particularly in high-risk or dense breasts. The non-invasive nature of MWI, which requires no breast compression, also enhances patient comfort, further

setting it apart from traditional mammography. Furthermore, MWI's cost-effectiveness, portability, and potential for use in resource-limited settings make it an attractive option for broadening access to breast cancer screening. By leveraging scattering parameters to distinguish between different tissue types, MWI holds the potential to significantly improve breast cancer detection, offering a safer, more efficient, and patient-friendly alternative to mammography [33,34,37–41].

In addition to its advantages over mammography, microwave imaging (MWI) offers several key benefits over MRI. These include lower cost, greater portability, and the ability to avoid both ionizing radiation and the use of contrast agents, making MWI an especially attractive option for breast cancer screening in resource-limited settings. While MRI is highly effective for detecting soft tissue abnormalities, gadolinium-enhanced MRI may struggle to differentiate between benign and malignant lesions in dense breast tissue due to the complexities of signal interpretation. In contrast, MWI leverages microwave scattering, which is particularly sensitive to the unique dielectric properties of cancerous tissues, enabling it to more effectively detect tumors in dense breasts [33,34,38,39,42–45].

Many research groups have made significant strides in the development and testing of microwave breast imaging (MBI) prototypes. For instance, Micrima Ltd. (University of Bristol, Bristol, UK) conducted clinical trials using their MBI system, MARIA, involving 225 patients and achieving a sensitivity rate of 76% [46]. Similarly, Umbria Bioengineering Technologies (Rivotorto, Italy) tested their MammoWave prototype on 58 patients, reporting a sensitivity of 74% and employing Support Vector Machine (SVM)-based algorithms for automated lesion classification. The MammoWave system achieved an accuracy of 91%, a sensitivity of 84.4%, and a specificity of 97.2% [47]. Microwave Vision Group (MVG) in France also validated their Wavelia MBI prototype on 24 patients, demonstrating its ability to distinguish between benign and malignant lesions with an accuracy of 88.5% [37]. While these studies are promising, most of them have used cross-validation techniques, which tend to involve smaller datasets and may limit the generalizability of the results. Additionally, many of these studies have focused on simpler tasks, such as distinguishing between benign and malignant lesions or between malignant tissue and healthy tissue. However, they have not addressed the more complex task of differentiating malignant lesions (positive) from both benign lesions and healthy tissues (negative) in a unified framework, which is essential for accurate and reliable breast cancer detection.

At Mitos Medical Technologies (Istanbul, Turkey), we developed the SAFE (Scan and Find Early) device, a microwave-based imaging system designed to enhance breast cancer detection. The device underwent initial testing on a cohort of 54 subjects, achieving a sensitivity of 63% [39]. Building on this, our most recent study expanded the cohort to 113 patients. The focus of this study was to evaluate SAFE's ability to differentiate benign and healthy tissues (negative) from malignant lesions (positive) [34]. This study achieved a sensitivity of 79% and specificity of 77% using 5-fold cross-validation. The primary objective was to assess SAFE's capability to accurately classify malignant lesions from benign and healthy tissues, ultimately reducing unnecessary biopsies for patients already flagged by conventional screening methods like mammography.

This represents a significant advancement over previous efforts, primarily due to the increase in cohort size and methodological rigor [34,39]. In our first preliminary study, we used microwave images, but following that study, we decided to consistently use scattering parameters. These parameters represent the dielectric properties of tissues, which have been shown to differ significantly between cancerous and non-cancerous tissues, providing a more effective method for distinguishing between the two [39].

In contrast to earlier studies, which predominantly relied on cross-validation, we employed a more robust train–test split methodology. Specifically, the model was trained

on data from 324 patients, validated on 112 patients, and tested on 15% (n = 80) of the unseen data. This approach provides a more reliable and independent evaluation of model performance, enhancing the generalizability of the results and reducing the risk of overfitting—an issue commonly associated with smaller training sets and cross-validation.

The aim of our study was to focus on a binary classification task, where we differentiate malignant lesions (positive) from both benign lesions and healthy tissues (negative). This is a more clinically relevant problem, as it combines benign and healthy tissues into a single negative class. This approach improves diagnostic accuracy, as it allows the device to focus specifically on identifying malignant lesions, while minimizing the confusion between benign and healthy tissues.

2. Materials and Methods

2.1. Patient Selection

The study received approval from the local ethics committee, and all protocols and procedures were conducted in accordance with institutional guidelines, ethical research standards, and the World Medical Association Declaration of Helsinki. Participants included patients aged 18 or older who visited the collaborating breast imaging clinic between May 2022 and October 2023 for screening, diagnostic assessment, or biopsy. Participation was entirely voluntary, with written informed consent obtained from each patient. Exclusion criteria included patients with a prior history of breast cancer and those with breast implants.

2.2. Device Description

SAFE is a novel microwave breast screening device, developed by Mitos Medical Technologies. The device has two antennas, one designated for transmitting and the other for receiving microwave signals. Both antennas function within a ceramic medium for impedance matching, operating across a frequency range of 0.6–8 GHz with a sampling interval of 200 MHz. These antennas are connected to a 2-port vector network analyzer (P9372A Keysight USB) and are housed within a cubical enclosure lined internally with microwave absorbers. The enclosure features a ceramic cylinder embedded in a cavity, designed to accommodate the patient's breast in a prone position. Adjustable ceramic cups are provided as part of the cylindrical setup to account for variations in breast size. The antennas rotate azimuthally, capturing signals in a configuration where each transmitting position corresponds to multiple receiving positions. In this arrangement, both the transmitting and receiving antennas have 36 positions, with each position separated by a 10° interval. The SAFE system completes data acquisition in approximately 5 min per breast. The 10° interval captures diverse angles, enhancing measurement variety and enabling accurate differentiation between tissue types (e.g., malignant vs. benign) using S11 and S21 parameters (Figure 1).

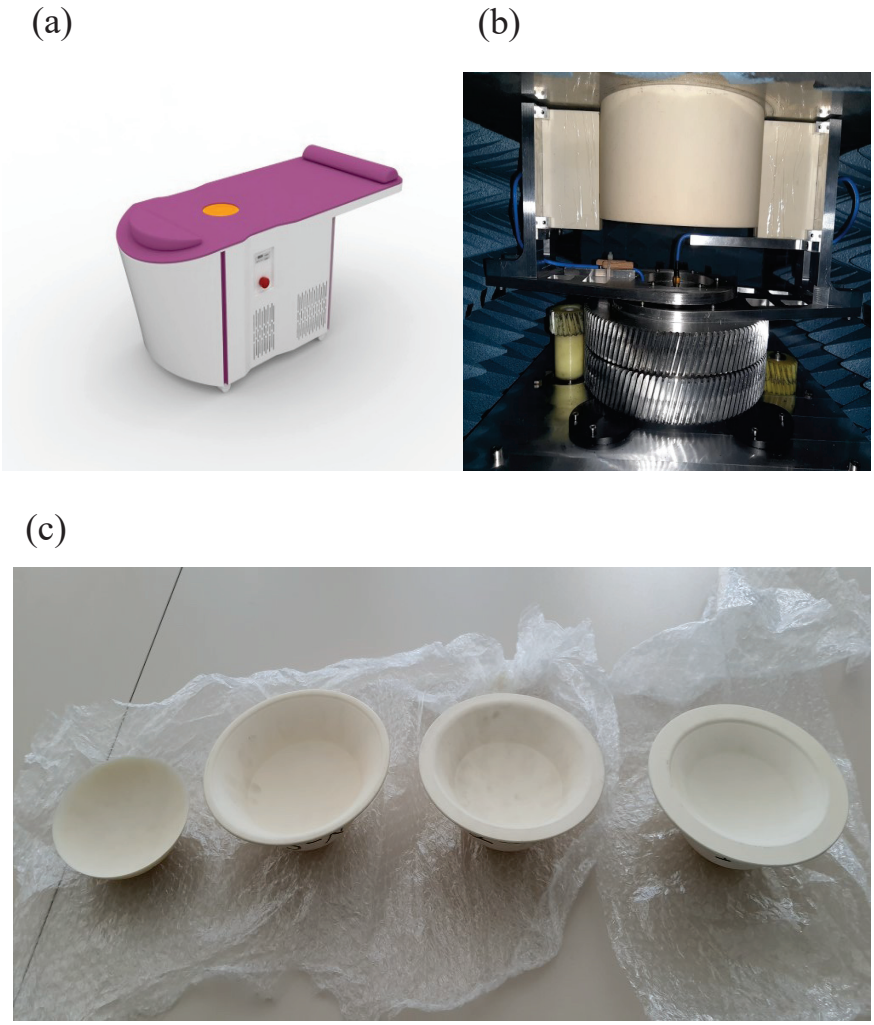


Figure 1. SAFE (Scan and Find Early) scanning system for clinical setup, system configuration and breast size cups. (**a**) The SAFE (Scan and Find Early) scanning bed system as used in a clinical setting, where data acquisition for each breast is completed in approximately 5 min; (**b**) the mechanically mobile bistatic system, where both transmitting and receiving antennas are positioned at 36 intervals of 10° each, generating 1296 measurements per frequency to ensure comprehensive data capture; and (**c**) patients are categorized according to breast size (1, 2, 3, and 4). Each patient is assigned a corresponding cup designed to fit their specific breast size for accurate measurements.

2.3. Data Categorization

Patients with BI-RADS mammogram readings of categories 1 and 2, as well as those who underwent biopsy with benign pathology results, were categorized as the negative group. Patients with malignant pathology findings confirmed by biopsy were categorized as the positive group. To ensure this categorization reflected meaningful distinctions, Welch's *t*-test was applied to compare the two groups for each frequency individually (Figure 2) and for the combined frequencies (Supplementary Figure S1a). A *p*-value threshold of 0.05 was used to determine statistical significance. The significance levels are indicated in the figures as follows: a single asterisk (*) for *p*-values between 0.01 and 0.05, two asterisks (**) for *p*-values between 0.001 and 0.01, three asterisks (***) for *p*-values between 0.0001 and 0.001, and four asterisks (****) for *p*-values less than 0.0001. This statistical framework provides robust support for the classification of patients with malignant pathology as positive and those with benign or no pathology as negative.

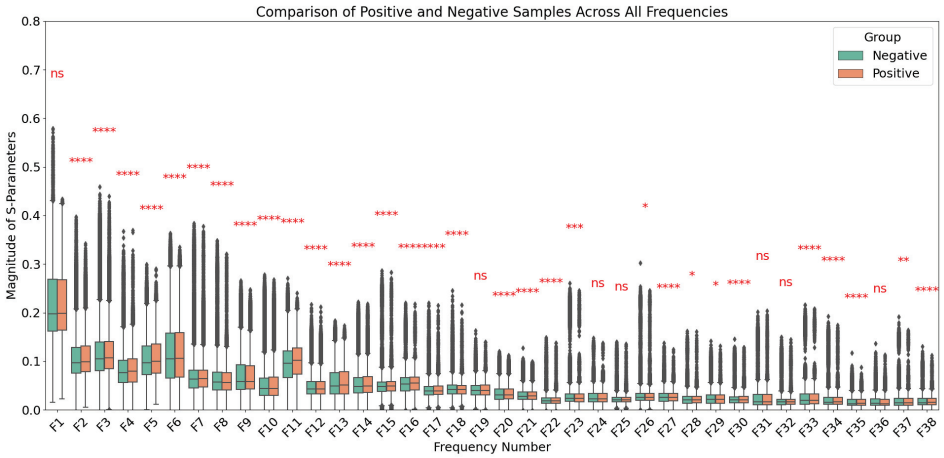


Figure 2. Comparison of S-parameters between positive (malignant) and negative (benign or no pathology) groups across frequencies ranging from 0.6 GHz to 8 GHz in 0.2 GHz increments. Welch's *t*-test was applied at each frequency to assess statistical significance, with *p*-value thresholds represented as follows: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***), and $p < 0.0001$ (****), with "ns" denoting not significant ($p > 0.05$). Data are presented as mean ± standard deviation for each group. Significant differences at specific frequencies highlight the distinct characteristics between the two groups.

Patients were also categorized into subgroups based on factors such as age, breast density, and breast size to check how the model performance alters across those groups. Breast density was recorded for those who underwent mammography or MRI before or after biopsy, while it remained unknown for patients examined solely by ultrasound. Density was assessed using the BI-RADS scale, as follows: A—predominantly fatty; B—scattered fibroglandular tissue; C—heterogeneously dense; and D—extremely dense [48]. Given the limited sample size, patients were divided into two density categories: non-dense (BI-RADS A and B) and dense (BI-RADS C and D). Age was divided into two groups: younger (18–47 years) and older (48 years and above), with the median age of participants as the cutoff. Breast size was also classified into two groups based on the ceramic cups used to support the breast during scanning: small (cup sizes 1 and 2) and large (cup sizes 3 and 4), with cup selection determined by the medical technician performing the scan. Breast lesions were categorized into three groups based on size: T1, for lesions measuring between 0 mm and 20 mm; T2, for lesions between 21 mm and 50 mm; and T3, for lesions larger than 50 mm. Due to the limitations of the dataset, T4 lesions, which had spread to adjacent tissues and were larger than 50 mm in size, were not included in the analysis.

2.4. Data Analysis

To enable automated breast lesion classification, both S11 and S21 parameters—proportional to the electromagnetic field emerging after the incident wave radiated by the transmitting antenna has interacted with breast tissue—were collected for each negative group (healthy, benign) and positive group (malignant) at 38 different frequency points. The measurements were conducted using vector network analysis (VNA), where the intermediate frequency (IF) bandwidth was adjusted to ensure a low noise ratio. The raw S11 and S21 data, denoised by the VNA, were directly used in their original time-series form without applying additional transformations.

For each patient sample, a 36 × 36 matrix was constructed, where each row represented the magnitude of S11 and S21 measurements across 36 receiver positions. The total number of S-parameters equaled 36 × 36 due to the 36 transmitter positions. This process was repeated for 38 distinct frequencies spanning the range of 0.6–8 GHz. Principal component analysis (PCA) was applied to the 36 × 36 matrices at each frequency. The dimensionality of the data was reduced by summing the singular values obtained from the decomposed

vectors. As a result, each patient was represented by 38 features, corresponding to the 38 frequency points. The features for all patients were compiled into a feature matrix.

To classify the negative and positive samples, the extreme gradient boosting (XGBoost) algorithm—a tree-based ensemble method—was utilized due to its ability to handle structured datasets, its resilience to overfitting through built-in regularization, and its effectiveness in processing imbalanced datasets. The algorithm was implemented using the Python programming language and the XGBoost library (xgboost v2.0.2). Histopathological results served as the ground truth for training and validation. XGBoost minimizes residuals and enhances predictive performance by combining multiple weak learner base estimators [49]. Prediction probabilities were calculated to determine malignancy, with a default threshold value of 0.5 used to classify cases as malignant. Patients with probabilities exceeding this threshold were classified as likely to have malignant breast cancer.

Previously, it was observed that breast size significantly impacts model performance [39]. Therefore, to prepare the dataset for training, samples were stratified by breast size to ensure balanced representation: 2× healthy samples, 1× benign sample, and 1× malignant sample from each breast size were selected. This strategy was designed to better reflect real-world clinical scenarios, where healthy cases are more prevalent. In total, the dataset comprised 80 breasts, of which 59 were negative (healthy + benign) and 21 were positive (malignant). The full dataset consisted of 526 samples, with a test set designed to include 15% of the total samples, adjusted to reflect the distribution observed in clinical practice. For training and validation, with the remaining 446 samples, we applied stratified splitting based on labels (negative/positive) and breast sizes using a 3:1 split ratio. The final split included 168 negative samples and 166 positive samples for the training set, and 59 negative samples and 53 positive samples for the validation set. Sensitivity, specificity, and accuracy metrics were calculated for each group, and the results were combined to evaluate the overall performance of the models in detecting malignant breast cancer cases. In addition to evaluating the overall performance of SAFE in classifying positive and negative samples, the performance of SAFE within specific subgroups was analyzed, focusing on breast density, breast size, and age groups. These subgroup analyses provided insights into how the model performed across varying sample characteristics. All results are presented using ROC curves, demonstrating the model's performance across different thresholds for both overall classification and subgroup-specific analyses.

To compare the negative and positive groups across age categories, Welch's *t*-test was applied to assess the differences in age both within the groups (negative vs. positive) and between the groups (young vs. older patients). Similarly, to evaluate the distribution of the negative and positive groups across different breast density and breast size categories, the chi-square test was utilized. Lastly, to determine whether lesion size differed between the training and test sets, Welch's *t*-test was performed. All statistical analyses were performed using the Python library statannot v.0.2.3. For the significance levels, *p*-values were categorized as follows: a *p*-value less than 0.0001 was considered highly significant (****), *p*-values between 0.0001 and 0.001 indicated strong significance (***), *p*-values ranging from 0.001 to 0.01 were categorized as moderately significant (**), and *p*-values between 0.01 and 0.05 represented mild significance (*).

3. Results

In this study, a total of 526 patients were included, with 286 classified as the negative group (healthy patients and benign patients) and 240, diagnosed with malignant breast cancer, classified as the positive group. The patients' ages ranged from 19 to 74 years, with a mean age of 47.2 ± 9.7 years. A slight but statistically significant difference was observed between the negative and positive samples in the older age group (*p*-value

between 0.05 and 0.01) (Supplementary Figure S1b). No significant differences were found when comparing the negative and positive samples within the breast density and breast size groups, nor were there statistically significant differences in the distribution of patients across these categories (Supplementary Figure S1c,d). Additionally, a significant difference was observed between the younger and older patient groups, confirming that the age threshold of 48 years effectively separated the two categories (Supplementary Figure S1b). Among the malignant cases, lesion sizes ranged from a minimum of 8 mm to a maximum of 46 mm, with an average size of 22.3 ± 9.2 mm.

Notably, the negative groups consisted of more patients than the positive group in both small breasts (202 negative, 174 positive) and large breasts (84 negative, 66 positive). After splitting the data into training, validation, and test sets, the total numbers were 334 for the training set, 112 for the validation set, and 80 for the test set. Of the 80 test samples, 21 were positive samples, and 59 were negative samples. Within the test set, 47 out of 80 breasts were categorized as non-dense (BI-RADS types A and B), while 23 were classified as dense (BI-RADS types C and D). Density measurements were unavailable for the remaining 10 breast samples due to the exclusive availability of ultrasonography (US) images. In the smaller-breast subgroup of the test data, the count of dense breasts (n = 20) closely paralleled that of non-dense breasts (n = 28). Conversely, in the larger-breast subgroup, the number of dense breasts (n = 3) was notably lower than that of non-dense breasts (n = 19). Additionally, the test dataset included 42 patients categorized as old (Age $\geq$ 48) and 38 patients categorized as young (Age < 48) (Supplementary Table S2).

Upon reviewing the histopathological findings of the lesions, 286 out of 491 lesions formed by positive and negative samples (malignant and benign lesions) were confirmed with specific histological diagnoses. The most common finding was invasive ductal carcinoma (IDC), which accounted for 82 cases (approximately 28.7%) of the total positive samples. Additionally, fibroepithelial lesions, fibrosis, and fibroadenomas were observed in 68 cases (approximately 23.8%). Other notable findings included chronic inflammation (n = 10), with additional histological diagnoses each occurring in fewer than 10 cases. These predominant histological findings were more common in the positive samples compared to the negative samples, which showed a less varied distribution of findings. In the test set, 21 positive (malignant) and 21 negative (benign) samples were analyzed for histological distribution. Of these, 14 samples were diagnosed with invasive ductal carcinoma (IDC), and 10 samples had either fibroepithelial lesions or fibrosis. The remaining cases included invasive apocrine carcinoma and metastatic carcinoma and chronic inflammation (Supplementary Figure S2a, Supplementary Table S1).

Following the classification of positive and negative samples, the model performance for all breasts was measured. SAFE was able to correctly classify 65 of the samples out of 80 patients in the test group (false positive: 11, false negative: 4). As a result, the sensitivity, specificity, and accuracy of the model performance for all breasts were measured as 80%, 81%, and 81%, respectively (Figure 3).

To assess performance variations relative to breast size, the model's efficacy was evaluated separately for large and small breasts. Notably, in small breasts, SAFE demonstrated an accuracy of 82%, while in large breasts, it achieved a lower accuracy rate of 78%. Moreover, the model applied to small breasts exhibited a superior specificity of 86% compared to the 81% specificity observed in large breasts. Correspondingly, the model's sensitivity in small breasts was measured at 81%, whereas in large breasts, it notably dropped to 71% (Figure 4).

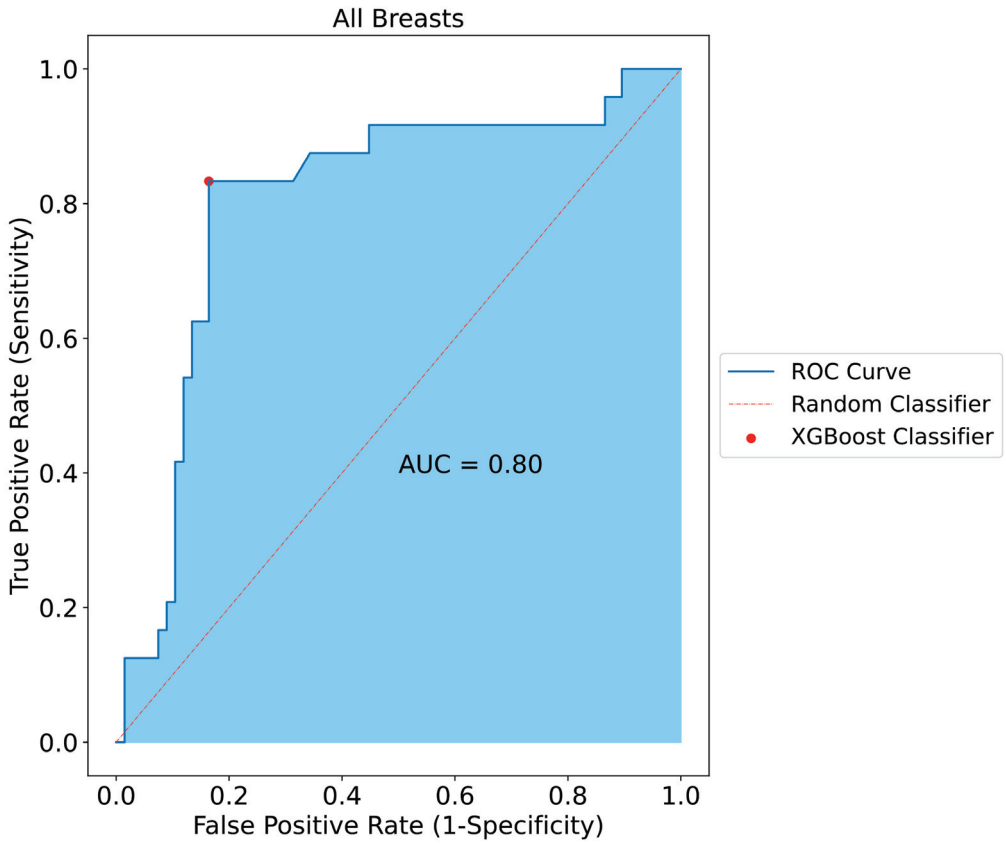


Figure 3. Classification results for all breasts included in the study. SAFE correctly classified 65 out of 80 tested subjects with 0.80, 0.81, and 0.81 for sensitivity, specificity, and accuracy, respectively.

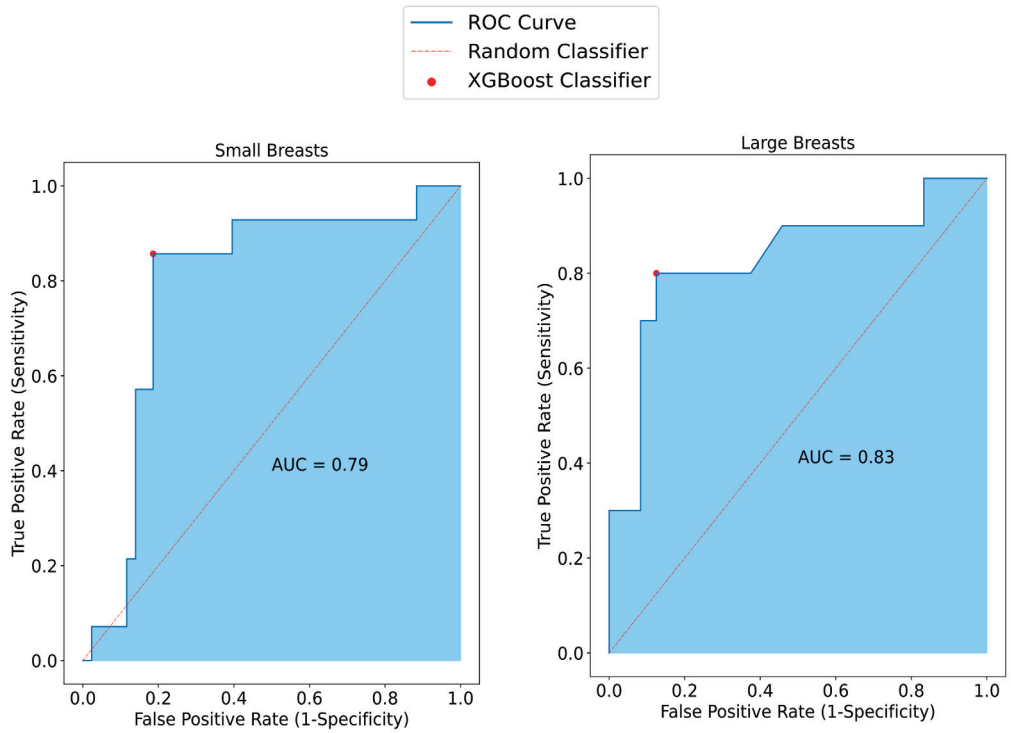


Figure 4. SAFE classification results vary with breast size. In smaller breasts, sensitivity reaches 81%, surpassing the 71% observed in larger breasts. Conversely, specificity decreases to 81% in larger breasts compared to 86% in smaller ones. Accuracy improves from 82% in larger breasts to 78% in smaller ones.

Considering the performance of the overall model in terms of breast density, in non-dense breasts, SAFE achieved an accuracy of 79%, while it was measured as 91% in dense breasts. Also, the model had higher specificity (92%) for dense breasts than the model for the non-dense breasts (82%). Similarly, the sensitivity of the model for the non-dense breasts (67%) was lower than for the dense breasts (91%). These performance results only considered the breast samples that had information on breast density (Figure 5).

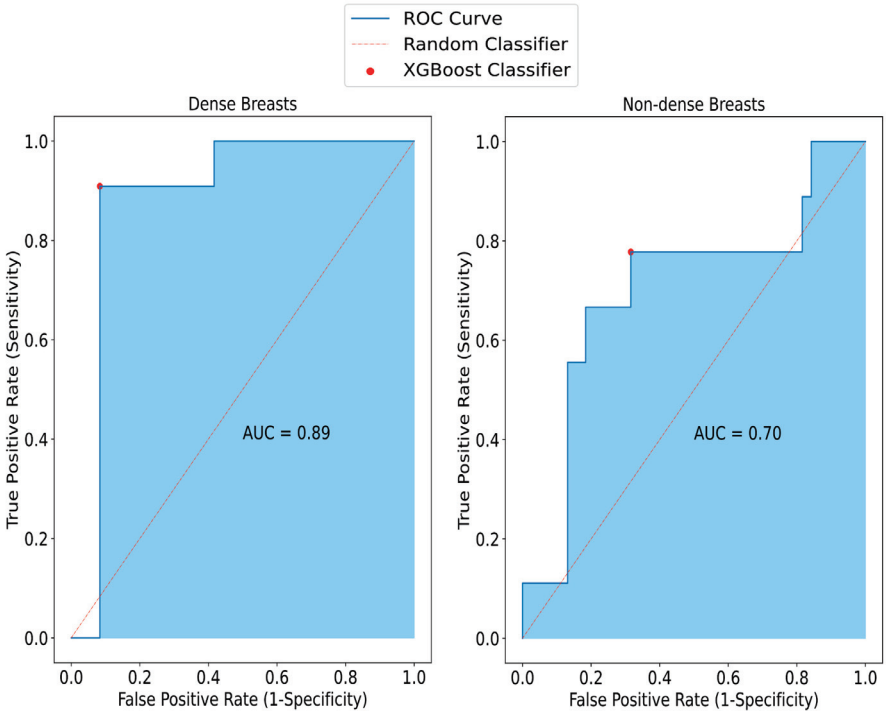


Figure 5. The impact of density on SAFE classification results. Sensitivity, specificity, and accuracy rates obtained from non-dense small breasts are 67%, 82%, and 79%, respectively, whereas in dense breasts, these results are 91%, 92%, and 91%, respectively.

To investigate the effect of the patient's age on the model, the model performance for both the old age group and the young age group was assessed. The sensitivity of SAFE was lower in the old patient group (80%) than in the young patient group (83%). Similarly, comparing the specificity of the patients, the specificity of SAFE for older patients (78%) was also lower than for the younger patients (84%). Lastly, the accuracy results for old and young patients were 79% and 84%, respectively (Figure 6).

The model's performance in classifying histological findings was evaluated on the test set. For invasive ductal carcinoma (IDC), the model accurately classified 11 out of 14 cases (78.6%) of this malignant lesion. For benign lesions, such as fibroepithelial lesions and fibrosis, the model demonstrated excellent performance, correctly classifying 90% of all cases. Due to the limited number of other histological findings in the test set, their classification performance was not considered. Finally, the distribution of lesion sizes between the training and test sets was examined. Statistical analysis using Welch's *t*-test revealed no significant difference in lesion size between the two groups (Welch's *t*-test, $p = 0.05$), indicating that the dataset is balanced with respect to lesion size (Supplementary Figure S2b).

The performance of the device was also assessed in terms of lesion size. SAFE successfully identified 8 out of 10 T1 lesions (smaller than 2 cm) and accurately detected 9 out of 11 larger T2 and T3 lesions (Figure 7).

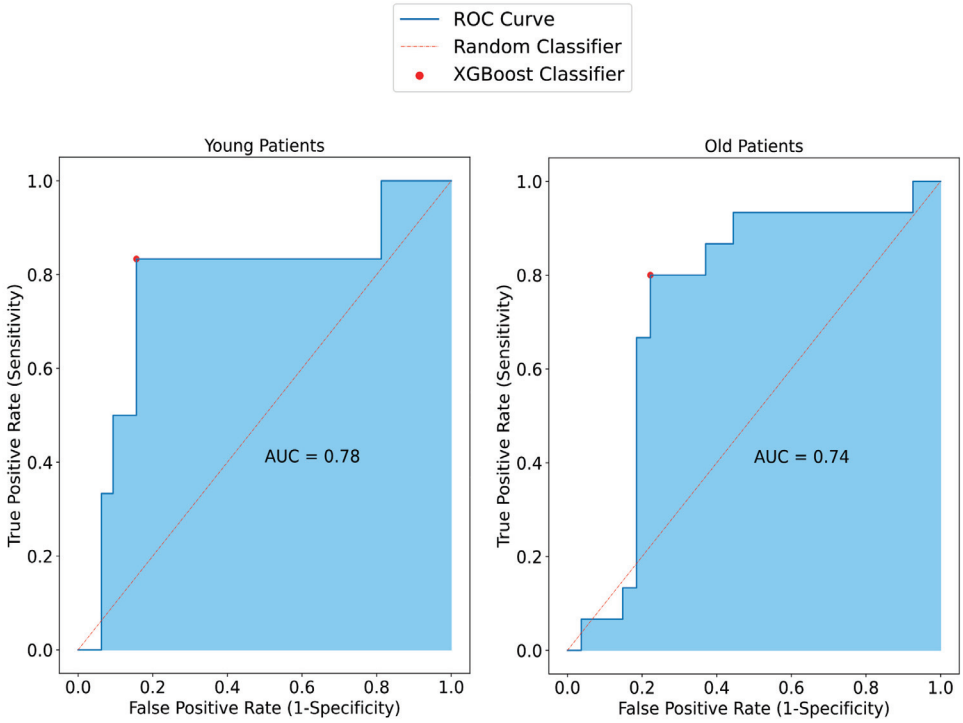


Figure 6. Representation of SAFE performance in relation to age among patients with small breasts. In the younger group, sensitivity, specificity, and accuracy were measured at 83%, 84%, and 84%, respectively. Conversely, in the older patient group, higher values were observed compared to the younger group, with sensitivity, specificity, and accuracy measured at 80%, 78%, and 79%, respectively.

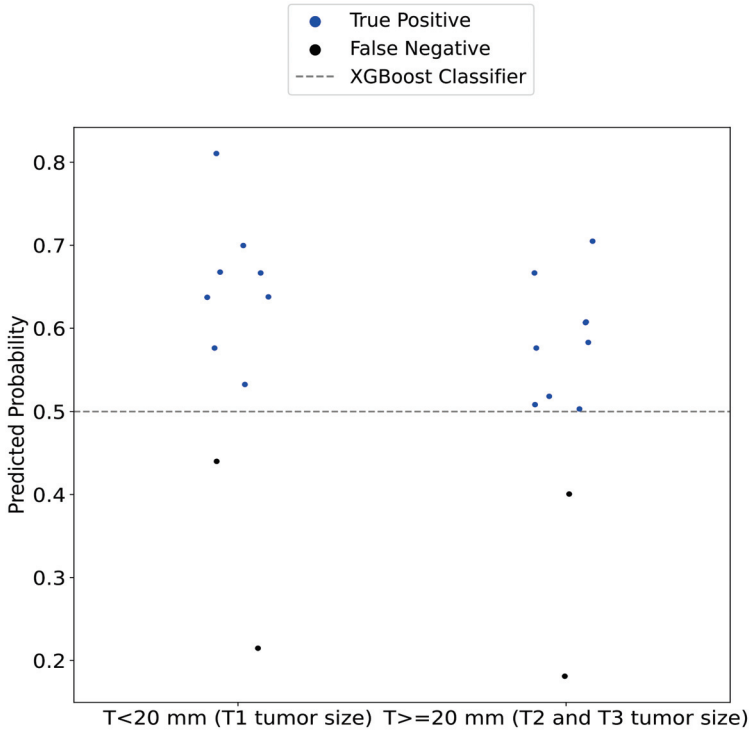


Figure 7. The performance of the SAFE system was evaluated across different cancerous lesion groups. In group 1, consisting of T1 tumor lesions, the system correctly identified 80% of the malignant tumors. For group 2, which includes T2 and T3, the sensitivity of the SAFE system was measured at 82%.

4. Discussion

The aim of our study was to focus on a binary classification task to differentiate malignant breast lesions (positive) from benign lesions and healthy tissues (negative). This approach is clinically significant because it combines both benign and healthy tissues into a single negative category, which more closely aligns with the diagnostic challenge in clinical settings. By adopting this binary classification framework, the device is able to focus on accurately identifying malignant lesions while minimizing the risk of misclassifying benign tissues. This approach improves diagnostic accuracy, helping to reduce the confusion between benign and healthy tissues, which is crucial for effective clinical decision making.

To achieve this classification, we utilized the SAFE microwave breast imaging device to gather data from patients with both malignant and non-cancerous breast tissues (comprising healthy and benign patients). The absolute values of the S11 and S21 complex matrices, representing the electromagnetic properties of the breast tissue, were used as feature variables in the predictive modeling. We applied the XGBoost machine learning algorithm, known for its robustness in handling complex, high-dimensional data, to classify non-cancerous individuals and malignant breast cancer patients [49].

In our analysis, we compared the negative and positive samples within each age group. A slight but statistically significant difference was observed in the older age group, with a p-value falling between 0.05 and 0.01. This indicates that the negative and positive samples were well-matched across age groups, ensuring a balanced representation of both healthy and unhealthy patients within similar age ranges. The balanced matching of samples is essential for the model's performance, as it strengthens the model's robustness by providing a consistent and diverse dataset for training. This approach ultimately leads to more reliable and generalizable predictions.

Further, when comparing the negative and positive samples within each breast density and breast size group, no significant differences were found, which suggests that the dataset maintains a balanced representation across both sample types within these categories. Similarly, when comparing the number of patients in each density and breast size group, no statistically significant differences were observed. This reinforces the well-distributed nature of the selected samples, which is crucial for effective model training.

In contrast, when comparing the younger and older patient groups, a significant difference was observed between the two age categories, indicating that the age threshold of 48 years effectively separates the groups. This finding confirms the appropriateness of the chosen age cutoff, ensuring a clear distinction between the two age categories for the further performance analysis of the SAFE system based on age.

Notably, in this study, the dataset consisted of T1 and T2 lesions, with only a very small number of T3 lesions (n = 1) across the training, validation, and test sets. This structure was intentionally designed to focus on small breasts, particularly in the test set, allowing for a more targeted evaluation of the SAFE model's performance on this subset. The overall number of T3 lesions was very low, accounting for only 23 lesions out of 491 lesions (240 malignant lesions and 251 lesions samples). Additionally, the number of small-breast samples was higher than that of large breasts, further emphasizing the focus on small-breast categories.

In our first preliminary study, we employed qualitative image algorithms for the detection of breast cancer patients, reconstructing the MWI images by the localization of breast cancer. We scanned a total of 115 patient breasts and had a sensitivity of 63% in the overall patient samples and concluded that breast size dramatically affects results, where the device had 51% sensitivity in smaller breasts, while in larger breasts it had a sensitivity of 74% [39].

In the current study, the performance of SAFE was evaluated across both small and large breasts, and the performance of the device was remarkably similar between the two groups. However, a slight advantage was observed in smaller breasts, which may be attributed to the sample size of the large-breast group. Smaller breasts may allow for more direct coupling between antennas, potentially reducing the interaction of electromagnetic waves with the breast tissue. In scenarios where antennas are closely coupled, detecting subtle variations in the electromagnetic field becomes particularly challenging [39]. This phenomenon could potentially affect SAFE's ability to detect breast cancer in patients with smaller breasts. However, the noise removal process implemented during preprocessing appeared to mitigate these challenges, leading to improved performance of SAFE in smaller breasts. Despite the differences in model performance between breast sizes, preliminary results indicate notable improvements in SAFE's ability to detect breast cancer in smaller breasts.

Breast density can significantly influence the performance of mammography. Mammography typically demonstrates higher sensitivity in non-dense breasts (75%), which decreases to 51.3% in dense breasts [39]. In our preliminary findings, we demonstrated that SAFE exhibits superior performance regardless of breast density [39]. In our current expanded study, sensitivity rates of 91% for dense breasts and 67% for non-dense breasts were observed. Notably, dense breasts were evenly distributed across both the large- and small-breast test sets. However, non-dense breasts were underrepresented, with only three samples in the large-breast test set and nine in the positive group test set out of 21 samples. SAFE accurately classified all three non-dense breasts within this subset. Given the impact of breast size on model performance, we plan to increase the number of non-dense breast samples, particularly within the large-breast subset, to further enhance the sensitivity beyond the current 67%. Until this time, the high performance in dense breasts is promising as a potential supplemental screening tool in dense breasts.

In mammography, screening is recommended for older patients due to the cumulative harmful effects of ionizing radiation [6–9]. However, SAFE, which operates using non-ionizing low electromagnetic frequencies, poses no such risks and achieves 83% sensitivity in younger patients. However, the cohort size for malignant patients in the younger age group (n = 6) was small, and increasing this sample size would allow for more precise predictions. The results for younger patients were consistent with SAFE's previous performance in our preliminary study [39]. In the older patient cohort, SAFE demonstrated enhanced performance, with sensitivity, specificity, and accuracy rates of 80%, 78%, and 79%, respectively.

The prognosis of breast cancer is greatly influenced by lesion size. Early diagnosis is crucial to prevent metastases and increase survival [50,51]. SAFE successfully identified 80% of T1 (smaller than 2 cm) malignant lesions. Additionally, SAFE demonstrated an 82% detection rate for larger lesions exceeding 2 cm in size. These results suggest that SAFE could be a promising tool for screening, capable of detecting cancerous tissue before significant progression occurs.

To evaluate the generalizability of our model and check for signs of overfitting, we also examined the training and validation loss curves. A typical indication of overfitting would be a scenario where the training loss continues to decrease while the validation loss increases over time. In contrast, our model demonstrated a consistent decrease in both training and validation loss throughout the training process, as shown in Supplementary Figure S3. This indicates that the model generalizes well to unseen data and has not overfitted to the training set, which is a key strength of the model's robustness.

In comparing our study's binary classification framework with previous microwave breast imaging (MBI) research, several key distinctions and advancements emerge that

highlight the significance of our work. Our study addresses a binary classification task that differentiates malignant lesions from both benign and healthy tissues, which not only aligns more closely with real-world diagnostic needs but also reduces confusion in clinical decision making. By grouping benign and healthy tissues together as a single negative category, our framework minimizes unnecessary complexity and allows clinicians to focus more precisely on the critical task of identifying malignant lesions. This approach is especially important in clinical settings where the primary concern is detecting cancer, and reducing the number of false positives can lead to fewer follow-up procedures.

One of the novel contributions of our study is the ability of SAFE to effectively classify negative samples (both benign and healthy tissue) from positive samples (malignant tissue). This binary classification framework has broad implications for early detection and pre-screening. Since SAFE distinguishes malignancies from healthy tissue early in the process, it can be utilized prior to traditional screening technologies like mammography or ultrasound. This is a crucial advancement, particularly for dense breasts, where traditional methods often fall short. By identifying suspicious lesions early, SAFE provides a first line of defense in breast cancer detection, potentially reducing the need for more invasive diagnostic procedures.

Furthermore, SAFE's ability to distinguish benign from malignant lesions helps facilitate the biopsy process. In current clinical practice, biopsies are often used to confirm the presence of cancer after initial imaging has identified suspicious areas [52]. SAFE's pre-screening capability can help refine this process by more accurately classifying malignant lesions from benign ones, ensuring that biopsies are performed on truly suspicious lesions, and reducing the number of unnecessary biopsies for benign cases. This could lead to fewer procedures, reduced patient anxiety, and more efficient use of medical resources.

There are many limitations in our study: one limitation is that while the cohort size is larger compared to previous preliminary studies [34,39], the choice to use a holdout validation method introduces certain limitations. Specifically, using 20% of the dataset as a test set limits the data available for model training. Even though we applied a stratified split to the data, this can potentially affect the model's ability to learn from the full range of patient data, especially if the test set contains rare or outlier cases that are underrepresented in the training set.

Intriguingly, although the model demonstrated good performance in classifying benign lesions, the classification of IDC was slightly more challenging. The lower accuracy for IDC (78.6%) may be attributed to several factors. First, IDC tumors are often heterogeneous, with varying tissue structures, cellularity, and vascularization across different regions of the tumor. This heterogeneity can lead to variations in the microwave signal patterns, making it harder for the model to consistently identify IDC samples [53].

Furthermore, the water content in IDC tumors could contribute to these challenges. Malignant tumors, particularly IDC, often have a different water content distribution compared to benign lesions. The presence of necrotic regions and increased blood supply in IDC could alter the water retention properties of the tissue, influencing the microwave signal scattering in ways that are less predictable compared to the more uniform patterns observed in benign lesions [54].

Future work could focus on enhancing the model's ability to discriminate between different tumor types by incorporating data from a wider variety of malignant subtypes. This would not only improve the model's robustness in identifying IDC but also help in better classifying other rare malignant tumors that were underrepresented in this study. Additionally, incorporating features related to water content and tissue composition could provide valuable insights into how these factors influence microwave signal interactions, potentially leading to improved classification accuracy for both malignant and benign lesions.

A significant limitation in this study arises from the subjective assessment of breast cup size. Since breast cup size is typically assessed based on visual or physical measurement, it is prone to variability between healthcare providers. This subjectivity introduces a potential source of mismatch error, particularly in categorizing patients into large- or small-breast size groups. Inconsistent categorization may affect the study's analysis, as breast size was found to influence the performance of the SAFE device. For instance, if a patient's breast size is categorized incorrectly, it could impact the model's sensitivity and specificity for different patient subgroups, leading to inaccurate conclusions. To address this, future studies could implement a more standardized or objective system for breast size classification, such as using specific anatomical measurements or imaging-based techniques, to minimize variability and ensure consistency across patients.

Another limitation of this study involves the missing breast density categorization in 10 test samples. These patients underwent ultrasound (US) examination rather than traditional mammography, which led to the absence of breast density data. Breast density is a critical factor in the performance of many breast cancer detection techniques, including SAFE. The lack of density categorization in a portion of the test dataset limits the ability to evaluate how SAFE performs across different breast densities. These missing data could introduce a bias in the study, as the device's sensitivity might vary depending on whether the breast tissue is dense or non-dense. To improve the accuracy of future studies, it is important to ensure that all test samples are categorized for breast density, either through mammography or alternative techniques such as ultrasound or MRI. This would provide a more comprehensive analysis of the device's performance across various tissue types and densities.

Although the current study offers valuable insights into the performance of the SAFE device, its generalizability to broader populations is limited. The cohort may have been drawn from a specific geographic region or demographic group, which may not fully represent the global population. For example, the sample might be biased towards particular ethnicities or risk groups, which could affect the model's performance in other populations. Differences in breast tissue characteristics, cancer prevalence, and genetic factors could influence the device's efficacy. To enhance generalizability, future studies should include a more diverse cohort from different geographical locations, ethnic backgrounds, and risk categories. A multi-center, international study could provide a broader evaluation of the device's performance and ensure that the findings are applicable to a wider range of patients.

5. Conclusions

The SAFE microwave breast imaging system demonstrates significant promise in the binary classification of malignant lesions versus benign and healthy tissues, aligning closely with real-world clinical needs. By combining benign and healthy tissues into a single negative category, the system simplifies the diagnostic process and reduces confusion, improving overall diagnostic accuracy. The system's ability to identify malignant lesions early, especially in dense breasts, offers a valuable tool in pre-screening and early detection, which can complement traditional screening methods like mammography.

In addition, SAFE's capacity to differentiate benign from malignant lesions helps in improving the biopsy process, reducing unnecessary procedures and associated patient anxiety. While SAFE performs well across varying breast sizes and densities, further enhancements in breast size classification and breast density categorization are needed to improve performance consistency. Larger and more diverse cohorts would also help enhance the generalizability of the system.

Overall, SAFE provides a non-invasive, non-ionizing approach to breast cancer detection with potential applications in early diagnosis, screening, and biopsy guidance, representing an important step forward in breast cancer care.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cancers17020214/s1, Figure S1. Statistical Comparisons of patient groups by age, breast density, and breast Size, including raw data analysis between patient groups. (a) Comparison of negative and positive patient groups for all frequencies combined. The **** symbol indicates that the *p*-value is less than 0.0001. (b) Comparison of negative and positive patient groups within older and younger age groups. The **** symbol indicates that the *p*-value for the comparison between older and younger groups, for both negative and positive patients, is less than 0.0001. A slight but statistically significant difference was observed between negative and positive groups within the older age group (*, *p*-value between 0.05 and 0.1), while no statistical difference was observed in the younger age group. (c,d) No significant differences were observed across or between groups in terms of breast size or breast density. For breast size, comparisons were made between small and large breasts for each patient group, as well as between negative and positive groups within each breast size. For breast density, comparisons were made between non-dense and dense breasts for each patient group, as well as between negative and positive groups within each density group. Welch's *t*-test was applied for the comparison of raw data and age groups, while the Chi-square test was used for comparisons of breast size and density categories. Figure S2. Histopathological findings for all groups of patients and lesion size comparisons between train and test groups. (a) The top 8 histopathological findings of lesions from positive and negative groups. A total of 286 confirmed histopathological results were analyzed. Among these, 82 cases (28.7%) were identified as invasive ductal carcinoma, while 23.8% were classified as fibroepithelial lesions, fibrosis, and fibroadenomas. Comparison of lesion sizes between the training and test groups. No significant differences were observed (ns). Welch's t-test was applied, with a *p*-value threshold of 0.5. Figure S3. Training and Validation Loss During Model Training. The plot illustrates the training and validation loss during model training. The x-axis represents the number of iterations through the entire training dataset, while the y-axis represents the log loss. As the training loss decreases, the validation loss also decreases, indicating consistent model performance. Table S1. Comprehensive List of Histological Findings for Lesions Across All Patient Groups. The table provides a detailed summary of all histological findings for lesions across the patient groups, along with their respective counts. A total of 286 histological findings were documented. Table S2. Distribution of Patient Samples Across Various Categories. This table provides the distribution of patient samples across different categories, including the number of healthy and unhealthy samples, as well as the breakdown of training, validation, and test samples. Additionally, it presents the categorization of test samples based on breast density (dense vs. non-dense) and age (young vs. old). All sample sizes were shown in relation to breast size.

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

Funding: This research was funded by the Scientific and Technology Research Council of Turkey (TUBITAK).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of Marmara University Hospital Breast Center (protocol code 70737436-050.06.04 and date of approval 9 June 2014).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data are not publicly available due to Governmental and Institutional (Marmara University Hospital Breast Center) regulations and non-violation of patients privacy.

Conflicts of Interest: Authors Ali Yurtseven, Aleksandar Janjic, Mehmet Cayoren, Mustafa Erkin Aribal and Ibrahim Akduman were employed by Mitos Medical Technologies. The remining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflicts of interest.

References

1. Wilkinson, L.; Gathani, T. Understanding breast cancer as a global health concern. *Br. J. Radiol.* **2022**, *95*, 20211033. [CrossRef] [PubMed]
2. Bhushan, A.; Gonsalves, A.; Menon, J.U. Current state of breast cancer diagnosis, treatment, and theranostics. *Pharmaceutics* **2021**, *13*, 723. [CrossRef] [PubMed]
3. Duffy, S.W.; Tabár, L.; Yen, A.M.F.; Dean, P.B.; Smith, R.A.; Jonsson, H.; Törnberg, S.; Chen, S.L.S.; Chiu, S.Y.H.; Fann, J.C.Y.; et al. Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women. *Cancer* **2020**, *126*, 2971–2979. [CrossRef] [PubMed]
4. Grimm, L.J.; Avery, C.S.; Hendrick, E.; Baker, J.A. Benefits and risks of mammography screening in women ages 40 to 49 years. *J. Prim. Care Community Health* **2022**, *13*, 21501327211058322. [CrossRef]
5. Hendrick, R.E.; Helvie, M.A. United States preventive services task force screening mammography recommendations: Science ignored. *Am. J. Roentgenol.* **2011**, *196*, W112–W116. CrossRef]
6. Armstrong, K.; Moye, E.; Williams, S.; Berlin, J.A.; Reynolds, E.E. Screening mammography in women 40 to 49 years of age: A systematic review for the American College of Physicians. *Ann. Intern. Med.* **2007**, *146*, 516–526. [CrossRef]
7. van den Ende, C.; Oordt-Speets, A.M.; Vroling, H.; van Agt, H.M. Benefits and harms of breast cancer screening with mammography in women aged 40–49 years: A systematic review. *Int. J. Cancer* **2017**, *141*, 1295–1306. [CrossRef]
8. Egeli, R.; Urban, J. Mammography in symptomatic women 50 years of age and under, and those over 50. *Cancer* **1979**, *43*, 878–882. [CrossRef]
9. Canelo-Aybar, C.; Ferreira, D.S.; Ballesteros, M.; Posso, M.; Montero, N.; Solà, I.; Saz-Parkinson, Z.; Lerda, D.; Rossi, P.G.; Duffy, S.W.; et al. Benefits and harms of breast cancer mammography screening for women at average risk of breast cancer: A systematic review for the European Commission Initiative on Breast Cancer. *J. Med. Screen.* **2021**, *28*, 389–404. [CrossRef]
10. Zhao, H.; Zou, L.; Geng, X.; Zheng, S. Limitations of mammography in the diagnosis of breast diseases compared with ultrasonography: A single-center retrospective analysis of 274 cases. *Eur. J. Med. Res.* **2015**, *20*, 49. [CrossRef]
11. Boyd, N.F.; Guo, H.; Martin, L.J.; Sun, L.; Stone, J.; Fishell, E.; Jong, R.A.; Hislop, G.; Chiarelli, A.; Minkin, S.; et al. Mammographic density and the risk and detection of breast cancer. *N. Engl. J. Med.* **2007**, *356*, 227–236. [CrossRef] [PubMed]
12. Nazari, S.S.; Mukherjee, P. An overview of mammographic density and its association with breast cancer. *Breast Cancer* **2018**, *25*, 259–267. [CrossRef] [PubMed]
13. Yaghjyan, L.; Colditz, G.A.; Rosner, B.; Tamimi, R.M. Mammographic breast density and subsequent risk of breast cancer in postmenopausal women according to the time since the mammogram. *Cancer Epidemiol. Biomarkers Prev.* **2013**, *22*, 1110–1117. [CrossRef] [PubMed]
14. Sak, M.A.; Littrup, P.J.; Duric, N.; Mullooly, M.; Sherman, M.E.; Gierach, G.L. Current and future methods for measuring breast density: A brief comparative review. *Breast Cancer Manag.* **2015**, *4*, 209–221. [CrossRef]
15. Lee, J.S.; Oh, M. Breast density of mammography is correlated with reproductive risk factors regardless of menopausal status: A cross-sectional study of the Korean National Screening Program. *Asian Pac. J. Cancer Prev. APJCP* **2020**, *21*, 1011. [CrossRef]
16. Tice, J.A.; Gard, C.C.; Miglioretti, D.L.; Sprague, B.L.; Tosteson, A.N.; Joe, B.N.; Ho, T.Q.H.; Kerlikowske, K. Comparing mammographic density assessed by digital breast tomosynthesis or digital mammography: The breast cancer surveillance consortium. *Radiology* **2022**, *302*, 286–292. [CrossRef]
17. Lucassen, A.; Watson, E.; Eccles, D. Advice about mammography for a young woman with a family history of breast cancer. *BMJ* **2001**, *322*, 1040–1042. [CrossRef]
18. Mao, X.; He, W.; Humphreys, K.; Eriksson, M.; Holowko, N.; Yang, H.; Tapia, J.; Hall, P.; Czene, K. Breast Cancer incidence after a false-positive mammography result. *JAMA Oncol.* **2024**, *10*, 63–70. [CrossRef]
19. Bijwaard, H.; Brenner, A.; Dekkers, F.; van Dillen, T.; Land, C.E.; Boice, J.D., Jr. Breast cancer risk from different mammography screening practices. *Radiat. Res.* **2010**, *174*, 367–376. [CrossRef]
20. Broeks, A.; Braaf, L.M.; Huseinovic, A.; Nooijen, A.; Urbanus, J.; Hogervorst, F.B.; Schmidt, M.K.; Klijn, J.G.; Russell, N.S.; Van Leeuwen, F.E.; et al. Identification of women with an increased risk of developing radiation-induced breast cancer: A case only study. *Breast Cancer Res.* **2007**, *9*, R26. [CrossRef]
21. de Gelder, R.; Draisma, G.; Heijnsdijk, E.A.; de Koning, H.J. Population-based mammography screening below age 50: Balancing radiation-induced vs prevented breast cancer deaths. *Br. J. Cancer* **2011**, *104*, 1214–1220. [CrossRef] [PubMed]

22. Goldfrank, D.; Chuai, S.; Bernstein, J.L.; Ramon y Cajal, T.; Lee, J.B.; Alonso, M.C.; Diez, O.; Baiget, M.; Kauff, N.D.; Offit, K.; et al. Effect of mammography on breast cancer risk in women with mutations in BRCA1 or BRCA2. *Cancer Epidemiol. Biomarkers Prev.* **2006**, *15*, 2311–2313. [CrossRef] [PubMed]
23. Gøtzsche, P.C.; Jørgensen, K.J. Screening for breast cancer with mammography. *Cochrane Database Syst. Rev.* **2013**, *2013*, CD001877. [CrossRef] [PubMed]
24. Jatoi, I.; Pinsky, P.F. Breast cancer screening trials: Endpoints and overdiagnosis. *JNCI J. Natl. Cancer Inst.* **2021**, *113*, 1131–1135. [CrossRef] [PubMed]
25. Srivastava, S.; Koay, E.J.; Borowsky, A.D.; De Marzo, A.M.; Ghosh, S.; Wagner, P.D.; Kramer, B.S. Cancer overdiagnosis: A biological challenge and clinical dilemma. *Nat. Rev. Cancer* **2019**, *19*, 349–358. [CrossRef]
26. Nicosia, L.; Battaglia, O.; Venturini, M.; Fontana, F.; Minenna, M.; Pesenti, A.; Budascu, D.; Pesapane, F.; Bozzini, A.C.; Pizzamiglio, M.; et al. Contrast-enhanced mammography BI-RADS: A case-based approach to radiology reporting. *Insights Imaging* **2024**, *15*, 37. [CrossRef]
27. Coffey, K.; Jochelson, M.S. Contrast-enhanced mammography in breast cancer screening. *Eur. J. Radiol.* **2022**, *156*, 110513. [CrossRef]
28. Van Nijnatten, T.J.; Lobbes, M.B.; Cozzi, A.; Patel, B.K.; Zuley, M.L.; Jochelson, M.S. Barriers to implementation of contrast-enhanced mammography in clinical practice: AJR expert panel narrative review. *Am. J. Roentgenol.* **2023**, *221*, 3–6. [CrossRef]
29. Scherer, L.D.; Lewis, C.L.; McCaffery, K.; Hersch, J.; Cappella, J.N.; Tate, C.; Morse, B.; Arnett, K.; Mosley, B.; Smyth, H.L.; et al. Mammography Screening Preferences Among Screening-Eligible Women in Their 40s: A National US Survey. *Ann. Intern. Med.* **2024**, *177*, 1069–1077. [CrossRef]
30. Golubicic, I.; Borojevic, N.; Pavlovic, T. Risk factors for breast cancer: Is ionizing radiation among them. *J. BUON* **2008**, *13*, 487–494.
31. Ronckers, C.M.; Erdmann, C.A.; Land, C.E. Radiation and breast cancer: A review of current evidence. *Breast Cancer Res.* **2004**, *7*, 21. [CrossRef] [PubMed]
32. Mumtaz, S.; Rana, J.N.; Choi, E.H.; Han, I. Microwave radiation and the brain: Mechanisms, current status, and future prospects. *Int. J. Mol. Sci.* **2022**, *23*, 9288. [CrossRef] [PubMed]
33. Kwon, S.; Lee, S. Recent advances in microwave imaging for breast cancer detection. *Int. J. Biomed. Imaging* **2016**, *2016*, 5054912. [CrossRef] [PubMed]
34. Janjic, A.; Akduman, I.; Cayoren, M.; Bugdayci, O.; Aribal, M.E. Microwave breast lesion classification–results from clinical investigation of the safe microwave breast cancer system. *Acad. Radiol.* **2023**, *30*, S1–S8. [CrossRef]
35. Cheng, Y.; Fu, M. Dielectric properties for non-invasive detection of normal, benign, and malignant breast tissues using microwave theories. *Thorac. Cancer* **2018**, *9*, 459–465. [CrossRef]
36. Gupta, S.S.; Mayrovitz, H.N. The breast edema enigma: Features, diagnosis, treatment, and recommendations. *Cureus* **2022**, *14*, e23797. [CrossRef]
37. Moloney, B.M.; McAnena, P.F.; Abd Elwahab, S.M.; Fasoula, A.; Duchesne, L.; Cano, J.D.G.; Glynn, C.; O'Connell, A.; Ennis, R.; Lowery, A.J.; et al. Microwave imaging in breast cancer–results from the first-in-human clinical investigation of the wavelia system. *Acad. Radiol.* **2022**, *29*, S211–S222. [CrossRef]
38. Wang, L. Microwave imaging and sensing techniques for breast cancer detection. *Micromachines* **2023**, *14*, 1462. [CrossRef]
39. Janjic, A.; Cayoren, M.; Akduman, I.; Yilmaz, T.; Onemli, E.; Bugdayci, O.; Aribal, M.E. SAFE: A novel microwave imaging system design for breast cancer screening and early detection—Clinical evaluation. *Diagnostics* **2021**, *11*, 533. [CrossRef]
40. Kaur, G.; Kaur, A. Breast tissue tumor detection using "S" parameter analysis with an UWB stacked aperture coupled microstrip patch antenna having a "+" shaped defected ground structure. *Int. J. Microw. Wirel. Technol.* **2020**, *12*, 635–651. [CrossRef]
41. AlSawaftah, N.; El-Abed, S.; Dhou, S.; Zakaria, A. Microwave imaging for early breast cancer detection: Current state, challenges, and future directions. *J. Imaging* **2022**, *8*, 123. [CrossRef]
42. Li, X.; Fan, Z.; Jiang, H.; Niu, J.; Bian, W.; Wang, C.; Wang, Y.; Zhang, R.; Zhang, H. Synthetic MRI in breast cancer: Differentiating benign from malignant lesions and predicting immunohistochemical expression status. *Sci. Rep.* **2023**, *13*, 17978. [CrossRef] [PubMed]
43. Penn, A.I.; Medved, M.; Dialani, V.; Pisano, E.D.; Cole, E.B.; Brousseau, D.; Karczmar, G.S.; Gao, G.; Reich, B.D.; Abe, H. Discrimination of benign from malignant breast lesions in dense breasts with model-based analysis of regions-of-interest using directional diffusion-weighted images. *BMC Med. Imaging* **2020**, *20*, 61. [CrossRef] [PubMed]
44. Moloney, B.M.; O'Loughlin, D.; Abd Elwahab, S.; Kerin, M.J. Breast cancer detection—A synopsis of conventional modalities and the potential role of microwave imaging. *Diagnostics* **2020**, *10*, 103. [CrossRef] [PubMed]
45. Adachi, M.; Nakagawa, T.; Fujioka, T.; Mori, M.; Kubota, K.; Oda, G.; Kikkawa, T. Feasibility of portable microwave imaging device for breast cancer detection. *Diagnostics* **2021**, *12*, 27. [CrossRef] [PubMed]

46. Shere, M.; Lyburn, I.; Sidebottom, R.; Massey, H.; Gillett, C.; Jones, L. MARIA® M5: A multicentre clinical study to evaluate the ability of the Micrima radio-wave radar breast imaging system (MARIA®) to detect lesions in the symptomatic breast. *Eur. J. Radiol.* **2019**, *116*, 61–67. [CrossRef]
47. Ghavami, N.; Sánchez-Bayuela, D.Á.; Sani, L.; Vispa, A.; Bigotti, A.; Badia, M.; Papini, L.; Raspa, G.; Rana, S.P.; Castellano, C.R.; et al. MammoWave Breast Imaging Device: An International and Multicentric Clinical Investigation. In Proceedings of the 2023 17th European Conference on Antennas and Propagation (EuCAP), Florence, Italy, 26–31 March 2023; pp. 1–5.
48. American College of Radiology. *Breast Imaging Reporting & Data System (BI-RADS®)*; American College of Radiology: Reston, VA, USA, 2013.
49. Wang, C.; Deng, C.; Wang, S. Imbalance-XGBoost: Leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. *Pattern Recognit. Lett.* **2020**, *136*, 190–197. [CrossRef]
50. Berry, D.A.; Cronin, K.A.; Plevritis, S.K.; Fryback, D.G.; Clarke, L.; Zelen, M.; Mandelblatt, J.S.; Yakovlev, A.Y.; Habbema, J.D.F.; Feuer, E.J. Effect of screening and adjuvant therapy on mortality from breast cancer. *N. Engl. J. Med.* **2005**, *353*, 1784–1792. [CrossRef]
51. Silverstein, M.J.; Skinner, K.A.; Lomis, T.J. Predicting axillary nodal positivity in 2282 patients with breast carcinoma. *World J. Surg.* **2001**, *25*, 767–772. [CrossRef]
52. Ho, C.P.; Gillis, J.E.; Atkins, K.A.; Harvey, J.A.; Nicholson, B.T. Interactive case review of radiologic and pathologic findings from breast biopsy: Are they concordant? How do I manage the results? *Radiographics* **2013**, *33*, E149–E152. [CrossRef]
53. Moloney, B.M.; McAnena, P.F.; Elwahab, S.M.; Fasoula, A.; Duchesne, L.; Gil Cano, J.D.; Glynn, C.; O'Connell, A.; Ennis, R.; Lowery, A.J.; et al. The Wavelia Microwave Breast Imaging system–tumour discriminating features and their clinical usefulness. *Br. J. Radiol.* **2021**, *94*, 20210907. [CrossRef]
54. Meaney, P.M.; Kaufman, P.A.; Muffly, L.S.; Click, M.; Poplack, S.P.; Wells, W.A.; Schwartz, G.N.; di Florio-Alexander, R.M.; Tosteson, T.D.; Li, Z.; et al. Microwave imaging for neoadjuvant chemotherapy monitoring: Initial clinical experience. *Breast Cancer Res.* **2013**, *15*, R35. [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.

Journal of
Clinical Medicine

MDPI

Article

Evaluation of Advanced Artificial Intelligence Algorithms' Diagnostic Efficacy in Acute Ischemic Stroke: A Comparative Analysis of ChatGPT-4o and Claude 3.5 Sonnet Models

Mustafa Koyun [1,*] and Ismail Taskent [2]

[1] Department of Radiology, Kastamonu Training and Research Hospital, Kastamonu 37150, Turkey
[2] Department of Radiology, Kastamonu University, Kastamonu 37150, Turkey; dr.taskent@gmail.com
* Correspondence: dr.mustafakoyun@gmail.com

Abstract: Background/Objectives: Acute ischemic stroke (AIS) is a leading cause of mortality and disability worldwide, with early and accurate diagnosis being critical for timely intervention and improved patient outcomes. This retrospective study aimed to assess the diagnostic performance of two advanced artificial intelligence (AI) models, Chat Generative Pre-trained Transformer (ChatGPT-4o) and Claude 3.5 Sonnet, in identifying AIS from diffusion-weighted imaging (DWI). **Methods:** The DWI images of a total of 110 cases (AIS group: n = 55, healthy controls: n = 55) were provided to the AI models via standardized prompts. The models' responses were compared to radiologists' gold-standard evaluations, and performance metrics such as sensitivity, specificity, and diagnostic accuracy were calculated. **Results:** Both models exhibited a high sensitivity for AIS detection (ChatGPT-4o: 100%, Claude 3.5 Sonnet: 94.5%). However, ChatGPT-4o demonstrated a significantly lower specificity (3.6%) compared to Claude 3.5 Sonnet (74.5%). The agreement with radiologists was poor for ChatGPT-4o (κ = 0.036; %95 CI: −0.013, 0.085) but good for Claude 3.5 Sonnet (κ = 0.691; %95 CI: 0.558, 0.824). In terms of the AIS hemispheric localization accuracy, Claude 3.5 Sonnet (67.2%) outperformed ChatGPT-4o (32.7%). Similarly, for specific AIS localization, Claude 3.5 Sonnet (30.9%) showed greater accuracy than ChatGPT-4o (7.3%), with these differences being statistically significant ($p < 0.05$). **Conclusions:** This study highlights the superior diagnostic performance of Claude 3.5 Sonnet compared to ChatGPT-4o in identifying AIS from DWI. Despite its advantages, both models demonstrated notable limitations in accuracy, emphasizing the need for further development before achieving full clinical applicability. These findings underline the potential of AI tools in radiological diagnostics while acknowledging their current limitations.

Keywords: ChatGPT; Claude; artificial intelligence; acute ischemic stroke; magnetic resonance imaging; radiology

1. Introduction

Stroke is defined by the American Heart Association/American Stroke Association (AHA/ASA) as a neurological deficit secondary to the acute focal injury of the central nervous system (CNS) caused by conditions such as cerebral infarction, intracerebral hemorrhage (ICH), and subarachnoid hemorrhage (SAH) [1]. It is reported that stroke is more common in elderly individuals, with the average age being 69 years [2]. Acute ischemic stroke (AIS) accounts for 65% of all stroke cases worldwide, with this proportion reaching up to 87% in some studies [3,4]. AIS, a significant cause of disability and mortality worldwide, affects approximately 57.93 million people annually [5]. However, AIS is a treatable

condition when addressed in its early stages [6]. Patients in the post-stroke period typically experience physical movement limitations, communication disorders, and cognitive function losses [6]. These multiple impairments cause patients to require comprehensive rehabilitation and intensive medical care. When AIS is intervened in the early period, it is a treatable condition, and, therefore, its early detection is critically important [7]. AIS can be treated with a medical approach called "thrombolysis" using alteplase (also known as tissue Plasminogen Activator (tPA)) or Tenecteplase (TNK) [2]. Another treatment option for AIS is mechanical thrombectomy, and studies are being conducted aiming to reduce the door-to-puncture time (DTPT) to enable rapid intervention [8].

While various radiological imaging techniques can be used to detect AIS, diffusion-weighted magnetic resonance imaging (DW-MRI) is the most accurate method [9]. However, recent studies suggest that non-contrast computed tomography (NCCT) and computed tomography angiography (CTA), which can be performed more quickly than DW-MRI, should be primarily implemented to detect AIS and rule out intracranial hemorrhage (ICH) so that treatment can be initiated without delay [2,8].

Artificial intelligence (AI) is a sophisticated branch of technology that mimics human intelligence and has brought about revolutionary changes across various fields. This technology can generate original content in multiple forms, including text, images, music, and video. Applicable in nearly every sector—from healthcare to finance, education to transportation—AI stands out for its ability to solve complex problems, perform data analysis, and make predictions. The use of AI in healthcare covers a broad area [10]. CardioAI, which analyzes cardiac magnetic resonance images, and computer-aided diagnosis (CAD) systems used in endoscopy–colonoscopy procedures [11] can be given as examples of AI applications in healthcare.

AI technology is divided into various specialized subdisciplines, including machine learning (ML), deep learning (DL), and computer vision [11]. ML identifies patterns using specific features and optimizes decision-making processes by applying this knowledge to similar situations. This technology has evolved into DL, which utilizes artificial neural networks (ANNs) mimicking human neural systems. Image processing technology encompasses the digital perception and analysis of visual data [11]. Two of the most prominent examples of AI are Chat Generative Pre-trained Transformer (ChatGPT) (OpenAI, San Francisco, CA, USA) and Claude (Anthropic, San Francisco, CA, USA), which have made groundbreaking advancements in the field of natural language processing. ChatGPT, developed by OpenAI and introduced on 30 November 2022, is one of the large language models (LLMs) capable of generating human-like text, responding to complex questions, and leveraging an extensive knowledge base [12]. On the other hand, Claude, launched by Anthropic in March 2023, is an LLM designed with a particular emphasis on ethics and safety, incorporating a more controlled and transparent approach [13]. Both models excel in interacting naturally with users, providing information on a wide range of topics, and generating creative content. ChatGPT-4o (ChatGPT-4 Omni) and Claude 3.5 Sonnet models have been recently developed and offer the ability to evaluate visual materials.

The integration of AI technology into radiological imaging systems holds significant potential for accelerating disease diagnosis processes, enhancing workflow efficiency, and minimizing errors caused by human factors [14]. Recent studies demonstrate the effectiveness of DL applications in detecting various pathologies from radiological imaging data [15–17]. DL applications are also coming to the forefront in AIS detection (Table 1) [18–21]. It is reported that sensitivity and specificity values above 80% have been achieved with these AI applications. However, there are only a few studies on the detection of pathologies from radiological images using ChatGPT [22–25], and, among these, only the study by Kuzan et al. focused on AIS detection from DW-MRI images [25].

However, to the best of our knowledge, no studies have been conducted on the detection of pathologies from radiological images using Claude.

Table 1. Summary of diagnostic performance metrics for deep-learning-based AI models in stroke diagnosis.

Author and Year	AI Model Used	DL-Based Approaches	Sample Size	Main Findings
Shinohara et al., 2020 [18]	Deep learning	DCNN	702 patients	AUC: 0.87 Accuracy: 0.82 Sensitivity: 0.82 Specificity: 0.81
Cui et al., 2021 [19]	Deep learning	DeepSym-3D-CNN	190 patients	AUC: 0.86 Accuracy: 0.85 Sensitivity: 0.85 Specificity: 0.84
Abedi et al., 2021 [20]	Deep learning	ANN	260 patients	Precision: 0.92 Sensitivity: 0.80 Specificity: 0.86
Zhang et al., 2018 [21]	Deep learning	3D-CNN	242 patients	Dice similarity coefficient: 0.79 Precision: 0.92

AI: artificial intelligence; ANN: artificial neural network; DCNN: deep convolutional neural network; AUC: area under the curve.

The aim of this study is to investigate the detectability of AIS from DW-MRI images using the two most popular AI models, ChatGPT-4o and Claude 3.5 Sonnet, and to compare their diagnostic performance.

2. Materials and Methods

2.1. Selection of Cases

A total of 1256 patients with DW-MRI images recorded in our hospital's information management system between May 2024 and September 2024 were retrospectively analyzed. These patients underwent systematic evaluation. Accordingly, patients over 18 years of age who presented to the emergency department with neurological symptoms suggestive of AIS (hemiparesis, aphasia, facial asymmetry, diplopia, and vision loss) and underwent DW-MRI on the same day with preliminary diagnosis of AIS were identified. Among the screened patients, 55 cases with diffusion restriction on DW-MRI were classified as the AIS group (Figure 1). Cases with clinical findings but no diffusion restriction on DW-MRI were not included in the AIS group (Figure 1).

The control group consisted of 55 healthy cases who underwent DW-MRI for various indications but showed no diffusion restriction (Figure 1).

The exclusion criteria for both groups were standardized as follows: the presence of artifacts limiting image evaluation, age under 18 years, a history of intracranial mass or recent cranial surgery, intracranial hemorrhage, or the presence of metallic objects (Figure 1). Cases were determined by consensus between two radiologists with 8 and 12 years of experience.

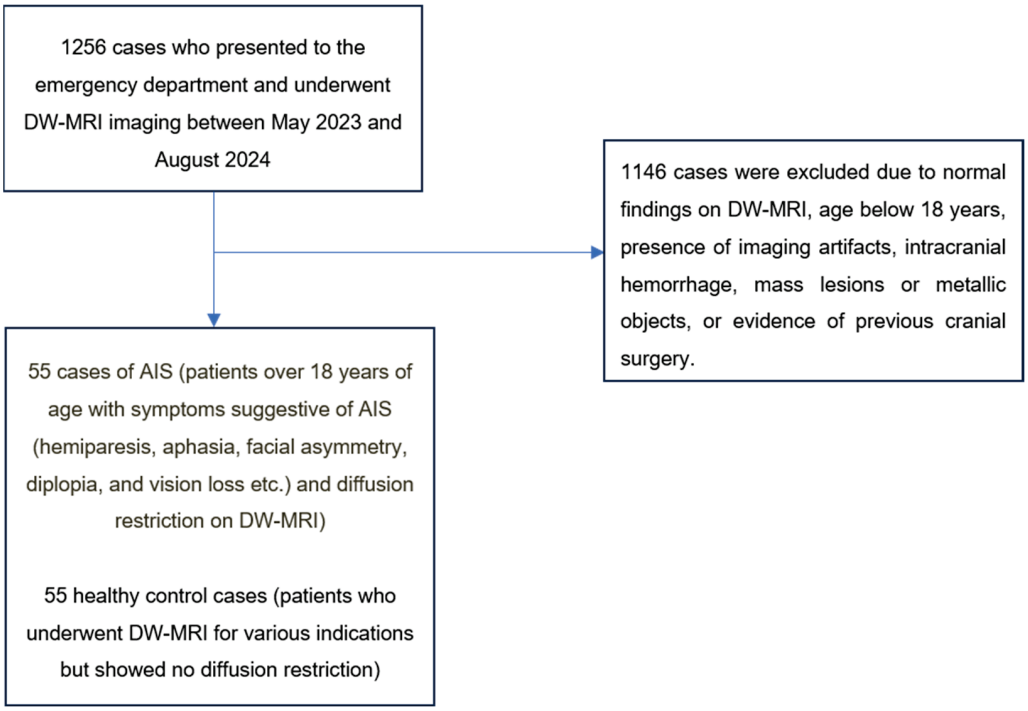


Figure 1. Patient selection process.

2.2. Brain DWI and ADC İmages

The brain DWI and ADC images for all cases were obtained using a 1.5T MRI scanner (GE Signa Explorer; GE Healthcare, Milwaukee, WI, USA).

2.3. AI Models Used in the Study

In our study, we used ChatGPT-4o (GPT-4 Omni), developed by OpenAI (OpenAI, San Francisco, CA, USA) and introduced on 13 May 2024, and Claude 3.5 Sonnet model, developed by Anthropic (Anthropic, San Francisco, CA, USA) and introduced on 20 June 2024, to analyze DW-MRI images.

2.4. Prompt Prepared for AI Models

A structured three-question prompt was developed for the images presented to the ChatGPT-4 and Claude 3.5 Sonnet models as follows:

1. In these DWI and ADC images, is there an acute ischemic stroke? Please answer with "Yes" or "No".
2. If there is an acute ischemic stroke, what is the localization according to the patient? Please answer with "Right cerebral hemisphere", "Left cerebral hemisphere", "Right cerebellar hemisphere", or "Left cerebellar hemisphere".
3. If there is acute ischemia, could you specify the affected specific lobe or region?

2.5. Image Selection, and Preparation and Presentation to Models

Image slice selection criteria for AIS cases were standardized with a methodological approach. DWI sequence slices showing the ischemic lesion at its maximum size and optimal visibility, along with their corresponding ADC maps, were designated as reference images for primary analysis. Parameters such as similar distribution numbers of anatomical lesions or similar types of clinical presentations were disregarded in image selection.

The selection of image slices for the healthy control group was systematically performed to show anatomical correlation with the reference slices in the AIS group. The spatial localization of image slices was standardized to match the slice levels in AIS cases as

closely as possible. This methodological approach aimed to ensure that AI models analyzed equivalent anatomical structures and localizations in both groups.

The original Digital Imaging and Communications in Medicine (DICOM) format DWI and ADC images of the cases were first anonymized using AW Volumeshare 7 software (AW version 4.7, GE Healthcare, Milwaukee, WI, USA). Subsequently, the images were converted to JPEG format through this software without altering basic parameters such as resolution, brightness, and contrast. All images were exported while preserving the original pixel size. During conversion, high-level JPEG quality was preferred to preserve diagnostically important details as much as possible. Later, DWI and ADC images were combined into a single image using Microsoft Paint software (Microsoft, Corp., Redmond, WA, USA), and irrelevant regions outside the brain were cropped from the newly created image. Additionally, orientation verification was performed by radiologists on the images before uploading them to the models. To minimize bias, the cases were presented to the models in a randomized order, accompanied by the three-question prompt mentioned earlier. The responses provided by the models were recorded (Figure 2). Additionally, to assess intra-model reliability, all cases were reanalyzed with the same models two weeks later.

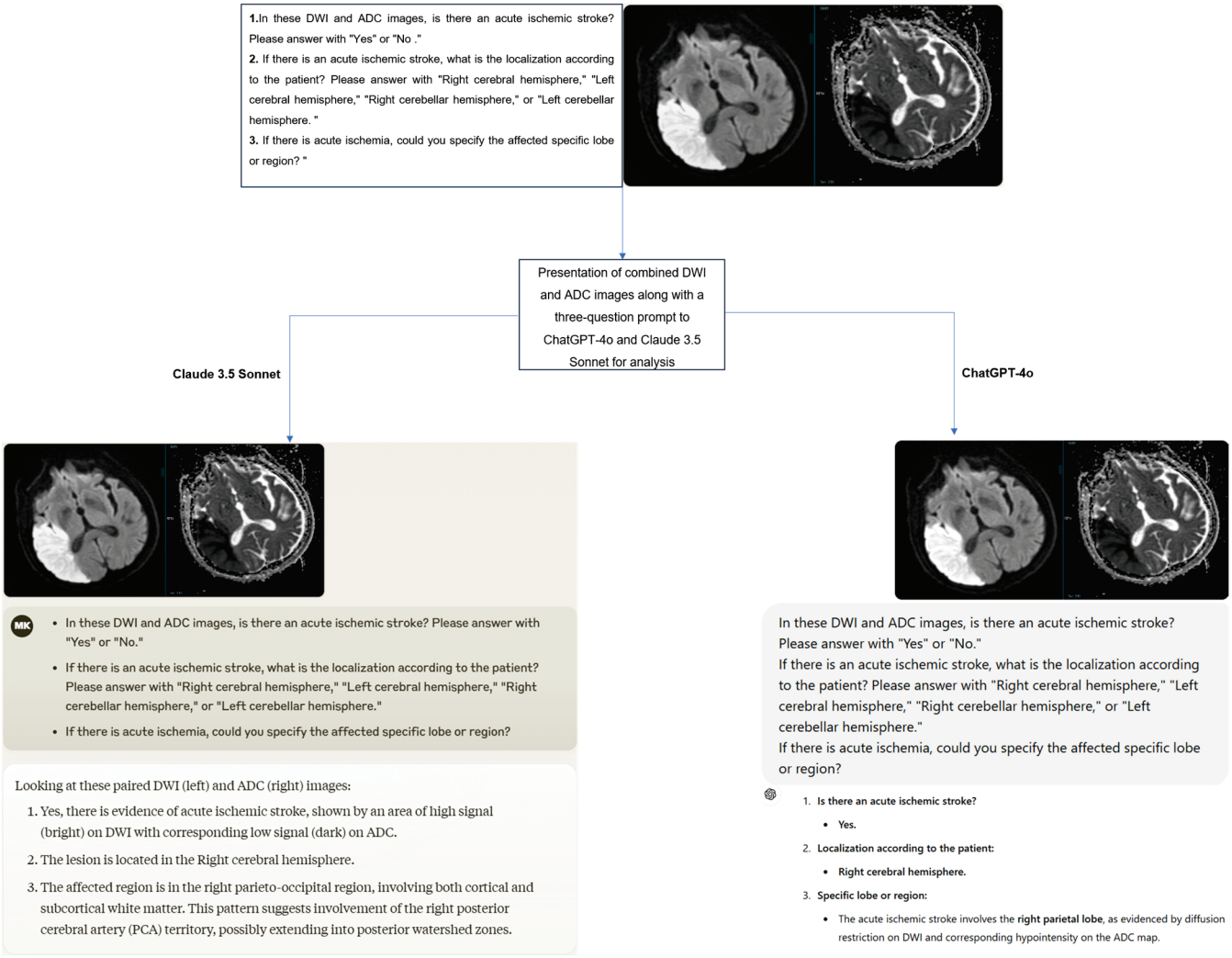


Figure 2. The process of presenting and evaluating combined DWI and ADC images in a single image with a three-question prompt to the models.

2.6. Evaluation of Outputs

To ensure standardization during the evaluation process, preliminary assessments were conducted on sample cases not included in the study to establish consensus on evaluation criteria among the radiologists. The response categorization criteria were defined as follows: a "correct" response required full agreement between the model's answers and the radiologists' reports, while a "wrong" response was defined as a mismatch in key findings or the presence of diagnostically critical errors.

Throughout the evaluation process, radiologists conducted their work using medical monitors on diagnostic imaging workstations.

2.7. Statistical Analysis

Statistical analyses were conducted using IBM SPSS 23 software (IBM Corporation, Armonk, NY, USA). The normality of data distribution was assessed with the Kolmogorov–Smirnov test. An Independent-Samples *t*-test was used to compare age between groups, while a Chi-square test was applied for gender distribution comparisons. Quantitative data with normal distribution were presented as mean and standard deviation (Mean ± SD), and categorical data were reported as frequency (n) and percentage (%).

The imaging analysis performance of the models was evaluated in comparison to radiologist assessments (gold standard). The performance of ChatGPT-4o and Claude 3.5 Sonnet was assessed by calculating sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and diagnostic accuracy rates. The agreement between the models and radiologists, as well as inter- and intra-model reliability, was assessed using Cohen's kappa coefficient.

The interpretation of the kappa coefficient was as follows: $\leq$0.20 as "slight agreement", 0.21–0.40 as "fair agreement", 0.41–0.60 as "moderate agreement", 0.61–0.80 as "substantial agreement", and 0.81–1.00 as "almost perfect agreement". The effect of the age variable on the models' AIS detection performance was evaluated using binary logistic regression analysis, while the effect of the gender variable on model performance was assessed using the Pearson Chi-square test. A binomial test was used to compare the models' predictions with random guessing. Statistical significance was set at $p < 0.05$.

2.8. Ethical Approval

This retrospective study was conducted in accordance with ethical guidelines and was approved by the Clinical Research Ethics Committee of Kastamonu University on 10 December 2024, under approval number 2024-KAEK-134. The study adhered to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) guidelines [26]. Due to the retrospective nature of the study, written informed consent was not obtained from the patients.

3. Results

3.1. Results Obtained from Demographic Data

A total of 110 cases were included in the study, with 55 cases in the AIS group and 55 in the healthy control group. Both groups consisted of 25 females (45.5%) and 30 males (54.5%). The mean age of the AIS group was 73.11 ± 11.43 years, while the mean age of the healthy control group was 72.75 ± 10.75 years. No significant differences were found between the AIS and healthy control groups in terms of age or gender ($p > 0.05$) (Table 2).

Table 2. Demographic data of AIS group and healthy control group cases.

	AIS Group (n = 55)	Healthy Control Group (n = 55)	p-Value
Gender ¶, n (%)			
Female	25 (45.5%)	25 (45.5%)	1.0
Male	30 (54.5%)	30 (54.5%)	
Age *, years, Mean ± SD	73.11 ± 11.43	72.75 ± 10.75	0.90

¶ Chi-square Test; * Independent-Samples t-test; AIS: acute ischemic stroke; SD: standard deviation.

3.2. Diagnostic Performance of the Models

In the diagnostic performance analysis of ChatGPT-4o for AIS detection, all cases in the AIS group (100%) yielded true positive results. However, in the control group, only two cases (3.6%) produced true negative results, while 53 cases (96.4%) resulted in false positive outcomes (Table 3).

Table 3. Results of ChatGPT-4o and Claude 3.5 Sonnet in assessing the presence of AIS.

	True Positive (n)	False Positive (n)	True Negative (n)	False Negative (n)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Diagnostic Accuracy (%)	F1 Score (%)
ChatGPT-4o	55	53	2	0	100	3.6	50.9	100	51.8	67.5
Claude 3.5 Sonnet	52	14	41	3	94.5	74.5	78.8	93.2	84.5	85.9

PPV: positive predictive value; NPV: negative predictive value.

For Claude 3.5 Sonnet, the diagnostic performance for AIS detection showed 52 cases (94.5%) with true positive results and 3 cases (5.5%) with false negative results in the AIS group. In the control group, 41 cases (74.5%) produced true negative results, while 14 cases (25.5%) resulted in false positive outcomes (Table 3).

Both models demonstrated a high sensitivity for AIS detection (ChatGPT-4o: 100%, Claude 3.5 Sonnet: 94.5%), but specificity values differed significantly (ChatGPT-4o: 3.6%, Claude 3.5 Sonnet: 74.5%). The diagnostic accuracy of Claude 3.5 Sonnet (84.5%) was higher compared to ChatGPT-4o (51.8%). The binomial test revealed that ChatGPT-4o's accuracy in detecting AIS (51.8%) was not statistically different from random guessing (50%) (p = 0.775). In contrast, the same test demonstrated that Claude 3.5 Sonnet's accuracy in detecting AIS was 84.5%, which is significantly higher than the random guessing rate of 50% (p = 0.000). The positive predictive value (PPV), negative predictive value (NPV), and F1 score for both models are presented in Table 3.

Both models answered all the questions in the prompt consistently with each other. When the models responded that AIS was present, they answered the other questions accordingly, and, when they responded that AIS was not present, they left the other questions unanswered. The number of cases where ChatGPT-4o answered all questions with complete accuracy was 4 (7.3%), while, for Claude 3.5 Sonnet, it was 17 (30.9%) (Table 4).

Table 4. Comparative analysis of full-accuracy responses given by ChatGPT-4o and Claude 3.5 Sonnet to all questions in the AIS group.

	Completely Correct, n (%)	Incorrect or at Least One Answer Correct, n (%)	Total, n (%)
ChatGPT-4o	4 (7.3)	51 (92.7)	55 (100)
Claude 3.5 Sonnet	17 (30.9)	38 (69.1)	55 (100)

The AIS detection performance of both models does not show any statistically significant differences between genders ($p > 0.05$). Likewise, no statistically significant differences were found between age and the models' AIS detection performance ($p > 0.05$).

3.3. Agreement Analysis Between Models and Radiologists

ChatGPT-4o demonstrated slight agreement with the radiologists' gold-standard assessment (κ = 0.036; %95 CI: −0.013, 0.085). In contrast, Claude 3.5 Sonnet showed substantial agreement with the radiologists (κ = 0.691; %95 CI: 0.558, 0.824) (Table 5).

Table 5. Evaluation of agreement between ChatGPT-4o and Claude 3.5 Sonnet with radiologists in detecting the presence of AIS.

		Radiologists		
		Negative (n)	**Positive (n)**	**Total (n)**
ChatGPT-4o	Negative (n)	2	0	2
	Positive (n)	53	55	108
	Total (n)	55	55	110
Cohen's Kappa	(κ = 0.036; %95 CI: −0.013, 0.085)			
Claude 3.5 Sonnet	Negative (n)	41	3	44
	Positive (n)	14	52	66
	Total (n)	55	55	110
Cohen's Kappa	(κ = 0.691; %95 CI: 0.558, 0.824)			

3.4. Performance of the Models in Detecting Hemispheric Localization of AIS

ChatGPT-4o correctly identified the hemisphere of AIS localization in 18 out of 55 cases (32.7%) in the AIS group (16 in the right cerebral hemisphere and 2 in the left cerebral hemisphere). However, in the healthy control group, it incorrectly made localization predictions in 53 out of 55 cases (96.4%) despite the absence of AIS (Table 6).

Table 6. ChatGPT-4o's performance in hemispheric localization assessment.

AIS Localization	Correct Answer (n)	Incorrect Answer (n)	Total (n)
No AIS	2	53	55
Right cerebral	16	14	30
Left cerebral	2	21	23
Both cerebral	0	1	1
Left cerebellar	0	1	1
Total (n)	20	90	110

In terms of hemispheric localization, ChatGPT-4o demonstrated slight agreement with radiologists (κ = 0.077; %95 CI: −0.151, −0.003).

Claude 3.5 Sonnet correctly identified the hemisphere of AIS localization in 37 out of 55 cases (67.3%) in the AIS group (17 in the right cerebral hemisphere and 20 in the left cerebral hemisphere). However, in the healthy control group, it incorrectly made localization predictions in 14 out of 55 cases (25.5%) despite the absence of AIS (Table 7).

Table 7. Claude 3.5 Sonnet's performance in hemispheric localization assessment.

AIS Localization	Correct Answer (n)	Incorrect Answer (n)	Total (n)
No AIS	41	14	55
Right cerebral	17	13	30
Left cerebral	20	3	23
Both cerebral	0	1	1
Left cerebellar	0	1	1
Total (n)	78	32	110

In detecting AIS localization by hemisphere, Claude 3.5 Sonnet demonstrated moderate agreement with the gold-standard radiologists (κ = 0.572; %95 CI: 0.456, 0.688). Additionally, Claude 3.5 Sonnet was statistically significantly more successful than ChatGPT-4o in accurately identifying AIS localization by hemisphere (p = 0.0006).

3.5. Performance of the Models in Detecting Specific AIS Localization

ChatGPT-4o correctly identified the specific localization of AIS in only 4 cases (7.3%) within the AIS group, while making incorrect determinations in the remaining 51 cases (92.7%). In contrast, Claude 3.5 Sonnet accurately identified the specific localization in 17 cases (30.9%) within the AIS group, but made incorrect determinations in 38 cases (69%) (Table 8).

Table 8. Specific lobe-region localizations of AIS accurately identified by ChatGPT-4o and Claude 3.5 Sonnet.

ChatGPT-4o	Correct Answer (n)	Claude 3.5 Sonnet	Correct Answer (n)
Right parietal lobe	2	Right frontal	2
Right periventricular WM	1	Right parietal	5
Left frontal lobe	1	Right temporal	1
		Right temporo-occipital	1
		Left parietal	3
		Left occipital	1
		Left parieto-occipital	2
		Left temporo-occipital	1
		Left thalamus	1
Total (n)	4		17

WM: White matter.

In detecting the specific lobe or region of AIS localization, ChatGPT-4o demonstrated slight agreement with radiologists (κ = 0.026; %95 CI: −0.015, 0.067), while Claude 3.5 Sonnet showed fair agreement (κ = 0.365; %95 CI: 0.277, 0.453). Claude 3.5 Sonnet was found to be statistically significantly more successful than ChatGPT-4o in accurately identifying the specific lobe or region of AIS localization (p = 0.0016).

3.6. Second Round of Model Evaluations and Intra-Model Agreement Analysis

After a two-week interval, the DWI-ADC images were re-evaluated by the models. In the second evaluation, ChatGPT-4o again showed 100% true positive results for all AIS cases. In the healthy control group, 48 cases (87.3%) resulted in false positives, while 7 cases (12.7%) produced true negative results (Table 9). Claude 3.5 Sonnet correctly identified 46 cases (83.6%) as true positive and 9 cases (16.4%) as false negative in the AIS group. In the healthy control group, 46 cases (83.6%) were identified as true negative, and 9 cases (16.4%) were false positive (Table 7).

Table 9. First and second round evaluation results of ChatGPT-4o and Claude 3.5 Sonnet in detecting the presence of AIS.

	True Positive (n)	False Positive (n)	True Negative (n)	False Negative (n)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Diagnostic Accuracy (%)
1. ChatGPT-4o	55	53	2	0	100	3.6	50.9	100	51.8
2. ChatGPT-4o	55	48	7	0	100	12.7	53.4	100	56.4
1. Claude 3.5 Sonnet	52	14	41	3	94.5	74.5	78.8	93.2	84.5
2. Claude 3.5 Sonnet	46	9	46	9	83.6	83.6	83.6	83.6	83.6

In the second evaluation, ChatGPT-4o showed an increase in specificity, positive predictive value (PPV), and diagnostic accuracy (Table 9). On the other hand, Claude 3.5 Sonnet showed a decrease in sensitivity, negative predictive value (NPV), and diagnostic accuracy, while the specificity and PPV values increased (Table 9).

Moderate agreement was found between the two evaluations for both ChatGPT-4o and Claude 3.5 Sonnet in detecting the presence of AIS ($\kappa = 0.428$; %95 CI: 0.026, 0.830 for ChatGPT-4o, $\kappa = 0.582$; %95 CI: 0.433, 0.731 for Claude 3.5 Sonnet).

In the assessment of the hemispheric localization of AIS, fair agreement was found between the first and second round results of ChatGPT-4o ($\kappa = 0.259$; %95 CI: 0.100, 0.418), while moderate agreement was observed between the first and second round results of Claude 3.5 Sonnet ($\kappa = 0.524$; %95 CI: 0.404, 0.644).

In the evaluation of the AIS localization in specific lobes or regions, moderate agreement was found between the first and second round results of ChatGPT-4o ($\kappa = 0.443$; %95 CI: 0.343, 0.543), while fair agreement was found between the results of Claude 3.5 Sonnet ($\kappa = 0.384$; %95 CI: 0.284, 0.484).

4. Discussion

In this study, we evaluated the ability of the ChatGPT-4o and Claude 3.5 Sonnet AI models to detect the presence and localization of AIS from DWI and ADC images. To the best of our knowledge, this is the first study in the literature that uses both ChatGPT-4o and Claude 3.5 Sonnet for AIS detection.

In this study, while Claude 3.5 Sonnet achieved a high diagnostic accuracy rate (84.5%), ChatGPT-4o had a low diagnostic accuracy rate (51.8%). No statistically significant difference was found between ChatGPT-4o's predictions and random guessing ($p > 0.05$). This suggests that ChatGPT-4o was making random predictions in image evaluation. In the study by Kuzan et al., ChatGPT showed a sensitivity of 79.57% and specificity of 84.87% in stroke detection, whereas, in our study, ChatGPT-4o exhibited a higher sensitivity (100%) but a significantly lower specificity (3.6%) [25]. Kuzan et al. excluded lacunar infarcts smaller than 1 cm in their study [25]. However, in our study, AIS cases of all sizes were included, which could affect the sensitivity and specificity values of the models. There are no existing literature data to compare the performance of Claude 3.5 Sonnet, and it is anticipated that our study will serve as a reference for future research.

In detecting the hemispheric localization of AIS, Claude 3.5 Sonnet (67.3%) demonstrated a statistically significantly better performance than ChatGPT-4o (32.3%) ($p = 0.0006$). ChatGPT-4o's low hemispheric localization success (32.3%) is similar to the findings in Kuzan et al.'s study (26.2%) [25]. Both models showed the lowest success in detecting the affected specific lobe or region. However, Claude 3.5 Sonnet demonstrated a statistically significantly higher performance compared to ChatGPT-4o (30.9% vs. 7.3%, $p = 0.0016$). In Kuzan et al.'s study, the lowest success rate (20.4%) was also observed in detecting the specific lobe or region [25]. These results indicate that Claude 3.5 Sonnet outperformed

ChatGPT-4o in localizing AIS. Additionally, our study reveals that both models show performance limitations in situations requiring complex assessments. The poor performance of models in detecting AIS localizations significantly limits their clinical use at present. In a clinical setting, the accurate localization of findings is crucial for determining correct diagnosis, treatment, and management strategies. Localization errors—for example, a false positive in an area without pathology or a false negative in an area with pathology—can adversely affect patient outcomes by leading to incorrect or delayed clinical interventions.

In terms of detecting the presence of AIS, ChatGPT-4o showed slight agreement with the radiologists' gold-standard assessments ($\kappa = 0.036$; %95 CI: -0.013, 0.085), while Claude 3.5 Sonnet demonstrated substantial agreement ($\kappa = 0.691$; %95 CI: 0.558, 0.824). Both ChatGPT-4o and Claude 3.5 Sonnet showed moderate agreement in their internal evaluations regarding the presence of AIS ($\kappa = 0.428$; %95 CI: 0.026, 0.830 and $\kappa = 0.582$; %95 CI: 0.433, 0.731, respectively). When assessing the hemispheric localization of AIS, ChatGPT-4o demonstrated fair internal agreement ($\kappa = 0.259$; %95 CI: 0.100, 0.418), while Claude 3.5 Sonnet showed moderate internal agreement ($\kappa = 0.524$; %95 CI: 0.404, 0.644). In evaluating the specific lobe or region localization of AIS, ChatGPT-4o exhibited moderate internal agreement ($\kappa = 0.443$; %95 CI: 0.343, 0.543), whereas Claude 3.5 Sonnet showed fair internal agreement ($\kappa = 0.384$; %95 CI: 0.284, 0.484). These findings indicate that the AI models demonstrate some level of consistency in AIS detection, with Claude 3.5 Sonnet showing relatively higher internal consistency. However, these results also suggest that both models need further development to optimize their diagnostic consistency.

LLMs like ChatGPT and Claude have stochastic output generation mechanisms [25]. This characteristic leads to different results being produced in repeated processes for the same input. This stochastic nature has the potential to enhance user experience by increasing diversity and dynamism in human–machine interactions. However, this becomes a significant limiting factor when applied in areas that require precision, such as medical image analysis. Since consistency and repeatability are critical in medical diagnosis and interpretation processes, this structural feature of the models limits their reliability and validity in clinical applications. Both models produced false positive and false negative results, and the stochastic nature of the models may be one of the reasons for these outcomes. Additionally, the lack of specific training for medical image analysis and the limited image processing capabilities of these models could also contribute to the incorrect results.

The use LLMs, especially ChatGPT, in radiology is steadily increasing. In the study by Akinci D'Antonoli and colleagues, both the benefits and challenges of using language models in radiology are highlighted [27]. Recent literature shows that these models are capable of tasks such as data mining from free-text radiology reports [28], generating structured reports [29], answering disease-related questions [30], and responding to radiology-board-style exam questions [31]. Furthermore, recent studies suggest that evaluating radiological images with these models may be feasible [23,24,32]. LLMs could assist radiologists in image assessment and clinicians in diagnosing at centers without radiologists, but it should be noted that these models are still in the early stages and can make errors.

Our study demonstrates that LLMs, such as ChatGPT and Claude, still face significant limitations in their application to medical image analysis. It was observed that neither model, particularly ChatGPT, achieved sufficient diagnostic sensitivity and accuracy in detecting AIS. ChatGPT's notably high false-positive results in the healthy control group and its predictions resembling random guesses pose a risk of unnecessary additional testing and treatment, raising significant clinical concerns.

Nevertheless, our study highlights the potential of these models in detecting AIS from DW-MRI images, contributing significantly to the theoretical and practical knowledge base in the fields of AI and neuroradiology. Despite their current inability to achieve the

desired success in medical image analysis, we anticipate that technological advancements will enable these models to reach high success rates in the coming years, playing a crucial role in optimizing the increasing workload of radiologists. In healthcare centers with insufficient radiologist staffing, healthcare personnel with limited clinical experience could utilize these LLMs for the preliminary evaluation of DW-MRI images, thereby expediting diagnostic processes. This innovative technology could also serve as a secondary control mechanism for junior radiologists during their professional development, minimizing the risk of overlooking critical findings and enhancing diagnostic reliability. Additionally, in studies implementing optimized protocols aimed at reducing door-to-puncture time (DTPT) in conditions requiring early intervention, such as AIS [8], AI models like LLMs could be incorporated into the image evaluation process by relevant specialists. However, it must be acknowledged that these models are still in their early stages and are prone to errors.

AI has found a wide range of applications in the healthcare sector, from drug development processes to remote patient monitoring [14]. While the use of AI in healthcare services offers innovative solutions in medical diagnosis and treatment, this also brings serious ethical concerns. The use of AI in healthcare is based on nine fundamental ethical principles: accountability, autonomy, equity, integrity, non-maleficence, privacy, security, transparency, and trust [33]. The protection of fundamental ethical principles such as autonomy, beneficence, non-maleficence, justice, and privacy is particularly critical in AI applications [34]. AI systems' influence on medical diagnosis and treatment decisions can affect the informed decision-making processes of patients and healthcare professionals and may limit autonomy. Furthermore, biases in AI algorithms can lead to unfair healthcare services, while non-transparent decision-making processes can cause trust issues. In this context, adopting principles of explainability and transparency, protecting patient privacy, and clarifying responsibility mechanisms should form the cornerstones of ethical AI development and implementation processes in the integration of AI into healthcare services.

The integration of AI technologies into healthcare services is directly related to public attitudes. Despite its developmental potential in this field, the widespread adoption of AI applications requires social acceptance. Trust in the collection of health data and the use of AI in diagnosis and treatment processes is critical for the integration of these technologies into the healthcare system [10]. Recent Pew Research Center surveys show that a large segment of society has concerns about the standalone use of AI in healthcare services [35]. Additionally, surveys reveal that, while public interest in AI use in the healthcare sector is increasing, certain concerns persist [35]. Furthermore, surveys indicate that, while public interest in AI use in the healthcare sector is increasing, certain concerns persist [35]. These concerns are thought to stem from an unfamiliarity with AI technologies, a lack of knowledge, and concerns about data security [10].

Our study has some limitations. The first is that not all DWI and ADC slices from the patients were uploaded to the models. Not analyzing all slices of patients may lead to missing important details, especially in cases where the lesion is small or located in peripheral regions. It may also prevent the complete mapping of the lesion, which can reduce the diagnostic sensitivity of the models. Particularly in cases where the lesion boundaries are not clear or show heterogeneous density, missing slices may lead to false negative results. Additionally, not analyzing all slices may make it difficult to determine the exact localization of the lesion. If the entire set of images had been uploaded, the diagnostic accuracy of both models might have increased.

The second limitation of the study is the necessity of converting the original radiological images in DICOM format to JPEG or PNG formats to enable the analysis by both models. This conversion process may potentially lead to the loss of diagnostically critical

image details and metadata, which could negatively affect the analytical capacity of the AI models and compromise the reliability of the results [24].

The third limitation of our study is that only two widely used LLMs (ChatGPT-4o and Claude 3.5 Sonnet) were evaluated for their performance, while others were not assessed. Future studies should evaluate the performance of specialized medical large vision language models (LVLMs) such as BiomedCLIP and LLaVA-Med, which could potentially demonstrate superior diagnostic capabilities in detecting AIS on DW-MRI images, as well as other emerging LLMs.

Compared to deep-learning-based studies, our small sample size can be considered another limitation of our study. Future studies could evaluate the performance of the models using larger sample sizes.

Another limitation of our study is the potential selection bias that may have occurred during the image selection process for DWI and ADC. The specific inclusion and exclusion criteria applied by the radiologists could have impacted the representativeness of the sample. Finally, the fact that the models were not specifically trained for medical image analysis can be considered another limitation of our study.

5. Conclusions

In light of the findings of our study, advanced language models such as ChatGPT and Claude AI can be considered as potential supportive tools in detecting the presence of AIS from DWI and ADC images. The low accuracy rates in localization detection highlight the need for the further development of these models in this regard. The performance demonstrated by these models complements traditional radiological evaluation processes and represents an innovative approach that could be integrated into clinical decision-making mechanisms. However, before these technologies are routinely used in clinical practice, they need to be validated and optimized through larger-scale, multicenter studies. Future studies should evaluate the overall diagnostic performance of these models using larger and more diverse datasets, and particularly focus on improving anatomical localization accuracy.

Author Contributions: Conceptualization, M.K.; methodology, M.K. and I.T.; validation, M.K. and I.T.; formal analysis, M.K.; investigation, M.K.; resources, M.K. and I.T.; data curation, M.K. and I.T.; writing—original draft preparation, M.K.; writing—review and editing, M.K. and I.T.; visualization, M.K.; supervision, I.T.; project administration, M.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 was approved by the Clinical Research Ethics Committee of Kastamonu University, Turkey (Decision No.: 2024-KAEK-134, Date: 10 December 2024).

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

Data Availability Statement: The data presented in this study are available upon request from the corresponding author. The data are not publicly available due to privacy.

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

References

1. Sacco, R.L.; Kasner, S.E.; Broderick, J.P.; Caplan, L.R.; Connors, J.J.; Culebras, A.; Elkind, M.S.; George, M.G.; Hamdan, A.D.; Higashida, R.T.; et al. An updated definition of stroke for the 21st century: A statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* **2013**, *44*, 2064–2089. [CrossRef] [PubMed]

2. Mirmozaffari, M.; Kamal, N. A data envelopment analysis model for optimizing transfer time of ischemic stroke patients under endovascular thrombectomy. *Healthc. Anal.* **2024**, *6*, 100364. [CrossRef]
3. Nentwich, L.M. Diagnosis of acute ischemic stoke. *Emerg. Med. Clin. N. Am.* **2016**, *34*, 837–859. [CrossRef] [PubMed]
4. Saini, V.; Guada, L.; Yavagal, D.R. Global epidemiology of stroke and access to acute ischemic stroke interventions. *Neurology* **2021**, *97* (Suppl. S2), S6–S16. [CrossRef]
5. Faletti, D.O.; Fakayode, O.O.; Adedara, V.O.; Kuteyi, A.O.; Adedara, C.A.; Ogunmoyin, T.E.; Chen, J.C.; Olasimbo, O.; Aina, S.A.; Alozie, G.U.; et al. Comparative efficacy and safety of thrombectomy versus thrombolysis for large vessel occlusion in acute ischemic stroke: A systemic review. *Cureus* **2024**, *16*, e72323. [CrossRef]
6. Pawluk, H.; Kołodziejska, R.; Grześk, G.; Kozakiewicz, M.; Woźniak, A.; Pawluk, M.; Kosinska, A.; Grześk, M.; Wojtasik, J.; Kozera, G. Selected mediators of inflammation in patients with acute ischemic stroke. *Int. J. Mol. Sci.* **2022**, *23*, 10614. [CrossRef]
7. Kurz, K.D.; Ringstad, G.; Odland, A.; Advani, R.; Farbu, E.; Kurz, M.W. Radiological imaging in acute ischaemic stroke. *Eur. J. Neurol.* **2016**, *23*, 8–17. [CrossRef]
8. Kansagra, A.P.; Wallace, A.N.; Curfman, D.R.; McEachern, J.D.; Moran, C.J.; Cross, D.T., III; Lee, J.M.; Ford, A.L.; Manu, S.G.; Panagos, P.D.; et al. Streamlined triage and transfer protocols improve door-to-puncture time for endovascular thrombectomy in acute ischemic stroke. *Clin. Neurol. Neurosurg.* **2018**, *166*, 71–75. [CrossRef]
9. Kunst, M.M.; Schaefer, P.W. Ischemic stroke. *Radiol. Clin.* **2011**, *49*, 1–26. [CrossRef]
10. Bekbolatova, M.; Mayer, J.; Ong, C.W.; Toma, M. Transformative potential of AI in healthcare: Definitions, applications, and navigating the ethical landscape and public perspectives. *Healthcare* **2024**, *12*, 125. [CrossRef]
11. Kaul, V.; Enslin, S.; Gross, S.A. History of artificial intelligence in medicine. *Gastrointest. Endosc.* **2020**, *92*, 807–812. [CrossRef] [PubMed]
12. ChatGPT. Available online: https://chatgpt.com/ (accessed on 18 December 2024).
13. Claude. Available online: https://claude.ai/new (accessed on 18 December 2024).
14. Hosny, A.; Parmar, C.; Quackenbush, J.; Schwartz, L.H.; Aerts, H.J. Artificial intelligence in radiology. *Nat. Rev. Cancer* **2018**, *18*, 500–510. [CrossRef] [PubMed]
15. Kundisch, A.; Hönning, A.; Mutze, S.; Kreissl, L.; Spohn, F.; Lemcke, J.; Sitz, M.; Sparenberg, P.; Goelz, L. Deep learning algorithm in detecting intracranial hemorrhages on emergency computed tomographies. *PLoS ONE* **2021**, *16*, e0260560. [CrossRef] [PubMed]
16. Dar, R.A.; Rasool, M.; Assad, A. Breast cancer detection using deep learning: Datasets, methods, and challenges ahead. *Comput. Biol. Med.* **2022**, *149*, 106073.
17. Lakshmipriya, B.; Pottakkat, B.; Ramkumar, G. Deep learning techniques in liver tumour diagnosis using CT and MR imaging-a systematic review. *Artif. Intell. Med.* **2023**, *141*, 102557. [CrossRef]
18. Shinohara, Y.; Takahashi, N.; Lee, Y.; Ohmura, T.; Umetsu, A.; Kinoshita, F.; Kuya, K.; Kato, A.; Kinoshita, T. Usefulness of deep learning-assisted identification of hyperdense MCA sign in acute ischemic stroke: Comparison with readers' performance. *Jpn. J. Radiol.* **2020**, *38*, 870–877. [CrossRef]
19. Cui, L.; Han, S.; Qi, S.; Duan, Y.; Kang, Y.; Luo, Y. Deep symmetric three-dimensional convolutional neural networks for identifying acute ischemic stroke via diffusion-weighted images. *J. X-Ray Sci. Technol.* **2021**, *29*, 551–566. [CrossRef]
20. Abedi, V.; Goyal, N.; Tsivgoulis, G.; Hosseinichimeh, N.; Hontecillas, R.; Bassaganya-Riera, J.; Elijovich, L.; Metter, J.E.; Alexandrov, A.W.; Liebeskind, D.S.; et al. Novel screening tool for stroke using artificial neural network. *Stroke* **2017**, *48*, 1678–1681. [CrossRef]
21. Zhang, R.; Zhao, L.; Lou, W.; Abrigo, J.M.; Mok, V.C.; Chu, W.C.; Wang, D.; Shi, L. Automatic segmentation of acute ischemic stroke from DWI using 3-D fully convolutional DenseNets. *IEEE Trans. Med. Imaging* **2018**, *37*, 2149–2160. [CrossRef]
22. Haver, H.L.; Bahl, M.; Doo, F.X.; Kamel, P.I.; Parekh, V.S.; Jeudy, J.; Yi, P.H. Evaluation of multimodal ChatGPT (GPT-4V) in describing mammography image features. *Can. Assoc. Radiol. J.* **2024**, *75*, 947–949. [CrossRef]
23. Mert, S.; Stoerzer, P.; Brauer, J.; Fuchs, B.; Haas-Lützenberger, E.M.; Demmer, W.; Giunta, R.E.; Nuernberger, T. Diagnostic power of ChatGPT 4 in distal radius fracture detection through wrist radiographs. *Arch. Orthop. Trauma Surg.* **2024**, *144*, 2461–2467. [CrossRef] [PubMed]
24. Dehdab, R.; Brendlin, A.; Werner, S.; Almansour, H.; Gassenmaier, S.; Brendel, J.M.; Nikolaou, K.; Afat, S. Evaluating ChatGPT-4V in chest CT diagnostics: A critical image interpretation assessment. *Jpn. J. Radiol.* **2024**, *42*, 1168–1177. [CrossRef] [PubMed]
25. Kuzan, B.N.; Meşe, İ.; Yaşar, S.; Kuzan, T.Y. A retrospective evaluation of the potential of ChatGPT in the accurate diagnosis of acute stroke. *Diagn. Interv. Radiol.* 2024, *online ahead of print*. [CrossRef]
26. Mongan, J.; Moy, L.; Kahn, C.E., Jr. Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers. *Radiol. Artif. Intell.* **2020**, *2*, e200029. [CrossRef]
27. D'Antonoli, T.A.; Stanzione, A.; Bluethgen, C.; Vernuccio, F.; Ugga, L.; Klontzas, M.E.; Cuocolo, R.; Cannella, R.; Koçak, B. Large language models in radiology: Fundamentals, applications, ethical considerations, risks, and future directions. *Diagn. Interv. Radiol.* **2024**, *30*, 80–90. [CrossRef]
28. Fink, M.A.; Bischoff, A.; Fink, C.A.; Moll, M.; Kroschke, J.; Dulz, L.; Heußel, C.P.; Kauczor, H.U.; Weber, T.F. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. *Radiology* **2023**, *308*, e231362. [CrossRef]

29. Adams, L.C.; Truhn, D.; Busch, F.; Kader, A.; Niehues, S.M.; Makowski, M.R.; Bressem, K.K. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: A multilingual feasibility study. *Radiology* **2023**, *307*, e230725. [CrossRef]
30. Rahsepar, A.A.; Tavakoli, N.; Kim GH, J.; Hassani, C.; Abtin, F.; Bedayat, A. How AI responds to common lung cancer questions: ChatGPT versus Google Bard. *Radiology* **2023**, *307*, e230922. [CrossRef]
31. Bhayana, R.; Krishna, S.; Bleakney, R.R. Performance of ChatGPT on a radiology board-style examination: Insights into current strengths and limitations. *Radiology* **2023**, *307*, e230582. [CrossRef]
32. Haver, H.L.; Yi, P.H.; Jeudy, J.; Bahl, M. Use of ChatGPT to Assign BI-RADS Assessment Categories to Breast Imaging Reports. *Am. J. Roentgenol.* **2024**, *223*, e2431093. [CrossRef]
33. Ning, Y.; Teixayavong, S.; Shang, Y.; Savulescu, J.; Nagaraj, V.; Miao, D.; Mertens, M.; Ting, D.S.W.; Ong, J.C.L.; Liu, M.; et al. Generative artificial intelligence and ethical considerations in health care: A scoping review and ethics checklist. *Lancet Digit. Health* **2024**, *6*, e848–e856. [CrossRef] [PubMed]
34. Galiana, L.I.; Gudino, L.C.; González, P.M. Ethics and artificial intelligence. *Rev. Clínica Española* **2024**, *224*, 178–186. [CrossRef]
35. Tyson, A.; Pasquini, G.; Spencer, A.; Funk, C. *60% of Americans Would Be Uncomfortable with Provider Relying on AI in Their Own Health Care*; Technical Report; Pew Research Center: Washington, DC, USA, 2023.

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

diagnostics

MDPI

Review

Artificial Intelligence-Empowered Radiology—Current Status and Critical Review

Rafał Obuchowicz [1], Julia Lasek [2,*], Marek Wodziński [3], Adam Piórkowski [4], Michał Strzelecki [5] and Karolina Nurzynska [6]

[1] Department of Diagnostic Imaging, Jagiellonian University Medical College, 30-663 Krakow, Poland; gechrzas@cyf-kr.edu.pl
[2] Faculty of Geology, Geophysics and Environmental Protection, AGH University of Krakow, 30-059 Krakow, Poland
[3] Department of Measurement and Electronics, AGH University of Krakow, 30-059 Krakow, Poland; wodzinski@agh.edu.pl
[4] Department of Biocybernetics and Biomedical Engineering, AGH University of Krakow, 30-059 Krakow, Poland
[5] Institute of Electronics, Lodz University of Technology, 93-590 Lodz, Poland; michal.strzelecki@p.lodz.pl
[6] Department of Algorithmics and Software, Silesian University of Technology, 44-100 Gliwice, Poland; karolina.nurzynska@polsl.pl
* Correspondence: julial@agh.edu.pl

Abstract: Humanity stands at a pivotal moment of technological revolution, with artificial intelligence (AI) reshaping fields traditionally reliant on human cognitive abilities. This transition, driven by advancements in artificial neural networks, has transformed data processing and evaluation, creating opportunities for addressing complex and time-consuming tasks with AI solutions. Convolutional networks (CNNs) and the adoption of GPU technology have already revolutionized image recognition by enhancing computational efficiency and accuracy. In radiology, AI applications are particularly valuable for tasks involving pattern detection and classification; for example, AI tools have enhanced diagnostic accuracy and efficiency in detecting abnormalities across imaging modalities through automated feature extraction. Our analysis reveals that neuroimaging and chest imaging, as well as CT and MRI modalities, are the primary focus areas for AI products, reflecting their high clinical demand and complexity. AI tools are also used to target high-prevalence diseases, such as lung cancer, stroke, and breast cancer, underscoring AI's alignment with impactful diagnostic needs. The regulatory landscape is a critical factor in AI product development, with the majority of products certified under the Medical Device Directive (MDD) and Medical Device Regulation (MDR) in Class IIa or Class I categories, indicating compliance with moderate-risk standards. A rapid increase in AI product development from 2017 to 2020, peaking in 2020 and followed by recent stabilization and saturation, was identified. In this work, the authors review the advancements in AI-based imaging applications, underscoring AI's transformative potential for enhanced diagnostic support and focusing on the critical role of CNNs, regulatory challenges, and potential threats to human labor in the field of diagnostic imaging.

Keywords: radiology; artificial intelligence; regulatory systems; AI applications; job threat

1. Introduction

Humanity currently stands at the threshold of a technological revolution, in which the application of decision-making systems based on microprocessors represents a transformation on par with the invention of the wheel or the harnessing of fire [1–3]. The

ability to transfer decision-making processes concerning the interpretation and outcomes of human-related data—traditionally reserved for the human brain—to artificial neural network systems marks a profound shift [4–6], representing a significant breakthrough in data processing and evaluation methods [7].

Decision-making in semiconductor-based neural networks is, to some extent, determined by humans as the problem-solving approach relies on patterns within the training data [8,9]. It is worth noting that deeper, multilayered networks have the capacity to develop autonomous decision pathways, gradually constructed based on their own iterative experiences through exposure to input data [10,11]; this process can be opaque, often resulting in solutions that surpass human intuition, as illustrated in the notable example of AI algorithms outperforming humans in the game of Go [12]. Given these capabilities, AI not only offers solutions to repetitive and time-intensive tasks but also presents novel approaches to long-standing problems, a potential that explains AI's extensive applications in medicine [13].

AI applications are apparent in fields such as materials biology, biochemistry, and genetics where multidimensional data analysis leads to breakthroughs in creating unique structures and compounds valuable for biomaterials and pharmaceuticals [14]. AI's utility is particularly evident in diagnostic imaging where physicians engage in the detection and classification of patterns; these tasks are time-consuming for specialized medical professionals and, due to the repetitive nature of pattern recognition, are prime candidates for AI-powered decision systems [15,16]. Consequently, developing such systems has become a priority within the scientific community, as reflected in the increasing number of algorithms aimed at detecting various anomalies in medical images [17,18]; these algorithms primarily rely on convolutional neural network architectures trained on specific image patterns [19].

The vast diversity in data encountered in modern imaging diagnostics presents substantial challenges. Algorithms must be tailored to specific tissues, organs, and imaging modalities, such as computed tomography, magnetic resonance imaging, mammography, and X-ray; this diversity is reflected in the wide array of algorithms designed to analyze various regions of the human body [20,21]. Depending on an algorithm's validated effectiveness and integration capabilities with radiological systems, it may or may not receive endorsements from certification bodies [22,23].

In this review, the authors present an overview of existing algorithms and outline the most critical techniques used in the classification of medical imaging data, subdivided in this article into the following sections: historical perspective, deep learning solution overview, human vs. machine interaction consideration, job risk overview, and current medical solutions review.

We aim to provide a comprehensive analysis of the advancements and challenges associated with AI applications in radiology, specifically examining the evolution of convolutional neural networks (CNNs) as foundational tools in diagnostic imaging, evaluating the regulatory frameworks influencing AI integration into clinical workflows, and exploring the potential impacts of these technologies on radiologists' roles and healthcare delivery. By synthesizing the current state of AI in radiology in this work, we seek to bridge existing knowledge gaps, propose directions for future research, and offer actionable insights for stakeholders in the medical imaging field.

This manuscript is organized as follows: In Section 2, we provide a historical background, highlighting key milestones in AI development and their impacts on medical imaging. In Section 3, we delve into contemporary deep learning models, discussing their architecture, applications, and limitations. In Section 4, we explore the comparative efficiency of AI and human radiologists, shedding light on their complementary roles in diagnostic workflows. In Section 5, we address concerns regarding the potential displacement

of radiologists and emphasize the importance of balanced human–AI collaboration. In Section 6, we examine the critical role of data preparation in enhancing AI model reliability; then, in Section 7, we introduce the significance of textural analysis in medical imaging. In Sections 8 and 9, we discuss emerging trends, challenges, and the current market landscape for AI in radiology. Finally, we conclude the manuscript with a summary of our findings and recommendations for future research.

2. Historical Background

The success of AI in image recognition is primarily driven by advancements in convolutional neural networks (CNNs), traced back to the application of GPU technology and modifications to the dropout technique, first implemented in 2012, which helped secure a win in the ImageNet contest. The ImageNet contest was conceived as a benchmark for evaluating the most effective approaches to image recognition, initially comprising over 10 million images across more than 22,000 categories. The breakthrough, in 2012, by Geoffrey Hinton's students, Ilya Sutskever and Alex Krizhevsky, marked a pivotal moment in computer vision and image recognition when these researchers, initially working in speech recognition, achieved significantly better results than other competitors using two NVIDIA GTX 580 GPUs (NVIDIA Corporation, Santa Clara, CA, USA), each with 3 GB memory, to train a model with 60 million parameters over 90 epochs, completing the process in two days and presenting significantly reduced error rates in comparison to their competitors [24,25].

The success of the above approach catalyzed the broad adoption of GPUs in deep learning, a shift that represented a turning point in computer vision, in which GPUs demonstrated their efficiency in increasing the effectiveness of computer-driven image recognition. The next significant advancement, inspired by AlexNet, was ResNet, which won the 2015 ImageNet competition [26,27]; ResNet's incorporation of residual connections (i.e., calculating output–input differences) enhanced network performance, solidifying CNNs as fundamental architecture in AI-driven image analysis [28]. The developments that followed highlighted the capabilities of CNNs and the power of GPU technology [29–31]. Building on these achievements, the first successful radiology-directed applications emerged, such as deep learning for lung nodule detection in the 2016 LUNA (Lung Nodule Analysis) challenge [32], which served as an early benchmark for the performance of deep learning algorithms in identifying lung nodules, thereby demonstrating the potential of AI in radiology. The introduction of deep learning models, inspired by AlexNet and encouraged in such competitions, paved the way for subsequent programs that rely on object detection within medical imaging, showcasing AI's potential in advancing radiological diagnostics and opening new, transformative possibilities for efficient image analysis in radiology, enabling the recognition of features in radiographs, CT scans, and MRI images; these breakthroughs enabled automated feature extraction and, if broadly used, may substantially improve the accuracy of image analysis in diagnostic imaging. This progress has led to the development of AI tools that could support radiologists by providing second opinions, identifying anomalies, and even predicting disease [33].

3. Deep Learning Models: A Short Introduction to Current Solutions

In radiology, several data analysis aspects should be considered, and some current deep learning solutions are depicted in Figure 1. First, a visual inspection of the myriad imagery data generated with equipment allows for insight into the human body structure. Here, the deep learning models are applied to solve the classification task between healthy and unhealthy patients, sometimes considering, more precisely, the severity stages of an illness; on the other hand, such general information might need to be more comprehensive

as the determination of the region of changes, or simply segmenting the data for further analysis, is expected, in which case, convolutional neural networks are considered. However, the problem is so complex in several cases that training only one model is insufficient, and a pipeline of various models is prepared to achieve the goal.

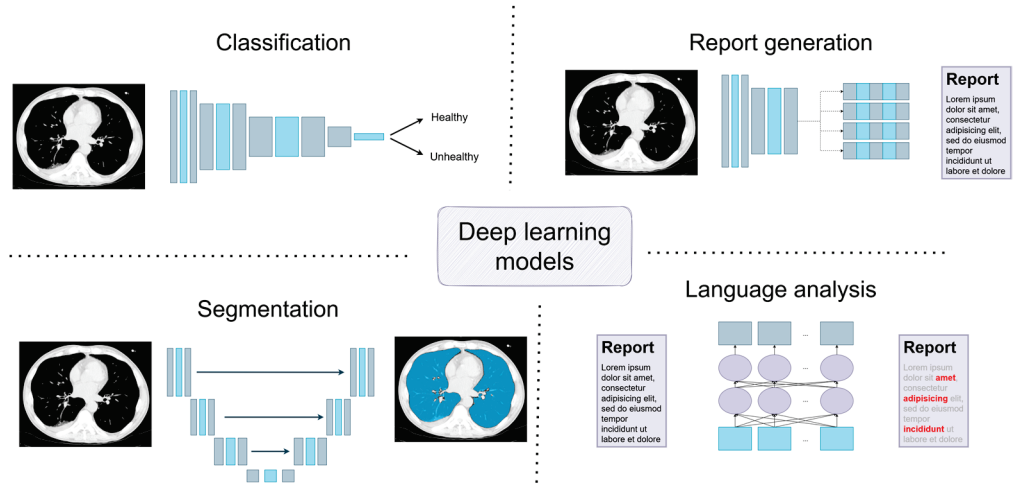


Figure 1. Deep learning models applicable for radiological data processing.

3.1. Classification

When visually analyzing radiographs (X-ray), computed tomography (CT), magnetic resonance imaging (MRI), or ultrasonography scans, convolutional neural networks (CNNs) are considered. ResNet [34] is one such CNN that has gained a lot of popularity; for instance, the smallest version, ResNet-18, used for histopathological images, CT/MRI scans, and genomic data, supports prognosis in clear-cell renal cell carcinoma [35], while the largest version, ResNet-152, was applied to detect pneumonia in chest X-rays [36] and constituted part of a model dedicated to diagnosing and predicting outcomes of COVID-19 pneumonia [37]. A modified ResNet architecture was exploited for attenuation correction in pelvic PET/MR images, significantly reducing voxel-based errors and improving the quantification of bone lesions compared to other existing methods [38]; meanwhile, the 3D version of this architecture helps predict the imaging characteristics, malignancy, and pathological subtypes of pulmonary nodules detected in CT scans [39]. Of course, other architectures also find application; for example, Albiol et al. [40] compared the outcomes of ResNet-50 [34], DenseNet-121 [41], Inception_v3 [42], and Inception-Resnet_v2 [43] against radiologists' interpretations of chest radiographs for the early detection of COVID-19, proving that the deep learning approach overcomes the experts, reaching an AUC of 0.85, compared with 0.71 in the reference group. Additionally, Fink et al. [44] showed that using Xception architecture for musculoskeletal radiographs through projection and body-side resulted in an accuracy of 0.97 in classification tasks.

Not all tasks demand large and pre-trained models; some models achieve better results with small designs but dedicated architectures. Arbabshirani et al. [45] used tailored network architecture to detect intracranial hemorrhage in head CT scans. Nguyen [46] deployed a model for accurately measuring 12 spinal alignment parameters from X-ray images, addressing the manual challenges of spinal misalignment assessment. Solak et al. [47] designed capsule networks, comprising capsules grouping neurons that collectively represent a visual entity, transforming the visual information into more representative features (e.g., height, width) and employing it to classify adrenal lesions in MR images with 0.98 accuracy.

3.2. Segmentation

Beyond the classification of a whole image, it is also possible to determine the detailed region where an organ or tissue of interest is depicted within an image; such a problem in computer science is called semantic segmentation, in which a mask is generated for the whole image, and each pixel is classified as belonging to some data class, consequently delineating specific regions. The encoder–decoder architecture in the U-Net structure is the most widely used for this task; for example, Gasulla et al. [48] used this approach to create a system to assess lung condition severity during the COVID-19 pandemic. Many other approaches concentrate on the segmentation of abdominal organs or the determination of body composition, in the context of muscle and fat distribution, as mentioned by Santhanam et al. [49] in their review. However, such solutions significantly impact medical imaging analysis as there are still imperfections in AI solutions when compared to human ones, as reported by Willemink et al. [50] who also claimed that applying transformer networks [51] should bring about improvements.

Although deep models have proven to be of high quality, more than one model is required in complex tasks as one usually cannot work well to meet several constraints simultaneously; therefore, in such cases, the entire process combines applying various models to achieve a goal. Larson et al. [52] designed a system composed of FastRCNN, based on ResNet-101, for landmark detection, followed by an EfficientNet-D0 model trained to recognize the presence or absence of hardware within extracted joint image patches; this system showed 99% correlation with radiologist measurements of leg length. Nurzynska et al. [53] created a pipeline, composed of two Inception_v3 networks, to determine whether the data were positive or negative for acid-fast (AF) mycobacteria; the first network removed the background, while the second network created a heatmap corresponding to the probability of AF presence. A sequence of three models was applied by Haji Maghsoudi [54] to discriminate breast cancer, in which the first model was responsible for removing the background, the second for finding and removing the pectoralis muscle—with both models based on U-Net [55] architecture—while the last one segmented dense tissue and classified it with the ResNet encoder.

As one can see, a combination of models is necessary when the determination of the main objectives requires the need to learn a system of changes or analyses; thus, usually, the first stages are responsible for removing unnecessary information, allowing the following models to concentrate more effectively on the main task.

Although the model architecture proves the solution's success, it is not the only possibility, nor always the most optimal solution. The training algorithm chosen, especially when labeled data are absent or insufficient [53], as well as the optimization algorithm used while training the network, may influence the final outcome [56].

3.3. Report Generation

The application of large language models (LLM) has recently found its place in the automation of radiologist imagery description analysis. Additionally, AI technology is mature enough, and the amount of gathered data is sufficient to prepare models that automatically generate radiological image descriptions, but this does not place radiologists out of work as AI systems cannot provide a diagnosis. AI should, however, facilitate radiologists' work with automatically generated image descriptions that they should verify and improve in detail; such an approach may improve early diagnosis and limit the radiologist's workload [57]. For example, Zhang et al. [58] developed a generative model to automate the generation of radiological reports from chest X-rays. Blind assessments by radiologists indicated that the generated reports were comparable in quality to those produced by human experts. Another approach was presented by PRAS Bassi et al. [59] who introduced

RadGPT, an anatomy-aware vision-language AI agent designed to generate detailed reports from CT scans. RadGPT segments tumors, including benign cysts and malignant growths, along with surrounding anatomical structures, transforming this information into both structured and narrative reports. These reports provide comprehensive details on tumor size, shape, location, attenuation, volume, and interactions with adjacent blood vessels and organs.

Conversely, Sun et al. [60] evaluated GPT-4's ability to generate the "Impressions" section of radiology reports from the "Findings" section, comparing its outputs to those of human radiologists. Radiologists rated human-generated impressions higher in coherence, comprehensiveness, and factual consistency and concluded that, despite GPT-4's potential to assist in report generation, its outputs currently do not match the quality of human radiologists' work.

3.4. Language Analysis

The functionality of LLMs in radiology is not restricted to generating reports but includes broader applications across the field. For instance, LLMs have shown potential in accurately identifying and classifying true and false laterality errors in radiology reports [61]. Enhancing patient understanding of their condition is crucial for improved outcomes, making the creation of more patient-friendly imaging reports an essential area for LLM implementation. As Butler et al. [62] demonstrated in the context of orthopedic radiology, AI-LLM may improve the readability of radiological reports across multiple imaging modalities. Moreover, LLMs can serve as effective tools for post hoc structured reporting in radiology, enabling significant time savings by automating the organization and structuring of radiological data for improved efficiency and accessibility [63]. LLMs have also been employed to classify unstructured radiology reports into standardized categories. For example, a feasibility study by Matute-González et al. developed LiverAI, a specialized large language model designed to automatically annotate free-text MRI reports with LI-RADS v2018 categories. The findings revealed that incorporating LiverAI into clinical processes could streamline workflows by reducing the radiologists' workload by 45% while maintaining high diagnostic accuracy [64].

These advancements underscore the versatility of LLMs in radiology, highlighting their capacity to improve diagnostic accuracy, patient communication, and workflow efficiency across various applications.

4. Are Machines More Efficient than Human Doctors?

Radiology involves the interpretation and classification of medical images based on characteristic features; a radiologist is trained to distinguish between features considered pathological and normal, identifying abnormalities and grouping them under specific diagnostic categories [65,66]. The process of extracting common features from sets of known images represents a combination of knowledge and expertise, along with the innate pattern recognition skills of the imaging specialist. This ability to recognize shared features across random datasets can be described mathematically and is often viewed as a measure of cognitive intelligence as the capability to achieve specific goals based on these generalizations is a classic marker of system intelligence [67,68]. Machine learning has shown promise in replacing aspects of this process as machines can emulate intelligence by generalizing and identifying similarities across varied models with limited connections because the ability to find similarities across patterns is understood as a background of artificial intelligence [69]. Furthermore, human knowledge dissemination is inherently slow, limited by communication bottlenecks—dictation or writing proceeds at only a few bits per second, even with templates, which remains considerably slower than machine

processing [70]. A significant challenge in radiology is the diversity in data, which vary not only among imaging modalities (e.g., CT, MRI, CR) but also within the same anatomical structures, which can convey differing information, thus requiring a deep understanding of imaging processes and extensive training [71]. Additionally, the vast range of pathologies is challenging to encode in numerical formats that are suitable for machine processing; however, recent advances show that machines with self-learning capabilities can efficiently extract and replicate patterns, demonstrating an impressive ability to recognize features and share learned patterns swiftly [72,73]—progress which has motivated various research groups to create specialized programs designed to extract specific pathological features. However, there is no comprehensive program that can accurately detect all pathological features in any given image (i.e., a generalized AI for pathology detection); currently, radiology relies on a suite of AI tools, each tailored to address a particular problem. Familiarity with these tools is essential for radiologists to work effectively, especially as the gap between available radiology professionals and the volume of medical images continues to grow. The widespread use of AI tools in radiology could enhance diagnostic accuracy, reduce error rates, and improve patient outcomes by standardizing processes and expediting diagnoses.

Starting with connections, radiologists, represented by the human brain, have approximately 80 trillion synaptic links, highlighting their complex neural network that enables deep adaptability and perceptual sensitivity. In contrast, AI models, with around 3 trillion parameters, have fewer connections, which points to their efficiency with structured data but reveals potential limitations in adaptability (Figure 2). Moving to data processing volume, radiologists handle a moderate amount of data in each session, reflecting their focus on quality over quantity and their ability to adapt to diverse cases [74].

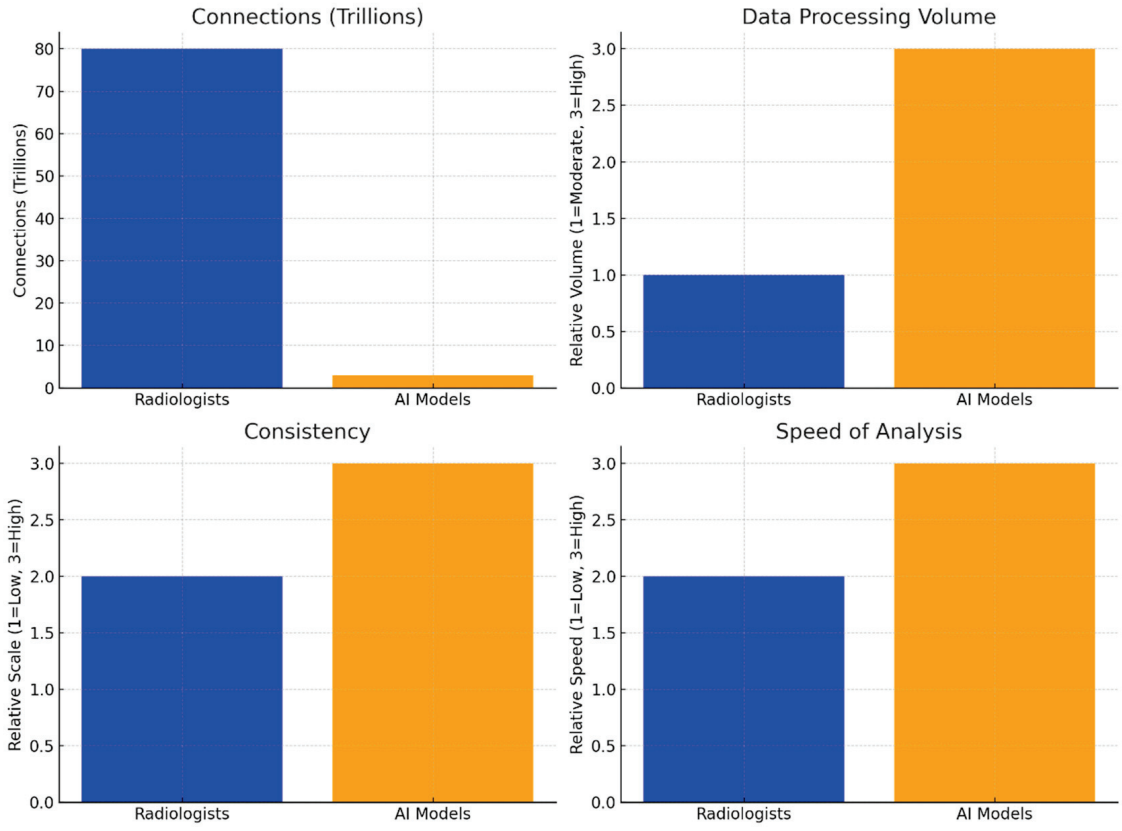


Figure 2. Detailed comparison of human radiologists and AI models, emphasizing the unique strengths each brings to medical imaging tasks.

However, AI models surpass radiologists in their ability to process large volumes of data quickly, which is advantageous for tasks requiring speed and consistency, but AI models may lack the nuanced adaptability seen in human analysis. When it comes to consistency, radiologists show moderate reliability, as their work is often subject to slight variability due to factors such as fatigue and cognitive biases; AI models, on the other hand, demonstrate high consistency, performing uniformly across datasets and maintaining precision without the effects of fatigue or bias [75].

The high consistency of AI models makes them suitable for tasks that require repetitive accuracy though they may not capture subtleties as effectively as humans. In terms of speed of analysis, radiologists typically exhibit a moderate pace, balancing thoroughness with the need for careful evaluation; their work involves a nuanced approach that can be time-consuming but essential for accuracy in complex cases. AI models, in contrast, excel in speed, rapidly processing images in high volumes, which proves useful for scenarios in which rapid diagnostics are critical though it often requires human validation for complex findings [76].

The above comparison highlights the complementary nature of radiologists and AI models: radiologists bring adaptability, perceptual depth, and expertise, while AI models contribute efficiency, speed, and consistent accuracy. Together, these strengths suggest that an integrated approach, combining human insight with AI efficiency, might offer the most comprehensive and effective path forward in medical imaging diagnostics.

5. Is the Job of a Radiologist at Risk?

Machines are often considered superior to humans in solving logical problems; however, in terms of perception—especially in complex medical image recognition—humans still outperform machines [77], and this remains true when we consider a radiologist's ability to adapt to diverse datasets and detect various pathologies [78–80]. The following two primary factors determine the performance of both biological and artificial networks: first, the number and architecture of connections, which dictate the network's ability to capture intricate patterns; and second, the volume and quality of data as better data typically yield better outcomes, a principle often summarized in the phrase "garbage in, garbage out" [81]. Additionally, there are other factors, unique to artificial networks, which are not directly comparable to human brain processes but significantly influence artificial network performance, including training techniques, optimization algorithms, hyperparameter tuning, model generalization, and transfer learning [82,83]. Biological systems, such as the human brain, contain an enormous number of connections—up to 80 trillion synaptic links—whereas large language models typically rely on around 3 trillion parameters [84–86] (Table 1), revealing a fundamental difference between the two: while humans operate with many connections and a relatively low data volume, machines possess vast data volumes but operate with relatively fewer connections. Machine learning models' information extraction abilities are proportional to their capacity to identify similarities across datasets, including visual patterns, which can be measured with data compression tests [69,87]; on the other hand, foundation models trained on extensive datasets demonstrate high proficiency in detecting pathologies and, when well-trained, can outperform human observers in terms of accuracy, sensitivity, consistency, and speed [88], without suffering from fatigue or biases that commonly limit humans [89–91]. Currently, radiologists, equipped with broad and adaptable knowledge, remain the most proficient operators of relevant tools, which addresses the pressing question regarding the future of radiology [77,92,93]; the unique insights and adaptability of radiologists suggest that their expertise will remain vital in integrating AI tools into medical practice [94].

Table 1. Detailed comparison of human radiologists vs. AI models [78–87].

Factor	Radiologists	AI Models
Data Processing Volume	Moderate	High
Connections (Trillions)	80	3
Adaptability	High	Low
Perception of Patterns	High	Moderate
Consistency	Moderate	High
Speed of Analysis	Moderate	High
Fatigue Resistance	No	Yes
Bias Resistance	No	Yes
Training Techniques Required	No	Yes

6. Importance of Data Preparation for Processing with AI

Preparing medical imaging data for machine learning (ML) requires a systematic approach to ensure model reliability, accuracy, and generalizability (Figure 3). First, clear project goals guide data preparation as classification, segmentation, and detection tasks each have unique requirements [95]. Standardization is essential, including converting images to a consistent format (e.g., DICOM or NIfTI) and normalizing resolution and intensity to account for differences across modalities such as X-ray, MRI, or CT [96]. Data cleaning addresses artifacts (e.g., motion blur) and ensures label accuracy, while quality assurance (QA) checks verify noise levels, contrast, and anatomical coverage. Accurate annotation by experienced radiologists is critical, with guidelines established to maintain consistency, especially for complex tasks such as segmentation. Data augmentation techniques (e.g., rotations, flips, contrast adjustments) expand limited datasets, increasing robustness. For data augmentation, techniques that are appropriate to the problem and logical should be used. It is crucial that data augmentation techniques align with the clinical context to avoid introducing unrealistic transformations. For example, while minor rotations (a few degrees left or right) can enhance the robustness of models, extreme rotations such as 90 degrees are inappropriate for X-ray imaging as they do not reflect how such images are typically analyzed by clinicians. After preparation, datasets should be split into training, validation, and test sets while preventing leakage by keeping patient images in only one subset. Balanced datasets are ideal, but, in smaller datasets, k-fold cross-validation can be helpful to maximize data use. Metadata management and privacy safeguards are essential, and patient information should be de-identified to comply with regulations (e.g., HIPAA or GDPR). Finally, pre-training checks, including basic statistical analysis and baseline model evaluation, help identify potential issues before training [97].

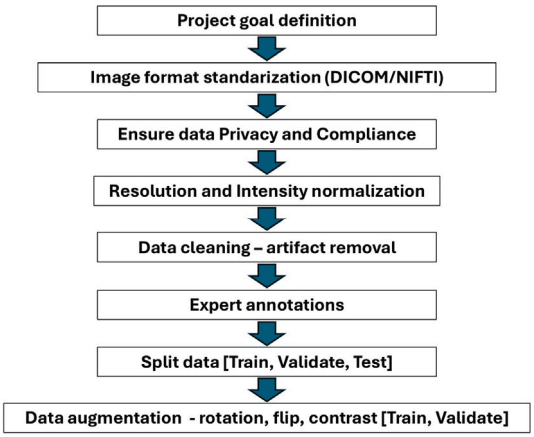


Figure 3. Necessary steps in medical imaging data preparation for processing with AI.

7. The Role of Textural Analysis in Image Preprocessing

Image texture is a crucial component of various image types, including medical images, as medical imaging modalities visualize the properties of internal organs and tissues through textural representation; for instance, the textures observed in tomographic cross-sections provide valuable diagnostic insights. Textural parameters are part of radiomics [98] and reflect the physiological characteristics of tissues, enabling applications such as organ segmentation, lesion detection, and the assessment of pathological changes. The importance of textural analysis in diagnostic imaging has been established across various modalities, including computed tomography (CT) [99], magnetic resonance imaging (MRI) [100], and ultrasound [101].

Furthermore, variations in acquisition parameters across different patient images often affect brightness and contrast in regions of interest (ROIs); these variations can occur between consecutive images, causing some textural features to depend not only on texture but also on factors such as average brightness and contrast. Consequently, features intended to describe tissue structure may inadvertently reflect scanner sensitivity inconsistencies within the analyzed region [102], potentially leading to inaccurate tissue characterization or misclassification.

In order to mitigate these issues, the normalization of ROIs is commonly applied, typically serving as an initial step and expanding the image histogram within the ROI to cover the entire available intensity range, a process that enhances the contrast between bright and dark structural elements within the texture and reduces the influence of local mean intensity, thereby improving the quality of extracted features. Different methods for ROI normalization, as well as their influence to the overall image analysis results, are presented in reference [103].

Texture feature maps also play important roles in the analysis of biomedical images. A feature map is an image of the distribution of values of a given textural parameter, determined for the entire image or its fragment; the value of a feature is determined for each image point in its neighborhood large enough to enable the correct characterization of the texture. A feature map allows us to observe how well a given feature distinguishes the analyzed textures; in the case of a properly selected parameter, a map of the feature transforms the textured areas into relatively uniform fragments differing in brightness (thereby encoding the values of the selected parameter).

The map—or, more often, maps—of features selected during the selection process are used as input data for the segmentation stage. An example feature map for an image of a brain cross-section, containing a fragment of a subarachnoid hemorrhage, is shown in Figure 4 where the map of the AngularSecondMoment feature (calculated based on the Gray Level Cooccurrence Matrix, GLCM) [104] contains the stroke area, which has a significantly different value than the rest of the brain; therefore, the image of this map allows us to isolate the stroke area using simple brightness thresholding.

Another example of using such maps for image segmentation is shown in Figure 5, determined for the SumAverage parameter from GLCM for a 15×15-pixel sliding window across the T1W MR image of a foot cross-section; bone regions are shown on this map as smooth areas with a small range of brightness levels, which allows them to be relatively easily distinguished from other regions. Texture feature maps can also be used for edge detection. Figure 5C,D shows such maps, obtained for the SumOfSquares (GLM) and Sigma (autoregressive model) features, respectively, in which the edges shown define bone tissue regions. An example of a texture map application combined with an active contour model for biomedical image segmentation is discussed in reference [105].

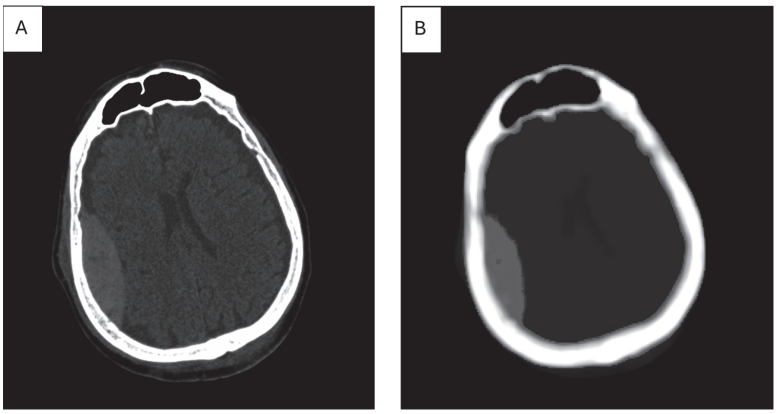


Figure 4. CT image of a brain cross-section with a marked stroke area (**A**); AngularSecondMoment feature map calculated for this image, in which a 15 × 15-pixel window was used (**B**).

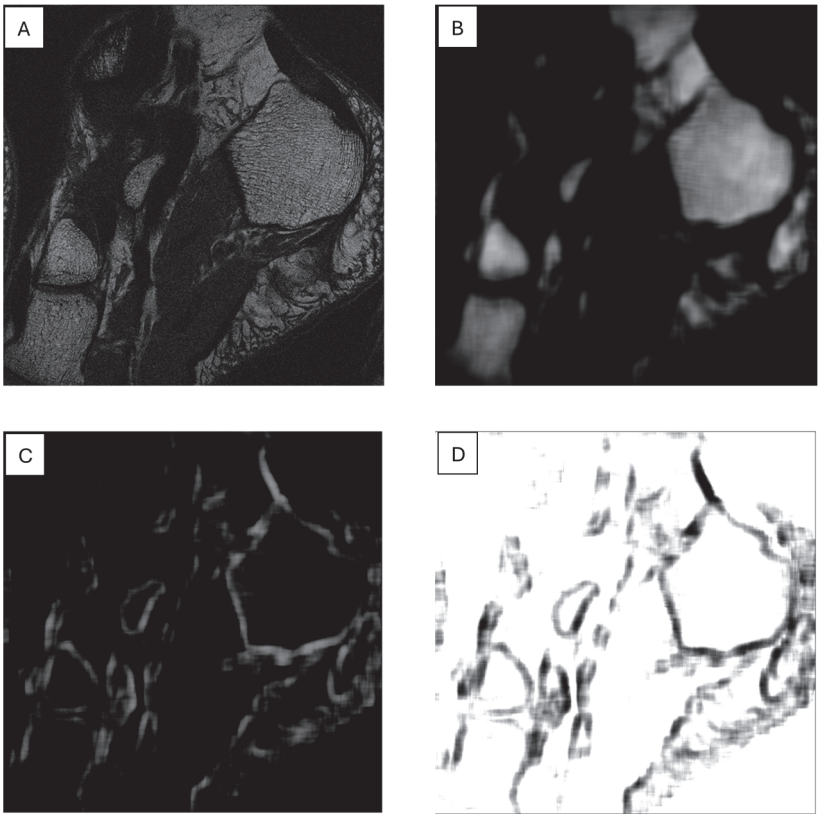


Figure 5. T1W MR image of the foot bone (**A**); feature maps calculated for this image: SumAverage (**B**); SumOfSquares (**C**); and Sigma (**D**).

8. AI Is Supportive but Must Be Used with Caution

Although AI-based solutions have numerous applications in radiology, e.g., they represent state-of-the-art technology used for image segmentation [106,107], there are several applications for which they are suboptimal or should be used with caution.

Firstly, deep learning-based contributions to medical image registration are still less accurate than methods based on classical numerical optimization [108–112]. In an experiment classifying the patient sex using automated analysis of computed tomography scans of vertebrae, standard machine learning, used for the classification of textural features (classical approach), achieved an accuracy of 69%, while deep convolutional networks for this task yielded a slightly lower accuracy of 59% [113]. AI-based algorithms tend to have problems with generalizability to previously unseen cases, requiring further instance-level

optimization in order to achieve optimal results; for example, in the large-scale Learn2Reg benchmark, all of the best-performing methods used classical optimization to further improve the learning-based results for both the brain MRI and abdominal CT tasks [108]. The same observations apply to the BraTS-Reg challenge [109], which was dedicated to comparing different algorithms against pre-operative and post-surgery registration of brain MRIs; all of the best-performing solutions used traditional numerical optimization and outperformed the AI-based contributions. Moreover, classical numerical optimization methods, such as PDE-constrained optimization, have been shown to be highly effective in medical image analysis, providing reliable and high-fidelity results compared to deep learning-based solutions, especially in applications requiring precise alignment, such as neuroimaging and cardiovascular imaging [114]. Therefore, while deep learning-based methods show promise, they still lag behind classical numerical optimization techniques in terms of accuracy and generalizability. Further research is needed to enhance the performance of AI-based algorithms and address their limitations [115].

Another area in which AI methods should be used with caution is image generation and reconstruction; nowadays, people attempt to use AI to generate new synthetic volumes for image augmentation [116] to transfer one modality to another (e.g., to generate synthetic CT from MR to avoid irradiation) [117], to directly create synthetic volumes using textual prompts [118], or to accelerate image reconstruction [119]. Although these applications are becoming increasingly successful, people should use these methods with special care because they often tend to generate non-existing structures that may misguide medical experts [120]. Deep learning models used for MR-to-CT synthesis have demonstrated impressive results, but careful validation is necessary in order to ensure clinical reliability [121]; reconstructed images should be compared with standard sequences to ensure no degradation or unintended alternations in image quality and anatomy have occurred [122].

AI-based contributions are often incorrectly evaluated [123], leading to incorrect conclusions; for example, most of the contributions in automatic image segmentation evaluate accuracy using annotations from just a single annotator rather than reporting confidence intervals [124,125]. However, if the ground truth is based on annotations from human experts, then any evaluation should be always performed using ground truths annotated by several radiologists and inter-variability should be evaluated. Moreover, deep learning methods inherently suffer from problems with standardization and explainability [126], making it difficult to compare results across different studies and ensure consistency in performance metrics [127]. Even though there are significant attempts to improve the interpretability and explainability of deep learning methods, we are still far from clear interpretations of their performed decisions and recommendations [128].

Lastly, one should remember the vulnerability of CNN-based architectures to adversarial attacks, a significant concern in the field of deep learning, as they involve making small, often imperceptible, changes to input images that can cause a CNN to make incorrect predictions. Even a single pixel modification or added learned adversarial pattern to the input may completely fool a deep network [129], a situation that may be dangerous because external modifications to the input volumes, completely invisible to radiologists, may completely change the neural network recommendations. Even though there are methods with which to detect adversarial attacks using feature response maps, the users of CNNs should be aware of the risk [130]; more recent architectures, based on vision transformers (ViTs), are more resistant to certain types of adversarial attacks compared to CNNs [131], but it is important to note that no model is entirely immune to such abuse [132].

9. Review of AI Products Used in Radiology: Status in 2024

The review of AI products used in radiology in 2024 was conducted utilizing the two following key databases: the Health AI Register [133] and AI Central [134]. The Health AI Register lists AI products available in the European market, all of which are CE-marked, indicating compliance with European regulatory standards; conversely, AI Central focuses on AI tools cleared by the FDA and commercially available in the United States. Our dual-database approach ensured comprehensive coverage of the radiology AI landscape, and the review process involved counting and cataloging available products, analyzing their distribution across radiological subspecialties, imaging modalities, and targeted diseases. Regulatory certifications (CE marking and FDA clearance) were systematically reviewed, alongside data on the year of market entry and FDA clearance. This review was conducted in October 2024 to capture the most current state of the field.

The current status of AI applications used in Europe was reviewed on the basis of the Health AI Register [133], which provides the names of applications along with their subspeciality modality and type of certification approved in the EU market. The same system was used for a similar analysis by Van Leuven in 2021 [20]; however, the landscape of the AI market has since significantly changed, and our analysis reflects the rate of development for AI solutions in the healthcare sector, revealing that, as of October 2024, there are 222 commercial AI-based products, representing an increase of 122%—among these, 213 products are reported to be certified, marking a 150% increase compared to the 85 certified products reported in 2021 (Figure 6).

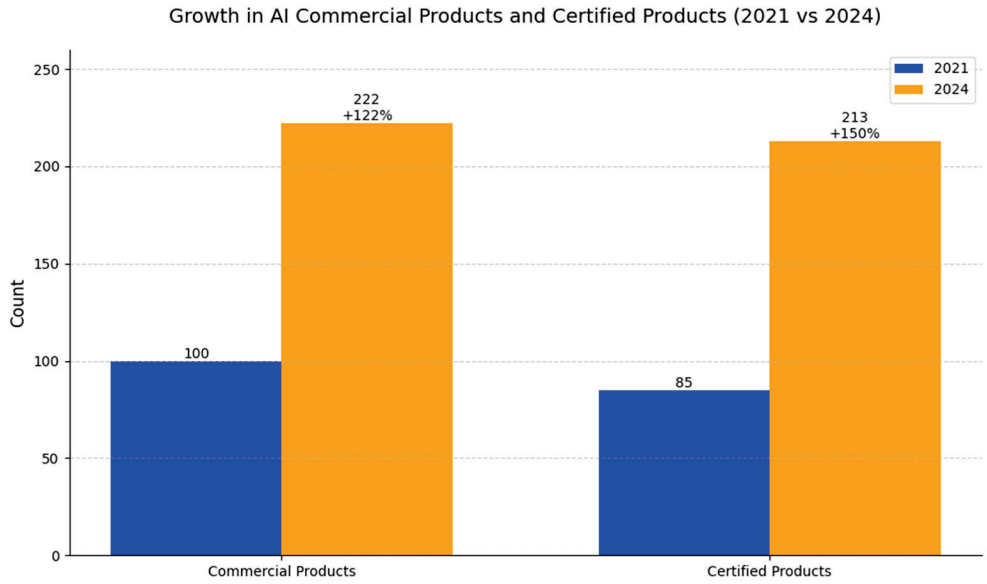


Figure 6. A bar graph illustrating the growth in the number of working AI algorithms and certified products between 2021 and October 2024. The annotations highlight the counts for both years and the percentage increases, emphasizing the rapid expansion of AI solutions in healthcare.

The focus on neuroimaging and chest imaging, with 73 and 71 AI products, respectively, suggests a strong emphasis on developing AI applications in these areas, which may be attributed to high clinical demand and the complexity of interpreting neuroimaging and chest images, making these subspecialties ideal for AI innovation [135,136]. Areas such as musculoskeletal (MSK), abdominal, cardiac, and breast imaging also show considerable development, with product counts between 20 and 28, indicating significant interest and activity; in contrast, subspecialties including vascular, head/neck, spine, thyroid, and FDG PET-CT imaging have relatively few AI products (between 1 and 2), possibly reflecting lower demand, limited dataset availability, or less complex imaging

challenges [137–139]. The above distributions highlight the concentrated development of AI tools in high-demand areas of medical imaging, with the potential for further growth as new needs and opportunities emerge across other subspecialties (Figure 7).

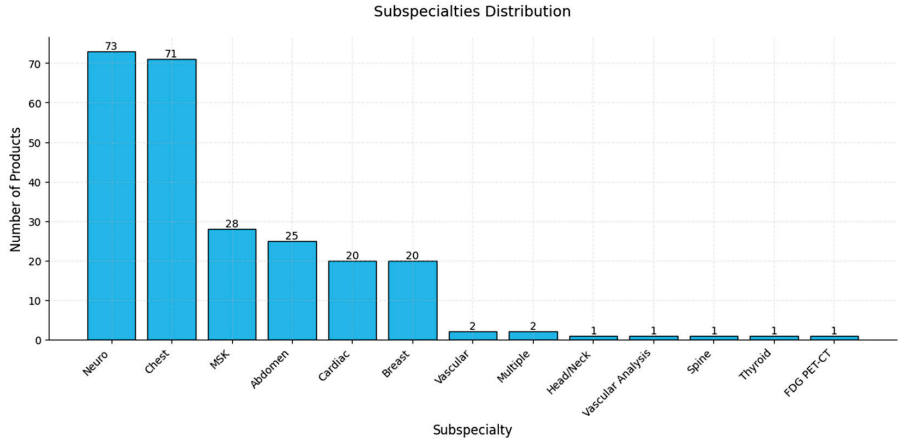


Figure 7. A bar chart illustrating the distribution of AI-based products across various medical subspecialties. The *x*-axis lists the different subspecialties, while the *y*-axis indicates the number of AI products available in each area.

In terms of imaging modalities, CT and MR stand out, with the highest numbers of AI products—89 and 66, respectively—likely due to the extensive use and high diagnostic value of these modalities, as well as the complexity of their data, which makes them particularly well-suited for AI applications [140]. X-ray imaging follows, with 46 products, emphasizing a strong focus on AI in this widely utilized modality [141]; meanwhile, mammography and ultrasound, with 16 and 10 products, respectively, exhibit moderate levels of AI development, likely due to the specialized nature of mammography and the inherent variability of ultrasound imaging, which present unique challenges for AI algorithms. PET and SPECT, with only 3 and 1 products, respectively, represent the lowest levels of AI development among the imaging modalities, possibly due to their lower usage rates and specialized applications in nuclear medicine [142]. Overall, this distribution reflects the alignment of AI product development with high-demand, complex imaging modalities within the healthcare sector (Figure 8).

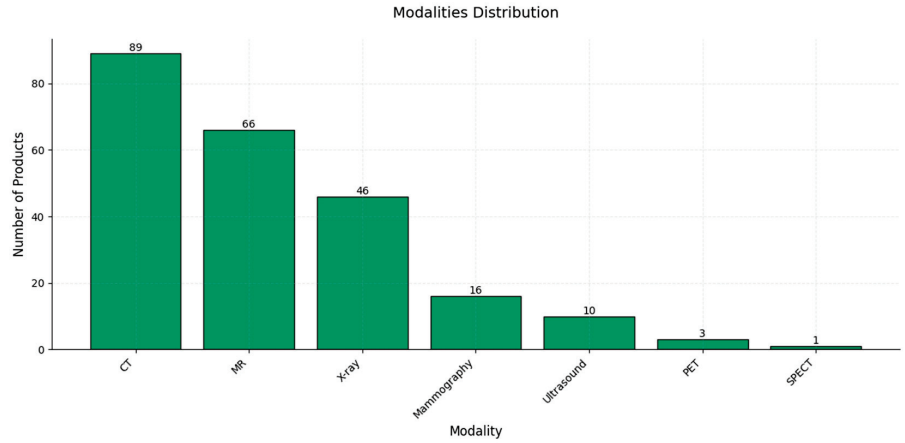


Figure 8. A bar chart illustrating the distribution of AI-based products across various imaging modalities. The *x*-axis represents different imaging modalities—CT, MR, X-ray, mammography, ultrasound, PET, and SPECT—while the *y*-axis shows the number of AI products available for each modality.

Regarding targeted diseases, lung cancer, stroke, and breast cancer lead, with 28, 24, and 19 AI products, respectively, underscoring the prevalence and clinical importance of these conditions [143] and suggesting that the complexity of diagnosis and the high clinical impacts of these diseases make them attractive areas for AI applications. Other notable areas of AI development include pneumothorax, with 16 products, and dementia, with 12 products, indicating growing interest in respiratory and neurodegenerative conditions. Moderate levels of AI product development are seen in diseases such as multiple sclerosis (11 products), emphysema (9 products), prostate cancer (9 products), pleural effusion (9 products), and tuberculosis (9 products), indicating valuable opportunities for AI applications in these areas as well. Diseases including pulmonary embolism and consolidation, pneumonia, COPD, COVID-19, and intracranial hemorrhage, with product counts ranging from 5 to 7, represent the least targeted conditions, potentially reflecting emerging areas in which AI applications are still in the early stages. Targeted disease distribution indicates a strong focus on high-impact, high-prevalence diseases [144,145] while also highlighting the potential for continued growth in AI-driven solutions for a broader range of diseases [146–151] (Figure 9).

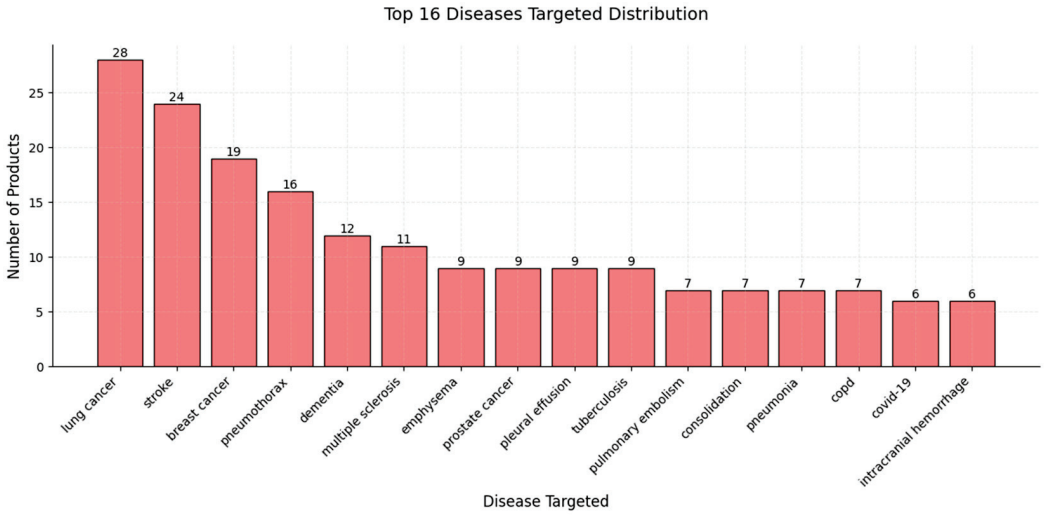


Figure 9. A bar chart illustrating the distribution of AI products with regard to their top 16 targeted diseases. The *x*-axis lists the targeted diseases, while the *y*-axis represents the number of AI products developed for each condition.

A majority of AI products are certified as Class IIa under the Medical Device Directive (MDD), totaling 66 products, followed by Class I under the MDD, with 58 products, and Class IIa under the Medical Device Regulation (MDR), with 49 products; these categories represent the bulk of certified AI products, indicating a high level of compliance with established regulatory standards. Class IIb under the MDR comprises 30 products, showing a significant but smaller representation, while there are 9 products that are not yet certified or have no certification, suggesting that they may be in development or awaiting regulatory approval. Smaller categories include "Certified, Class unknown" (8 products), "Certified, Class I, MDR" (1 product), and "Certified, Class IIb, MDD" (1 product). The concentration of AI products in Class IIa and I certifications under both the MDD and MDR reflects a strong effort to comply with regulatory frameworks, especially for products targeting moderate-risk classifications; this distribution underscores the critical role of regulatory certification in the development and deployment of AI products in healthcare (Figure 10).

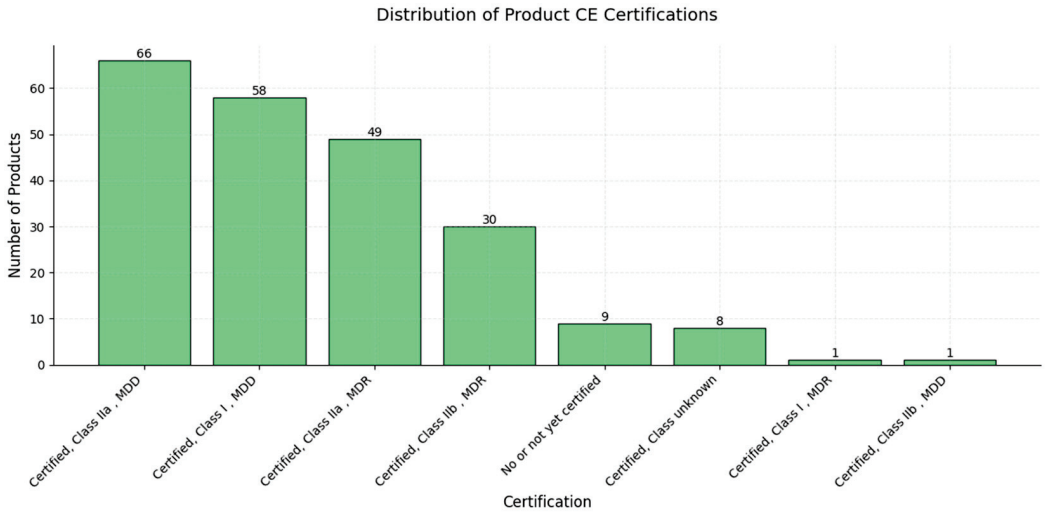


Figure 10. A bar chart displaying the distribution of CE certifications across various AI products. The *x*-axis represents different certification categories, while the *y*-axis shows the number of products in each category.

In terms of the market entry year, AI product releases before 2014 represent a notable group, with 18 entries, reflecting early AI market activity; from 2014 to 2016, there was a gradual increase in market entries, though still with small numbers each year; however, beginning in 2017, a steady rise occurred, with 15 products in 2017 and 21 in 2018, an upward trend that reached a peak in 2020, with 50 products, indicating a surge in AI adoption, likely driven by technological advancements and growing market demand [152]. Although 2021 saw a slight decline, to 31 products, it still represented high market activity; from 2022 to 2024, the number of entries continued to decrease, with 8, 13, and 3 products, respectively, possibly reflecting a maturing market, saturation in product development, or a shift in focus toward other healthcare areas—a trend which demonstrates the rapid expansion and growth of the AI market in healthcare, with a potential move toward saturation in recent years [153–157]. The peak in AI radiology product approvals in 2020 can be attributed to several key factors: during this period, a wealth of COVID-19 imaging datasets became publicly available, offering researchers the crucial data needed to train and validate AI models effectively; additionally, significant increases in funding and grants were directed toward AI research, particularly in response to the pandemic [158] (Figure 11).

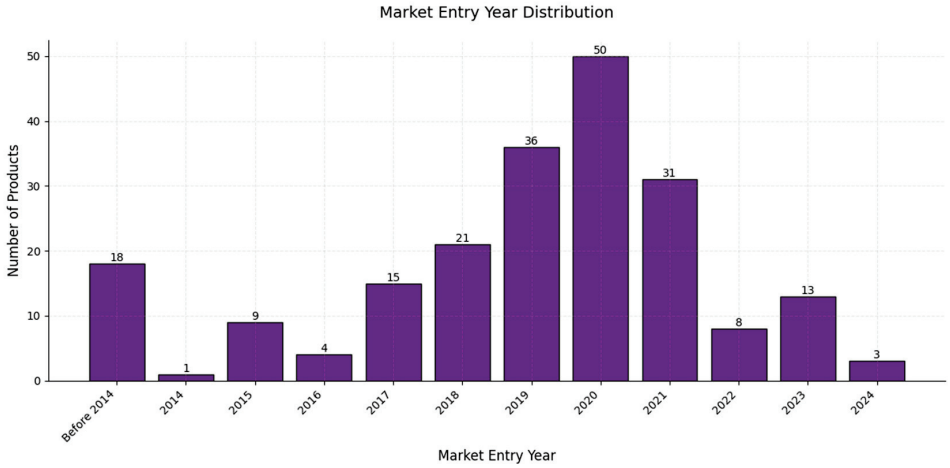


Figure 11. A bar chart illustrating the distribution of AI products by the market entry year. The *x*-axis represents the market entry years, ranging from "Before 2014" to 2024, while the *y*-axis shows the number of products that entered the market in each year.

The chart in Figure 11 depicts the market entry year distribution for AI products, while that in Figure 12 presents the annual number of AI products cleared for use in radiology from 2008 to 2024. A comparison of these data provides insights into the dynamics of AI adoption in radiology as we can observe a significant growth phase starting around 2017, with surges in product entries and clearances. The peak in market entries occurred in 2020, with 50 new products entering the market, while product clearances peaked slightly later, in 2023—a lag that suggests that products entering the market require time for regulatory approval [159,160].

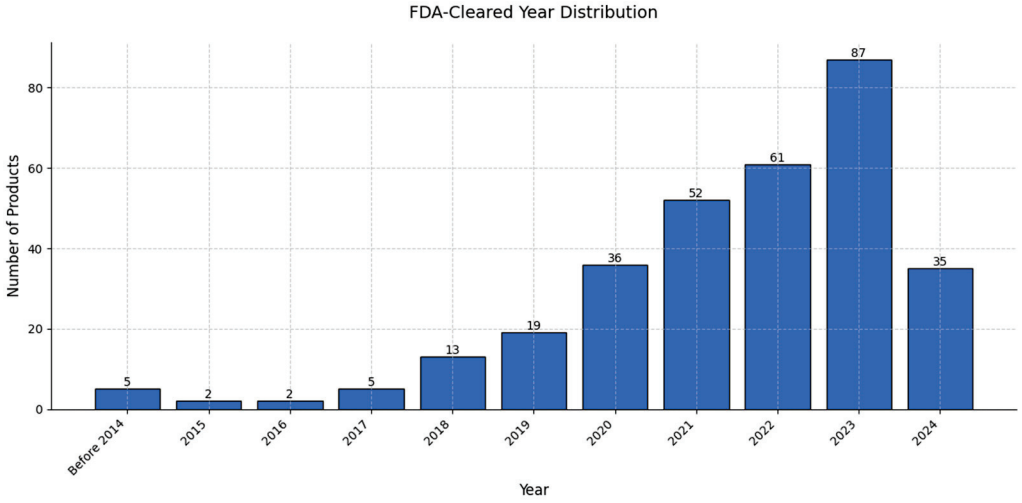


Figure 12. A bar chart illustrating the annual number of AI products cleared for use in radiology between 2008 and 2024. The trend reached its peak in 2023, with over 80 products cleared that year, reflecting the rapid growth and adoption of AI technologies in healthcare during this period. The year 2024 shows a slight decline, potentially signaling market stabilization or shifts in regulatory processes. The progression depicted here highlights the increasing integration of AI solutions into clinical practice, particularly in medical imaging.

After 2020, the number of new product entries declined steadily, with only three entries recorded in 2024; this trend reflects a possible stabilization or maturation of the AI radiology market in which fewer groundbreaking entries were being introduced. The slight decline in product clearances in 2024, following a peak in 2023, suggests market saturation. The regulatory landscape may be stabilizing and the focus might be shifting, from introducing new products to refining and deploying existing solutions [161].

The initial growth phase and subsequent stabilization observed in Figures 11 and 12 highlight the natural lifecycle of an emerging technology market, in which the early years are marked by innovation and rapid expansion, leading to a peak, followed by market consolidation and saturation as competition and regulatory frameworks mature [162,163].

In this work, we emphasize the rapid development in the AI medical market in recent years. Three years ago, Van Leeuwen et al. [20] described the landscape of artificial intelligence (AI) products in radiology, identifying 100 commercially available CE-marked solutions in the European market. In the current work, on the basis of the same vendor-supplied noncommercial database, we were able to distinguish over 220 products (www.radiology.healthairegister.com (accessed on 28 August 2024)), a result that is unlikely to reflect all available products, which becomes evident when different databases are compared, for instance, AI and radiology databases (https://aiandradiology.com (accessed on 30.08.2024)), used to collect and review products present in the domestic Polish market. The Health AI Register database is the largest and most reflective as it is the most comprehensive database, including the process of certification. Three years ago, 85 out of 100 reported products were validated, while, currently, 213 out of 222 registered products

to date are certified among the different classes (CE I–III), representing an important shift toward their acceptance by medical boards and easing their implementation into different medical institutions (Figure 3). Wu [164] investigated the adoption and usage of over 500 FDA-approved AI medical devices in the U.S. in 2023, focusing on 16 specific procedures that are billable through AI-specific CPT codes, spanning a variety of medical domains, including cardiology, ophthalmology, radiology, and liver health, with the most prevalent applications being coronary artery disease and diabetic retinopathy. Recently, in the U.S., 700 cleared AI algorithms were reported by Fornell, with 76% present in radiology but covering various specialties (radiology: 527; cardiology: 71; neurology: 16; hematology: 14; gastroenterology and urology: 10; clinical chemistry: 7; ophthalmic: 7; general and plastic surgery: 5; anesthesiology: 5; pathology: 4; microbiology: 4; general hospital: 3; orthopedic: 3; ear, nose, and throat: 2; and dental: 1). The above comparison illustrates the differences between the European and U.S. markets and reflects adoption obstacles within the European market, indicating a significantly larger size of the U.S. market and forecasting rapid development in the future [165,166].

Varghese [167] described several key challenges in the clinical adoption of artificial intelligence in medicine, organizing them under the RISE framework. Regulatory approval and sufficient evidence remain significant hurdles as many AI systems lack the rigorous prospective and multicenter clinical trials necessary for validation [22,168,169]; additionally, the majority of studies focus on retrospective or theoretical aspects, delaying their translation into clinical practice. Interpretability also poses a barrier as high-performing AI models, particularly those utilizing deep learning architectures, often lack transparency in their decision-making processes although clinicians are more likely to adopt AI systems that provide human-understandable explanations for their outputs. Additionally, interoperability challenges hinder the integration of AI systems into clinical workflows as effective communication between diverse hospital information systems, while preserving the meaning and context of medical data, is required. Finally, AI systems depend on high-quality and structured data, which are often limited, and reliance on unstructured sources, such as free-text medical records, introduces ambiguity and inaccuracies. The need to address these challenges through robust validation processes explains why AI adoption, despite a need for efficient automated systems in healthcare, is still limited. The development of integrating platforms and the formation of universal algorithms will, therefore, boost further AI implementation into medical practice.

10. Examples of Practical Implementation of AI Models

The safe, effective, and high-quality deployment of AI technologies is critical for advancing medical practice. Several medical environments provide exemplary use cases of robust AI models in detecting various pathologies and improving diagnostic workflows. In 2020, a multihospital experiment in Moscow [170] tested AI solutions for chest X-ray analysis across 178 state healthcare centers. AI frameworks analyzed redirected X-rays to detect abnormalities without prior training data. A top-performing framework employed advanced techniques such as EfficientNets and DenseNet, analyzing 17,888 cases over one month with an overall AUC of 0.77, ranging from 0.55 for herniation to 0.90 for pneumothorax. Robert et al. [171] evaluated the impact of AI as a second reader in detecting and localizing lung nodules on chest radiographs (CXRs) from 40 hospitals across the U.S. The study showed that AI assistance significantly improved diagnostic performance, with the mean AFROC increasing from 0.73 to 0.81 and AUROC from 0.77 to 0.84, along with a sensitivity improvement from 72.8% to 83.5%.

These results highlight AI's potential as a tool to enhance diagnostic accuracy without increasing false positives. Sacha et al. [172] assessed an AI system for detecting clinically

significant prostate cancer on MRI using a retrospective cohort of 10,207 MRI examinations and a multi-reader study with 62 radiologists. The AI system demonstrated superior performance compared to radiologists (AUROC: 0.91 vs. 0.86) and detected 6.8% more significant cancers at the same specificity, emphasizing its potential as a diagnostic support tool. In a meta-analysis by Wang [173], the effectiveness of deep learning algorithms for detecting and segmenting brain metastases on MRI was evaluated. The study identified 42 relevant studies and assessed them using QUADAS-2 and CLAIM tools. The results showed a pooled lesion-wise dice score of 79% and sensitivities of 86% (patient-wise) and 87% (lesion-wise). U-Net models performed best, with accuracy influenced by MRI hardware diversity and slice thickness. These findings underscore deep learning's promise in brain metastasis diagnostics. Lu et al. [174] conducted a randomized, multi-reader, multi-case study to assess AI-assisted auto-contouring for brain tumor stereotactic radiosurgery (SRS). Nine professionals contoured brain tumors in assisted and unassisted modes. AI significantly improved inter-reader agreement (DSC from 0.86 to 0.90, $p < 0.001$) and lesion detection sensitivity (91.3% vs. 82.6%, $p = 0.030$). Additionally, AI assistance enhanced contouring accuracy and reduced time spent by 30.8%, particularly benefiting less-experienced clinicians. Salehi et al. [175] performed a meta-analysis on AI algorithms detecting primary bone tumors, comparing their diagnostic performance to clinicians. Internal validation showed AI sensitivity and specificity at 84% and 86%, respectively, compared to clinicians' 76% and 64%. In external validation, AI achieved 84% sensitivity and 91% specificity, while clinicians reached 85% and 94%. With AI assistance, clinicians improved sensitivity to 95% but experienced reduced specificity (57%). This study highlights AI's potential and emphasizes the need for further optimization. Bachmann et al. [176] evaluated the impact of an AI tool on nonspecialist readers detecting traumatic fractures in appendicular skeleton radiographs. Using a multi-reader, multi-case design with 340 radiographic exams, sensitivity increased from 72% to 80% and specificity from 81% to 85% with AI assistance ($p < 0.05$). Missed fractures decreased by 29%, and false positives by 21%, without affecting reading time. The greatest improvement was in detecting nonobvious fractures, demonstrating AI's potential to enhance diagnostic performance efficiently. Jalal et al. [177] highlighted AI's role in emergency radiology, emphasizing its ability to enhance imaging analysis, workflow efficiency, and patient care quality, and mitigate radiologist burnout. Ketola et al. [178] proposed a structured evaluation process for AI applications in radiology, including pre-evaluation, retrospective testing, and prospective clinical integration, to ensure safety, effectiveness, and compatibility with clinical standards.

11. Summary

Artificial intelligence has shown transformative potential in the field of radiology, revolutionizing how medical imaging is interpreted and utilized in clinical practice as AI technologies have demonstrated the ability to automate a wide range of tasks, thereby significantly enhancing efficiency and diagnostic accuracy. Tasks ranging from image segmentation, abnormality detection, and classification of imaging data to more advanced processes, such as report analysis and automated case triaging, are reshaping traditional radiological workflows, thus allowing radiologists to focus on complex cases and critical decision-making, while routine processes are efficiently handled using AI.

In recent years, the field of radiology has experienced an explosion of innovative AI applications as the research has introduced numerous methodologies that leverage deep learning, machine learning, and other AI techniques to address specific challenges in medical imaging; concurrently, the market has witnessed the emergence of many commercial AI products, some of which have obtained regulatory certifications, indicating

their readiness for clinical deployment and underscoring the growing roles of AI in both research and practice.

Despite the potential of AI, it is imperative to recognize the ongoing importance of standard image analysis techniques, such as textural analysis and radiomics; these methods, which are rooted in well-established statistical and geometric principles, remain valuable tools for extracting meaningful features from medical images. The advantage of these features is their partial medical interpretability, which is not represented in deep learning models [113]; combining these standard approaches with AI techniques has the potential to create hybrid systems that leverage the strengths of both methods, thus yielding robust, interpretable, and clinically useful outcomes.

Moreover, ethical considerations and adherence to regulatory standards are critical for the successful integration of AI into radiology as protecting patient privacy and ensuring data security must remain fundamental priorities. AI developers and healthcare institutions must comply with the stringent regulations governing the use of medical data in order to maintain trust and integrity in healthcare systems. Additionally, transparency in AI development, from dataset selection to model validation, is vital for addressing potential biases and ensuring equitable outcomes.

In conclusion, while AI offers unprecedented opportunities to advance radiology, its integration requires a balanced approach; acknowledging the continued relevance of standard methods, addressing ethical and regulatory challenges, and fostering a collaborative relationship between radiologists and AI are essential for realizing its full potential. By combining the strengths of AI and traditional methodologies, radiology has the opportunity to achieve new levels of precision, efficiency, and impact in medical imaging.

The combined insights from the data reviewed in this manuscript suggest that the AI radiology market has transitioned from a phase of rapid growth to one of stabilization and potential saturation, a trend that reflects the increasing maturity of the field, in which regulatory processes and product deployment are catching up with the initial surge in innovation. Future efforts may focus more on optimizing existing solutions, rather than introducing entirely new products.

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

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable due to the absence of additional data beyond those included in the content of the work.

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

References

1. 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]

2. Morris, M.A.; Saboury, B.; Burkett, B.; Gao, J.; Siegel, E.L. Reinventing Radiology: Big Data and the Future of Medical Imaging. *J. Thorac. Imaging* **2018**, *33*, 4–16. [CrossRef] [PubMed]
3. Qureshi, J.; Iqbal, S.; Ahmad, N. The Radiological Renaissance: Transformative Advances in Emergency Care. *Cosm. J. Biol.* **2024**, *2*, 62–69.
4. Phillips-Wren, G. Ai Tools in Decision Making Support Systems: A Review. *Int. J. Artif. Intell. Tools* **2012**, *21*, 1240005. [CrossRef]
5. Soori, M.; Jough, F.K.G.; Dastres, R.; Arezoo, B. AI-Based Decision Support Systems in Industry 4.0, A Review. *J. Econ. Technol.* **2024**. [CrossRef]
6. Pierce, R.L.; Van Biesen, W.; Van Cauwenberge, D.; Decruyenaere, J.; Sterckx, S. Explainability in Medicine in an Era of AI-Based Clinical Decision Support Systems. *Front. Genet.* **2022**, *13*, 903600. [CrossRef] [PubMed]
7. Raparthi, M.; Gayam, S.R.; Kasaraneni, B.P.; Kondapaka, K.K.; Pattyam, S.P.; Putha, S.; Kuna, S.S.; Nimmagadda, V.S.P.; Sahu, M.K.; Thuniki, P. AI-Driven Decision Support Systems for Precision Medicine: Examining the Development and Implementation of AI-Driven Decision Support Systems in Precision Medicine. *J. Artif. Intell. Res.* **2021**, *1*, 11–20.
8. Mintz, Y.; Brodie, R. Introduction to Artificial Intelligence in Medicine. *Minim. Invasive Ther. Allied Technol.* **2019**, *28*, 73–81. [CrossRef] [PubMed]
9. Lai, V.; Chen, C.; Liao, Q.V.; Smith-Renner, A.; Tan, C. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. *arXiv* **2021**, arXiv:2112.11471.
10. De Vreede, T.; Raghavan, M.; De Vreede, G.-J. Design Foundations for AI Assisted Decision Making: A Self Determination Theory Approach. In Proceedings of the 54th Hawaii International Conference on System Sciences, Kauai, HI, USA, 5–8 January 2021.
11. Yang, X.; Aurisicchio, M. Designing Conversational Agents: A Self-Determination Theory Approach. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, ACM, Yokohama, Japan, 6 May 2021; pp. 1–16.
12. Wang, F.-Y.; Zhang, J.J.; Zheng, X.; Wang, X.; Yuan, Y.; Dai, X.; Zhang, J.; Yang, L. Where Does AlphaGo Go: From Church-Turing Thesis to AlphaGo Thesis and Beyond. *IEEE/CAA J. Autom. Sin.* **2016**, *3*, 113–120. [CrossRef]
13. Hamet, P.; Tremblay, J. Artificial Intelligence in Medicine. *Metab. Clin. Exp.* **2017**, *69*, S36–S40. [CrossRef] [PubMed]
14. Amisha; Malik, P.; Pathania, M.; Rathaur, V. Overview of Artificial Intelligence in Medicine. *J. Fam. Med. Prim. Care* **2019**, *8*, 2328. [CrossRef] [PubMed]
15. Rezazade Mehrizi, M.H.; Van Ooijen, P.; Homan, M. Applications of Artificial Intelligence (AI) in Diagnostic Radiology: A Technography Study. *Eur. Radiol.* **2021**, *31*, 1805–1811. [CrossRef]
16. 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]
17. Tariq, A.; Purkayastha, S.; Padmanaban, G.P.; Krupinski, E.; Trivedi, H.; Banerjee, I.; Gichoya, J.W. Current Clinical Applications of Artificial Intelligence in Radiology and Their Best Supporting Evidence. *J. Am. Coll. Radiol.* **2020**, *17*, 1371–1381. [CrossRef] [PubMed]
18. Kapoor, N.; Lacson, R.; Khorasani, R. Workflow Applications of Artificial Intelligence in Radiology and an Overview of Available Tools. *J. Am. Coll. Radiol.* **2020**, *17*, 1363–1370. [CrossRef]
19. Yamashita, R.; Nishio, M.; Do, R.K.G.; Togashi, K. Convolutional Neural Networks: An Overview and Application in Radiology. *Insights Imaging* **2018**, *9*, 611–629. [CrossRef] [PubMed]
20. Van Leeuwen, K.G.; Schalekamp, S.; Rutten, M.J.C.M.; Van Ginneken, B.; De Rooij, M. Artificial Intelligence in Radiology: 100 Commercially Available Products and Their Scientific Evidence. *Eur. Radiol.* **2021**, *31*, 3797–3804. [CrossRef]
21. Wichmann, J.L.; Willemink, M.J.; De Cecco, C.N. Artificial Intelligence and Machine Learning in Radiology: Current State and Considerations for Routine Clinical Implementation. *Invest. Radiol.* **2020**, *55*, 619–627. [CrossRef]
22. Pesapane, F.; Summers, P. Ethics and Regulations for AI in Radiology. In *Artificial Intelligence for Medicine*; Elsevier: Amsterdam, The Netherlands, 2024; pp. 179–192, ISBN 978-0-443-13671-9.
23. Maleki, F.; Muthukrishnan, N.; Ovens, K.; Reinhold, C.; Forghani, R. Machine Learning Algorithm Validation. *Neuroimaging Clin. N. Am.* **2020**, *30*, 433–445. [CrossRef] [PubMed]
24. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
25. Sutskever, I. Training Recurrent Neural Networks. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2013.
26. Wu, Z.; Shen, C.; Van Den Hengel, A. Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. *Pattern Recognit.* **2019**, *90*, 119–133. [CrossRef]
27. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Kai, L.; Li, F.-F. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, June 20-25 June 2009; pp. 248–255.
28. Witt, S. How Jensen Huang's Nvidia Is Powering the A.I. Revolution. *The New Yorker*. 2023. Available online: https://www.newyorker.com/magazine/2023/12/04/how-jensen-huangs-nvidia-is-powering-the-ai-revolution (accessed on 1 December 2024).

29. Goriparthi, R.; Luqman, S. Deep Learning Architectures for Real-Time Image Recognition: Innovations and Applications. *J. Data Sci. Intell. Syst.* **2024**, *15*, 880–907.
30. Khan, A.; Sohail, A.; Zahoora, U.; Qureshi, A.S. A Survey of the Recent Architectures of Deep Convolutional Neural Networks. *Artif. Intell. Rev.* **2020**, *53*, 5455–5516. [CrossRef]
31. Cheng, J.; Wang, P.; Li, G.; Hu, Q.; Lu, H. Recent Advances in Efficient Computation of Deep Convolutional Neural Networks. *Front. Inf. Technol. Electron. Eng.* **2018**, *19*, 64–77. [CrossRef]
32. Ali, I.; Hart, G.R.; Gunabushanam, G.; Liang, Y.; Muhammad, W.; Nartowt, B.; Kane, M.; Ma, X.; Deng, J. Lung Nodule Detection via Deep Reinforcement Learning. *Front. Oncol.* **2018**, *8*, 108. [CrossRef]
33. Obuchowicz, R.; Strzelecki, M.; Piórkowski, A. Clinical Applications of Artificial Intelligence in Medical Imaging and Image Processing—A Review. *Cancers* **2024**, *16*, 1870. [CrossRef] [PubMed]
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
35. Schulz, S.; Woerl, A.-C.; Jungmann, F.; Glasner, C.; Stenzel, P.; Strobl, S.; Fernandez, A.; Wagner, D.-C.; Haferkamp, A.; Mildenberger, P.; et al. Multimodal Deep Learning for Prognosis Prediction in Renal Cancer. *Front. Oncol.* **2021**, *11*, 788740. [CrossRef] [PubMed]
36. Yi, P.H.; Kim, T.K.; Yu, A.C.; Bennett, B.; Eng, J.; Lin, C.T. Can AI Outperform a Junior Resident? Comparison of Deep Neural Network to First-Year Radiology Residents for Identification of Pneumothorax. *Emerg. Radiol.* **2020**, *27*, 367–375. [CrossRef] [PubMed]
37. Chamberlin, J.H.; Aquino, G.; Nance, S.; Wortham, A.; Leaphart, N.; Paladugu, N.; Brady, S.; Baird, H.; Fiegel, M.; Fitzpatrick, L.; et al. Automated Diagnosis and Prognosis of COVID-19 Pneumonia from Initial ER Chest X-Rays Using Deep Learning. *BMC Infect. Dis.* **2022**, *22*, 637. [CrossRef]
38. Abrahamsen, B.S.; Knudtsen, I.S.; Eikenes, L.; Bathen, T.F.; Elschot, M. Pelvic PET/MR Attenuation Correction in the Image Space Using Deep Learning. *Front. Oncol.* **2023**, *13*, 1220009. [CrossRef] [PubMed]
39. Wang, C.; Shao, J.; Xu, X.; Yi, L.; Wang, G.; Bai, C.; Guo, J.; He, Y.; Zhang, L.; Yi, Z.; et al. DeepLN: A Multi-Task AI Tool to Predict the Imaging Characteristics, Malignancy and Pathological Subtypes in CT-Detected Pulmonary Nodules. *Front. Oncol.* **2022**, *12*, 683792. [CrossRef] [PubMed]
40. Albiol, A.; Albiol, F.; Paredes, R.; Plasencia-Martínez, J.M.; Blanco Barrio, A.; Santos, J.M.G.; Tortajada, S.; González Montaño, V.M.; Rodríguez Godoy, C.E.; Fernández Gómez, S.; et al. A Comparison of COVID-19 Early Detection between Convolutional Neural Networks and Radiologists. *Insights Imaging* **2022**, *13*, 122. [CrossRef] [PubMed]
41. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
42. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
43. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2016.
44. Fink, A.; Tran, H.; Reisert, M.; Rau, A.; Bayer, J.; Kotter, E.; Bamberg, F.; Russe, M.F. A Deep Learning Approach for Projection and Body-Side Classification in Musculoskeletal Radiographs. *Eur. Radiol. Exp.* **2024**, *8*, 23. [CrossRef] [PubMed]
45. Arbabshirani, M.R.; Fornwalt, B.K.; Mongelluzzo, G.J.; Suever, J.D.; Geise, B.D.; Patel, A.A.; Moore, G.J. Advanced Machine Learning in Action: Identification of Intracranial Hemorrhage on Computed Tomography Scans of the Head with Clinical Workflow Integration. *NPJ Digit. Med.* **2018**, *1*, 9. [CrossRef] [PubMed]
46. Nguyen, T.P.; Jung, J.W.; Yoo, Y.J.; Choi, S.H.; Yoon, J. Intelligent Evaluation of Global Spinal Alignment by a Decentralized Convolutional Neural Network. *J. Digit. Imaging* **2022**, *35*, 213–225. [CrossRef]
47. Solak, A.; Ceylan, R.; Bozkurt, M.A.; Cebeci, H.; Koplay, M. Adrenal Lesion Classification with Abdomen Caps and the Effect of ROI Size. *Phys. Eng. Sci. Med.* **2023**, *46*, 865–875. [CrossRef]
48. Gasulla, Ó.; Ledesma-Carbayo, M.J.; Borrell, L.N.; Fortuny-Profitós, J.; Mazaira-Font, F.A.; Barbero Allende, J.M.; Alonso-Menchén, D.; García-Bennett, J.; Del Río-Carrrero, B.; Jofré-Grimaldo, H.; et al. Enhancing Physicians' Radiology Diagnostics of COVID-19's Effects on Lung Health by Leveraging Artificial Intelligence. *Front. Bioeng. Biotechnol.* **2023**, *11*, 1010679. [CrossRef] [PubMed]
49. Santhanam, P.; Nath, T.; Peng, C.; Bai, H.; Zhang, H.; Ahima, R.S.; Chellappa, R. Artificial Intelligence and Body Composition. *Diabetes Metab. Syndr.: Clin. Res. Rev.* **2023**, *17*, 102732. [CrossRef] [PubMed]
50. Willemink, M.J.; Roth, H.R.; Sandfort, V. Toward Foundational Deep Learning Models for Medical Imaging in the New Era of Transformer Networks. *Radiol. Artif. Intell.* **2022**, *4*, e210284. [CrossRef]
51. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Guyon, I.,

Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Newry, UK, 2017; Volume 30.
52. Larson, N.; Nguyen, C.; Do, B.; Kaul, A.; Larson, A.; Wang, S.; Wang, E.; Bultman, E.; Stevens, K.; Pai, J.; et al. Artificial Intelligence System for Automatic Quantitative Analysis and Radiology Reporting of Leg Length Radiographs. *J. Digit. Imaging* **2022**, *35*, 1494–1505. [CrossRef] [PubMed]
53. Nurzynska, K.; Li, D.; Walts, A.E.; Gertych, A. Multilayer Outperforms Single-Layer Slide Scanning in AI-Based Classification of Whole Slide Images with Low-Burden Acid-Fast Mycobacteria (AFB). *Comput. Methods Programs Biomed.* **2023**, *234*, 107518. [CrossRef] [PubMed]
54. Haji Maghsoudi, O.; Gastounioti, A.; Scott, C.; Pantalone, L.; Wu, F.-F.; Cohen, E.A.; Winham, S.; Conant, E.F.; Vachon, C.; Kontos, D. Deep-LIBRA: An Artificial-Intelligence Method for Robust Quantification of Breast Density with Independent Validation in Breast Cancer Risk Assessment. *Med. Image Anal.* **2021**, *73*, 102138. [CrossRef]
55. 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.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2015; Volume 9351, pp. 234–241, ISBN 978-3-319-24573-7.
56. Stember, J.N.; Shalu, H. Deep Neuroevolution Squeezes More Out of Small Neural Networks and Small Training Sets: Sample Application to MRI Brain Sequence Classification. In *International Symposium on Intelligent Informatics*; Thampi, S.M., Mukhopadhyay, J., Paprzycki, M., Li, K.-C., Eds.; Smart Innovation, Systems and Technologies; Springer Nature: Singapore, 2023; Volume 333, pp. 153–167, ISBN 978-981-19-8093-0.
57. Seker, M.E.; Koyluoglu, Y.O.; Ozaydin, A.N.; Gurdal, S.O.; Ozcinar, B.; Cabioglu, N.; Ozmen, V.; Aribal, E. Diagnostic Capabilities of Artificial Intelligence as an Additional Reader in a Breast Cancer Screening Program. *Eur. Radiol.* **2024**, *34*, 6145–6157. [CrossRef] [PubMed]
58. Zhang, S.; Xin, X.; Wang, Y.; Guo, Y.; Hao, Q.; Yang, X.; Wang, J.; Zhang, J.; Zhang, B.; Wang, W. Automated Radiological Report Generation For Chest X-Rays With Weakly-Supervised End-to-End Deep Learning. *arXiv* **2020**, arXiv:2006.10347.
59. Bassi, P.R.A.S.; Yavuz, M.C.; Wang, K.; Chen, X.; Li, W.; Decherchi, S.; Cavalli, A.; Yang, Y.; Yuille, A.; Zhou, Z. RadGPT: Constructing 3D Image-Text Tumor Datasets. *arXiv* **2025**, arXiv:2501.04678.
60. Sun, Z.; Ong, H.; Kennedy, P.; Tang, L.; Chen, S.; Elias, J.; Lucas, E.; Shih, G.; Peng, Y. Evaluating GPT4 on Impressions Generation in Radiology Reports. *Radiology* **2023**, *307*, e231259. [CrossRef] [PubMed]
61. Kathait, A.S.; Garza-Frias, E.; Sikka, T.; Schultz, T.J.; Bizzo, B.; Kalra, M.K.; Dreyer, K.J. Assessing Laterality Errors in Radiology: Comparing Generative Artificial Intelligence and Natural Language Processing. *J. Am. Coll. Radiol.* **2024**, *21*, 1575–1582. [CrossRef]
62. Butler, J.J.; Harrington, M.C.; Tong, Y.; Rosenbaum, A.J.; Samsonov, A.P.; Walls, R.J.; Kennedy, J.G. From Jargon to Clarity: Improving the Readability of Foot and Ankle Radiology Reports with an Artificial Intelligence Large Language Model. *Foot Ankle Surg.* **2024**, *30*, 331–337. [CrossRef]
63. Adams, L.C.; Truhn, D.; Busch, F.; Kader, A.; Niehues, S.M.; Makowski, M.R.; Bressem, K.K. Leveraging GPT-4 for Post Hoc Transformation of Free-Text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. *Radiology* **2023**, *307*, e230725. [CrossRef] [PubMed]
64. Matute-González, M.; Darnell, A.; Comas-Cufí, M.; Pazó, J.; Soler, A.; Saborido, B.; Mauro, E.; Turnes, J.; Forner, A.; Reig, M.; et al. Utilizing a Domain-Specific Large Language Model for LI-RADS V2018 Categorization of Free-Text MRI Reports: A Feasibility Study. *Insights Imaging* **2024**, *15*, 280. [CrossRef] [PubMed]
65. Zhou, S.K.; Greenspan, H.; Davatzikos, C.; Duncan, J.S.; Van Ginneken, B.; Madabhushi, A.; Prince, J.L.; Rueckert, D.; Summers, R.M. A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises. *Proc. IEEE* **2021**, *109*, 820–838. [CrossRef] [PubMed]
66. Najjar, R. Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging. *Diagnostics* **2023**, *13*, 2760. [CrossRef] [PubMed]
67. A. Semary, N.; Ahmed, W.; Amin, K.; Pławiak, P.; Hammad, M. Enhancing Machine Learning-Based Sentiment Analysis through Feature Extraction Techniques. *PLoS ONE* **2024**, *19*, e0294968. [CrossRef]
68. Davis, E. Benchmarks for Automated Commonsense Reasoning: A Survey. *ACM Comput. Surv.* **2024**, *56*, 81. [CrossRef]
69. Huang, Y.; Zhang, J.; Shan, Z.; He, J. Compression Represents Intelligence Linearly. *arXiv* **2024**, arXiv:2404.09937.
70. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
71. Alshqaqeeq, F.; McGuire, C.; Overcash, M.; Ali, K.; Twomey, J. Choosing Radiology Imaging Modalities to Meet Patient Needs with Lower Environmental Impact. *Resour. Conserv. Recycl.* **2020**, *155*, 104657. [CrossRef]
72. Wang, P.; Vasconcelos, N. A Machine Teaching Framework for Scalable Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 11–17 October 2021; pp. 4945–4954.
73. Raina, R.; Battle, A.; Lee, H.; Packer, B.; Ng, A.Y. Self-Taught Learning: Transfer Learning from Unlabeled Data. In Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, USA, 20–24 June 2007; pp. 759–766.

74. Pfob, A.; Lu, S.-C.; Sidey-Gibbons, C. Machine Learning in Medicine: A Practical Introduction to Techniques for Data Pre-Processing, Hyperparameter Tuning, and Model Comparison. *BMC Med. Res. Methodol.* **2022**, *22*, 282. [CrossRef] [PubMed]
75. Borgli, R.J.; Kvale Stensland, H.; Riegler, M.A.; Halvorsen, P. Automatic Hyperparameter Optimization for Transfer Learning on Medical Image Datasets Using Bayesian Optimization. In Proceedings of the 2019 13th International Symposium on Medical Information and Communication Technology (ISMICT), Oslo, Norway, 8–10 May 2019; pp. 1–6.
76. Bradshaw, T.J.; Huemann, Z.; Hu, J.; Rahmim, A. A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging. *Radiol. Artif. Intell.* **2023**, *5*, e220232. [CrossRef] [PubMed]
77. Coppola, F.; Faggioni, L.; Gabelloni, M.; De Vietro, F.; Mendola, V.; Cattabriga, A.; Cocozza, M.A.; Vara, G.; Piccinino, A.; Lo Monaco, S.; et al. Human, All Too Human? An All-Around Appraisal of the "Artificial Intelligence Revolution" in Medical Imaging. *Front. Psychol.* **2021**, *12*, 710982. [CrossRef] [PubMed]
78. Jun-Yong, H.; Hyun, P.S.; Young-Jin, J. Artificial Intelligence Based Medical Imaging: An Overview. *J. Radiol. Sci. Technol.* **2020**, *43*, 195–208. [CrossRef]
79. Tran, K.; Bøtker, J.P.; Aframian, A.; Memarzadeh, K. Artificial Intelligence for Medical Imaging. In *Artificial Intelligence in Healthcare*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 143–162, ISBN 978-0-12-818438-7.
80. Hasani, N.; Morris, M.A.; Rahmim, A.; Summers, R.M.; Jones, E.; Siegel, E.; Saboury, B. Trustworthy Artificial Intelligence in Medical Imaging. *PET Clin.* **2022**, *17*, 1–12. [CrossRef]
81. Jiang, C.; Huang, Z.; Pedapati, T.; Chen, P.-Y.; Sun, Y.; Gao, J. Network Properties Determine Neural Network Performance. *Nat. Commun.* **2024**, *15*, 5718. [CrossRef]
82. Kaur, P.; Singh, R.K. A Review on Optimization Techniques for Medical Image Analysis. *Concurr. Comput.* **2023**, *35*, e7443. [CrossRef]
83. Kim, Y.J.; Kim, K.G. Development of an Optimized Deep Learning Model for Medical Imaging. *J. Korean Soc. Radiol.* **2020**, *81*, 1274. [CrossRef]
84. Centeno-Telleria, M.; Zulueta, E.; Fernandez-Gamiz, U.; Teso-Fz-Betoño, D.; Teso-Fz-Betoño, A. Differential Evolution Optimal Parameters Tuning with Artificial Neural Network. *Mathematics* **2021**, *9*, 427. [CrossRef]
85. Rajbhandari, S.; Rasley, J.; Ruwase, O.; He, Y. ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. In Proceedings of the SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, Atlanta, GA, USA, 9–19 November 2019.
86. Ji, J.L.; Spronk, M.; Kulkarni, K.; Repovš, G.; Anticevic, A.; Cole, M.W. Mapping the Human Brain's Cortical-Subcortical Functional Network Organization. *NeuroImage* **2019**, *185*, 35–57. [CrossRef]
87. Maguire, P.; Moser, P.; Maguire, R. Understanding Consciousness as Data Compression. *J. Cogn. Sci.* **2016**, *17*, 63–94. [CrossRef]
88. Waqas, A.; Bui, M.M.; Glassy, E.F.; El Naqa, I.; Borkowski, P.; Borkowski, A.A.; Rasool, G. Revolutionizing Digital Pathology With the Power of Generative Artificial Intelligence and Foundation Models. *Lab. Investig.* **2023**, *103*, 100255. [CrossRef]
89. Busby, L.P.; Courtier, J.L.; Glastonbury, C.M. Bias in Radiology: The How and Why of Misses and Misinterpretations. *RadioGraphics* **2018**, *38*, 236–247. [CrossRef]
90. Tee, Q.X.; Nambiar, M.; Stuckey, S. Error and Cognitive Bias in Diagnostic Radiology. *J. Med. Imag. Rad. Onc* **2022**, *66*, 202–207. [CrossRef] [PubMed]
91. Dujmović, M.; Malhotra, G.; Bowers, J.S. What Do Adversarial Images Tell Us about Human Vision? *eLife* **2020**, *9*, e55978. [CrossRef] [PubMed]
92. Rashid, A.B.; Kausik, M.A.K. AI Revolutionizing Industries Worldwide: A Comprehensive Overview of Its Diverse Applications. *Hybrid. Adv.* **2024**, *7*, 100277. [CrossRef]
93. Yousaf Gill, A.; Saeed, A.; Rasool, S.; Husnain, A.; Khawar Hussain, H. Revolutionizing Healthcare: How Machine Learning Is Transforming Patient Diagnoses—A Comprehensive Review of AI's Impact on Medical Diagnosis. *JWS* **2023**, *2*, 1638–1652. [CrossRef]
94. Recht, M.P.; Dewey, M.; Dreyer, K.; Langlotz, C.; Niessen, W.; Prainsack, B.; Smith, J.J. Integrating Artificial Intelligence into the Clinical Practice of Radiology: Challenges and Recommendations. *Eur. Radiol.* **2020**, *30*, 3576–3584. [CrossRef] [PubMed]
95. Willemink, M.J.; Koszek, W.A.; Hardell, C.; Wu, J.; Fleischmann, D.; Harvey, H.; Folio, L.R.; Summers, R.M.; Rubin, D.L.; Lungren, M.P. Preparing Medical Imaging Data for Machine Learning. *Radiology* **2020**, *295*, 4–15. [CrossRef]
96. Li, X.; Morgan, P.S.; Ashburner, J.; Smith, J.; Rorden, C. The First Step for Neuroimaging Data Analysis: DICOM to NIfTI Conversion. *J. Neurosci. Methods* **2016**, *264*, 47–56. [CrossRef]
97. Tovino, S.A. The HIPAA Privacy Rule and the EU GDPR: Illustrative Comparisons. *Seton Hall. Law. Rev.* **2017**, *47*, 973–993. [PubMed]
98. 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]
99. Vliegenthart, R.; Fouras, A.; Jacobs, C.; Papanikolaou, N. Innovations in Thoracic Imaging: CT, Radiomics, AI and X-ray Velocimetry. *Respirology* **2022**, *27*, 818–833. [CrossRef]

100. Kunimatsu, A.; Yasaka, K.; Akai, H.; Sugawara, H.; Kunimatsu, N.; Abe, O. Texture Analysis in Brain Tumor MR Imaging. *MRMS* **2022**, *21*, 95–109. [CrossRef]
101. Obuchowicz, R.; Kruszyńska, J.; Strzelecki, M. Classifying Median Nerves in Carpal Tunnel Syndrome: Ultrasound Image Analysis. *Biocybern. Biomed. Eng.* **2021**, *41*, 335–351. [CrossRef]
102. Materka, A.; Strzelecki, M. On the Importance of MRI Nonuniformity Correction for Texture Analysis. In Proceedings of the IEEE International Conference on Signal Processing, Algorithms, Architectures, Arrangements and Applications, SPA, Poznan, Poland, 26–28 September 2013.
103. Kociołek, M.; Strzelecki, M.; Obuchowicz, R. Does Image Normalization and Intensity Resolution Impact Texture Classification? *Comput. Med. Imaging Graph.* **2020**, *81*, 101716. [CrossRef] [PubMed]
104. Szczypiński, P.M.; Strzelecki, M.; Materka, A.; Klepaczko, A. Mazda—The Software Package for Textural Analysis of Biomedical Images. In *Computers in Medical Activity*; Kącki, E., Rudnicki, M., Stempczyńska, J., Eds.; Advances in Soft Computing; Springer: Berlin/Heidelberg, Germany, 2009; Volume 65, pp. 73–84, ISBN 978-3-642-04461-8.
105. Reska, D.; Kretowski, M. GPU-Accelerated Image Segmentation Based on Level Sets and Multiple Texture Features. *Multimed. Tools Appl.* **2021**, *80*, 5087–5109. [CrossRef]
106. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]
107. Li, W.; Qu, C.; Chen, X.; Bassi, P.R.A.S.; Shi, Y.; Lai, Y.; Yu, Q.; Xue, H.; Chen, Y.; Lin, X.; et al. AbdomenAtlas: A Large-Scale, Detailed-Annotated, & Multi-Center Dataset for Efficient Transfer Learning and Open Algorithmic Benchmarking. *Med. Image Anal.* **2024**, *97*, 103285. [CrossRef] [PubMed]
108. Hering, A.; Hansen, L.; Mok, T.C.W.; Chung, A.C.S.; Siebert, H.; Hager, S.; Lange, A.; Kuckertz, S.; Heldmann, S.; Shao, W.; et al. Learn2Reg: Comprehensive Multi-Task Medical Image Registration Challenge, Dataset and Evaluation in the Era of Deep Learning. *IEEE Trans. Med. Imaging* **2023**, *42*, 697–712. [CrossRef] [PubMed]
109. Baheti, B.; Chakrabarty, S.; Akbari, H.; Bilello, M.; Wiestler, B.; Schwarting, J.; Calabrese, E.; Rudie, J.; Abidi, S.; Mousa, M.; et al. The Brain Tumor Sequence Registration (BraTS-Reg) Challenge: Establishing Correspondence Between Pre-Operative and Follow-up MRI Scans of Diffuse Glioma Patients. *arXiv* **2021**, arXiv:2112.06979.
110. Chen, X.; Diaz-Pinto, A.; Ravikumar, N.; Frangi, A. Deep Learning in Medical Image Registration. *Prog. Biomed. Eng.* **2020**, *3*, 012003. [CrossRef]
111. Xiao, H.; Teng, X.; Liu, C.; Li, T.; Ren, G.; Yang, R.; Shen, D.; Cai, J. A Review of Deep Learning-Based Three-Dimensional Medical Image Registration Methods. *Quant. Imaging Med. Surg.* **2021**, *11*, 4895–4916. [CrossRef]
112. Andrade, N.; Faria, F.A.; Cappabianco, F.A.M. A Practical Review on Medical Image Registration: From Rigid to Deep Learning Based Approaches. In Proceedings of the 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Parana, Brazil, 29 October–1 November 2018; pp. 463–470.
113. Nurzynska, K.; Piórkowski, A.; Strzelecki, M.; Kociołek, M.; Banyś, R.P.; Obuchowicz, R. Differentiating Age and Sex in Vertebral Body CT Scans—Texture Analysis versus Deep Learning Approach. *Biocybern. Biomed. Eng.* **2024**, *44*, 20–30. [CrossRef]
114. Mang, A.; Gholami, A.; Davatzikos, C.; Biros, G. PDE-Constrained Optimization in Medical Image Analysis. *Optim. Eng.* **2018**, *19*, 765–812. [CrossRef]
115. Chen, J.; Liu, Y.; Wei, S.; Bian, Z.; Subramanian, S.; Carass, A.; Prince, J.L.; Du, Y. A Survey on Deep Learning in Medical Image Registration: New Technologies, Uncertainty, Evaluation Metrics, and Beyond. *Med. Image Anal.* **2025**, *100*, 103385. [CrossRef]
116. Friedrich, P.; Frisch, Y.; Cattin, P.C. Deep Generative Models for 3D Medical Image Synthesis. *arXiv* **2024**, arXiv:2410.17664.
117. Huijben, E.M.C.; Terpstra, M.L.; Galapon, A., Jr.; Pai, S.; Thummerer, A.; Koopmans, P.; Afonso, M.; Van Eijnatten, M.; Gurney-Champion, O.; Chen, Z.; et al. Generating Synthetic Computed Tomography for Radiotherapy: SynthRAD2023 Challenge Report. *Med. Image Anal.* **2024**, *97*, 103276. [CrossRef] [PubMed]
118. Cho, J.; Zakka, C.; Kaur, D.; Shad, R.; Wightman, R.; Chaudhari, A.; Hiesinger, W. MediSyn: Text-Guided Diffusion Models for Broad Medical 2D and 3D Image Synthesis. *arXiv* **2024**, arXiv:2405.09806.
119. Muckley, M.J.; Riemenschneider, B.; Radmanesh, A.; Kim, S.; Jeong, G.; Ko, J.; Jun, Y.; Shin, H.; Hwang, D.; Mostapha, M.; et al. Results of the 2020 fastMRI Challenge for Machine Learning MR Image Reconstruction. *IEEE Trans Med Imaging* **2021**, *40*, 2306–2317. [CrossRef]
120. Bhadra, S.; Kelkar, V.A.; Brooks, F.J.; Anastasio, M.A. On Hallucinations in Tomographic Image Reconstruction. *IEEE Trans. Med. Imaging* **2021**, *40*, 3249–3260. [CrossRef] [PubMed]
121. Roh, J.; Ryu, D.; Lee, J. CT Synthesis with Deep Learning for MR-Only Radiotherapy Planning: A Review. *Biomed. Eng. Lett.* **2024**, *14*, 1259–1278. [CrossRef] [PubMed]
122. Chung, T.; Dillman, J.R. Deep Learning Image Reconstruction: A Tremendous Advance for Clinical MRI but Be Careful. . .. *Pediatr. Radiol.* **2023**, *53*, 2157–2158. [CrossRef] [PubMed]

123. Reinke, A.; Tizabi, M.D.; Baumgartner, M.; Eisenmann, M.; Heckmann-Nötzel, D.; Kavur, A.E.; Rädsch, T.; Sudre, C.H.; Acion, L.; Antonelli, M.; et al. Understanding Metric-Related Pitfalls in Image Analysis Validation. *Nat. Methods* **2024**, *21*, 182–194. [CrossRef] [PubMed]
124. Rädsch, T.; Reinke, A.; Weru, V.; Tizabi, M.D.; Heller, N.; Isensee, F.; Kopp-Schneider, A.; Maier-Hein, L. Quality Assured: Rethinking Annotation Strategies in Imaging AI. In *Computer Vision—ECCV 2024*; Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G., Eds.; Lecture Notes in Computer Science; Springer Nature Switzerland: Cham, Switzerland, 2025; Volume 15136, pp. 52–69, ISBN 978-3-031-73228-7.
125. Christodoulou, E.; Reinke, A.; Houhou, R.; Kalinowski, P.; Erkan, S.; Sudre, C.H.; Burgos, N.; Boutaj, S.; Loizillon, S.; Solal, M.; et al. Confidence Intervals Uncovered: Are We Ready for Real-World Medical Imaging AI? In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2024*; Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A., Eds.; Lecture Notes in Computer Science; Springer Nature Switzerland: Cham, Switzerland, 2024; Volume 15010, pp. 124–132, ISBN 978-3-031-72116-8.
126. Fernandez-Quilez, A. Deep Learning in Radiology: Ethics of Data and on the Value of Algorithm Transparency, Interpretability and Explainability. *AI Ethics* **2023**, *3*, 257–265. [CrossRef]
127. Keel, B.; Quyn, A.; Jayne, D.; Relton, S.D. State-of-the-Art Performance of Deep Learning Methods for Pre-Operative Radiologic Staging of Colorectal Cancer Lymph Node Metastasis: A Scoping Review. *BMJ Open* **2024**, *14*, e086896. [CrossRef] [PubMed]
128. Marey, A.; Arjmand, P.; Alerab, A.D.S.; Eslami, M.J.; Saad, A.M.; Sanchez, N.; Umair, M. Explainability, Transparency and Black Box Challenges of AI in Radiology: Impact on Patient Care in Cardiovascular Radiology. *Egypt. J. Radiol. Nucl. Med.* **2024**, *55*, 183. [CrossRef]
129. Tsai, M.-J.; Lin, P.-Y.; Lee, M.-E. Adversarial Attacks on Medical Image Classification. *Cancers* **2023**, *15*, 4228. [CrossRef] [PubMed]
130. Amirian, M.; Schwenker, F.; Stadelmann, T. Trace and Detect Adversarial Attacks on CNNs Using Feature Response Maps. In *Artificial Neural Networks in Pattern Recognition*; Pancioni, L., Schwenker, F., Trentin, E., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11081, pp. 346–358, ISBN 978-3-319-99977-7.
131. Shao, R.; Shi, Z.; Yi, J.; Chen, P.-Y.; Hsieh, C.-J. On the Adversarial Robustness of Vision Transformers. *arXiv* **2021**, arXiv:2103.15670.
132. Mo, Y.; Wu, D.; Wang, Y.; Guo, Y.; Wang, Y. When Adversarial Training Meets Vision Transformers: Recipes from Training to Architecture. In Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; Curran Associates Inc.: Red Hook, NY, USA, 2024.
133. Health AI Register 2024. Available online: https://www.HealthAIregister.com (accessed on 28 August 2024).
134. AI Central. Available online: https://aicentral.acrdsi.org/ (accessed on 3 September 2024).
135. Grenier, P.A.; Brun, A.L.; Mellot, F. The Potential Role of Artificial Intelligence in Lung Cancer Screening Using Low-Dose Computed Tomography. *Diagnostics* **2022**, *12*, 2435. [CrossRef]
136. Lauritzen, A.D.; Lillholm, M.; Lynge, E.; Nielsen, M.; Karssemeijer, N.; Vejborg, I. Early Indicators of the Impact of Using AI in Mammography Screening for Breast Cancer. *Radiology* **2024**, *311*, e232479. [CrossRef]
137. Roca, P.; Attye, A.; Colas, L.; Tucholka, A.; Rubini, P.; Cackowski, S.; Ding, J.; Budzik, J.-F.; Renard, F.; Doyle, S.; et al. Artificial Intelligence to Predict Clinical Disability in Patients with Multiple Sclerosis Using FLAIR MRI. *Diagn. Interv. Imaging* **2020**, *101*, 795–802. [CrossRef]
138. Van Leeuwen, K.G.; Becks, M.J.; Grob, D.; De Lange, F.; Rutten, J.H.E.; Schalekamp, S.; Rutten, M.J.C.M.; Van Ginneken, B.; De Rooij, M.; Meijer, F.J.A. AI-Support for the Detection of Intracranial Large Vessel Occlusions: One-Year Prospective Evaluation. *Heliyon* **2023**, *9*, e19065. [CrossRef] [PubMed]
139. Adams, S.J.; Madtes, D.K.; Burbridge, B.; Johnston, J.; Goldberg, I.G.; Siegel, E.L.; Babyn, P.; Nair, V.S.; Calhoun, M.E. Clinical Impact and Generalizability of a Computer-Assisted Diagnostic Tool to Risk-Stratify Lung Nodules With CT. *J. Am. Coll. Radiol.* **2023**, *20*, 232–242. [CrossRef] [PubMed]
140. Asif, A.; Charters, P.F.P.; Thompson, C.A.S.; Komber, H.M.E.I.; Hudson, B.J.; Rodrigues, J.C.L. Artificial Intelligence Can Detect Left Ventricular Dilatation on Contrast-Enhanced Thoracic Computer Tomography Relative to Cardiac Magnetic Resonance Imaging. *Br. J. Radiol.* **2022**, *95*, 20210852. [CrossRef] [PubMed]
141. Arasu, V.A.; Habel, L.A.; Achacoso, N.S.; Buist, D.S.M.; Cord, J.B.; Esserman, L.J.; Hylton, N.M.; Glymour, M.M.; Kornak, J.; Kushi, L.H.; et al. Comparison of Mammography AI Algorithms with a Clinical Risk Model for 5-Year Breast Cancer Risk Prediction: An Observational Study. *Radiology* **2023**, *307*, e222733. [CrossRef] [PubMed]
142. Gräfe, D.; Beeskow, A.B.; Pfäffle, R.; Rosolowski, M.; Chung, T.S.; DiFranco, M.D. Automated Bone Age Assessment in a German Pediatric Cohort: Agreement between an Artificial Intelligence Software and the Manual Greulich and Pyle Method. *Eur. Radiol.* **2023**, *34*, 4407–4413. [CrossRef]
143. Eom, H.J.; Cha, J.H.; Choi, W.J.; Cho, S.M.; Jin, K.; Kim, H.H. Mammographic Density Assessment: Comparison of Radiologists, Automated Volumetric Measurement, and Artificial Intelligence-Based Computer-Assisted Diagnosis. *Acta Radiol.* **2024**, *65*, 708–715. [CrossRef] [PubMed]

144. Habuza, T.; Navaz, A.N.; Hashim, F.; Alnajjar, F.; Zaki, N.; Serhani, M.A.; Statsenko, Y. AI Applications in Robotics, Diagnostic Image Analysis and Precision Medicine: Current Limitations, Future Trends, Guidelines on CAD Systems for Medicine. *Inform. Med. Unlocked* **2021**, *24*, 100596. [CrossRef]
145. Tadavarthi, Y.; Vey, B.; Krupinski, E.; Prater, A.; Gichoya, J.; Safdar, N.; Trivedi, H. The State of Radiology AI: Considerations for Purchase Decisions and Current Market Offerings. *Radiol. Artif. Intell.* **2020**, *2*, e200004. [CrossRef]
146. Wenderott, K.; Krups, J.; Luetkens, J.A.; Gambashidze, N.; Weigl, M. Prospective Effects of an Artificial Intelligence-Based Computer-Aided Detection System for Prostate Imaging on Routine Workflow and Radiologists' Outcomes. *Eur. J. Radiol.* **2024**, *170*, 111252. [CrossRef]
147. Jacobs, C.; Schreuder, A.; Van Riel, S.J.; Scholten, E.T.; Wittenberg, R.; Wille, M.M.W.; De Hoop, B.; Sprengers, R.; Mets, O.M.; Geurts, B.; et al. Assisted versus Manual Interpretation of Low-Dose CT Scans for Lung Cancer Screening: Impact on Lung-RADS Agreement. *Radiol. Imaging Cancer* **2021**, *3*, e200160. [CrossRef] [PubMed]
148. Jimenez-Pastor, A.; Lopez-Gonzalez, R.; Fos-Guarinos, B.; Garcia-Castro, F.; Wittenberg, M.; Torregrosa-Andrés, A.; Marti-Bonmati, L.; Garcia-Fontes, M.; Duarte, P.; Gambini, J.P.; et al. Automated Prostate Multi-Regional Segmentation in Magnetic Resonance Using Fully Convolutional Neural Networks. *Eur. Radiol.* **2023**, *33*, 5087–5096. [CrossRef]
149. Nampewo, I.; Ariana, P.; Vijayan, S. Benefits of Artificial Intelligence versus Human-Reader in Chest X-Ray Screening for Tuberculosis in the Philippines. *Int. J. Health Sci. Res.* **2024**, *14*, 277–287. [CrossRef]
150. Pemberton, H.G.; Goodkin, O.; Prados, F.; Das, R.K.; Vos, S.B.; Moggridge, J.; Coath, W.; Gordon, E.; Barrett, R.; Schmitt, A.; et al. Automated Quantitative MRI Volumetry Reports Support Diagnostic Interpretation in Dementia: A Multi-Rater, Clinical Accuracy Study. *Eur. Radiol.* **2021**, *31*, 5312–5323. [CrossRef]
151. Van Leeuwen, K.G.; Schalekamp, S.; Rutten, M.J.C.M.; Huisman, M.; Schaefer-Prokop, C.M.; De Rooij, M.; Van Ginneken, B.; Maresch, B.; Geurts, B.H.J.; Van Dijke, C.F.; et al. Comparison of Commercial AI Software Performance for Radiograph Lung Nodule Detection and Bone Age Prediction. *Radiology* **2024**, *310*, e230981. [CrossRef] [PubMed]
152. Cooper, R.G. The AI Transformation of Product Innovation. *Ind. Mark. Manag.* **2024**, *119*, 62–74. [CrossRef]
153. Gao, S.; Xu, Z.; Kang, W.; Lv, X.; Chu, N.; Xu, S.; Hou, D. Artificial Intelligence-Driven Computer Aided Diagnosis System Provides Similar Diagnosis Value Compared with Doctors' Evaluation in Lung Cancer Screening. *BMC Med. Imaging* **2024**, *24*, 141. [CrossRef] [PubMed]
154. Savage, C.H.; Elkassem, A.A.; Hamki, O.; Sturdivant, A.; Benson, D.; Grumley, S.; Tzabari, J.; Junck, K.; Li, Y.; Li, M.; et al. Prospective Evaluation of Artificial Intelligence Triage of Incidental Pulmonary Emboli on Contrast-Enhanced CT Examinations of the Chest or Abdomen. *Am. J. Roentgenol.* **2024**, *223*, e2431067. [CrossRef]
155. Rothenberg, S.A.; Savage, C.H.; Abou Elkassem, A.; Singh, S.; Abozeed, M.; Hamki, O.; Junck, K.; Tridandapani, S.; Li, M.; Li, Y.; et al. Prospective Evaluation of AI Triage of Pulmonary Emboli on CT Pulmonary Angiograms. *Radiology* **2023**, *309*, e230702. [CrossRef]
156. Chien, H.-W.C.; Yang, T.-L.; Juang, W.-C.; Chen, Y.-Y.A.; Li, Y.-C.J.; Chen, C.-Y. Pilot Report for Intracranial Hemorrhage Detection with Deep Learning Implanted Head Computed Tomography Images at Emergency Department. *J. Med. Syst.* **2022**, *46*, 49. [CrossRef] [PubMed]
157. Schaffter, T.; Buist, D.S.M.; Lee, C.I.; Nikulin, Y.; Ribli, D.; Guan, Y.; Lotter, W.; Jie, Z.; Du, H.; Wang, S.; et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. *JAMA Netw. Open* **2020**, *3*, e200265. [CrossRef] [PubMed]
158. Lauri, C.; Shimpo, F.; Sokołowski, M.M. Artificial Intelligence and Robotics on the Frontlines of the Pandemic Response: The Regulatory Models for Technology Adoption and the Development of Resilient Organisations in Smart Cities. *J. Ambient. Intell. Human. Comput.* **2023**, *14*, 14753–14764. [CrossRef] [PubMed]
159. Justo-Hanani, R. The Politics of Artificial Intelligence Regulation and Governance Reform in the European Union. *Policy Sci.* **2022**, *55*, 137–159. [CrossRef]
160. Smuha, N.A. From a 'Race to AI' to a 'Race to AI Regulation': Regulatory Competition for Artificial Intelligence. *Law Innov. Technol.* **2021**, *13*, 57–84. [CrossRef]
161. Tanguay, W.; Acar, P.; Fine, B.; Abdolell, M.; Gong, B.; Cadrin-Chênevert, A.; Chartrand-Lefebvre, C.; Chalaoui, J.; Gorgos, A.; Chin, A.S.-L.; et al. Assessment of Radiology Artificial Intelligence Software: A Validation and Evaluation Framework. *Can. Assoc. Radiol. J.* **2023**, *74*, 326–333. [CrossRef] [PubMed]
162. Wimpfheimer, O.; Kimmel, Y. Artificial Intelligence in Medical Imaging: An Overview of a Decade of Experience. *Isr. Med. Assoc. J.* **2024**, *26*, 122–125.
163. Mehta, S.S. *Commercializing Successful Biomedical Technologies: Basic Principles for the Development of Drugs, Diagnostics and Devices*; Cambridge University Press: Cambridge, NY, USA, 2008; ISBN 978-0-521-87098-6.
164. Wu, K.; Wu, E.; Theodorou, B.; Liang, W.; Mack, C.; Glass, L.; Sun, J.; Zou, J. Characterizing the Clinical Adoption of Medical AI through U.S. Insurance Claims. *NEJM AI* **2023**, *1*, AIoa2300030. [CrossRef]

165. Fornell, D. FDA Has Now Cleared 700 AI Healthcare Algorithms, More Than 76% in Radiology. Health Imaging. 2023. Available online: https://healthimaging.com/topics/artificial-intelligence/fda-has-now-cleared-700-ai-healthcare-algorithms-more-76-radiology (accessed on 1 December 2024).
166. MarketsandMarkets Artificial Intelligence (AI) in Healthcare Market by Offering (Hardware, Software, Services), Technology (Machine Learning, NLP, Context-Aware Computing, Computer Vision), Application, End User and Region—Global Forecast to 2028. 2023. Available online: https://www.marketsandmarkets.com/Market-Reports/ai-toolkit-market-252755052.html (accessed on 30 November 2024).
167. Varghese, J. Artificial Intelligence in Medicine: Chances and Challenges for Wide Clinical Adoption. *Visc. Med.* **2020**, *36*, 443–449. [CrossRef]
168. Chau, M. Ethical, Legal, and Regulatory Landscape of Artificial Intelligence in Australian Healthcare and Ethical Integration in Radiography: A Narrative Review. *J. Med. Imaging Radiat. Sci.* **2024**, *55*, 101733. [CrossRef] [PubMed]
169. Zhu, S.; Gilbert, M.; Chetty, I.; Siddiqui, F. The 2021 Landscape of FDA-Approved Artificial Intelligence/Machine Learning-Enabled Medical Devices: An Analysis of the Characteristics and Intended Use. *Int. J. Med. Inform.* **2022**, *165*, 104828. [CrossRef] [PubMed]
170. Ibragimov, B.; Arzamasov, K.; Maksudov, B.; Kiselev, S.; Mongolin, A.; Mustafaev, T.; Ibragimova, D.; Evteeva, K.; Andreychenko, A.; Morozov, S. A 178-Clinical-Center Experiment of Integrating AI Solutions for Lung Pathology Diagnosis. *Sci. Rep.* **2023**, *13*, 1135. [CrossRef]
171. Robert, D.; Sathyamurthy, S.; Singh, A.K.; Matta, S.A.; Tadepalli, M.; Tanamala, S.; Bosemani, V.; Mammarappallil, J.; Kundnani, B. Effect of Artificial Intelligence as a Second Reader on the Lung Nodule Detection and Localization Accuracy of Radiologists and Non-Radiology Physicians in Chest Radiographs: A Multicenter Reader Study. *Acad. Radiol.* **2024**, *in press*. [CrossRef]
172. Saha, A.; Bosma, J.S.; Twilt, J.J.; van Ginneken, B.; Bjartell, A.; Padhani, A.R.; Bonekamp, D.; Villeirs, G.; Salomon, G.; Giannarini, G.; et al. Artificial Intelligence and Radiologists in Prostate Cancer Detection on MRI (PI-CAI): An International, Paired, Non-Inferiority, Confirmatory Study. *Lancet Oncol.* **2024**, *25*, 879–887. [CrossRef]
173. Wang, T.-W.; Hsu, M.-S.; Lee, W.-K.; Pan, H.-C.; Yang, H.-C.; Lee, C.-C.; Wu, Y.-T. Brain Metastasis Tumor Segmentation and Detection Using Deep Learning Algorithms: A Systematic Review and Meta-Analysis. *Radiother. Oncol.* **2024**, *190*, 110007. [CrossRef]
174. Lu, S.-L.; Xiao, F.-R.; Cheng, J.C.-H.; Yang, W.-C.; Cheng, Y.-H.; Chang, Y.-C.; Lin, J.-Y.; Liang, C.-H.; Lu, J.-T.; Chen, Y.-F.; et al. Randomized Multi-Reader Evaluation of Automated Detection and Segmentation of Brain Tumors in Stereotactic Radiosurgery with Deep Neural Networks. *Neuro Oncol.* **2021**, *23*, 1560–1568. [CrossRef] [PubMed]
175. Salehi, M.A.; Mohammadi, S.; Harandi, H.; Zakavi, S.S.; Jahanshahi, A.; Shahrabi Farahani, M.; Wu, J.S. Diagnostic Performance of Artificial Intelligence in Detection of Primary Malignant Bone Tumors: A Meta-Analysis. *J. Imaging Inform. Med.* **2024**, *37*, 766–777. [CrossRef]
176. Bachmann, R.; Gunes, G.; Hangaard, S.; Nexmann, A.; Lisouski, P.; Boesen, M.; Lundemann, M.; Baginski, S.G. Improving Traumatic Fracture Detection on Radiographs with Artificial Intelligence Support: A Multi-Reader Study. *BJR|Open* **2023**, *6*, tzae011. [CrossRef]
177. Jalal, S.; Parker, W.; Ferguson, D.; Nicolaou, S. Exploring the Role of Artificial Intelligence in an Emergency and Trauma Radiology Department. *Can. Assoc. Radiol. J.* **2021**, *72*, 167–174. [CrossRef]
178. Ketola, J.H.J.; Inkinen, S.I.; Mäkelä, T.; Syväranta, S.; Peltonen, J.; Kaasalainen, T.; Kortesniemi, M. Testing Process for Artificial Intelligence Applications in Radiology Practice. *Phys. Med.* **2024**, *128*, 104842. [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.

cancers

Reply

Reply to Keles et al. Comment on "Dalboni da Rocha et al. Artificial Intelligence for Neuroimaging in Pediatric Cancer. *Cancers* 2025, *17*, 622"

Josue Luiz Dalboni da Rocha [1,*], Jesyin Lai [1], Pankaj Pandey [1], Phyu Sin M. Myat [1], Zachary Loschinskey [1,2,3], Asim K. Bag [1] and Ranganatha Sitaram [1,*]

[1] Department of Radiology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA; jesyin.lai@stjude.org (J.L.); pankaj.pandey@stjude.org (P.P.); phyu.sin.m.myat@vanderbilt.edu (P.S.M.M.); zfl@bu.edu (Z.L.); asim.bag@stjude.org (A.K.B.)
[2] Department of Chemical and Biomedical Engineering, University of Missouri-Columbia, Columbia, MO 65211, USA
[3] Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA
* Correspondence: josueluiz.dalbonidarocha@stjude.org (J.L.D.d.R.); ranganatha.sitaram@stjude.org (R.S.)

We thank the readers for their thoughtful reading of our review, "Artificial Intelligence for Neuroimaging in Pediatric Cancer", and for their interest in discussing aspects of segmentation performance metrics [1].

Our intention was to provide a concise and accessible overview of key developments in this rapidly evolving field. To enhance readability, we summarized some of the most relevant recent approaches in a single table. As noted by the commenters, performance metrics across studies may differ in scope, measurement types, and validation strategies. We appreciate the opportunity to clarify that the values in Table 1 were intended as representative highlights—a simplified snapshot of recent work—rather than for direct cross-study comparison.

These distinctions—such as those between whole-tumor and subregion segmentation, metric types, or validation cohorts—reflect the range of methodological approaches found in the literature. We appreciate the readers' attention to these aspects and agree that interpretive comparisons by a general audience require careful consideration of each study's specific context. We will take these suggestions into account in future publications, where a more comprehensive and multi-dimensional presentation of results can be provided.

Nevertheless, the observations raised do not affect the overall conclusions of our review, which emphasized the promising role of AI-based neuroimaging tools in pediatric oncology. We appreciate the constructive nature of this exchange.

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

Reference

1. Keles, E.; Colakoglu, M.N.; Bengtsson, M. Comment on Dalboni da Rocha et al. Artificial Intelligence for Neuroimaging in Pediatric Cancer. *Cancers* 2025, *17*, 622. *Cancers* **2025**, *17*, 1776

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.

Academic Editor: Samuel C. Mok

Received: 8 May 2025
Accepted: 22 May 2025
Published: 26 May 2025

Citation: Rocha, J.L.D.d.; Lai, J.; Pandey, P.; Myat, P.S.M.; Loschinskey, Z.; Bag, A.K.; Sitaram, R. Reply to Keles et al. Comment on "Dalboni da Rocha et al. Artificial Intelligence for Neuroimaging in Pediatric Cancer. *Cancers* 2025, *17*, 622". *Cancers* **2025**, *17*, 1778. https://doi.org/10.3390/cancers17111778

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

cancers

MDPI

Review

Artificial Intelligence for Neuroimaging in Pediatric Cancer

Josue Luiz Dalboni da Rocha [1,*], Jesyin Lai [1], Pankaj Pandey [1], Phyu Sin M. Myat [1], Zachary Loschinskey [1,2,3], Asim K. Bag [1] and Ranganatha Sitaram [1,*]

[1] Department of Radiology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA; jesyin.lai@stjude.org (J.L.); pankaj.pandey@stjude.org (P.P.); phyu.sin.m.myat@vanderbilt.edu (P.S.M.M.); zfl@bu.edu (Z.L.); asim.bag@stjude.org (A.K.B.)
[2] Department of Chemical and Biomedical Engineering, University of Missouri-Columbia, Columbia, MO 65211, USA
[3] Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA
* Correspondence: josueluiz.dalbonidarocha@stjude.org (J.L.D.d.R.); ranganatha.sitaram@stjude.org (R.S.)

Simple Summary: Artificial intelligence (AI) is transforming how doctors use brain imaging to diagnose and treat diseases. While there has been significant progress in using AI for adult brain scans, less is known about its benefits for children with brain cancer. Our review examines how AI can improve pediatric brain imaging to detect and treat cancer more effectively. We found that AI can make imaging faster and safer for children by reducing the time they spend in scanners and lowering their exposure to radiation and contrast dyes. AI also helps doctors identify tumors more accurately and predict how well treatments might work. However, challenges like limited data from children and the need for AI tools that doctors can easily understand still exist. We suggest ways to overcome these hurdles so that AI can better assist in caring for children with cancer in the future.

Abstract: Background/Objectives: Artificial intelligence (AI) is transforming neuroimaging by enhancing diagnostic precision and treatment planning. However, its applications in pediatric cancer neuroimaging remain limited. This review assesses the current state, potential applications, and challenges of AI in pediatric neuroimaging for cancer, emphasizing the unique needs of the pediatric population. Methods: A comprehensive literature review was conducted, focusing on AI's impact on pediatric neuroimaging through accelerated image acquisition, reduced radiation, and improved tumor detection. Key methods include convolutional neural networks for tumor segmentation, radiomics for tumor characterization, and several tools for functional imaging. Challenges such as limited pediatric datasets, developmental variability, ethical concerns, and the need for explainable models were analyzed. Results: AI has shown significant potential to improve imaging quality, reduce scan times, and enhance diagnostic accuracy in pediatric neuroimaging, resulting in improved accuracy in tumor segmentation and outcome prediction for treatment. However, progress is hindered by the scarcity of pediatric datasets, issues with data sharing, and the ethical implications of applying AI in vulnerable populations. Conclusions: To overcome current limitations, future research should focus on building robust pediatric datasets, fostering multi-institutional collaborations for data sharing, and developing interpretable AI models that align with clinical practice and ethical standards. These efforts are essential in harnessing the full potential of AI in pediatric neuroimaging and improving outcomes for children with cancer.

Keywords: artificial intelligence; deep learning; machine learning; neuroimaging; cancer; medulloblastoma; craniopharyngioma; low-grade glioma

1. Introduction

Pediatric neuroimaging plays a pivotal role in diagnosing and managing neurological conditions affecting children with cancer. As technological advancements, particularly in artificial intelligence (AI), continue to reshape medical practices, it becomes imperative to assess the evolving landscape of AI applications in this specialized field. Children present unique challenges due to ongoing brain development, distinct pathologies, and the necessity for child-friendly imaging protocols. Clinicians often face practical challenges such as limited imaging time due to patient discomfort and the need for sedation, which can impact image quality and diagnostic accuracy. From an AI expert's perspective, technological limitations include the scarcity of large, high-quality datasets specific to pediatric populations, which hampers the development of robust AI models.

Patients and their families are central to this discussion. Improved diagnostic tools can lead to earlier detection, less invasive procedures, and more personalized treatments, significantly impacting patient experiences and outcomes. Therefore, integrating AI into pediatric neuroimaging is not just a technological advancement but a patient-centric imperative.

This review aims to address the scarcity of comprehensive assessments focusing on the intersection of AI, neuroimaging, and pediatric cancer, providing an understanding of the current state, potential applications, and limitations in this evolving field. In alignment with widely accepted clinical and research classifications, we define pediatric patients in this review as individuals aged 0–18 years. However, we also include select studies based on adult or general population data when their findings provide relevant translational insights into pediatric cancer. While existing reviews on AI in pediatric neuroradiology [1] and on AI in pediatric imaging [2] workflows provide valuable insights, they primarily focus on general pediatric neuroradiology and broader imaging applications. In contrast, our manuscript examines the specific challenges and advancements in applying AI to pediatric cancer neuroimaging, a field with distinct needs and complexities. By addressing this gap, we aim to highlight critical opportunities for improving diagnostic precision, treatment planning, and patient outcomes in pediatric cancer care.

In the landscape of pediatric neuroimaging, this review stands apart by directing its focus toward a specific and critical realm: AI applications in neuroimaging for pediatric cancer. While existing reviews offer comprehensive insights into general pediatric neuroradiology [1,3–5], our manuscript takes a distinctive approach, homing in on the unique challenges and advancements within the realm of catastrophic conditions in pediatric patients. This focus allows us to unravel intricacies, highlight advancements, and identify opportunities that are particularly pertinent to the complex landscape of pediatric cancer.

AI refers to the computational automatization of tasks that mimic human intelligence. AI may even surpass human intelligence in areas such as memory storage, working memory capacity, parallel multitasking, steadfast decision-making criteria, and pattern recognition [6]. Among the various branches of AI, machine learning stands out for its ability to enable computational systems to learn, make decisions, and adapt to new tasks based on data inputs, without explicit reprogramming [7].

Machine learning can be categorized as supervised, unsupervised, reinforcement learning, or evolutionary. In supervised learning, algorithms are trained on labeled data for which the input and corresponding output are provided, and the goal is to learn a mapping between inputs and outputs [8]. This category includes classification, regression, and time series forecasting. In unsupervised learning, algorithms are trained on unlabeled data, and their goal is to discover patterns and structures within the data without specific guidance [9]. This category includes clustering, dimensionality reduction, and anomaly detection. In reinforcement learning, agents interact with an environment and learn to take

actions to maximize a cumulative reward, which guides them toward achieving specific goals [10]. This category includes model-based, model-free, and deep learning. As a more recent type of machine learning, deep learning incorporates artificial neural networks that simulate the learning process and structure of the human brain (Figure 1). In evolutionary learning, the concepts of natural selection are represented in code to iterate a population of "organisms" to obtain an optimal solution. The steps involved in evolutionary learning are parent selection, progeny generation, and population evaluation [11]. Such technologies are rapidly reshaping the landscape of medical imaging by enabling more accurate, efficient, and personalized diagnostic and treatment approaches.

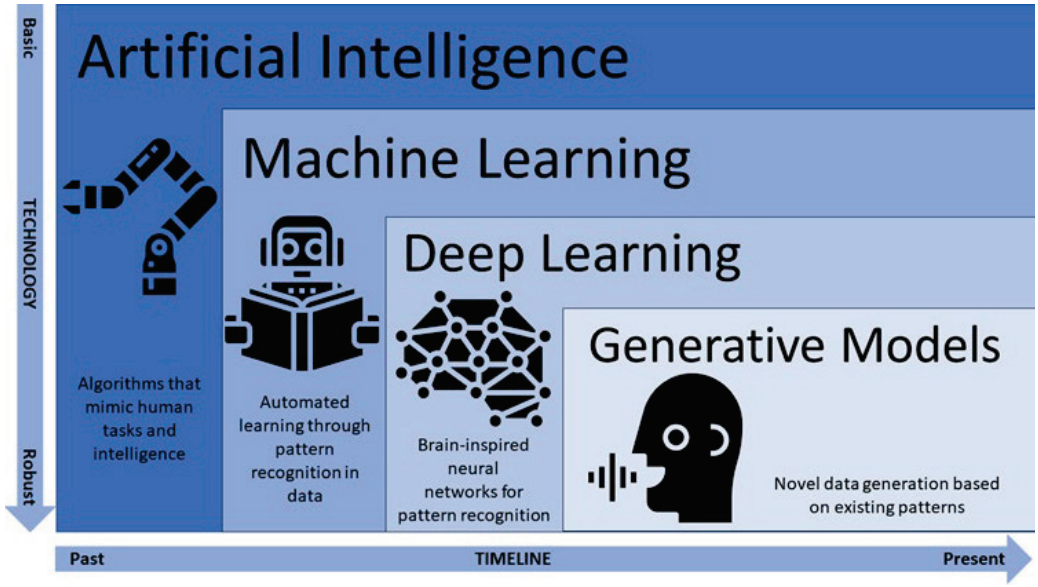


Figure 1. Chronological development of sub-categories and approaches in artificial intelligence.

The use of AI in pediatric neuroimaging has increased in recent years (Figure 2). Some common deep learning algorithms applied in neuroimaging include convolutional neural networks (CNNs), generative adversarial networks (GANs), and autoencoders. CNNs are deep learning approaches consisting of multiple neural layers [12]. When the volume of a sample, for example, three-dimensional (3D) images, is used as input data to train a CNN, the developed network is a 3D-CNN. Recently, 3D-CNNs have been used to identify patterns in neuroimaging data [13]. AlexNet, with deeper and stacked convolutional layers [14], and GoogLeNet, with 22 layers developed by Google researchers [15], are CNN models commonly used in object detection and image classification. Moreover, the residual neural network (ResNet), which is also a CNN architecture containing deeper layers that learn residual functions with reference to the input layer [16], has gained popularity in image recognition after AlexNet and GoogLeNet. Conversely, GANs can be especially applicable when the training data are limited, i.e., 110 images per class, because they can create new data resembling the training data [17]. A typical GAN has two components: a generator (a CNN that learns to generate data) and a discriminator (a classifier that can tell how "realistic" the generated data are). Autoencoders are effective at reconstructing images because they are feedforward CNNs that are trained to first encode the image then transform it to a different representation and eventually decode it to generate a reconstructed image. Autoencoders are a special subset of encoder–decoder models with identical input and output domains.

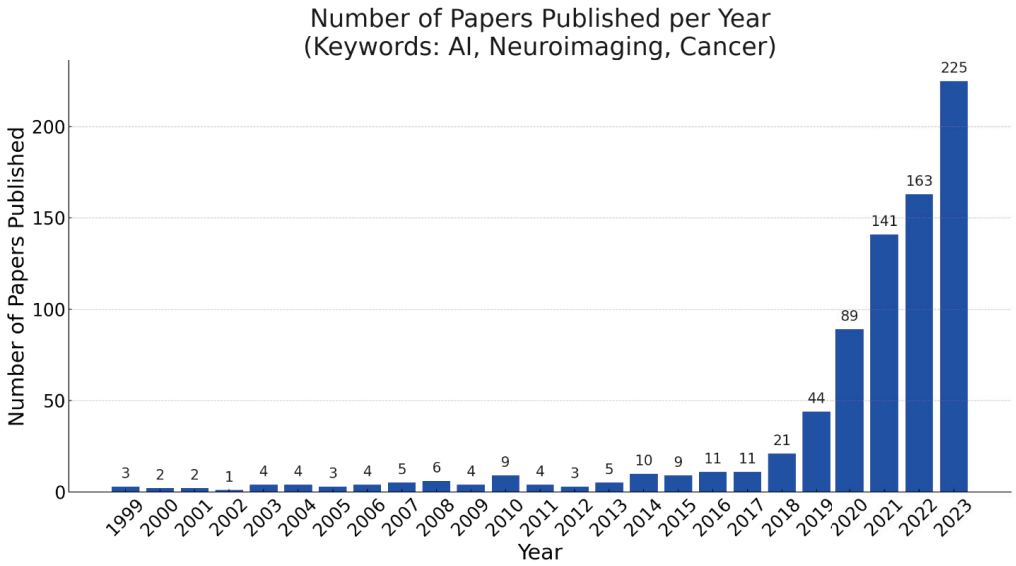


Figure 2. Number of papers published per year containing the keywords "artificial intelligence", "neuroimaging", and "cancer", according to the ScienceDirect search tool (www.sciencedirect.com/search (accessed on 31 October 2024).

Neuroimaging is a branch of medical imaging that maps the brain in a noninvasive manner to understand its structural, functional, biochemical, and pathological features. Neuroimaging has led to significant advances in diagnosing and treating neurological abnormalities. Some neuroimaging modalities are magnetic resonance imaging (MRI) (which includes functional, structural, diffusion, angiographic, and spectroscopic imaging), electroencephalography (EEG), magnetoencephalography (MEG), near-infrared spectroscopy (NIRS), positron emission tomography (PET), and computed tomography (CT). Functional neuroimaging refers to imaging the brain while participants perform set tasks to study task-related functions of the brain, rather than its structure.

Recently developed deep learning algorithms can perform object segmentation and recognition in a short time with high accuracy. These algorithms offer the advantages of saving time and optimizing the use of resources. As a result, AI has gradually been integrated into daily medical practice and has made considerable contributions to medical image processing.

Neuroimaging has seen substantial integration of AI across various domains encompassing a wide spectrum of functions (Figure 3). For example, CNNs have been effectively used to segment medulloblastomas from pediatric MRI data, enabling precise tumor delineation critical for treatment planning. A study [18] demonstrated the application of CNN-based models to automatically segment pediatric low-grade glioma (LGG) tumors, achieving high accuracy and significantly reducing the time required for manual segmentation by clinicians. This widely encompassing AI use spans improving radiology workflow [19], real-time image acquisition adjustments [20], image acquisition enhancement [21], and reconstruction tasks such as bolstering image signal-to-noise ratios (SNRs), refining image sharpness, and expediting reconstruction processes [22]. AI has also found applications in noise-reduction efforts [23], postprocessing of images, predicting specific absorption rates [24], prioritizing time-sensitive image interpretations [25], predicting and classifying types and subtypes of brain lesions based on molecular markers [26,27], and tissue characterization [28–30], as well as in anticipating treatment responses [31,32] and survival outcomes [33]. Importantly, these AI applications serve to elevate the efficiency of individual neuroscientists and physicians, optimize departmental workflows, enhance

institutional-level operations, and ultimately enrich the experiences of pediatric patients and their families.

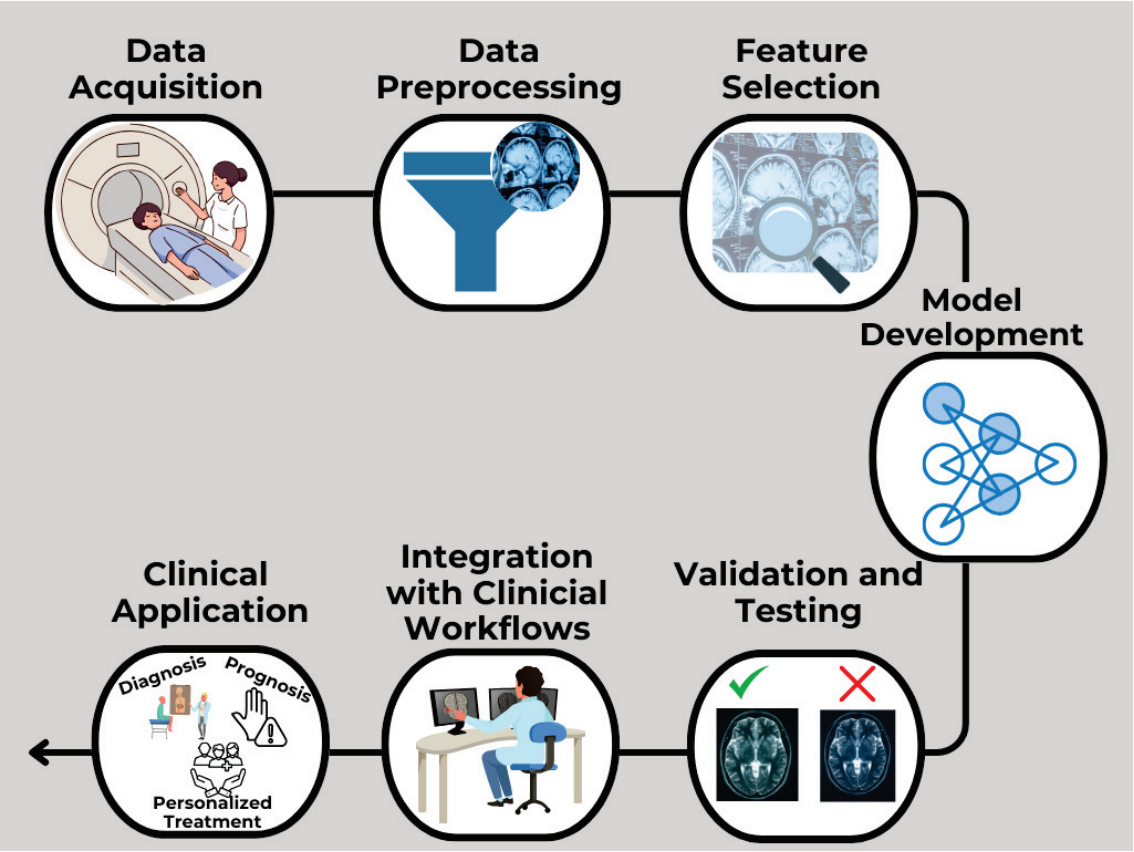


Figure 3. Workflow for artificial intelligence based pediatric neuroimaging.

As AI has demonstrated considerable usefulness and potential in neuroimaging, applying AI to pediatric neuroimaging for cancer could have a significant impact on patient outcomes and healthcare. Deep learning, a prominent branch of AI, has significantly transformed healthcare applications, particularly in medical imaging and diagnostics, by enabling faster, more accurate, and noninvasive approaches to disease detection and management [34]. AI holds immense promise in elucidating and addressing challenges in pediatric neuroimaging. Pediatric cancer encompasses many severe and life-threatening conditions that affect children from infancy through adolescence. This disease is characterized by its devastating impact on a child's overall health, development, and well-being. Children with cancer may experience chronic pain, physical disabilities, cognitive impairments, and disruptions of their normal growth and development [35]. The pediatric cancers addressed in this review include craniopharyngioma, LGG, and medulloblastoma. Craniopharyngiomas are rare, with an annual incidence of approximately 0.5–2 cases per million children and accounting for 5–10% of all pediatric brain tumors [36]. LGG is the most common pediatric brain tumor, comprising 30–40% of all childhood central nervous system tumors [37]. Medulloblastomas, the most common malignant pediatric brain tumors, represents 26% of all pediatric central nervous system tumors [38]. Together, these cancers pose significant clinical challenges due to their impact on neurodevelopment and treatment complexity.

Pediatric cancers and their treatments are often associated with cognitive impairment, affecting various aspects of a child's development and quality of life. Craniopharyngioma is a common, mostly benign, congenital tumor of the central nervous system that arises

near the pituitary gland and hypothalamus. These tumors often lead to visual disturbances due to their proximity to the optic chiasm and can also affect hormonal regulation, resulting in endocrine dysfunctions such as growth hormone deficiency and hypothyroidism [39]. While pediatric patients with craniopharyngioma typically demonstrate intact intellectual functioning, impairments of memory, executive function, attention, and processing speed are frequently observed. These cognitive deficits are often exacerbated by the tumor's impact on adjacent brain structures and the long-term effects of surgical and radiation treatments [40,41].

Gliomas, which originate from the glial cells of the brain [42], are categorized into four grades based on their histological characteristics and growth rates. Low-grade gliomas (LGGs; grades I and II) are generally less aggressive, whereas high-grade gliomas (HGGs; grades III and IV) are highly malignant [43]. Neurocognitive deficits in children with LGGs can range from memory impairment and attention problems to decreased processing speed, depending on the location of the glioma [44]. Tumors in the temporal or frontal lobes, for example, may result in memory impairments, attention problems, and reduced processing speed, while lesions in the parietal lobe can affect visuospatial skills and numerical cognition. Treatment modalities such as surgery and chemotherapy may further contribute to these deficits, underscoring the need for precise and less invasive diagnostic approaches.

Medulloblastoma is a malignant, invasive embryonal tumor that originates in the cerebellum or posterior fossa and spreads throughout the brain via the cerebrospinal fluid [45]. This highly invasive tumor is most common in children and presents a significant challenge due to its rapid progression and the long-term consequences of treatment. Survivors of medulloblastoma exhibit a progressive decrease in cognitive performance over time [46], in addition to deficits in attention, processing speed, and memory. This decline is compounded by deficits in attention, processing speed, memory, and executive function, which are likely related to both the tumor itself and the neurotoxic effects of craniospinal irradiation and chemotherapy. Furthermore, the cerebellar origin of medulloblastoma may uniquely impair motor coordination and cognitive processes associated with cerebellar function.

These cancers highlight the critical need for innovative diagnostic tools, such as AI-based neuroimaging, to address the limitations of traditional methods, mitigate treatment-associated cognitive deficits, and improve outcomes for pediatric patients. The unique characteristics of the developing pediatric brain introduce complexities that necessitate tailored approaches for accurate diagnoses and effective treatment strategies. This review explores the evolving landscape where AI intersects with pediatric neuroimaging, emphasizing the distinctive considerations and challenges inherent to the pediatric population. As we delve into the intricacies of AI applications in pediatric neuroimaging, it is crucial to recognize that this field necessitates a multidisciplinary understanding. Clinicians, neuroimaging specialists, and AI experts converge to navigate the complexities posed by the developing brain.

Our objective is to provide insights accessible to AI experts, neuroimaging specialists, and clinical practitioners alike. The convergence of these domains is pivotal in fostering collaborative solutions that enhance the understanding and treatment of pediatric neurological conditions. By establishing this common ground, we aim to propel the field forward, leveraging the potential of AI in pediatric neuroimaging to improve diagnostics, treatment planning, and outcomes for our young patients.

This review is organized into the following sections: challenges in pediatric neuroimaging, AI for improving image acquisition and preprocessing, AI for tumor detection and classification, AI for functional neuroimaging and neuromodulation, AI for selecting

and modifying personalized therapies, and AI applications for specific pediatric diseases, followed by a discussion and conclusions.

2. Challenges in Pediatric Neuroimaging

Multiple challenges and risks are associated with pediatric image acquisition, which vary significantly depending on the age and developmental stage of the child. Younger children, particularly those under 5 years old, are generally unable to stay still for extended periods, are more prone to movement during scans, and have limited tolerance for long scanning durations. This often necessitates general anesthesia to mitigate motion artifacts, ensure patient comfort, and maintain image quality, though this approach introduces concerns about anesthesia-related risks and costs. While less likely to require anesthesia, older children may still experience challenges related to motion and discomfort, albeit to a lesser extent.

Innovative motion correction techniques and child-friendly imaging protocols are needed to mitigate the impact of motion artifacts on pediatric neuroimaging data [47]. AI has emerged as a valuable tool for addressing these challenges. Automated motion correction algorithms, guided by machine learning, enhance the quality of acquired images by mitigating the impact of motion artifacts, particularly prevalent in younger pediatric populations. Scanning time remains a critical challenge across all age groups, as the need to re-scan motion-corrupted data can prolong imaging procedures. In addition to these challenges, concerns about ionizing radiation exposure and the side effects of contrast agents are universal among children and their parents. Obtaining high-quality images, ideally with high contrast, good spatial resolution, and high SNRs, is critical to ensure accurate diagnoses and suitable treatment plans. However, reducing the dose of contrast agents, while beneficial in minimizing risks, may sacrifice image quality and decrease SNRs.

In addition to the challenges faced during image acquisition, limitations are also encountered in image reconstruction. For example, MR images are acquired in the Fourier or spatial-frequency domain (also known as k-space); hence, image reconstruction is important in transforming the raw data into clinically interpretable images [48]. An inverse Fourier-transform operation is required to reconstruct an image, which is collected based on the Nyquist sampling theorem [49]. However, because of the considerations regarding minimizing scan time, optimizing patient comfort, and ensuring safety in pediatric populations, only a limited number of measurements are acquired by the scanners, which leads to difficulties in solving the inverse transform [50]. The fundamental physics, practical engineering aspects, and biological tissue response factors underlying the image-acquisition process make fully sampled MRI acquisitions very slow [48]. Furthermore, patient motion, system noise, and other imperfections during scanning can corrupt the collected raw data.

Navigating the landscape of pediatric neuroimaging is inherently fraught with challenges, each intricately tied to the unique characteristics of the developing brain. AI emerges as a transformative ally, offering tailored solutions to overcome these hurdles and enhance diagnostic precision and treatment efficacy in pediatric populations. Pediatric neuroimaging faces a substantial challenge due to the scarcity of comprehensive datasets specific to this demographic. Several resources address the scarcity of pediatric neuroimaging data, including the NIH Pediatric MRI Dataset [51] and the Haskins Pediatric Atlas [52] for mapping developmental changes, which provides open-access neuroimaging data with a pediatric focus. These datasets enable standardized analysis across studies and institutions. AI interventions can play a pivotal role in overcoming this limitation through innovative techniques like transfer learning, enabling models trained on larger datasets to be fine-tuned for pediatric applications. Collaborative efforts for data sharing

among institutions and research centers further amplify the potential for developing robust AI models.

The diverse and dynamic nature of pediatric brain development introduces a layer of complexity in image analysis. AI algorithms, particularly those equipped with advanced learning capabilities, hold the promise of deciphering intricate patterns associated with different stages of brain maturation. This adaptability ensures that neuroimaging analyses account for the nuances of evolving pediatric neuroanatomy.

A major challenge in pediatric neuroimaging is the variability in brain structure, function, and chemistry across ages due to rapid brain development during childhood. The developing brain undergoes quick structural and functional changes over time, which presents unique challenges in pediatric neuroimaging. For example, in early childhood, the rapid change in the myelination of brain white matter creates variability that complicates the interpretation of imaging data. Pediatric brains are not merely scaled-down versions of the adult brain; there are age-dependent loco-regional variations that require specialized approaches. Variability in myelination in patients younger than 6 months often results in poor gray–white matter myelination contrast, impacting the quality of conventional automatic segmentation methods [53]. Additionally, age-related motion artifacts during imaging are particularly prevalent in younger children who struggle to remain still during scans, necessitating the use of motion correction algorithms or general anesthesia to ensure high-quality imaging.

As the brain matures, there are significant structural changes, such as the formation of new sulci during infancy and the deepening of older sulci. Fractional anisotropy (FA) values on diffusion MRI, for instance, increase with developmental age, reflecting the maturation of white matter tracts [54]. These changes continue through adolescence and into early adulthood, with the brain evolving in shape and white/gray matter composition until the fourth decade of life [55,56]. Accurately interpreting these age-related differences requires careful consideration of age-appropriate image acquisition and analysis methods [57]. These examples highlight how different developmental stages, from infancy through adolescence, demand tailored neuroimaging approaches to account for the variability and complexity of childhood brain development.

Anatomical metrics of normal brain development involve identifying myelination and gyrification patterns in individuals or in a specific population and comparing them to the patterns in the brains of controls of the same age [58–60]. Accordingly, age-specific templates and atlases are essential to account for variations in brain morphology across different developmental stages [61]. Spatial normalization for MRI is based on a standard template that defines a common coordinate system for group analysis. ICBM152 and MNI305 are among the most used templates for adults; however, MNI305 is suboptimal for normalization and segmentation of pediatric brain images because of the previously discussed age-related changes [61]. Some pediatric templates are available, such as the custom age templates produced by Template-O-Matic [62] and neonate templates [63,64]. However, there is a lack of corresponding atlases for those templates. The Haskins pediatric atlas [52] labels 113 cortical and subcortical regions, but only 72 brains were used in its development. Longitudinal imaging studies that track brain development over time are essential for understanding neurodevelopmental trajectories and for identifying early signs of neurological disorders. Longitudinal studies provide valuable insights into the dynamic changes in the developing brain and offer opportunities for early interventions [65].

Another challenge in pediatric neuroimaging is the small sample sizes of the currently available pediatric data, which significantly impacts the training and effectiveness of AI algorithms. Large sample sizes are essential for robust statistical analyses and generalizability of results. However, recruiting sufficient pediatric participants for neuroimaging

studies can be challenging because of consent issues, logistical constraints, and the potential impact of imaging on young patients. These limitations are further compounded when obtaining reliable labeled data, as labeling often requires expert annotation and is particularly complex when integrating multimodal sources like structural MRI, functional MRI, and diffusion tensor imaging (DTI). Multi-site collaborations and data-sharing initiatives could help address this challenge [66], enabling the pooling of data across institutions to enhance both dataset size and diversity [67]. Additionally, integrating data from multiple neuroimaging modalities provides a comprehensive view of brain development and connectivity, allowing researchers to examine brain changes at both the macroscopic and microscopic levels, which is crucial for training AI models to extract meaningful patterns on the developing brain [68].

The concept of multi-institutional data sharing in the realm of healthcare and medical research is indeed complex and multifaceted. It involves exchanging sensitive information across different organizations, which inevitably gives rise to a range of legal, ethical, and privacy concerns. From a legal standpoint, sharing medical data across institutions often requires navigating a complex web of regulations and compliance standards, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in Europe. These regulations are designed to safeguard patient privacy and data security, making it imperative for institutions to ensure that any data-sharing practices are in strict compliance with these legal frameworks.

There are significant ethical considerations surrounding multi-institutional data sharing. These encompass questions related to patient consent, data ownership, and the potential for unintended consequences. Institutions must grapple with issues such as obtaining informed consent from patients for data sharing, ensuring that their rights are respected, and addressing concerns about how their data will be used, especially in research contexts.

Moreover, privacy concerns are paramount when sharing medical data. Patient information must be de-identified and protected rigorously to prevent data breaches or the identification of individuals. Striking the correct balance between sharing data for the common good of medical research and preserving individual privacy is a critical ethical challenge. The notion of federated learning holds promise as a potential solution to these barriers [69]. It operates on the principle of decentralized AI training, whereby machine learning models are developed collaboratively across multiple institutions without the need to centrally pool sensitive data. Instead, the models are trained locally on each institution's data, and only aggregated model updates are shared, thus preserving data privacy. However, it is important to note that federated learning is still an evolving concept and that it faces its own set of challenges, including technical complexities and standardization issues.

Research involving pediatric populations also raises ethical concerns, particularly regarding informed consent and the vulnerability of child participants. Balancing the potential benefits of neuroimaging research with protecting the rights of children is crucial. Ethical guidelines and careful consent procedures must be implemented to ensure the well-being and privacy of pediatric participants [70].

3. AI for Improving Image Acquisition and Preprocessing

The optimization of image acquisition and preprocessing is crucial in neuroimaging, significantly impacting the accuracy and reliability of subsequent analyses. In this section, we delve into the challenges and advancements related to these processes, highlighting the pivotal role of AI.

AI plays a relevant role in preprocessing steps, encompassing image denoising, normalization, and registration. These processes are essential for creating a standardized

and comparable dataset. AI algorithms contribute significantly to noise reduction, enhancing SNRs and subsequently improving the accuracy of downstream analyses [66]. Registration and normalization are critical steps in aligning imaging data across individuals or different imaging modalities. In pediatric neuroimaging, challenges arise due to the dynamic changes in brain anatomy during development. Conventional methods may struggle with the anatomical variability present in pediatric populations, necessitating advanced solutions. Accurate registration and normalization are fundamental for creating population-based templates and atlases, facilitating a common spatial framework for analysis. AI-based registration algorithms, capable of adapting to the unique anatomical features of pediatric brains, contribute to improved spatial normalization, ensuring more precise comparisons across subjects [71].

3.1. Accelerating Image Acquisition

In MRI, GAN sensing (CS) is used to accelerate scanning and reduce the resources required by acquiring MR images with good in-plane resolution but poor through-plane resolution [72]. CS assumes that suitably compressed under-sampled signals can be reconstructed accurately [73] without the need for full sampling. However, this technique limits interpretation to single directions and can introduce aliasing artifacts [74]. A deep learning-based algorithm called synthetic multi-orientation resolution enhancement (SMORE) has been applied to adult compressed scans in real time to reduce aliasing and improve spatial resolution [72]. A similar approach can, perhaps, be applied in pediatric MRI to reduce scan time and improve image resolution.

Moreover, applying GANs has been shown to improve the SNR of images [75]. However, despite their potential, GANs present several limitations that must be considered when applied to pediatric neuroimaging. One key challenge is the instability of adversarial training, where the generator and discriminator engage in a dynamic optimization process that can lead to oscillatory behavior and difficulty in convergence. Another major issue is mode collapse, where the generator learns to produce only a limited variety of outputs, thereby reducing the diversity of synthetic images and potentially biasing the training set. Additionally, GANs require high computational resources, making their widespread adoption in clinical settings more challenging. These limitations are particularly relevant in pediatric neuroimaging, where the scarcity of high-quality datasets may exacerbate mode collapse and impact generalizability.

Several studies have examined the use of GANs in medical imaging, highlighting both their advantages and constraints. GANs have been successfully used for image denoising, super-resolution enhancement, and data augmentation in adult neuroimaging [76], but their application to pediatric imaging remains limited due to dataset constraints and the increased variability in brain development across different ages [77]. Furthermore, studies quantifying GAN performance across different dataset sizes indicate that smaller datasets significantly impact training stability and generalization, further reinforcing the need for large-scale multi-institutional data sharing [18].

Despite these challenges, GANs remain a promising tool in pediatric neuroimaging. Combining GANs with fast imaging techniques could lead to faster image acquisition for children who are unlikely or unwilling to remain still for an extended period. By enhancing image quality from under-sampled data, GAN-based approaches may even help reduce the need for general anesthesia and/or sedation in younger pediatric patients, offering significant clinical benefits [1]. Addressing the existing limitations while leveraging the advantages of GANs will be crucial for integrating these models into real-world pediatric imaging workflows.

3.2. Reducing Radiation Exposure or Contrast Doses

Children are particularly vulnerable to the harmful effects of ionizing radiation, with the risk of cancer induction approximately 10 times higher in children compared to adults [78]. Radiation treatment can induce secondary malignancies and is associated with higher radiosensitivity of developing tissues. Consequently, the ALARA (as low as reasonably achievable) principle has become a cornerstone in pediatric imaging practices, emphasizing the need to minimize radiation doses while maintaining diagnostic quality.

AI-based approaches, such as low-dose imaging techniques and radiation-free modalities like MRI, play a critical role in advancing this goal. For example, deep learning based on an encoder–decoder CNN has been applied to generate high-quality post-contrast MRI from pre-contrast MRI and low-dose post-contrast MRI [79]. The above study showed that the gadolinium dosage for brain MRI can be reduced 10-fold while preserving image contrast information and avoiding significant image quality degradation. For CT, deep learning has great potential for image denoising based on its use in realizing low-dose CT imaging for pediatric populations. Chen et al. [80] trained another encoder–decoder CNN model to learn feature mapping from low-/normal-dose CT images. This model improved noise reduction when compared with other denoising methods. A deep CNN model using directional wavelets effectively removes complex noise patterns for low-dose CT reconstruction [81]. An autoencoder CNN, which was used to train pairs of standard-dose and ultra-low-dose CT images, could filter streak artifacts (i.e., artifacts appearing between metal or bone because of beam hardening and scatter) and other noise for ultra-low-dose CT images [82]. Furthermore, it is possible to interpolate data from one modality with a CNN trained on data from a completely different modality. Zaharchuk [83] trained a CNN with simultaneously acquired PET/MRI images to improve the resolution of low-dose PET imaging. These developments in deep learning show promise in terms of moving pediatric neuroimaging toward imaging with low or no radiation exposure.

3.3. Removing Artifacts

Accelerated MRI techniques such as CS and parallel imaging offer significant reductions in scan time. However, reconstructing images from the under-sampled data involves computationally intensive algorithms, thus posing a notable challenge because of the high computational costs incurred. Lee et al. [84] proposed deep ResNets that are much better at removing the aliasing artifacts from subsampled k-space data when compared with current CS and parallel reconstructions. Their deep learning framework provides high-quality reconstruction with shorter computational time than is required for CS methods. In addition to accelerating reconstruction, AI approaches can correct artifacts, such as those arising from MRI denoising [85] and motion correction [86,87] during image acquisition and reconstruction. Singh et al. [88] built two neural network layer structures by incorporating convolutions on both the frequency and image space features to remove noise, correct motion, and accelerate reconstruction. Deep learning-based approaches have also been investigated to reduce metal artifacts [89,90] that are common in CT imaging.

4. AI for Tumor Detection and Classification

AI holds great promise for enhancing the interpretability and utility of pediatric neuroimaging data. Methods such as automated image analysis can aid in identifying subtle brain alterations associated with neurodevelopmental disorders and can facilitate early diagnosis and personalized treatment strategies [91].

Deep learning neural networks, such as CNN, have proved successful in automatically segmenting infant brains and delineating gray–white boundaries in neonates [52]. The use of deep learning in segmenting individual cortical and subcortical regions of interest (ROIs)

based on features such as local shape, myelination, gyrification patterns, and clusters of functional activation is a promising field for exploration. AI can also be applied to tumor detection and categorization.

In children with cancer predisposition syndromes, surveillance screening increases the chance of early tumor detection and, therefore, the survival rate [92]. Deep learning algorithms may help in identifying imaging protocols to achieve optimal accuracy for early cancer diagnosis [93]. Soltaninejad et al. [94] used CNN for automated brain tumor detection in MRI scans. CNNs have been used to detect and characterize tumors [95] and to find associations between genotypes and imaging tumor patterns. Machine learning has also been used to classify pediatric brain tumors [96–98].

CNNs have the potential to identify the best model for categorizing pediatric brain abnormalities by combining brain features extracted from different mathematically derived measures, such as the spherical harmonic description method (SPHARM) [99], the multivariate concavity amplitude index (MCAI) [100], and Pyradiomics [101]. Moreover, CNNs can help in identifying the best approach for feature selection in pediatric brain data, such as a superpixel technique based on a simple linear iterative clustering (SLIC) method [94], combining spatial distance, intensity distance, spatial covariance, and mutual information.

4.1. AI in Tumor Segmentation

Early detection and precise classification of brain tumors are important for effective treatment [102–105]. The two main types of classification that can be performed based on brain images are classification into normal and abnormal tissues (i.e., whether a tumor is detected in the brain image) and classification into different classes of brain tumor (e.g., low-grade vs. high-grade) [106].

Segmenting tumors with AI methods has recently attracted much attention as a possible means of achieving more precise treatment, as described in Table 1. AI accomplishes brain tumor segmentation by identifying the class of each voxel (e.g., normal brain, glioma, or edema). Frameworks like pathomic fusion highlight the power of integrating multimodal data, including imaging and histopathology, to enhance the precision of cancer diagnosis and prognosis [107]. Additionally, the nnU-Net framework exemplifies a self-configuring deep learning approach that has set benchmarks for segmentation tasks across diverse biomedical datasets, further advancing the accuracy and reliability of tumor segmentation methods [108]. Two main AI methods have been reported in the literature: (1) hand-engineered features used with older classification methods (e.g., support vector machine [SVM] classifiers) and (2) deep learning using CNNs. CNNs and their variations can self-learn from a hierarchy of complex features to perform image segmentation [109].

Table 1. Summary of key studies evaluating AI-based tumor segmentation models in pediatric neuroimaging, including dataset size details and performance metrics.

Study	Dataset Size	Performance
[102]	Training: 184 samples/testing: 100 samples	Dice coefficient: 0.877
[103]	Training: 261 samples/testing: 112 samples	Dice coefficient: 0.657–0.967
[104]	Training: 261 samples/testing: 30 samples	Dice coefficient: 0.642
[105]	Training: 364 samples/testing: 125 samples	Dice coefficient: 0.98

4.2. AI in Tumor Margin Detection

Many brain tumors exhibit a distinctively infiltrative nature, which often results in poorly defined tumor margins. Compounding this challenge, the edema surrounding these tumors frequently manifests imaging characteristics similar to those of the tumor itself, further complicating the accurate delineation of the true tumor boundary. The

precise identification of this boundary holds immense significance, as it serves as a guiding principle for neurosurgeons aiming not only for gross total resection but also for margin-negative surgery, which is a critical factor in significantly enhancing patient survival rates.

To address this complex issue, advanced MRI techniques and PET have been employed in attempts to refine the delineation of tumor margins, albeit with varying degrees of success. More recently, AIs based solely on MR images or on a combination of MRI and PET data have exhibited potential for substantially improving the accuracy of true tumor margin detection [69]. For example, convolutional neural networks (CNNs) have been trained on multi-modal imaging datasets, such as MRI and PET, to generate highly accurate tumor boundary predictions. The nnU-Net (no-new-net) framework, which adapts its architecture to a given dataset, has shown state-of-the-art performance in segmenting brain tumor margins, achieving a dice similarity coefficient (DSC) of up to 0.9 [105]. Similarly, a deep learning approach combining MRI features with radiomic data has been used to differentiate infiltrative tumor boundaries from surrounding edema, enabling a more precise identification of tumor margins [110]. Moreover, GANs have been applied to enhance tumor margin visualization by synthesizing high-resolution images from low-quality scans, further aiding neurosurgical planning [76].

These innovative AI-driven techniques represent a significant leap forward in the field of neuroimaging, offering new avenues for improving the precision of brain tumor surgery and ultimately enhancing patient outcomes.

4.3. AI in Tumor Characterization

As high-throughput computing facilitates converting multimodal medical images into mineable high-dimensional data, there has been increased interest in using radiomics and radiogenomics to detect and classify tumors. Radiomics refers to studies or approaches that extract quantitative high-throughput features, often imperceptible to the naked eye, from radiographic images. These features are typically analyzed using machine learning techniques for tumor characterization, incorporating AI in neuro-oncological imaging [111]. Radiomics has multifaceted applications, with one of its primary roles being tumor diagnosis and classification. By elucidating distinct radiomic signatures, it empowers radiologists to augment diagnostic accuracy and timeliness, in addition to predicting disease trajectories, thereby enabling tailored treatment strategies. By tracking changes in radiomic features over the course of treatment, it can play a pivotal role in evaluating therapeutic efficacy and making timely adjustments when indicated. In oncology, this capability is of paramount importance for optimizing patient outcomes. The true promise of radiomics lies in the realm of personalized medicine, as it integrates imaging data with clinical parameters, genomics, and other patient-specific variables. This convergence provides the foundation for precision medicine by tailoring treatments to individual patients, thereby improving therapeutic outcomes while mitigating adverse effects.

Radiogenomics uses such radiographic image features to detect relationships specifically with genomic patterns [112]. Machine learning-based radiomics offers critical advantages in treating brain tumors that are genetically heterogeneous, and it can provide better predictions by linking genomics to extraordinarily complex imaging phenotypes.

Applying deep learning in radiomics is helpful in assessing higher-order features to improve the accuracy of prediction of brain tumor progression. Taking advantage of the 3D nature of MRI data, Casamitjana et al. [113] and Urban et al. [114] showed that a 3D-CNN could perform well for brain tumor segmentation. In addition, a CNN-based classification system can be used to segment and classify brain tumors [115].

4.4. Radiomics and Radiogenomics for Specific Pediatric Brain Tumors

The 2021 World Health Organization classification of central nervous system tumors highlights the pivotal role of molecular classification in diagnosing and treating pediatric brain tumors [116]. These tumors are currently characterized by using molecular markers. However, the techniques required for such molecular subgrouping, encompassing immunohistochemistry and genetic testing, are often not consistently available and are associated with significant delays, even in well-equipped healthcare settings. These delays introduce complexities into patient care, affecting various aspects ranging from prognosis and surgical strategies to treatment choices and participation in clinical trials. Given this pressing need for a more efficient approach, machine learning methods hold great promise for rapidly and accurately predicting molecular markers in pediatric neuro-oncology practice [117]. For example, a machine learning toll called CHARM can provide AI-based diagnoses from microscope images in less than a second, offering a faster alternative to current molecular sequencing methods that can take days to weeks [118]. While these approaches show promise, more research is needed to improve performance for rare diagnoses and to assess generalizability across diverse populations [116].

4.4.1. Posterior Fossa Tumors

Posterior fossa tumors represent a challenging group of central nervous system tumors that primarily affect children. Traditionally, diagnosis involves imaging-based tumor grading using modalities like MRI and CT, followed by invasive tissue biopsies for histopathological analysis. These manual processes, although reliable, are time-intensive and carry risks of complications, especially in pediatric patients. Visual system evaluations, including assessments of oculomotor function and visual acuity, are often performed to identify secondary effects of tumor compression on nearby structures. Neurocognitive assessments may also be used to evaluate the impact of the tumor on cognitive functions such as memory, attention, and processing speed. These evaluations provide critical insights but require significant time and expertise.

Machine learning has emerged as a promising avenue for noninvasive diagnosis, having demonstrated its ability to attain exceptional diagnostic accuracy with an impressive area under the receiver operating characteristics curve (AUROC) exceeding 0.99, on par with that of experienced pediatric neuroradiologists, in distinguishing among the three most prevalent posterior fossa tumors: astrocytoma, medulloblastoma, and ependymoma [119,120]. These techniques have the potential to make tissue biopsies unnecessary in the future and to lead to better treatments [121].

4.4.2. Craniopharyngiomas

Craniopharyngiomas are characterized by their diverse shapes and heterogeneous growth patterns, often complicating manual diagnosis. Imaging-based approaches, primarily using MRI, are the cornerstone of diagnosis and involve tumor localization, size measurement, and morphological analysis. Visual system evaluations are frequently conducted to assess the impact of the tumor on nearby optic pathways, as craniopharyngiomas commonly lead to visual disturbances. Neurocognitive assessments are employed to evaluate memory, executive function, and attention, which can be affected by the tumor's pressure on adjacent brain structures.

However, these manual processes are time-consuming and prone to inter-observer variability [122]. Such limitations can be overcome by using deep learning methods. Prince et al. [123] applied deep learning models for CT, MRI, and combined CT and MRI datasets to pinpoint parameters for identifying pediatric craniopharyngioma. They demonstrated

high test accuracies and exceptional improvement in the performance of their baseline model. Therefore, AI may help to improve the accuracy of diagnosis.

Mutations in the BRAF and CTNNB1 genes in craniopharyngiomas were predicted by applying radiomics and machine learning to MRI data [124]. An SVM model identified 11 optimal features with radiological features and was used to predict preoperative craniopharyngioma invasiveness [125].

4.4.3. Low-Grade Gliomas

Low-grade gliomas (LGGs) present a diagnostic challenge due to their slow progression and variable behavior. Manual grading relies on preoperative imaging and histopathological evaluation, which may not fully capture tumor complexity. AI has significantly advanced LGG diagnosis through radiomics research, where SVM and hybrid approaches have shown superior accuracy compared to manual assessments. Hybrid approaches achieved high accuracy in classifying LGG subtypes [26,126]. Another approach used a custom deep neural network and MRI data to categorize the gliomas as grade II, III, or IV [127]. Machine learning techniques using diffusion parameters have been used to predict the progression of optic pathway gliomas [128].

4.4.4. High-Grade Gliomas

HGGs are among the most aggressive pediatric brain tumors. Manual diagnosis typically involves imaging-based assessments using MRI to evaluate tumor location, size, and infiltration into surrounding tissues, followed by histopathological analysis for molecular marker identification [129]. Among the various subtypes of pediatric high-grade gliomas, machine learning techniques have found their most extensive application in characterizing diffuse midline glioma H3 K27-altered [45]. This tumor typically affects central brain structures that are nearly always non-resectable, resulting in a high fatality rate. Different machine learning techniques have been developed to predict this specific tumor marker with very high accuracy [130–133]. Radiomic analysis based on MRI data also shows promise in predicting progression-free survival among pediatric patients diagnosed with diffuse midline glioma or diffuse intrinsic pontine glioma [134].

4.4.5. Ependymomas

Ependymomas are typically diagnosed using imaging-based approaches, including T2-weighted and post-contrast T1-weighted MRI. These manual processes involve assessing tumor location, size, and enhancement patterns to differentiate between subtypes and guide treatment decisions [135]. AI has enhanced the diagnosis of ependymomas by leveraging radiomic features extracted from T2-weighted MRI and post-contrast T1-weighted MRI through machine learning approaches. In this way, it becomes feasible to differentiate between MRI phenotypes corresponding to two distinct types of posterior fossa ependymoma and to identify high-risk individuals within these groups [24]. Radiomic features can also differentiate supratentorial ependymomas from high-grade gliomas [136]. Moreover, a deep-learning model has been developed to segment posterior fossa ependymomas and predict molecular subtypes [135].

4.5. Hybrid Models

Improving the quality of the input data and carefully selecting the most pertinent features are both crucial steps in training AI models to achieve optimal performance. Conventional types of machine learning algorithms, such as the SVM and K-nearest neighbors (KNN) algorithms, are capable of quantifying and visualizing latent information contained within images [122,137]. Meanwhile, most recent deep learning algorithms, such as CNN, AlexNet, and GoogLeNet, excel at extracting features to derive comprehensive deep or

high-order features [138]. Hence, hybrid models that use both methods may work better in complex cases with multi-source, heterogeneous medical data. This concept was put into practice through a recent development for classifying brain tumors, leading to improved classification accuracy. In the model, a modified GoogLeNet was employed to extract deep features that were subsequently used to train SVM and KNN classifiers [139]. By combining radiomics and deep features extracted by a CNN from medical images and selecting the optimal feature subset as input for an SVM, Ning et al. [140] demonstrated that integrating radiomics and deep features can be used to grade gliomas. Raza et al. [141] proposed a hybrid deep learning model by changing the last 5 layers of GoogLeNet to 15 new layers, thereby obtaining performance that was better overall than that of other pre-trained models.

5. AI for Functional Imaging and Neuromodulation

Functional imaging is used in brain tumor studies in two major ways: for preoperative mapping and to assess postoperative outcomes. In preoperative mapping, functional imaging helps surgeons to understand the spatial relations between lesions and functional areas and enables surgical planning that reduces long-term neurological deficits. One retrospective, propensity-matched study found that patients with LGG who underwent preoperative fMRI subsequently underwent more aggressive surgeries when compared with other patients. Although these surgeries did not significantly change the survival outcomes, non-significant trends of higher postoperative functional improvement were observed in those patients who underwent preoperative fMRI and aggressive surgeries. Postoperative functional imaging studies investigate the functional brain changes sustained by survivors of brain tumors with the goal of improving treatment regimens to reduce cognitive inhibition. For example, one fMRI-based study of pediatric patients with medulloblastoma found evidence of long-term effects of prophylactic reading intervention, including significantly increased sound awareness [142]. Notably, a longitudinal study of medulloblastoma survivors revealed that support vector machine classification of functional MRI data indicated a progressive divergence in brain activity patterns compared to healthy controls over time, suggesting delayed effects of cancer treatment on brain function [143]. Alterations in brain regions involved in visual processing and orthographic recognition during rapid naming tasks were correlated with performance in tasks involving sound awareness, reading fluency, and word attack, highlighting the dynamic nature of post-treatment neurofunctional alterations. Additionally, a functional imaging study showed that adults who had experienced childhood craniopharyngioma exhibited cognitive interference processing abilities on a par with those of the control group, as fMRI of these survivors showed no compensatory activity within the cingulo-fronto-parietal attention network when they were compared to the control group [144].

Traditional deep learning models often lack transparency, making it challenging to understand which features contribute to their classification decisions. The eXplainable AI fNIRS (xAI-fNIRS) system is an innovative approach that addresses the issue of explainability in deep learning methods for classifying fNIRS data. This is achieved by incorporating an explanation module that can decompose the output of the deep learning model into interpretable input features [145].

The goal of neuromodulation is to restore normal neural function in areas affected by neurobiological abnormalities or to facilitate compensatory mechanisms by stimulating alternative neural networks. Once an association is established between a functional neurobiological pattern and a neurocognitive alteration in pediatric catastrophic disease survivors, neuromodulation becomes a potentially relevant and promising strategy for intervention. In the context of pediatric catastrophic disease survivors, who may experience

long-term neurocognitive deficits due to the disease or its treatment, neuromodulation offers several potential applications in neuroplasticity promotion, symptom management, cognitive enhancement, and personalized medicine.

Epilepsy is a neurological disorder characterized by recurrent seizures, and it can occur in pediatric patients with cancer, such as LGGs [146]. The association between epilepsy and these conditions is often related to the location of the underlying disease within the brain and its impact on neural circuits. For epilepsy, DBS is applied to the anterior thalamic nucleus to decrease brain excitement, which in turn decreases the frequency or duration of seizures [147].

Regarding neuromodulation, AI enables faster data collection and monitoring, which can aid in early diagnosis, treatment, patient monitoring, and disease prevention [32]. Machine learning can analyze large datasets to improve the efficiency of neuromodulation. A reinforcement learning paradigm intended to optimize a neuromodulation strategy for epilepsy treatment found a stimulation strategy that both reduces the frequency of seizures and minimizes the amount of stimulation applied [148].

Even though there has been extensive research on using AI in neuromodulation, only a very small subset of this research has pertained to pediatrics. However, a machine learning technique to monitor and predict epileptic seizures in pediatric patients has been reported [149]. The application of scalp EEG data to create a treatment prediction model for vagus nerve stimulation in pediatric epilepsy, using brain functional connectivity features, has also been reported [150].

Incorporating AI into the data collection and monitoring stages of neurofeedback offers the potential for early detection and precise non-pharmacological management of neurological conditions. AI facilitates the analysis of extensive patient data, enhancing the effectiveness and efficiency of neurofeedback processes [151,152]. Hence, there is a need for additional research to explore comprehensively and expand the use of brain–computer interfaces that incorporate AI [153]. Incorporating AI into neurofeedback holds promise for unlocking fresh avenues by which to enhance substantially the effectiveness of these therapeutic approaches for neurological disorders [153].

6. AI in Monitoring Treatment Response

Monitoring the response of pediatric brain tumors to therapy can pose challenges, particularly because of the intense inflammation that may occur in the early phases of treatments such as radiation therapy and immunotherapy. This inflammation can be transient, often improving over time—a phenomenon referred to as pseudoprogression. Importantly, pseudoprogression shares imaging features with true tumor progression, making it crucial to distinguish between the two, as their clinical management strategies differ significantly. Although advanced MRI and PET techniques have been applied to address this issue in pediatric neuro-oncology, it remains a challenge. Encouragingly, machine learning techniques have been deployed to differentiate between these two conditions, demonstrating promising success in this endeavor [154,155].

7. AI in Predicting Survival for Patients with Pediatric Brain Tumors

Despite extensive research, some pediatric brain tumors still carry a grim prognosis, with overall survival often being less than 2 years from the time of diagnosis. The molecular characterization of these tumors was a significant leap forward, offering clinicians valuable prognostic insights. However, considerable variability persists within specific tumor subgroups, and a universally applicable tool for accurately predicting survival in all patients with pediatric brain tumors remains elusive. Machine learning-based techniques present a promising avenue, particularly when it comes to predicting survival at the time of diagnosis

and especially for tumors that are not amenable to surgical resection. A substantial collective effort is now focused on leveraging machine learning to develop predictive tools for these challenging cases. For instance, one notable approach involves a subregion-based survival prediction framework tailored for gliomas using multi-sequence MRI data, achieving an area under the receiver operating characteristic curve (AUC) of 0.98 in predicting survival outcomes [150]. By using radiomic features derived from T1-weighted post-contrast imaging, progression-free survival can be predicted with concordance indices of up to 0.7 [156]. Similarly, a multiparametric MRI-based radiomics signature, integrated with machine learning, demonstrated strong potential for preoperative prognosis stratification in pediatric medulloblastoma, achieving an AUC of up to 0.835 in the validation set [157].

8. AI for Transparent Explanations in Cancer Neuroimaging

Research on explainable AI (XAI) for pediatric cancer neuroimaging is still limited, though recent efforts have focused on developing XAI models for neuroimaging in cancer populations, with opportunities to adapt these models for pediatric cases. These models aim to improve brain tumor detection, localization, and classification while providing interpretable results for clinicians.

A recent study [158] investigates the use of deep learning for automated recognition of pediatric posterior fossa tumors (PFT) in brain MRIs. It explored CNN models, including VGG16, VGG19, and ResNet50, for PFT detection and classification, using a dataset of 300,000 images from 500 patients. The study also analyzed model behavior using local interpretable model-agnostic explanations (LIME), Shapley additive explanations (SHAP), and individual conditional expectation (ICE). LIME [159] and SHAP [160] were specifically used to identify the importance of imaging features, such as tumor intensity and spatial patterns, in the model's predictions. For example, SHAP visualizations highlighted that some specific regions in T2-weighted MRI scans significantly influenced classification decisions, helping clinicians understand why a particular tumor type was predicted.

A practical case study includes the application of SHAP in a pediatric medulloblastoma cohort [77], where SHAP plots demonstrated the contribution of tumor shape and contrast-enhanced MRI features to distinguishing molecular subgroups of tumors. This analysis provided critical insights into how specific features aligned with clinical and pathological findings, bridging the gap between AI models and clinical interpretation.

A lightweight CNN with gradient-weighted class activation mapping (Grad-CAM) visualization achieved high accuracy in brain cancer detection and localization [141]. Grad-CAM provided heatmaps overlaying the most relevant tumor regions identified by the model, offering clinicians a direct view of the areas driving predictions.

The importance of XAI lies in its ability to visualize model features, improve interpretability, and enable human–machine interactions that are crucial for clinical adoption [161]. Recent developments, such as NeuroXAI, a framework implementing multiple XAI methods for MRI analysis of brain tumors, demonstrate the potential of XAI to enhance transparency and reliability in neuroimaging, facilitating the adoption of AI models in clinical practice for cancer diagnosis and treatment planning [162].

Another study explored the adaptive aquila optimizer with XAI for effective colorectal and osteosarcoma cancer classification, combining faster SqueezeNet for feature extraction, adaptive optimization for tuning, ensemble DL classifiers for diagnosis, and LIME for interpretability [163].

9. Discussion

AI has the potential to transform medicine by enabling the analysis of large quantities of patient data to provide faster, more accurate diagnoses and to reduce the need for

invasive procedures. AI can also help healthcare providers to optimize clinical workflows by automating repetitive tasks and reducing the administrative burden on clinicians.

A variety of machine learning techniques has been applied to address key challenges in pediatric neuroimaging for cancer. Table 2 provides an overview of these key issues, their descriptions, and the ML solutions currently being explored. This structured summary highlights limitations in dataset availability, tumor detection, scan time reduction, and treatment outcome prediction, emphasizing the role of AI-driven solutions such as deep learning, radiomics, and federated learning.

Table 2. Key issues and respective machine learning solutions in AI-based pediatric neuroimaging for cancer.

Key Issue	Description	Machine Learning Solutions
Limited pediatric data	Scarcity of large, high-quality datasets impedes AI model training.	Transfer learning, federated learning, GANs for data augmentation
Motion artifacts in young patients	Children struggle to stay still during scans, leading to degraded image quality.	AI-based motion correction, CNNs for image denoising
Prolonged scan times	Long MRI scans increase discomfort and require sedation.	Compressed sensing, AI-accelerated MRI reconstruction (e.g., deep CNNs)
Radiation and contrast agent exposure	Pediatric patients are more sensitive to radiation and contrast dyes.	Low-dose imaging with deep learning, contrast-free MRI enhancement with CNNs
Tumor detection and segmentation	Early and precise tumor localization is essential for treatment planning.	CNNs (3D-CNN, ResNet, U-Net, nnU-Net), GANs for data augmentation
Tumor classification and molecular subtyping	Differentiating between tumor types and predicting molecular markers.	Radiomics radiogenomics, CNNs, SVMs, deep learning classifiers
Functional neuroimaging and cognitive deficits	Pediatric brain tumors affect cognitive function, requiring fMRI and other functional imaging.	AI-based functional MRI analysis, machine learning for cognitive pattern recognition (e.g., SVM, CNNs)
Explainability and interpretability of AI	Clinicians need interpretable AI models for trust and adoption.	Explainable AI (XAI), SHAP, LIME, Grad-CAM
Personalized treatment planning	Individualized therapy selection based on imaging and genetic markers.	Multi-modal deep learning (integrating MRI, PET, clinical data)
Predicting treatment outcomes and survival	AI models need to distinguish between real tumor progression and pseudoprogression.	Radiomic feature analysis, deep learning prognostic models, survival prediction frameworks

9.1. Data-Related Issues

A current limitation on using AI for neuroimaging applications in pediatric cancer is the lack of large datasets, as AI models require a substantial amount of data to learn. One way to overcome this limitation is to encourage collaboration and data sharing among laboratories across universities, research organizations, hospitals, and other healthcare institutions. This can be achieved by developing standardized protocols for data collection, processing, and analysis, as well as by establishing data sharing agreements that ensure data privacy and ethical use of data. An alternative approach to address the problem of

limited datasets is transfer learning, in which pre-trained models from other domains or populations can be fine-tuned for use with smaller pediatric datasets.

Another limitation of pediatric neuroimaging is the lack of standardization in imaging protocols and quality control measures across different research laboratories. If non-standardized datasets are combined in an AI model, such differences can decrease the accuracy and reliability of the model. To streamline image processing and analysis, researchers and clinicians can leverage tools such as SPM [164] for statistical analysis, FSL [165] for multi-modal imaging, and FreeSurfer [166] for automated cortical and subcortical segmentation. These tools address common challenges like motion artifacts and variability in pediatric brain development.

9.2. Technical Limitations

There are several knowledge gaps concerning AI for neuroimaging applications in pediatric cancer. First, although AI models have shown promising results in diagnosing and predicting outcomes of some pediatric neurological diseases, such as brain tumors, there has been only limited research on the generalizability of these models across different patient populations.

Second, there is a need for more research on the interpretability of AI models in pediatric neuroimaging. Interpretability of AI models is critical, especially in healthcare applications such as pediatric neuroimaging. Although AI models may provide accurate predictions, the black-box nature of some AI models can be a concern, as it may limit the ability of the user to understand how the model arrives at its decisions or predictions, thereby potentially limiting their clinical utility. Developing methods to explain the logic behind AI model predictions would help to improve their transparency, trustworthiness, and scientific contribution [145].

Third, there is a need for more research on integrating AI models into clinical practice. The development of AI models is only the first step in their implementation into clinical workflows, and more research is needed to determine how these models can be integrated into clinical decision making and how they can have an impact on patient outcomes.

Explainable AI is paramount in pediatric neuroimaging for several reasons. In clinical applications, understanding the decision-making process of AI models is crucial for gaining trust among healthcare professionals and facilitating seamless integration into diagnostic workflows. Moreover, in cancer, where decisions can have profound implications, explainability ensures that clinicians and researchers comprehend how AI arrives at its predictions or classifications. The application of AI in pediatric neuroimaging demands a high level of interpretability. This transparency fosters a collaborative environment between AI tools and healthcare practitioners. As we navigate the future of AI in pediatric neuroimaging, emphasis should be placed on furthering the development of explainable AI techniques. This not only aligns with the growing demand for transparency in AI applications but also positions pediatric catastrophic disease research at the forefront of responsible AI implementation.

9.3. Ethical Considerations

Ethical concerns such as data privacy, bias, and transparency need to be addressed in the development and implementation of AI models for pediatric neuroimaging. Ensuring that these models are developed and validated in an ethical and responsible manner is essential to avoid potential harm to patients and maintain public trust in AI applications. Practical solutions to address privacy concerns include frameworks for multi-institutional data sharing and the potential of federated learning, which allows for collaborative model training without transferring raw data, and blockchain [167]. Establishing guidelines and

ethical frameworks for the responsible use of AI in patient care and research will be crucial as AI becomes more prominent in neuroimaging.

9.4. Future Directions

AI can facilitate the integration and analysis of multimodal neuroimaging data from various sources and research centers [168,169]. Collaborative efforts with AI-based tools can accelerate discoveries and improve data sharing within the scientific community. AI can help identify novel biomarkers associated with various neurological and psychiatric conditions by analyzing large datasets of functional imaging data. These biomarkers could aid in early diagnosis and personalized treatment strategies. Advanced AI algorithms can predict the progression of neurological disorders and their response to specific neuromodulation treatments. This predictive capability can help clinicians to make more informed decisions about patient care and long-term management.

The use of AI for pediatric neuroimaging is a rapidly growing field. Some of the key areas of focus for future developments in this field include large-scale data collection, multimodal integration of neuroimaging modalities (such as MRI, fMRI, resting-state fMRI, MRS, DTI, fNIRS, EEG, and MEG), early detection of pathologies, real-time neuroimage processing and analysis, personalized treatment planning, prediction of treatment outcomes, longitudinal neuroimaging analysis, and the theoretical explanation of pathological phenomena.

When studying brains that were previously affected by solid tumors, the approaches used should consider (1) mass effect, with reference to displacement and compression indices; (2) edema, with reference to morphometry, density, and composition; and (3) maps of tissue damage, in terms of volume, morphometry, density, and structural/functional connectivity. Density can be extracted from T1 MRI images, whereas composition is derived from PET scans. Structural connectivity analysis may consider diffusion imaging features extracted as voxel-based measures, such as the apparent diffusion coefficient, fractional anisotropy, mean diffusivity, radial diffusivity, and axial diffusivity, and tensor-based measures, such as tensor connectivity maps.

AI can further refine and enhance the analysis of functional imaging data, such as fMRI. Advanced AI algorithms can extract more precise information from complex brain activity patterns and lead to better insights into brain function and connectivity. For pediatric neuroimaging to truly benefit patient care, AI applications must extend beyond research settings to real-time clinical relevance. AI-driven tools can expedite data analysis, aiding in early detection, diagnosis, and treatment planning. The development of closed-loop systems, guided by real-time functional imaging data, showcases the potential of AI to dynamically adapt neuromodulation strategies, ensuring personalized and efficient interventions for pediatric patients.

AI can enable closed-loop systems in which real-time functional imaging data is used to adapt and optimize neuromodulation in real time. This dynamic approach can ensure that stimulation parameters are continuously adjusted to suit the changing state of the brain, making treatments more efficient and effective. AI can be used in designing and optimizing neuromodulation devices, such as DBS systems. AI-driven simulations can assist in developing more precise and targeted stimulation paradigms. Moreover, AI can assist in identifying the most effective neuromodulation techniques for individual patients, based on their unique brain activity patterns. By considering patient-specific features, such as neural network connectivity, AI can optimize treatment parameters for better outcomes.

10. Conclusions

Although AI has seen significant advances in its application to neuroimaging in adult populations, its implementation in children with cancer has been limited by several factors, including the scarcity of available datasets and the unique challenges of applying adult-focused AI methods to pediatric populations. The limited number of pediatric neuroimaging datasets poses a significant challenge in training AI models specifically for children with cancer. AI algorithms rely heavily on large, diverse datasets to learn patterns and make accurate predictions. Additionally, the neuroimaging characteristics of children are different from those of adults because of ongoing brain development, size differences, and diverse neurological conditions that may manifest differently in pediatric patients. Therefore, direct translation of AI approaches from adult neuroimaging to children might not be feasible without appropriate adjustments and validation. Conversely, AI models trained on a restricted pool of pediatric data might be less effective than those trained on larger and more comprehensive datasets derived from adults.

To address these challenges and to foster the development of AI in pediatric neuroimaging, there is a need to build larger and more diverse pediatric neuroimaging databases. Collaboration among institutions and data sharing initiatives are essential to ensure the responsible and effective use of AI in pediatric populations. As collaborations and data sharing initiatives begin to standardize pediatric data collection, the development of AI approaches tailored specifically to pediatric applications will increase significantly.

Author Contributions: Conceptualization, R.S.; methodology, J.L.D.d.R., J.L. and R.S.; investigation, J.L.D.d.R., J.L., P.P., P.S.M.M., Z.L., A.K.B. and R.S.; resources, R.S.; data curation, J.L.D.d.R.; writing—original draft preparation, J.L.D.d.R., J.L., P.P., P.S.M.M., Z.L., A.K.B. and R.S.; writing—review and editing, J.L.D.d.R., J.L., P.P., P.S.M.M., Z.L., A.K.B. and R.S.; supervision, R.S.; project administration, R.S.; funding acquisition, A.K.B. and R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported, in part, by the American Lebanese Syrian Associated Charities (ALSAC) in the form of salaries to all of the authors, equipment, and other expenses. The RSNA R&E Foundation and Treovir Inc. provided support in the form of grants to A.B.

Acknowledgments: The authors thank Keith A. Laycock, for scientific editing of the manuscript.

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

References

1. Martin, D.; Tong, E.; Kelly, B.; Yeom, K.; Yedavalli, V. Current perspectives of artificial intelligence in pediatric neuroradiology: An overview. *Front. Radiol.* **2021**, *1*, 713681. [CrossRef] [PubMed]
2. Singh, S.B.; Sarrami, A.H.; Gatidis, S.; Varniab, Z.S.; Chaudhari, A.; Daldrup-Link, H.E. Applications of Artificial Intelligence for Pediatric Cancer Imaging. *Am. J. Roentgenol.* **2024**, *223*, e2431076. [CrossRef] [PubMed]
3. Pringle, C.; Kilday, J.P.; Kamaly-Asl, I.; Stivaros, S.M. The role of artificial intelligence in paediatric neuroradiology. *Pediatr. Radiol.* **2022**, *52*, 2159–2172. [CrossRef]
4. Wang, J.; Wang, J.; Wang, S.; Zhang, Y. Deep learning in pediatric neuroimaging. *Displays* **2023**, *80*, 102583. [CrossRef]
5. Bhatia, A.; Khalvati, F.; Ertl-Wagner, B.B. Artificial Intelligence in the Future Landscape of Pediatric Neuroradiology: Opportunities and Challenges. *Am. J. Neuroradiol.* **2024**, *45*, 549–553. [CrossRef]
6. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 3rd ed.; Pearson: London, UK, 2016.
7. Helm, J.M.; Swiergosz, A.M.; Harberle, H.S.; Karnuta, J.M.; Schaffer, J.L.; Krebs, V.E.; Spitzer, A.I.; Ramkumar, P.N. Machine learning and artificial intelligence: Definitions, applications, and future directions. *Curr. Rev. Musculoskelet. Med.* **2020**, *13*, 69–76. [CrossRef]
8. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2009; ISBN 978-0387848570.
9. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006; ISBN 978-0387310732.

10. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018; ISBN 978-0262039246.
11. McCall, J. Genetic algorithms for modelling and optimisation. *J. Comput. Appl. Math.* **2005**, *184*, 205–222. [CrossRef]
12. Valueva, M.V.; Nagornov, N.N.; Lyakhov, P.A.; Valuev, G.V.; Chervyakov, N.I. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. *Math. Comput. Simul.* **2020**, *177*, 232–243. [CrossRef]
13. Mzoughi, H.; Njeh, I.; Wali, A.; Ben Slima, M.; BenHamida, A.; Mhiri, C.; Ben Mahfoudhe, K. Deep multi-scale 3D convolutional neural network (CNN) for MRI gliomas brain tumor classification. *J. Digit. Imaging* **2020**, *33*, 903–915. [CrossRef]
14. Lu, D. AlexNet-based digital recognition system for handwritten brushes. In Proceedings of the 7th International Conference on Cyber Security and Information Engineering (ICCSIE'22), Brisbane, Australia, 23–25 September 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 509–514. [CrossRef]
15. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A.; Liu, W.; et al. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]
16. Wen, L.; Dong, Y.; Gao, L. A new ensemble residual convolutional neural network for remaining useful life estimation. *Math. Biosci. Eng.* **2019**, *16*, 862–880. [CrossRef]
17. Jonnalagedda, P.; Weinberg, B.; Allen, J.; Min, T.; Bhanu, S.; Bhanu, B. SAGE: Sequential attribute generator for analyzing glioblastomas using limited dataset. *arXiv* **2020**, arXiv:2005.07225. [CrossRef]
18. Namdar, K.; Wagner, M.W.; Kudus, K.; Hawkins, C.; Tabori, U.; Ertl-Wagner, B.B.; Khalvati, F. Improving Deep Learning Models for Pediatric Low-Grade Glioma Tumours Molecular Subtype Identification Using MRI-based 3D Probability Distributions of Tumour Location. *Can. Assoc. Radiol. J.* **2024**, *2024*, 8465371241296834. [CrossRef] [PubMed]
19. Ranschaert, E.; Topff, L.; Pianykh, O. Optimization of radiology workflow with artificial intelligence. *Radiol. Clin. N. Am.* **2021**, *59*, 955–966. [CrossRef] [PubMed]
20. Ravi, K.S.; Nandakumar, G.; Thomas, N.; Lim, M.; Qian, E.; Jimeno, M.M.; Poojar, P.; Jin, Z.; Quarterman, P.; Srinivasan, G.; et al. Accelerated MRI using intelligent protocolling and subject-specific denoising applied to Alzheimer's disease imaging. *Front. Neuroimaging* **2023**, *2*, 1072759. [CrossRef]
21. Bash, S.; Wang, L.; Airriess, C.; Zaharchuk, G.; Gong, E.; Shankaranarayanan, A.; Tanenbaum, L.N. Deep learning enables 60% accelerated volumetric brain MRI while preserving quantitative performance: A prospective, multicenter, multireader trial. *Am. J. Neuroradiol.* **2021**, *42*, 2130–2137. [CrossRef]
22. Chen, Y.; Schönlieb, C.B.; Liò, P.; Leiner, T.; Dragotto, P.L.; Wang, G.; Rueckert, D.; Firmin, D.; Yang, G. AI-based reconstruction for fast MRI—A systematic review and meta-analysis. *Proc. IEEE* **2022**, *110*, 224–245. [CrossRef]
23. MacDougall, R.D.; Zhang, Y.; Callahan, M.J.; Perez-Rosselli, J.; Breen, M.A.; Johnston, P.R.; Yu, H. Improving low-dose pediatric abdominal CT by using convolutional neural networks. *Radiol. Artif. Intell.* **2019**, *1*, e180087. [CrossRef]
24. Gokyar, S.; Robb, F.; Kainz, W.; Chaudhari, A.; Winkler, S. MRSaiFE: An AI-based approach towards the real-time prediction of specific absorption rate. *IEEE Access* **2021**, *9*, 140824–140834. [CrossRef]
25. Dyer, T.; Chawda, S.; Alkilani, R.; Morgan, T.N.; Hughes, M.; Rasalingham, S. Validation of an artificial intelligence solution for acute triage and rule-out normal of non-contrast CT head scans. *Neuroradiology* **2022**, *64*, 735–743. [CrossRef]
26. Tian, Q.; Yan, L.; Zhang, X.; Hu, Y.; Han, Y.; Liu, Z.; Nan, H.; Sun, Q.; Sun, Y.; Yang, Y.; et al. Radiomics strategy for glioma grading using texture features from multiparametric MRI. *J. Magn. Reson. Imaging* **2018**, *48*, 1518–1528. [CrossRef]
27. Zhang, M.; Wong, S.W.; Wright, J.N.; Wagner, M.W.; Toescu, S.; Han, M.; Tam, L.T.; Zhou, Q.; Ahmadian, S.S.; Shpanskaya, K.; et al. MRI radiogenomics of pediatric medulloblastoma: A multicenter study. *Radiology* **2022**, *304*, 406–416. [CrossRef] [PubMed]
28. Chiu, F.Y.; Yen, Y. Efficient radiomics-based classification of multi-parametric MR images to identify volumetric habitats and signatures in glioblastoma: A machine learning approach. *Cancers* **2022**, *14*, 1475. [CrossRef] [PubMed]
29. Khalili, N.; Kazerooni, A.F.; Familiar, A.; Haldar, D.; Kraya, A.; Foster, J.; Koptyra, M.; Storm, P.B.; Resnick, A.C.; Nabavizadeh, A. Radiomics for characterization of the glioma immune microenvironment. *NPJ Precis. Oncol.* **2023**, *7*, 59. [CrossRef]
30. Park, J.E.; Kim, H.S.; Goh, M.J.; Kim, N.; Park, S.Y.; Kim, Y.H.; Kim, J.H. Pseudoprogression in patients with glioblastoma: Assessment by using volume-weighted voxel-based multiparametric clustering of MR imaging data in an independent test set. *Radiology* **2015**, *275*, 792–802. [CrossRef]
31. Yang, Y.; Zhao, Y.; Liu, X.; Huang, J. Artificial intelligence for prediction of response to cancer immunotherapy. *Semin. Cancer Biol.* **2022**, *87*, 137–147. [CrossRef]
32. Patel, U.K.; Anwar, A.; Saleem, S.; Malik, P.; Rasul, B.; Patel, K.; Yao, R.; Seshadri, A.; Yousufuddin, M.; Arumaithurai, K. Artificial intelligence as an emerging technology in the current care of neurological disorders. *J. Neurol.* **2021**, *268*, 1623–1642. [CrossRef]
33. Fathi Kazerooni, A.; Saxena, S.; Toorens, E.; Tu, D.; Bashyam, V.; Akbari, H.; Mamourian, E.; Sako, C.; Koumenis, C.; Verginadis, I.; et al. Clinical measures, radiomics, and genomics offer synergistic value in AI-based prediction of overall survival in patients with glioblastoma. *Sci. Rep.* **2022**, *12*, 8784. [CrossRef]

34. Yeh, F.C.; Tseng, W.Y. Advances in connectomics and tractography using diffusion MRI data. *NeuroImage* **2021**, *244*, 118720. [CrossRef]
35. McCarthy, M.L.; Ding, R.; Hsu, H.; Dausey, D.J. Developmental and behavioral comorbidities of asthma in children. *J. Dev. Behav. Pediatr.* **2018**, *39*, 293–304.
36. Zhang, L.-Y.; Du, H.-Z.; Lu, T.-T.; Song, S.-H.; Xu, R.; Jiang, Y.; Pan, H. Long-term outcome of childhood and adolescent patients with craniopharyngiomas: A single center retrospective experience. *BMC Cancer* **2024**, *24*, 1555. [CrossRef]
37. Fangusaro, J.; Jones, D.T.; Packer, R.J.; Gutmann, D.H.; Milde, T.; Witt, O.; Mueller, S.; Fisher, M.J.; Hansford, J.R.; Tabori, U.; et al. Pediatric low-grade glioma: State-of-the-art and ongoing challenges. *Neuro-Oncology* **2024**, *26*, 25–37. [CrossRef] [PubMed]
38. Gajjar, A.; Mahajan, A.; Abdelbaki, M.; Anderson, C.; Antony, R.; Bale, T.; Bindra, R.; Bowers, D.C.; Cohen, K.; Cole, B.; et al. Pediatric Central Nervous System Cancers, Version 2.2023. *J. Natl. Compr. Cancer Netw.* **2022**, *20*, 1339–1362.
39. Müller, H.L. Craniopharyngioma. *Endocr. Rev.* **2014**, *35*, 513–543. [CrossRef] [PubMed]
40. Fjalldal, S.; Holmer, H.; Rylander, L.; Elfving, M.; Ekman, B.; Osterberg, K.; Erfurth, E.M. Hypothalamic involvement predicts cognitive performance and psychosocial health in long-term survivors of childhood craniopharyngioma. *J. Clin. Endocrinol. Metab.* **2013**, *98*, 3253–3262. [CrossRef]
41. Jale Ozyurt, M.S.; Thiel, C.M.; Lorenzen, A.; Gebhardt, U.; Calaminus, G.; Warmuth-Metz, M.; Muller, H.L. Neuropsychological outcome in patients with childhood craniopharyngioma and hypothalamic involvement. *J. Pediatr.* **2014**, *164*, 876–881.e4. [CrossRef]
42. DeAngelis, L.M. Brain tumours. *N. Engl. J. Med.* **2001**, *344*, 114–122. [CrossRef]
43. Alksas, A.; Shehata, M.; Atef, H.; Sherif, F.; Alghamdi, N.S.; Ghazal, M.; Abdel Fattah, S.; El-Serougy, L.G.; El-Baz, A. A novel system for precise grading of glioma. *Bioengineering* **2022**, *9*, 532. [CrossRef] [PubMed] [PubMed Central]
44. Weyer-Jamora, C.; Brie, M.S.; Luks, T.L.; Smith, E.M.; Braunstein, S.E.; Villanueva-Meyer, J.E.; Bracci, P.M.; Chang, S.; Hervey-Jumper, S.L.; Taylor, J.W. Cognitive impact of lower-grade gliomas and strategies for rehabilitation. *Neurooncol. Pract.* **2021**, *8*, 117–128. [CrossRef]
45. Louis, D.N.; Ohgaki, H.; Wiestler, O.D.; Cavenee, W.K.; Burger, P.C.; Jouvet, A.; Scheithauer, B.W.; Kleihues, P. The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol.* **2017**, *114*, 97–109. [CrossRef]
46. Speville, E.D.; Kieffer, V.; Dufour, C.; Grill, J.; Noulhiane, M.; Hertz-Pannier, L.; Chevignard, M. Neuropsychological consequences of childhood medulloblastoma and possible interventions. *Neurochirurgie* **2021**, *67*, 90–98. [CrossRef]
47. Power, J.D.; Barnes, K.A.; Snyder, A.Z.; Schlaggar, B.L.; Petersen, S.E. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage* **2015**, *59*, 2142–2154. [CrossRef] [PubMed]
48. Beauferris, Y.; Teuwen, J.; Karkalousos, D.; Moriakov, N.; Caan, M.; Yiasemis, G.; Rodrigues, L.; Lopes, A.; Pedrini, H.; Rittner, L.; et al. Multi-coil MRI reconstruction challenge—Assessing brain MRI reconstruction models and their generalizability to varying coil configurations. *Front. Neurosci.* **2022**, *16*, 919186. [CrossRef] [PubMed]
49. Lustig, M.; Donoho, D.L.; Santos, J.M.; Pauly, J.M. Compressed sensing MRI. *IEEE Signal Process. Mag.* **2008**, *25*, 72–82. [CrossRef]
50. Singh, N.M.; Harrod, J.B.; Subramanian, S.; Robinson, M.; Chang, K.; Cetin-Karayumak, S.; Dalca, A.V.; Eickhoff, S.; Fox, M.; Franke, L.; et al. How machine learning is powering neuroimaging to improve brain health. *Neuroinformatics* **2022**, *20*, 943–964. [CrossRef]
51. Evans, A.C.; Brain Development Cooperative Group. The NIH MRI study of normal brain development. *NeuroImage* **2006**, *30*, 184–202. [CrossRef]
52. Molfese, P.J.; Glen, D.; Mesite, L.; Cox, R.W.; Hoeft, F.; Frost, S.J.; Mencl, W.E.; Pugh, K.R.; Bandettini, P.A. The Haskins pediatric atlas: A magnetic-resonance-imaging-based pediatric template and atlas. *Pediatr. Radiol.* **2021**, *51*, 628–639. [CrossRef]
53. Wang, L.; Gao, Y.; Shi, F.; Li, G.; Gilmore, J.H.; Lin, W.; Shen, D. LINKS: Learning-based multi-source integration framework for segmentation of infant brain images. *Neuroimage* **2015**, *108*, 160–172. [CrossRef]
54. Barkovich, A.J.; Raybaud, C. *Pediatric Neuroimaging*, 5th ed.; Lippincott Williams & Wilkins, Wolters Kluwer Health: Philadelphia, PA, USA, 2012.
55. Slater, D.A.; Melie-Garcia, L.; Preisig, M.; Kherif, F.; Lutti, A.; Draganski, B. Evolution of white matter tract microstructure across the lifespan. *Hum. Brain Mapp.* **2019**, *40*, 2252–2268. [CrossRef]
56. Pfefferbaum, A.; Mathalon, D.H.; Sullivan, E.V.; Rawles, J.M.; Zipursky, R.B.; Lim, K.O. A quantitative magnetic resonance imaging study of changes in brain morphology from infancy to late adulthood. *Arch. Neurol.* **1994**, *51*, 874–887. [CrossRef]
57. Blockmans, L.; Golestani, N.; Dalboni da Rocha, J.L.; Wouters, J.; Ghesquière, P.; Vandermosten, M. Role of family risk and of pre-reading auditory and neurostructural measures in predicting reading outcome. *Neurobiol. Lang.* **2023**, *4*, 474–500. [CrossRef]
58. Kim, S.H.; Lyu, I.; Fonov, V.S.; Vachet, C.; Hazlett, H.C.; Smith, R.G.; Piven, J.; Dager, S.R.; McKinstry, R.C.; Pruett, J.R., Jr.; et al. Development of cortical shape in the human brain from 6 to 24 months of age via a novel measure of shape complexity. *NeuroImage* **2016**, *135*, 163–176. [CrossRef] [PubMed]
59. da Rocha, J.L.D.; Schneider, P.; Benner, J.; Santoro, R.; Atanasova, T.; Van De Ville, D.; Golestani, N. TASH: Toolbox for the Automated Segmentation of Heschl's gyrus. *Sci. Rep.* **2020**, *10*, 3887. [CrossRef]

60. Maghsadhagh, S.; Dalboni da Rocha, J.L.; Benner, J.; Schneider, P.; Golestani, N.; Behjat, H. A discriminative characterization of Heschl's gyrus morphology using spectral graph features. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Virtual, 1–5 November 2021; pp. 3577–3581. [CrossRef]
61. Fonov, V.; Evans, A.C.; Botteron, K.; Almli, C.R.; McKinstry, R.C.; Collins, D.L.; the Brain Development Cooperative Group. Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage* **2011**, *54*, 313–327. [CrossRef] [PubMed]
62. Wilke, M.; Holland, S.K.; Altaye, M.; Gaser, C. Template-O-Matic: A toolbox for creating customized pediatric templates. *Neuroimage* **2008**, *41*, 903–913. [CrossRef]
63. Kazemi, K.; Moghaddam, H.A.; Grebe, R.; Gondry-Jouet, C.; Wallois, F. A neonatal atlas template for spatial normalization of whole-brain magnetic resonance images of newborns: Preliminary results. *NeuroImage* **2007**, *37*, 463–473. [CrossRef]
64. Shi, F.; Yap, P.-T.; Wu, G.; Jia, H.; Gilmore, J.H.; Lin, W.; Shen, D. Infant brain atlases from neonates to 1- and 2-year-olds. *PLoS ONE* **2011**, *6*, e18746. [CrossRef]
65. Giedd, J.N. Structural magnetic resonance imaging of the adolescent brain. *Ann. N.Y. Acad. Sci.* **2004**, *1021*, 77–85. [CrossRef]
66. Mohsen, F.; Ali, H.; El Hajj, N.; Shah, Z. Artificial intelligence-based methods for fusion of electronic health records and imaging data. *Sci. Rep.* **2022**, *12*, 17981. [CrossRef]
67. Nichols, T.E.; Das, S.; Eickhoff, S.B.; Evans, A.C.; Glatard, T.; Hanke, M.; Yeo, B.T. Best practices in data analysis and sharing in neuroimaging using MRI. *Nat. Neurosci.* **2017**, *20*, 299–303. [CrossRef]
68. Calhoun, V.D.; Miller, R.; Pearlson, G.; Adalı, T. The chronnectome: Time-varying connectivity networks as the next frontier in fMRI data discovery. *Neuron* **2012**, *84*, 262–274. [CrossRef]
69. Pati, S.; Baid, U.; Edwards, B.; Sheller, M.H.; Wang, S.H.; Reina, G.A.; Foley, P.; Gruzdev, A.; Karkada, D.; Davatzikos, C.; et al. Federated learning enables big data for rare cancer boundary detection. *Nat. Commun.* **2022**, *13*, 7346. [CrossRef] [PubMed]
70. Karkazis, K.; Cho, M.K.; Radhakrishnan, R. Human subjects protections and research on the developing brain: Drawing lines and shifting boundaries. *Neuron* **2017**, *95*, 221–228.
71. Huang, Y.; Ahmad, S.; Fan, J.; Shen, D.; Yap, P. Difficulty-aware hierarchical convolutional neural networks for deformable registration of brain MR images. *Med. Image Anal.* **2020**, *67*, 101817. [CrossRef]
72. Zhao, C.; Shao, M.; Carass, A.; Li, H.; Dewey, B.E.; Ellingsen, L.M.; Woo, J.; Guttman, M.A.; Blitz, A.M.; Stone, M.; et al. Applications of a deep learning method for anti-aliasing and super-resolution in MRI. *Magn. Reson. Imaging* **2019**, *64*, 132–141. [CrossRef]
73. Fair, M.J.; Gatehouse, P.D.; DiBella, E.V.R.; Firmin, D.N. A review of 3D first-pass, whole-heart, myocardial perfusion cardiovascular magnetic resonance. *J. Cardiovasc. Magn. Reson.* **2015**, *17*, 68. [CrossRef]
74. Monsour, R.; Dutta, M.; Mohamed, A.Z.; Borkowski, A.; Viswanadhan, N.A. Neuroimaging in the era of artificial intelligence: Current applications. *Fed. Pract.* **2022**, *39* (Suppl. S1), S14–S20. [CrossRef]
75. Mardani, M.; Gong, E.; Cheng, J.Y.; Vasanawala, S.; Zaharchuk, G.; Alley, M.; Thakur, N.; Han, S.; Dally, W.; Pauly, J.M.; et al. Deep generative adversarial networks for compressed sensing automates MRI. *arXiv* **2017**, arXiv:1706.00051.
76. Ng, C.K.C. Generative Adversarial Network (Generative Artificial Intelligence) in Pediatric Radiology: A Systematic Review. *Children* **2023**, *10*, 1372. [CrossRef]
77. Familiar, A.M.; Kazerooni, A.F.; Kraya, A.; Rathi, K.; Khalili, N.; Gandhi, D.; Anderson, H.; Mahtabfar, A.; Ware, J.B.; Vossough, A.; et al. IMG-13. Prediction of pediatric medulloblastoma subgroups using clinico-radiomic analysis. *Neuro-Oncology* **2024**, *26* (Suppl. S4), 103–104. [CrossRef]
78. Hall, E.J. Lessons we have learned from our children: Cancer risks from diagnostic radiology. *Pediatr. Radiol.* **2002**, *32*, 700–706. [CrossRef] [PubMed]
79. Gong, E.; Pauly, J.M.; Wintermark, M.; Zaharchuk, G. Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI. *J. Magn. Reson. Imaging* **2018**, *48*, 330–340. [CrossRef] [PubMed]
80. Chen, H.; Zhang, Y.; Zhang, W.; Liao, P.; Li, K.; Zhou, J.; Wang, G. Low-dose CT via convolutional neural network. *Biomed. Opt. Express* **2017**, *8*, 679–694. [CrossRef] [PubMed]
81. Kang, J.; Rancati, T.; Lee, S.; Oh, J.H.; Kerns, S.L.; Scott, J.G.; Schwartz, R.; Kim, S.; Rosenstein, B.S. Machine learning and radiogenomics: Lessons learned and future directions. *Front. Oncol.* **2018**, *8*, 228. [CrossRef] [PubMed]
82. Nishio, M.; Nagashima, C.; Hirabayashi, S.; Ohnishi, A.; Sasaki, K.; Sagawa, T.; Hamada, M.; Yamashita, T. Convolutional auto-encoder for image denoising of ultra-low-dose CT. *Heliyon* **2017**, *3*, e00393. [CrossRef] [PubMed]
83. Zaharchuk, G. Next-generation research applications for hybrid PET/MR and PET/CT imaging using deep learning. *Eur. J. Nucl. Med. Mol. Imaging* **2019**, *46*, 2700–2707. [CrossRef]
84. Lee, D.; Yoo, J.; Tak, S.; Ye, J.C. Deep residual learning for accelerated MRI using magnitude and phase networks. *IEEE Trans. Biomed. Eng.* **2018**, *65*, 1985–1995. [CrossRef]
85. Anand, C.S.; Sahambi, J.S. Wavelet domain non-linear filtering for MRI denoising. *Magn. Reson. Imaging* **2010**, *28*, 842–861. [CrossRef]

86. Chun, S.Y.; Reese, T.G.; Ouyang, J.; Guerin, B.; Catana, C.; Zhu, X.; Alpert, N.M.; El Fakhri, G. MRI-based nonrigid motion correction in simultaneous PET/MRI. *J. Nucl. Med.* **2012**, *53*, 1284–1291. [CrossRef]
87. Haskell, M.W.; Cauley, S.F.; Wald, L.L. Targeted motion estimation and reduction (TAMER): Data consistency based motion mitigation for MRI using a reduced model joint optimization. *IEEE Trans. Med. Imaging* **2018**, *37*, 1253–1265. [CrossRef]
88. Singh, N.M.; Iglesias, J.E.; Adalsteinsson, E.; Dalca, A.V.; Golland, P. Joint frequency and image space learning for Fourier imaging. *arXiv* **2020**, arXiv:200701441.
89. Gjesteby, L.; Yang, Q.; Xi, Y.; Shan, H.; Claus, B.; Jin, Y.; Man, B.D.; Wang, G. Deep learning methods for CT image-domain metal artifact reduction. In *Developments in X-Ray Tomography XI*; International Society for Optics and Photonics: Bellingham, WC, USA, 2017; p. 103910W. [CrossRef]
90. Hu, Z.; Jiang, C.; Sun, F.; Zhang, Q.; Ge, Y.; Yang, Y.; Liu, X.; Zheng, H.; Liang, D. Artifact correction in low-dose dental CT imaging using Wasserstein generative adversarial networks. *Med. Phys.* **2019**, *46*, 1686–1696. [CrossRef] [PubMed]
91. Zhu, D.; Zhang, T.; Jiang, X.; Hu, X.; Chen, H. Opportunities and challenges of artificial intelligence in pediatric neuroimaging. *Front. Neurosci.* **2020**, *14*, 204.
92. Schooler, G.R.; Davis, J.T.; Daldrup-Link, H.E.; Frush, D.P. Current utilization and procedural practices in pediatric whole-body MRI. *Pediatr. Radiol.* **2018**, *48*, 1101–1107. [CrossRef] [PubMed]
93. Daldrup-Link, H. Artificial intelligence applications for pediatric oncology imaging. *Pediatr. Radiol.* **2019**, *49*, 1384–1390. [CrossRef]
94. Soltaninejad, M.; Yang, G.; Lambrou, T.; Allinson, N.; Jones, T.L.; Barrick, T.R.; Howe, F.A.; Ye, X. Automated brain tumor detection and segmentation using super-pixel-based extremely randomized trees in FLAIR MRI. *Int. J. Comput. Assist. Radiol. Surg.* **2017**, *12*, 183–203. [CrossRef]
95. Irmak, E. Multi-classification of brain tumor MRI images using deep convolutional neural network with fully optimized framework. *Iran. J. Sci. Technol. Trans. Electr. Eng.* **2021**, *45*, 1015–1036. [CrossRef]
96. Fetit, A.E.; Novak, J.; Rodriguez, D.; Auer, D.P.; Clark, C.A.; Grundy, R.G.; Peet, A.C.; Arvanitis, T.N. Radiomics in paediatric neuro-oncology: A multicentre study on MRI texture analysis. *NMR Biomed.* **2018**, *31*, e3781. [CrossRef]
97. Kickingereder, P.; Bonekamp, D.; Nowosielski, M.; Kratz, A.; Sill, M.; Burth, S.; Wick, A.; Eidel, O.; Schlemmer, H.P.; Radbruch, A.; et al. Radiogenomics of glioblastoma: Machine learning-based classification of molecular characteristics by using multiparametric and multiregional MR imaging features. *Radiology* **2016**, *281*, 907–918. [CrossRef]
98. Zarinabad, N.; Wilson, M.; Gill, S.K.; Manias, K.A.; Davies, N.P.; Peet, A.C. Multiclass imbalance learning: Improving classification of pediatric brain tumors from magnetic resonance spectroscopy. *Magn. Reson. Med.* **2017**, *77*, 2114–2124. [CrossRef]
99. Styner, M.; Oguz, I.; Xu, S.; Brechbuehler, C.; Pantazis, D.; Levitt, J.J.; Shenton, M.E.; Gerig, G. Framework for the statistical shape analysis of brain structures using SPHARM-PDM. *Insight J.* **2006**, *1071*, 242–250. [CrossRef]
100. Dalboni da Rocha, J.L.; Kepinska, O.; Schneider, P.; Benner, J.; Degano, G.; Schneider, L.; Golestani, N. Multivariate Concavity Amplitude Index (MCAI) for characterizing Heschl's gyrus shape. *NeuroImage* **2023**, *272*, 120052. [CrossRef] [PubMed]
101. Van Griethuysen, J.J.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.; Fillion-Robin, J.C.; Pieper, S.; Aerts, H.J. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **2017**, *77*, e104–e107. [CrossRef] [PubMed]
102. Boyd, A.; Ye, Z.; Prabhu, S.; Tjong, M.C.; Zha, Y.; Zapaishchykova, A.; Vajapeyam, S.; Hayat, H.; Chopra, R.; Liu, K.X.; et al. Expert-level pediatric brain tumor segmentation in a limited data scenario with stepwise transfer learning. *medRxiv* **2023**. [CrossRef]
103. Mulvany, T.; Griffiths-King, D.; Novak, J.; Rose, H. Segmentation of pediatric brain tumors using a radiologically informed, deep learning cascade. *arXiv* **2024**, arXiv:2410.14020. [CrossRef]
104. Bengtsson, M.; Keles, E.; Durak, G.; Anwar, S.; Velichko, Y.S.; Linguraru, M.G.; Waanders, A.J.; Bagci, U. A new logic for pediatric brain tumor segmentation. *arXiv* **2024**, arXiv:2411.01390. [CrossRef]
105. Vossough, A.; Khalili, N.; Familiar, A.M.; Gandhi, D.; Viswanathan, K.; Tu, W.; Haldar, D.; Bagheri, S.; Anderson, H.; Haldar, S.; et al. Training and Comparison of nnU-Net and DeepMedic Methods for Autosegmentation of Pediatric Brain Tumors. *Am. J. Neuroradiol.* **2024**, *45*, 1081–1089. [CrossRef]
106. Gu, X.; Shen, Z.; Xue, J.; Fan, Y.; Ni, T. Brain tumor MR image classification using convolutional dictionary learning with local constraint. *Front. Neurosci.* **2021**, *15*, 679847. [CrossRef]
107. Chen, R.J.; Lu, M.Y.; Wang, J.; Williamson, D.F.K.; Rodig, S.J.; Lindeman, N.I.; Mahmood, F. Pathomic Fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans. Med. Imaging* **2022**, *41*, 757–770. [CrossRef]
108. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]

109. Bangalore Yogananda, C.G.; Shah, B.R.; Vejdani-Jahromi, M.; Nalawade, S.S.; Murugesan, G.K.; Yu, F.F.; Pinho, M.C.; Wagner, B.C.; Emblem, K.E.; Bjørnerud, A.; et al. A fully automated deep learning network for brain tumor segmentation. *Tomography* **2020**, *6*, 186–193. [CrossRef]
110. Florez, E.; Nichols, T.; Parker, E.E.; Lirette, S.T.; Howard, C.M.; Fatemi, A. Multiparametric magnetic resonance imaging in the assessment of primary brain tumors through radiomic features: A metric for guided radiation treatment planning. *Cureus* **2018**, *10*, e3426. [CrossRef]
111. Obuchowicz, R.; Lasek, J.; Wodziński, M.; Piórkowski, A.; Strzelecki, M.; Nurzynska, K. Artificial Intelligence-Empowered Radiology—Current Status and Critical Review. *Diagnostics* **2025**, *15*, 282. [CrossRef]
112. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images are more than pictures, they are data. *Radiology* **2016**, *278*, 563–577. [CrossRef]
113. Casamitjana, A.; Puch, S.; Aduriz, A.; Vilaplana, V. 3D convolutional neural networks for brain tumor segmentation: A comparison of multi-resolution architectures. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*; Crimi, A., Menze, B., Maier, O., Reyes, M., Winzeck, S., Handels, H., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2016; Volume 10154. [CrossRef]
114. Urban, G.; Bendszus, M.; Hamprecht, F.A.; Kleesiek, J. *Multi-Modal Brain Tumor Segmentation Using Deep Convolutional Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 31–35.
115. Sajjad, M.; Khan, S.; Muhammad, K.; Wu, W.; Ullah, A.; Baik, S.W. Multi-grade brain tumor classification using deep CNN with extensive data augmentation. *J. Comput. Sci.* **2019**, *30*, 174–182. [CrossRef]
116. Richter, V.; Ernemann, U.; Bender, B. Novel Imaging Approaches for Glioma Classification in the Era of the World Health Organization 2021 Update: A Scoping Review. *Cancers* **2024**, *16*, 1792. [CrossRef]
117. Moassefi, M.; Faghani, S.; Erickson, B. Artificial Intelligence in Neuro-Oncology: Predicting molecular markers and response to therapy. *Med. Res. Arch.* **2024**, *12*, 80. [CrossRef]
118. Gorden, D.V. Machine learning tool identifies molecular subtypes of brain cancer. *Oncol. Times* **2023**, *45*, 6. [CrossRef]
119. Yearley, A.G.; Blitz, S.E.; Patel, R.V.; Chan, A.; Baird, L.C.; Friedman, G.K.; Arnaout, O.; Smith, T.R.; Bernstock, J.D. Machine learning in the classification of pediatric posterior fossa tumors: A systematic review. *Cancers* **2022**, *14*, 5608. [CrossRef]
120. Quon, J.L.; Bala, W.; Chen, L.C.; Wright, J.; Kim, L.H.; Han, M.; Shpanskaya, K.; Lee, E.H.; Tong, E.; Iv, M.; et al. Deep learning for pediatric posterior fossa tumor detection and classification: A multi-institutional study. *Am. J. Neuroradiol.* **2020**, *41*, 1718–1725. [CrossRef]
121. Iv, M.; Zhou, M.; Shpanskaya, K.; Perreault, S.; Wang, Z.; Tranvinh, E.; Lanzman, B.; Vajapeyam, S.; Vitanza, N.; Fisher, P.; et al. MR imaging–based radiomic signatures of distinct molecular subgroups of medulloblastoma. *Am. J. Neuroradiol.* **2019**, *40*, 154–161. [CrossRef]
122. Qin, C.; Hu, W.; Wang, X.; Ma, X. Application of artificial intelligence in diagnosis of craniopharyngioma. *Front. Neurol.* **2022**, *12*, 752119. [CrossRef]
123. Prince, E.W.; Whelan, R.; Mirsky, D.M.; Stence, N.; Staulcup, S.; Klimo, P.; Anderson, R.C.E.; Niazi, T.N.; Grant, G.; Souweidane, M.; et al. Robust deep learning classification of adamantinomatous craniopharyngioma from limited preoperative radiographic images. *Sci. Rep.* **2020**, *10*, 16885. [CrossRef]
124. Chen, X.; Tong, Y.; Shi, Z.; Chen, H.; Yang, Z.; Wang, Y.; Chen, L.; Yu, J. Noninvasive molecular diagnosis of craniopharyngioma with MRI-based radiomics approach. *BMC Neurol.* **2019**, *19*, 6. [CrossRef]
125. Ma, G.; Kang, J.; Qiao, N.; Zhang, B.; Chen, X.; Li, G.; Gao, Z.; Gui, S. Non-invasive radiomics approach predicts invasiveness of adamantinomatous craniopharyngioma before surgery. *Front. Oncol.* **2021**, *10*, 599888. [CrossRef]
126. Ertosun, M.G.; Rubin, D.L. Automated grading of gliomas using deep learning in digital pathology images: A modular approach with ensemble of convolutional neural networks. *AMIA Annu. Symp. Proc.* **2015**, *2015*, 1899–1908. [PubMed] [PubMed Central]
127. Sultan, H.H.; Salem, N.M.; Al-Atabany, W. Multi-classification of brain tumor images using deep neural network. *IEEE Access* **2019**, *7*, 69215–69225. [CrossRef]
128. Pisapia, J.M.; Akbari, H.; Rozycki, M.; Thawani, J.P.; Storm, P.B.; Avery, R.A.; Vossough, A.; Fisher, M.J.; Heuer, G.G.; Davatzikos, C. Predicting pediatric optic pathway glioma progression using advanced magnetic resonance image analysis and machine learning. *Neurooncol. Adv.* **2020**, *2*, vdaa090. [CrossRef]
129. Sturm, D.; Pfister, S.M.; Jones, D.T. Pediatric Gliomas: Current Concepts on Diagnosis, Biology, and Clinical Management. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 35* **2017**, *21*, 2370–2377. [CrossRef]
130. Guo, W.; She, D.; Xing, Z.; Lin, X.; Wang, F.; Song, Y.; Cao, D. Multiparametric MRI-based radiomics model for predicting H3 K27M mutant status in diffuse midline glioma: A comparative study across different sequences and machine learning techniques. *Front. Oncol.* **2022**, *12*, 796583. [CrossRef]
131. Huang, B.; Zhang, Y.; Mao, Q.; Ju, Y.; Liu, Y.; Su, Z.; Lei, Y.; Ren, Y. Deep learning-based prediction of H3K27M alteration in diffuse midline gliomas based on whole-brain MRI. *Cancer Med.* **2023**, *12*, 17139–17148. [CrossRef]

132. Su, X.; Chen, N.; Sun, H.; Liu, Y.; Yang, X.; Wang, W.; Zhang, S.; Tan, Q.; Gong, Q.; Yue, Q. Automated machine learning based on radiomics features predicts H3 K27M mutation in midline gliomas of the brain. *Neuro Oncol.* **2020**, *22*, 393–401. [CrossRef]
133. Kandemirli, S.G.; Kocak, B.; Naganawa, S.; Ozturk, K.; Yip, S.S.F.; Chopra, S.; Rivetti, L.; Aldine, A.S.; Jones, K.; Cayci, Z.; et al. Machine learning-based multiparametric magnetic resonance imaging radiomics for prediction of H3K27M mutation in midline gliomas. *World Neurosurg.* **2021**, *151*, e78–e85. [CrossRef]
134. Wagner, M.W.; Namdar, K.; Napoleone, M.; Hainc, N.; Amirabadi, A.; Fonseca, A.; Laughlin, S.; Shroff, M.M.; Bouffet, E.; Hawkins, C.; et al. Radiomic features based on MRI predict progression-free survival in pediatric diffuse midline glioma/diffuse intrinsic pontine glioma. *Can. Assoc. Radiol. J.* **2023**, *74*, 119–126. [CrossRef]
135. Cheng, D.; Zhuo, Z.; Du, J.; Weng, J.; Zhang, C.; Duan, Y.; Sun, T.; Wu, M.; Guo, M.; Hua, T.; et al. A fully automated deep-learning model for predicting the molecular subtypes of posterior fossa ependymomas using T2-weighted images. *Clin. Cancer Res.* **2024**, *30*, 150–158. [CrossRef]
136. Safai, A.; Shinde, S.; Jadhav, M.; Chougule, T.; Indoria, A.; Kumar, M.; Santosh, V.; Jabeen, S.; Beniwal, M.; Konar, S.; et al. Developing a radiomics signature for supratentorial extra-ventricular ependymoma using multimodal MR imaging. *Front. Neurol.* **2021**, *12*, 648092. [CrossRef]
137. Lai, J.; Zou, P.; da Rocha, J.L.D.; Heitzer, A.M.; Patni, T.; Li, Y.; Scoggins, M.A.; Sharma, A.; Wang, W.C.; Helton, K.J.; et al. Hydroxyurea maintains working memory function in pediatric sickle cell disease. *PLoS ONE* **2024**, *19*, e0296196. [CrossRef]
138. Abdel-Jaber, H.; Devassy, D.; Al Salam, A.; Hidaytallah, L.; EL-Amir, M. A review of deep learning algorithms and their applications in healthcare. *Algorithms* **2022**, *15*, 71. [CrossRef]
139. Deepak, S.; Ameer, P.M. Brain tumor classification using deep CNN features via transfer learning. *Comput. Biol. Med.* **2019**, *111*, 103345. [CrossRef]
140. Ning, Z.; Luo, J.; Xiao, Q.; Cai, L.; Chen, Y.; Yu, X.; Wang, J.; Zhang, Y. Multi-modal magnetic resonance imaging-based grading analysis for gliomas by integrating radiomics and deep features. *Ann. Transl. Med.* **2021**, *9*, 298. [CrossRef]
141. Raza, A.; Ayub, H.; Khan, J.A.; Ahmad, I.; Salama, S.; Daradkeh, Y.I.; Javeed, D.; Rehman, A.U.; Hamam, H. A hybrid deep learning-based approach for brain tumor classification. *Electronics* **2022**, *11*, 1146. [CrossRef]
142. Zou, P.; Conklin, H.M.; Scoggins, M.A.; Li, Y.; Li, X.; Jones, M.M.; Palmer, S.L.; Gajjar, A.; Ogg, R.J. Functional MRI in medulloblastoma survivors supports prophylactic reading intervention during tumor treatment. *Brain Imaging Behav.* **2016**, *10*, 258–271. [CrossRef] [PubMed] [PubMed Central]
143. Dalboni da Rocha, J.L.; Zou Stinnett, P.; Scoggins, M.A.; McAfee, S.S.; Conklin, H.M.; Gajjar, A.; Sitaram, R. Functional MRI assessment of brain activity patterns associated with reading in medulloblastoma survivors. *Brain Sci.* **2024**, *14*, 904. [CrossRef] [PubMed]
144. Svärd, D.; Follin, C.; Fjalldal, S.; Hellerstedt, R.; Mannfolk, P.; Mårtensson, J.; Sundgren, P.; Erfurth, E.M. Cognitive interference processing in adults with childhood craniopharyngioma using functional magnetic resonance imaging. *Endocrine* **2021**, *74*, 714–722. [CrossRef] [PubMed]
145. Shibu, C.J.; Sreedharan, S.; Arun, K.M.; Kesavadas, C.; Sitaram, R. Explainable artificial intelligence model to predict brain states from fNIRS signals. *Front. Hum. Neurosci.* **2023**, *16*, 1029784. [CrossRef]
146. Budnick, H.C.; Baygani, S.; Easwaran, T.; Vortmeyer, A.; Jea, A.; Desai, V.; Raskin, J. Predictors of seizure freedom in pediatric low-grade gliomas. *Cureus* **2022**, *14*, e31915. [CrossRef]
147. Nune, G.; DeGiorgio, C.; Heck, C. Neuromodulation in the treatment of epilepsy. *Curr. Treat. Options Neurol.* **2015**, *17*, 43. [CrossRef]
148. Pineau, J.; Guex, A.; Vincent, R.; Panuccio, G.; Avoli, M. Treating epilepsy via adaptive neurostimulation: A reinforcement learning approach. *Int. J. Neural Syst.* **2009**, *19*, 227–240. [CrossRef]
149. Azriel, R.; Hahn, C.D.; Cooman, T.D.; Van Huffel, S.; Payne, E.T.; McBain, K.L.; Eytan, D.; Behar, J.A. Machine learning to support triage of children at risk for epileptic seizures in the pediatric intensive care unit. *Physiol. Meas.* **2022**, *43*, 095003. [CrossRef]
150. Chen, H.; Wang, Y.; Ji, T.; Jiang, Y.; Zhou, X.H. Brain functional connectivity-based prediction of vagus nerve stimulation efficacy in pediatric pharmacoresistant epilepsy. *CNS Neurosci. Ther.* **2023**, *29*, 3259–3268. [CrossRef]
151. Rana, M.; Gupta, N.; Dalboni da Rocha, J.L.; Lee, S.; Sitaram, R. A toolbox for real-time subject-independent and subject-dependent classification of brain states from fMRI signals. *Front. Neurosci.* **2013**, *7*, 170. [CrossRef]
152. Kim, S.; Dalboni da Rocha, J.L.; Birbaumer, N.; Sitaram, R. Self-regulation of the posterior–frontal brain activity with real-time fMRI neurofeedback to influence perceptual discrimination. *Brain Sci.* **2024**, *14*, 713. [CrossRef] [PubMed]
153. Ponce, H.; Martínez-Villaseñor, L.; Chen, Y. Editorial: Artificial intelligence in brain-computer interfaces and neuroimaging for neuromodulation and neurofeedback. *Front. Neurosci.* **2022**, *16*, 974269. [CrossRef] [PubMed]
154. Jang, B.S.; Jeon, S.H.; Kim, I.H.; Kim, I.A. Prediction of pseudoprogression versus progression using machine learning algorithm in glioblastoma. *Sci. Rep.* **2018**, *8*, 12516. [CrossRef]

155. Kim, J.Y.; Park, J.E.; Jo, Y.; Shim, W.H.; Nam, S.J.; Kim, J.H.; Yoo, R.E.; Choi, S.H.; Kim, H.S. Incorporating diffusion- and perfusion-weighted MRI into a radiomics model improves diagnostic performance for pseudoprogression in glioblastoma patients. *Neuro Oncol.* **2019**, *21*, 404–414. [CrossRef]
156. Liu, Z.M.; Zhang, H.; Ge, M.; Hao, X.L.; An, X.; Tian, Y.J. Radiomics signature for the prediction of progression-free survival and radiotherapeutic benefits in pediatric medulloblastoma. *Childs Nerv. Syst.* **2022**, *38*, 1085–1094. [CrossRef]
157. Luo, Y.; Zhuang, Y.; Xiao, Q.; Cai, L.; Chen, Y.; Yu, X.; Wang, J.; Zhang, Y. Multiparametric MRI-based radiomics signature with machine learning for preoperative prediction of prognosis stratification in pediatric medulloblastoma. *Acad. Radiol.* **2023**, *31*, 1629–1642. [CrossRef]
158. Ashry, E.R.; Maghraby, F.A.; El-Latif, Y.M.A.; Agag, M. Pediatric Posterior Fossa Tumors Classification and Explanation-Driven with Explainable Artificial Intelligence Models. *Int. J. Comput. Intell. Syst.* **2024**, *17*, 166. [CrossRef]
159. Kumarakulasinghe, N.B.; Blomberg, T.; Liu, J.; Leao, A.S.; Papapetrou, P. Evaluating Local Interpretable Model-Agnostic Explanations on Clinical Machine Learning Classification Models. In Proceedings of the 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, MN, USA, 28–30 July 2020; pp. 7–12.
160. Mosca, E.; Szigeti, F.; Tragianni, S.; Gallagher, D.; Groh, G.L. SHAP-Based Explanation Methods: A Review for NLP Interpretability. In Proceedings of the International Conference on Computational Linguistics 2022, Gyeongju, Republic of Korea, 12–17 October 2022.
161. Esmaeili, M.; Vettukattil, R.; Banitalebi, H.; Krogh, N.R.; Geitung, J.T. Explainable artificial intelligence for human-machine interaction in brain tumor localization. *J. Pers. Med.* **2021**, *11*, 1213. [CrossRef]
162. Zeineldin, R.A.; Karar, M.E.; Elshaer, Z.; Coburger, J.; Wirtz, C.R.; Burgert, O.; Mathis-Ullrich, F. Explainability of deep neural networks for MRI analysis of brain tumors. *Int. J. Comput. Assist. Radiol. Surg.* **2022**, *17*, 1673–1683. [CrossRef]
163. Alkhalaf, S.; Alturise, F.; Bahaddad, A.A.; Elnaim, B.M.E.; Shabana, S.; Abdel-Khalek, S.; Mansour, R.F. Adaptive Aquila Optimizer with explainable artificial intelligence-enabled cancer diagnosis on medical imaging. *Cancers* **2023**, *15*, 1492. [CrossRef]
164. Friston, K.J. Statistical Parametric Mapping. In *Neuroscience Databases*; Kötter, R., Ed.; Springer: Boston, MA, USA, 2003. [CrossRef]
165. Smith, S.M.; Jenkinson, M.; Woolrich, M.W.; Beckmann, C.F.; Behrens, T.E.; Johansen-Berg, H.; Bannister, P.R.; De Luca, M.; Drobnjak, I.; Flitney, D.E.; et al. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* **2004**, *23* (Suppl. S1), S208–S219. [CrossRef] [PubMed]
166. Fischl, B. FreeSurfer. *NeuroImage* **2012**, *62*, 774–781. [CrossRef] [PubMed]
167. Cheng, A.S.; Guan, Q.; Su, Y.; Zhou, P.; Zeng, Y. Integration of machine learning and blockchain technology in the healthcare field: A literature review and implications for cancer care. *Asia-Pac. J. Oncol. Nurs.* **2021**, *8*, 720–724. [CrossRef] [PubMed]
168. Dalboni da Rocha, J.L.; Coutinho, G.; Bramati, I.; Moll, F.T.; Sitaram, R. Multilevel diffusion tensor imaging classification technique for characterizing neurobehavioral disorders. *Brain Imaging Behav.* **2020**, *14*, 641–652. [CrossRef]
169. Zurita, M.; Montalba, C.; Labbé, T.; Cruz, J.P.; Dalboni da Rocha, J.; Tejos, C.; Ciampi, E.; Cárcamo, C.; Sitaram, R.; Uribe, S. Characterization of relapsing-remitting multiple sclerosis patients using support vector machine classifications of functional and diffusion MRI data. *Neuroimage Clin.* **2018**, *20*, 724–732. [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.

Journal of
Clinical Medicine

MDPI

Article

The Role of Baseline Total Kidney Volume Growth Rate in Predicting Tolvaptan Efficacy for ADPKD Patients: A Feasibility Study

Hreedi Dev [1,†], Zhongxiu Hu [1,†], Jon D. Blumenfeld [2,3], Arman Sharbatdaran [1], Yelynn Kim [1], Chenglin Zhu [1], Daniil Shimonov [2,3], James M. Chevalier [2,3], Stephanie Donahue [2], Alan Wu [4], Arindam RoyChoudhury [4], Xinzi He [1] and Martin R. Prince [1,5,*]

[1] Department of Radiology, Weill Cornell Medicine, New York, NY 10021, USA; hrd2001@med.cornell.edu (H.D.); zsh4001@med.cornell.edu (Z.H.); ars4017@med.cornell.edu (A.S.); yek4004@med.cornell.edu (Y.K.); chz4009@med.cornell.edu (C.Z.); xih4004@med.cornell.edu (X.H.)
[2] Rogosin Institute, New York, NY 10021, USA; jdblume@nyp.org (J.D.B.); das2041@nyp.org (D.S.); jac9014@nyp.org (J.M.C.); sld9001@nyp.org (S.D.)
[3] Department of Medicine, Weill Cornell Medicine, New York, NY 10021, USA
[4] Division of Biostatistics, Department of Population Health, Weill Cornell Medicine, New York, NY 10021, USA; alw4001@med.cornell.edu (A.W.); arr2014@med.cornell.edu (A.R.)
[5] Department of Radiology, Columbia Vagelos College of Physicians and Surgeons, New York, NY 10032, USA
* Correspondence: map2008@med.cornell.edu; Tel.: 1-646-962-7007
† These authors contributed equally to this work.

Abstract: Background/Objectives: Although tolvaptan efficacy in ADPKD has been demonstrated in randomized clinical trials, there is no definitive method for assessing its efficacy in the individual patient in the clinical setting. In this exploratory feasibility study, we report a method to quantify the change in total kidney volume (TKV) growth rate to retrospectively evaluate tolvaptan efficacy for individual patients. Treatment-related changes in estimated glomerular filtration rate (eGFR) are also assessed. **Methods:** MRI scans covering at least 1 year prior to and during treatment with tolvaptan were performed, with deep learning facilitated kidney segmentation and fitting multiple imaging timepoints to exponential growth in 32 ADPKD patients. Clustering analysis differentiated tolvaptan treatment "responders" and "non-responders" based upon the magnitude of change in TKV growth rate. Differences in rate of eGFR decline, urine osmolality, and other parameters were compared between responders and non-responders. **Results:** Eighteen (56%) tolvaptan responders (mean age 42 ± 8 years) were identified by k-means clustering, with an absolute reduction in annual TKV growth rate of >2% (mean = −5.1% ± 2.5% per year). Thirteen (44%) non-responders were identified, with <1% absolute reduction in annual TKV growth rate (mean = +2.4% ± 2.7% per year) during tolvaptan treatment. Compared to non-responders, tolvaptan responders had significantly higher mean TKV growth rates prior to tolvaptan treatment (7.1% ± 3.6% per year vs. 3.7% ± 2.4% per year; $p = 0.003$) and higher median pretreatment spot urine osmolality (Uosm, 393 mOsm/kg vs. 194 mOsm/kg, $p = 0.03$), confirmed by multivariate analysis. Mean annual rate of eGFR decline was less in responders than in non-responders (−0.25 ± 0.04, CI: [−0.27, −0.23] mL/min/1.73 m^2 per year vs. −0.40 ± 0.06, CI: [−0.43, −0.37] mL/min/1.73 m^2 per year, $p = 0.036$). **Conclusions**: In this feasibility study designed to assess predictors of tolvaptan treatment efficacy in individual patients with ADPKD, we found that high pretreatment levels of annual TKV growth rate and higher pretreatment spot urine osmolality were associated with a responder phenotype.

Keywords: ADPKD; tolvaptan; MRI; urine osmolality; total kidney volume; vasopressin; Mayo imaging classification; artificial intelligence; deep learning; treatment response

1. Introduction

Tolvaptan, a selective type 2 vasopressin receptor inhibitor, was FDA-approved in 2018 for treatment of autosomal dominant polycystic kidney disease (ADPKD) patients at risk for rapidly progressive disease, based upon randomized clinical trials demonstrating that it reduced the annual rates of total kidney volume (TKV) growth and attenuated the rate of decline in estimated glomerular filtration rate (eGFR) [1–7]. In those clinical trials, TKV was estimated either by stereology or an ellipsoidal solid approximation with 5% to 18% measurement variability, respectively, making it difficult to evaluate the efficacy of tolvaptan on TKV growth rate in individual subjects [8,9]. Subsequent advances in deep learning techniques have enabled more accurate quantification of TKV, with measurement variability as low as 1.3% [10–16]. The assessment of TKV growth rate, based on multiple TKV measurements, has also become more precise due to the development of machine learning methodologies, which are more accurate than single measurement-based methods used in previous clinical trials [1,17–20]. Consequently, the efficacy of tolvaptan in individual patients with ADPKD can be more reliably quantified by comparing TKV growth rates prior to and during treatment.

In TEMPO 3:4, a randomized, placebo-controlled trial in which the efficacy of tolvaptan in treating ADPKD was first established, a parallel design was used [3]. However, TEMPO 3:4 did not determine factors that predicted a favorable therapeutic response in individual subjects. At that time, assessing tolvaptan efficacy for individual subjects was not practical due to the limitations in the measurements of TKV and estimation of eGFR. Ideally, multiple TKV measurements prior to and during tolvaptan treatment are required to establish whether a response to tolvaptan has occurred.

Tolvaptan treatment is expensive (about EUR 40,000 per year [21]), and has significant renal (e.g., thirst, polyuria) and extrarenal (e.g., liver toxicity) side effects that require controlled access and mandatory regularly scheduled liver function tests. Therefore, it would be useful to identify the ADPKD patients who are most likely to benefit from this treatment and to confirm treatment benefit after initiating tolvaptan therapy, thereby personalizing its use to maximize its benefit and limit harm.

Machine learning techniques improve the precision and reproducibility of MRI assessment of TKV and allow accurate calculation of TKV growth rates in ADPKD patients [17,18]. Furthermore, the deep learning methods are semi-automatic, reducing subjectivity and reducing the effort required for analyzing multiple studies. We hypothesize that machine learning measurements of TKV to calculate and compare baseline TKV growth rates with during tolvaptan TKV growth rates will identify individual patients responding to tolvaptan treatment.

In this study we evaluated TKV growth rates, and other aspects of the ADPKD phenotype, prior to and during treatment with tolvaptan to identify predictors of its efficacy in individual patients.

2. Materials and Methods

This retrospective review of existing data from ADPKD subjects prior to and during tolvaptan treatment is HIPAA-compliant and approved by the Weill Medical College IRB. The requirement for written informed consent was waived.

Inclusion criteria were: (1) ADPKD patients with symmetrical cystic kidney disease and at least 2 abdominal MRI scans, separated by at least 1 year, before starting tolvaptan treatment, to establish the baseline TKV growth rate and (2) at least two abdominal MRI scans performed after initiating tolvaptan, separated by at least 1 year, to establish the kidney growth rate during treatment. Exclusion criteria were: (1) any interruption in tolvaptan treatment of more than 1 month, (2) non-adherence with tolvaptan management, and (3) events that affect kidney growth such as cyst rupture/aspiration.

2.1. Clinical Data Collection

Patient demographic, laboratory and clinical data were extracted from the electronic medical record. Laboratory measurements were collected within one month prior to tolvaptan initiation, including serum creatinine, liver function tests (alanine aminotransferase (ALT), aspartate aminotransferase (AST)), and spot urine osmolality (Uosm). The 2021 CKD-EPI equation was used to calculate eGFR [22].

Patient encounter notes were reviewed to identify systolic and diastolic blood pressure, hypertension medication, and tolvaptan dose.

2.2. Response Status Classification

2.2.1. Total Kidney Volume and Kidney Growth Rate Calculation

Left and right kidneys were contoured semi-automatically on abdominal MR scans by one expert (HD) for all scans as described previously using a 3D nnU-net deep learning model (www.traceorg.com). TKV was then computed as the sum of right kidney volume and left kidney volume. TKVs were calculated from all available pulse sequences that included both kidneys in their entirety, and the mean TKV across all available sequences, together with the age at the time of the scan were used to calculate kidney growth rates prior to and during treatment using the 2-parameter least squares fitting [18]. See Supplementary Figure S1 for kidney trajectories of each subject.

2.2.2. Interobserver Variability

Three expert observers (H.D., Z.H., Y.K.) independently contoured kidneys and calculated kidney volumes of all available scans in a subgroup of 9 (28%) subjects. TKV growth rates were then calculated (as per Section 2.2.1) using the kidney volumes measured by each expert observer. Interobserver variability in kidney volume was assessed by calculating interclass correlation coefficients.

2.2.3. Clustering Analysis

ADPKD subjects were designated as tolvaptan "responders" and "non-responders" based on the differences in TKV growth rates prior to, and during tolvaptan treatment. K-means clustering analysis was used to identify a natural separation between responders (in whom a larger magnitude of absolute decrease in the TKV growth rate occurred during treatment) and non-responders (in whom minimal or no decrease in the TKV growth rate occurred during treatment).

2.3. Estimated Kidney Function Trajectory

Serum creatinine was measured at monthly intervals for the first 18 months and then quarterly as routine monitoring of all patients receiving tolvaptan treatment, per standard care at our institution. eGFR was assessed over time for all subjects to determine whether the responder status for TKV growth rate corresponded to the rate of eGFR decline.

2.4. Statistics

Central tendency for continuous variables was calculated as mean ± standard deviation if they were normally distributed, as determined by the Shapiro–Wilk test. For non-normally distributed variables, median and interquartile range (IQR 25%/75%) are reported.

Student's *t*-test was used to assess the statistical significance of differences between responders and non-responders for normally distributed continuous variables and the Mann–Whitney U/Wilcoxon rank sum test was used for the continuous variables that were not normally distributed. The Chi-squared test was used to assess the statistical significance of differences between responders and non-responders in the distribution of gender and Mayo Imaging Classification. Benjamini–Hochberg corrections was applied to *p*-values in the univariate analysis. Multivariate logistic regression models were constructed using the variables identified in the univariate analysis with uncorrected $p < 0.05$ together with age and gender to assess their association with responder and non-responder status where slope, standard error, and confidence interval of the slope, and the Variable Inflation Factor was calculated for each variable. The final model was informed based on a hybrid approach and the relative values of Akaike Information Criterion.

Based on TEMPO 3:4 results [3], we expected a 2.7% absolute reduction in annual TKV growth rate from a mean of 5.5% to 2.8% with 3% variability. Using an effective size = 0.9 as calculated by Cohen's d for paired data, 90% power, alpha = 0.01, an estimated sample size of 18 subjects was required for this study to have adequate statistical power.

3. Results

Figure 1 indicates, 32 of 172 subjects initiated on tolvaptan fulfilled inclusion and exclusion criteria, with a sufficient number of MRI scans prior to and during tolvaptan treatment to determine the tolvaptan effect on TKV growth rate. Demographic and laboratory data are shown in Table 1, and a comparison to subjects not meeting study criteria show they have similar characteristics, Supplementary Table S1. Lower maximum tolvaptan median morning and evening doses reflect the fact that many of the excluded patients had stopped tolvaptan therapy before titrating up the maximum dose. The mean creatinine follow-up time on tolvaptan was 4 years and the mean imaging follow-up time was 3 years.

Table 1. Baseline demographic and laboratory data.

	All Subjects (*n* = 32)	**Responder (*n* = 18)**	**Non-Responder (*n* = 14)**	***p*-Value * Responder vs. Non-Responder**
Age (years)	42 ± 9	42 ± 8	42 ± 11	0.93
Sex (male:female, %male)	17:15, 53%	10:8, 56%	7:7, 50%	1.00
Height (m)	1.7 ± 0.1	1.7 ± 0.1	1.7 ± 0.1	0.59
Weight (kg)	75 ± 16	76 ± 19	73 ± 13	0.61
BMI [1] (kg/m^2)	26 ± 4	26 ± 5	25 ± 3	0.51
TKV (mL)	1622 [1067, 2715]	1616 [918, 2569]	1819 [1172, 2698]	0.69
htTKV (mL/m)	975 [589, 1578]	954 [538, 1519]	1019 [700, 1596]	0.66
Mayo Imaging Classification				
1A	0 (0%)	0 (0%)	0 (0%)	0.58
1B	2 (6%)	2 (11%)	0 (0%)	
1C	16 (50%)	9 (50%)	7 (50%)	
1D	8 (25%)	3 (17%)	5 (36%)	
1E	6 (19%)	4 (22%)	2 (14%)	

Table 1. *Cont.*

	All Subjects (n = 32)	Responder (n = 18)	Non-Responder (n = 14)	p-Value * Responder vs. Non-Responder
Blood Pressure (mmHg) [2]				
Systolic	126 ± 13	127 ± 14	124 ± 12	0.51
Diastolic	77 ± 8	77 ± 7	78 ± 8	0.86
Liver Function Tests				
AST [3] (U/L)	22 [19, 27]	22 [18, 28]	22 [21, 25]	0.82
ALT [4] (U/L)	22 [18, 27]	23 [20, 28]	20 [18, 26]	0.45
Bilirubin Total (mg/dL)	0.6 [0.5, 0.7]	0.6 [0.6, 0.7]	0.6 [0.5, 0.7]	0.75
Serum Creatinine (mg/dL)	1.3 ± 0.5	1.3 ± 0.4	1.3 ± 0.5	0.83
eGFR [5] (mL/min/1.73 m^2)	67 ± 27	68 ± 26	67 ± 29	0.96
Urine Specific Gravity	1.009 [1.005, 1.012]	1.010 [1.008, 1.013]	1.006 [1.005, 1.009]	0.22
Spot Uosm (mOsm/kg)	256 [168, 460]	393 [256, 519]	194 [167, 239]	0.03
Tolvaptan Morning Dose	90 [90, 90]	90 [68, 90]	90 [90, 90]	0.19
Tolvaptan Evening Dose	30 [30, 30]	30 [30, 30]	30 [30, 30]	0.73
Pre-tolvaptan TKV Growth Rate (% per year)	5.7 ± 3.5	7.1 ± 3.6	3.7 ± 2.4	0.003

[1] Body mass index; [2] Blood pressure was not recorded in four subjects at tolvaptan initiation; [3] Aspartate aminotransferase; [4] Alanine transaminase; [5] Estimated glomerular filtration rate; * None of these uncorrected p-values were significant after Benjamini–Hochberg corrections.

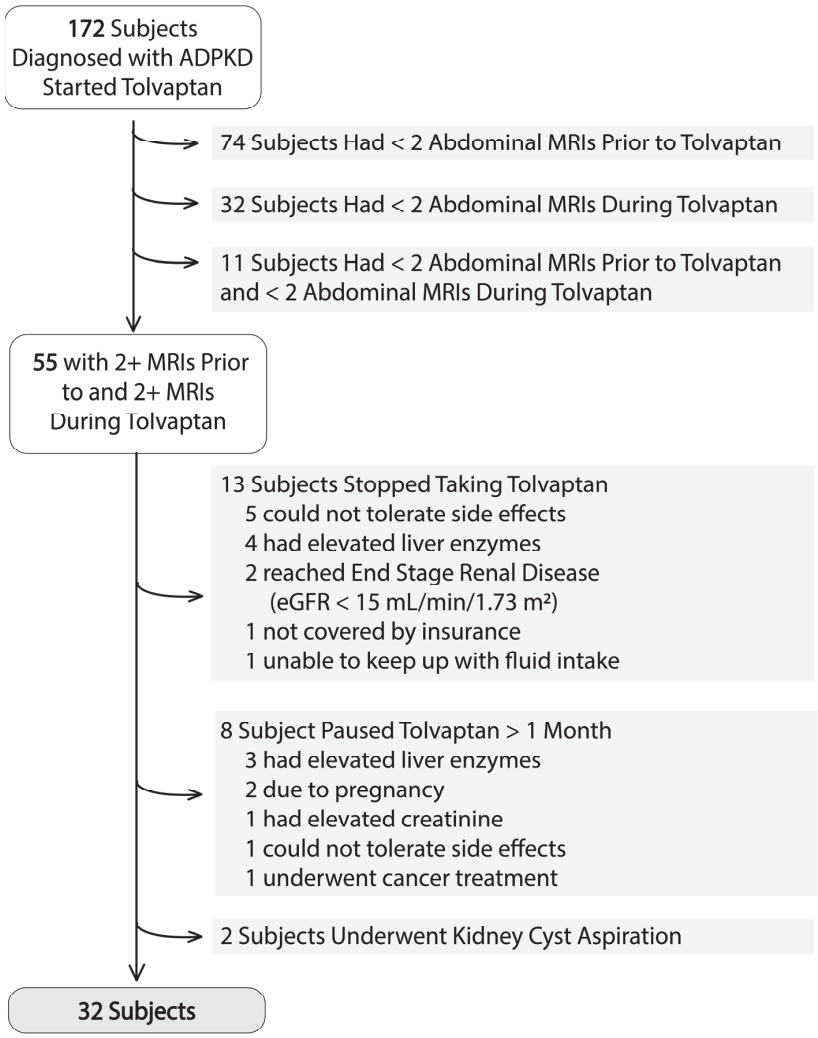


Figure 1. Patient flowchart.

3.1. Interobserver Variability

For 28% of cases with kidney labels corrected independently by three expert observers, there was excellent label agreement with ICC = 0.998.

3.2. Majority of Subjects Responded to Tolvaptan

K-means clustering analysis identified one group (n = 14) with an absolute TKV growth rate change of <1% and another group (n = 18) with an absolute TKV growth rate change >2% (Figure 2). Accordingly, all 14 subjects with <1% absolute reduction in annual absolute TKV growth rate during treatment were designated "non-responders". Eighteen exhibiting an absolute decrease in TKV growth rate >2% and were classified as "responders". Demographic and laboratory characteristics of subjects grouped by responder status are shown in Table 1.

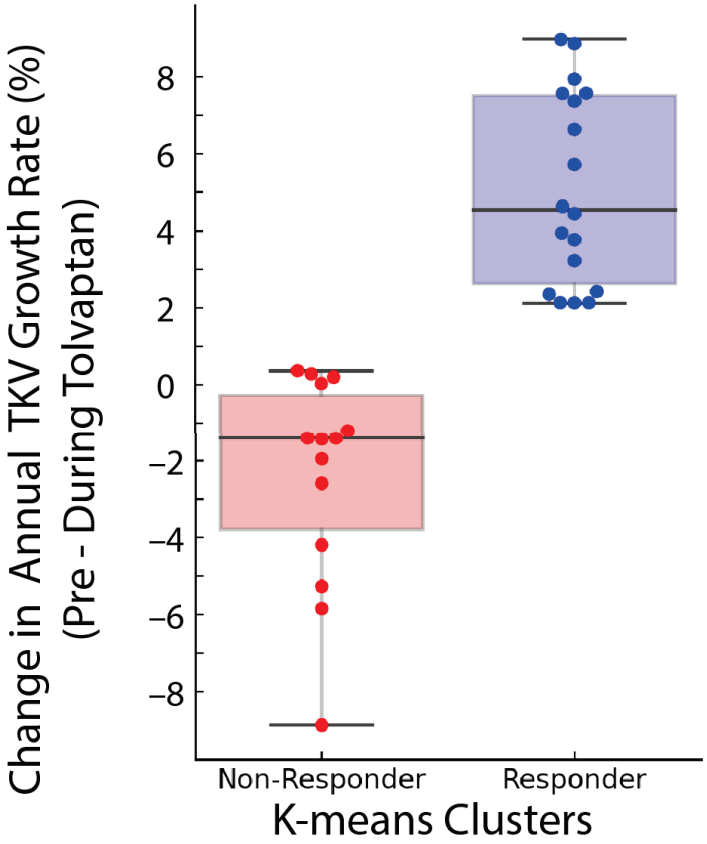


Figure 2. Clustering analysis of absolute change in TKV growth rate before and during tolvaptan treatment.

The mean annual TKV growth rate of all 32 subjects before tolvaptan was 5.7% ± 3.5% per year, decreasing to 3.8% ± 3.9% per year during tolvaptan treatment, as shown in Figure 3A, indicating that the cohort of 32 patients had a significant therapeutic response to tolvaptan therapy ($p = 0.03$). However, when the cohort was dichotomized as responders (Figure 3B) and non-responders (Figure 3C), the results diverged. The 18 responders had a mean pre-treatment annualized TKV growth rate of 7.1% ± 3.6%, which decreased to 2.0% ± 4.0% ($p < 1 \times 10^{-6}$) during tolvaptan treatment. In contrast, the 14 non-responders showed a mean pre-treatment annualized TKV growth rate of 3.7% ± 2.4%, which increased to 6.1% ± 2.3% ($p = 0.006$) during tolvaptan treatment. Examples of a responder and a non-responder are shown in Figure 4A,B, respectively.

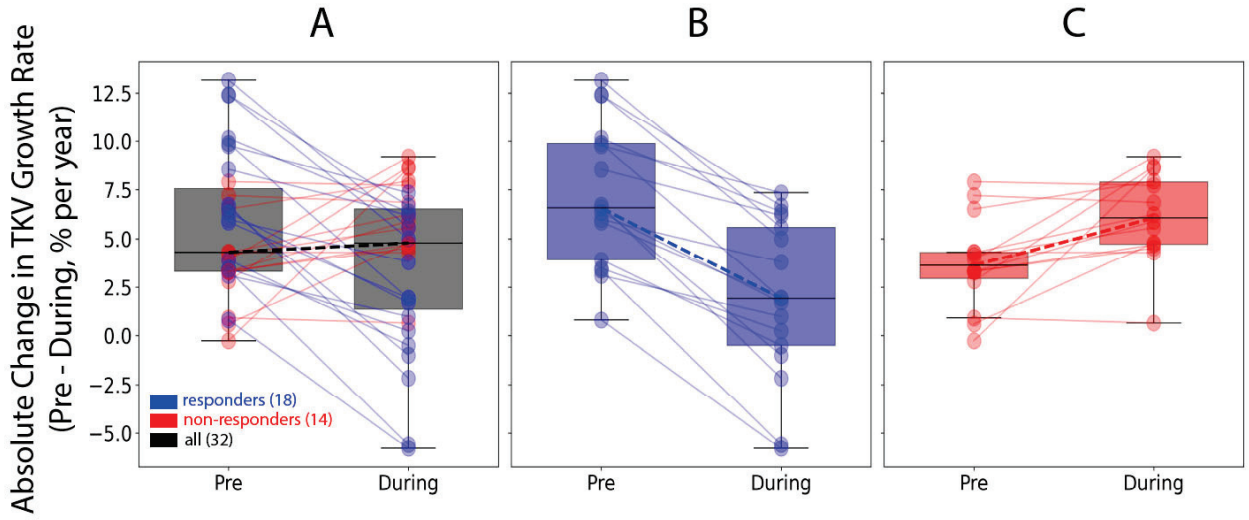


Figure 3. Change in TKV growth rate from pre to during tolvaptan for the entire cohort ((**A**), grey), responders ((**B**), blue), and non-responders ((**C**), red). Note that all responders (n = 18) had at least 2% absolute reduction in the TKV growth rate while non-responders (n = 14) all had less than 1% absolute reduction in TKV growth rate.

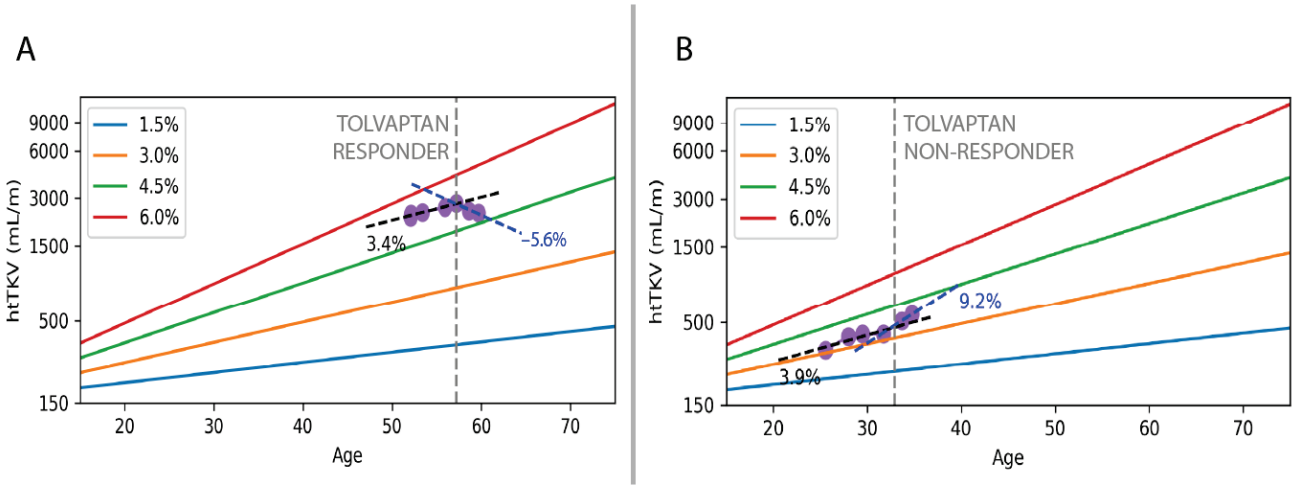


Figure 4. Change in TKV growth rate during tolvaptan treatment in a responder ((**A**), 3.4% to −5.6% TKV absolute change per year) and non-responder ((**B**), 3.9% to 9.2% TKV absolute change per year). Time of tolvaptan initiation is indicated by the vertical grey dashed line. Pre-tolvaptan TKV growth rate (black dashed line) and during-tolvaptan TKV growth rate (blue dashed line) were calculated using two-parameter least squares fitting of TKV measured on MRIs acquired before and after tolvaptan initiation respectively. Blue, orange, green and red lines correspond to Mayo Imaging Classification trajectories with annual TKV growth rates in the legend.

3.3. Features of Tolvaptan Responders vs. Non-Responders

The clinical, laboratory and imaging features at the start of tolvaptan treatment are shown in Table 1. Responders are associated with a higher baseline TKV growth rate (7.1% ± 3.6% per year, p = 0.003) before starting tolvaptan than non-responders (3.7% ± 2.4% per year). Baseline spot Uosm was higher in responders (393 [256, 519] mOsm/kg, p = 0.03) compared to non-responders (194 [167, 239] mOsm/kg).

On multivariate analysis, baseline spot Uosm and baseline TKV growth rate were both associated with response status to tolvaptan (Table 2 and Supplementary Table S2).

Table 2. Multivariate logistic regression assessment of parameters associated with tolvaptan responder status. This model is selected based on a hybrid process, where all multivariate models assessed can be found in Supplementary Table S2.

Variable	Coefficient	Standard Error	95% Confidence Interval	*p*-Value	VIF [1]
(Intercept)	4.2	1.6	[1.06, 7.3]	0.01 *	-
Baseline TKV Growth Rate	−0.43	0.19	[−0.81, −0.06]	0.02 *	1.14
Baseline Spot Uosm	−0.006	0.004	[0.014, −0.001]	0.03 *	1.14

[1] Variable Inflation Factor; * $p \leq 0.05$.

There were no significant differences in the dosing of tolvaptan between responders and non-responders; 13 of 18 (72%) Responders and 13 of 14 (93%) non-responders were on the highest daily tolvaptan dose (90 mg AM/30 mg PM).

3.4. Responders Showed Slower Decline in eGFR than Non-Responders

During the first 4 years of tolvaptan treatment, eGFR, declined less rapidly in responders than in non-responders (−0.25 ± 0.04, CI: [−0.27, −0.23] mL/min/1.73 m^2 per year vs. −0.40 ± 0.06, CI: [−0.43, −0.37] mL/min/1.73 m^2 per year, p = 0.036; Figure 5).

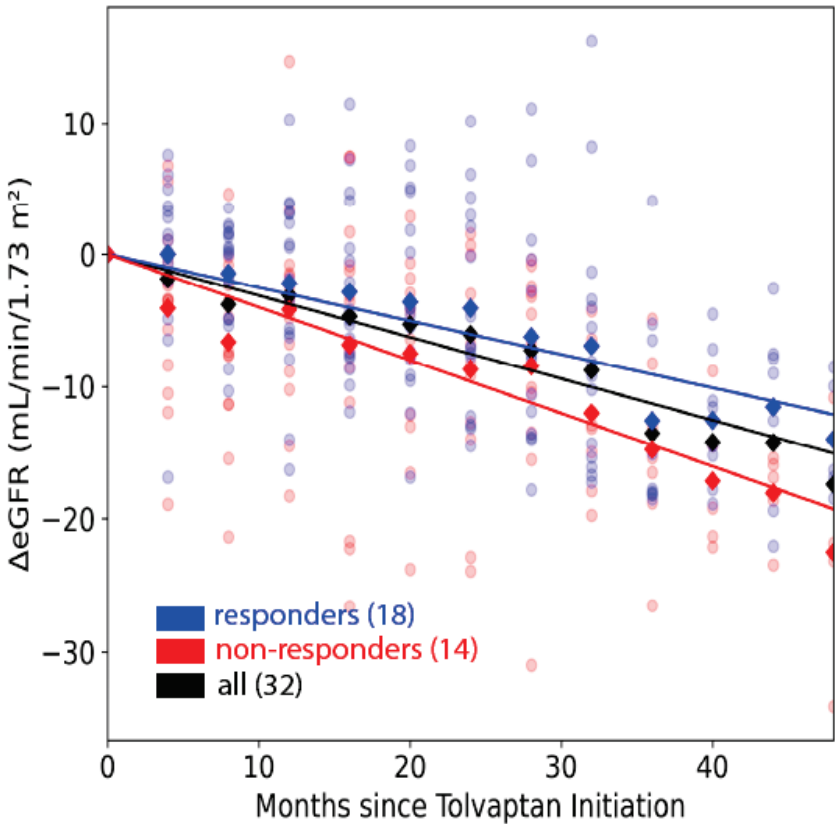


Figure 5. Trajectory of estimated glomerular filtration rate (eGFR) in 32 subjects. Change in eGFR (ΔeGFR) measurements of responders (blue dots) and non-responders (red dots) every 4 month since tolvaptan initiation to month 48 overlayed by fitted trendlines for all patients (black line), responders (blue line), non-responders (red line). Rate of eGFR decline in responders is −0.25 ± 0.04, CI: [−0.27, −0.23] mL/min/1.73 m^2 per year, and rate of eGFR decline in non-responders is −0.40 ± 0.06, CI: [−0.43, −0.37] mL/min/1.73 m^2 per year, which is 60% greater than responders.

4. Discussion

TEMPO 3:4 and REPRISE trials demonstrated that tolvaptan, a Type 2 vasopressin receptor antagonist, slows both TKV growth rate and the rate of eGFR decline in patients at high risk of rapidly progressing ADPKD [1–7]. However, these clinical trials did not differ-

entiate the individual subject's responses to tolvaptan. Moreover, sufficiently accurate and reproducible methods for measuring TKV, which is a biomarker of ADPKD progression, and for calculating TKV growth rate based on multiple imaging timepoints before and during tolvaptan within individual subjects were not available when those Phase 3 studies were performed. In the current feasibility study of 32 patients at high risk of rapidly progressing ADPKD, measurements of TKV growth rate were calculated from multiple timepoints using a more reproducible deep learning-based MRI TKV measurement technique [17,18]. We found that tolvaptan reduced the annual TKV growth rate in 56% of subjects (responders) and failed to slow the TKV growth rate in 44% (non-responders). This change in TKV growth rate response corresponded with the change in the eGFR response; those patients with a decrease in TKV growth rate during tolvaptan treatment also had a decrease in the rate of eGFR decline. Conversely, tolvaptan non-responders had a more rapid decline in eGFR compared to responders.

Responders were indistinguishable from non-responders by their baseline Mayo Clinic Imaging Classification (MCIC), a predictor of rate of progressive kidney disease in ADPKD that is based on height-adjusted kidney volume and patient age [1]. However, in our study, responders were associated with higher TKV growth rates prior to tolvaptan treatment compared to non-responders. Therefore, our data suggest that the MCIC is an inadequate marker for predicting therapeutic efficacy of tolvaptan in ADPKD. Accordingly, we propose that establishing a baseline TKV growth rate prior to treatment by using multiple imaging timepoints would appear to be a useful strategy for predicting the potential benefit of tolvaptan in individual ADPKD patients. Similarly, annual MRI scanning during tolvaptan with TKV measurements by a highly reproducible measurement technique appears to confirm whether a treatment response has occurred. If confirmed by prospective studies, this would have major prognostic implications, as the TKV growth rate response to tolvaptan corresponds with the trajectory of the eGFR response to tolvaptan treatment [3–5].

High Uosm is indicative of a high vasopressin level and has been identified as a potential biomarker for the long-term efficacy of tolvaptan in ADPKD by predicting the effects on TKV and eGFR [23]. A post hoc analysis of the TEMPO 3:4 trial reported that short-term reductions in Uosm correlated with slower rates of decline in kidney function, similar to the responders in our study [24]. Moreover, we recently showed in ADPKD patients that high water intake, sufficient to decrease mean 24 h Uosm and serum copeptin levels [25], decreased the height-adjusted TKV annual growth rate [26]. Vasopressin (via cyclic AMP) stimulates both urinary concentrating ability and kidney cyst growth and most likely accounts for the significant association of pretreatment TKV growth rate with the decrease in TKV growth rate during tolvaptan treatment. Spot Uosm measurements are variable and dependent upon the level of hydration and other factors at the time of sample collection. The broad confidence interval for the effect of baseline spot Uosm on determination of tolvaptan response status suggests further investigation of this effect in larger studies is needed, preferably measuring 24 h Uosm, which is a more stable indicator of chronic vasopressin activity than a spot Uosm. By contrast, TKV is a stable and reproducible measurement at any time point and is not affected acutely by transient factors such as hydration and solute intake.

Although urine specific gravity is sometimes used in place of Uosm because of their correlation, in our study, the baseline urine specific gravity showed only a very weak trend toward being higher in responders compared to non-responders ($p = 0.22$). This may reflect the fact that urine specific gravity is affected by the type of solute present, whereas Uosm is not affected by the type of solute. Thus, urine specific gravity should not be used in place of baseline Uosm for assessing the likelihood of response to tolvaptan.

Several other factors may contribute to the heterogeneity in responsiveness to tolvaptan. Although non-adherence to treatment is a consideration, it is unlikely that these subjects would be adherent to extensive imaging procedures and laboratory testing, but not to their treatment. Another consideration to account for the tolvaptan response phenotype is that kidney cysts originate from renal tubule epithelium from all nephron segments and are heterogeneous in size, appearance, and cellular characteristics. Although cortical collecting duct cysts, which are responsive to vasopressin, are predominant in ADPKD, cysts also originate from other nephron segments that are not responsive to vasopressin [27]. Therefore, it is conceivable that tolvaptan efficacy is related to the prevalence of vasopressin-responsive cysts in the individual patient.

We included patient age and sex in our multivariate analysis of tolvaptan response predictors because these factors are traditionally considered in assessing ADPKD progression [28]. However, neither age nor sex was a significant predictor of tolvaptan response in either univariate or multivariate analyses. This emphasizes that objective biomarkers such as baseline TKV growth rate and Uosm may be more relevant for predicting treatment efficacy than demographic characteristics.

The strengths of this study include: (i) multiple MRI scan data available both prior to and during tolvaptan treatment, (ii) MRI TKV measurement techniques that employed machine learning methodology that has been validated previously in large cohorts of ADPKD patients by experienced observers, and (iii) long periods of follow-up prior to and during tolvaptan treatment.

Important limitations are the relatively small sample size and retrospective data collection. This limited our study to reviewing MRI scans, the main imaging technique used at our institution to monitor ADPKD patients, because there were no subjects with sufficient CT scans to measure TKV growth rate both prior to and during tolvaptan. Ultrasound could not be used to track kidney volume because the variability of TKV measurement by the ultrasound equipment available at our center was too large to obtain meaningful TKV growth rate information. Another limitation is the lack of cross-validation or external/independent validation. Also, there was no adjustment for potential confounders in the assessment of eGFR decline, such as blood pressure, glycemic control, or concomitant medications. Although we were unable to use laboratory testing to assess adherence with their tolvaptan prescription, all patients were compliant with monthly clinically indicated laboratory testing and regular follow-up appointments with their nephrologist, which were required to maintain ongoing tolvaptan treatment. Patients fulfilling the inclusion/exclusion criteria (32 out of 172, 18%) were on a higher median tolvaptan dose compared to excluded patients on tolvaptan. This could have caused an increase in the proportion of responders, if patients on higher doses have a better treatment response, and is recommended to be explored in future studies.

5. Conclusions

Careful measurements of total kidney volume, using a highly reproducible MRI-based technique and evaluation of TKV growth rates prior to and during tolvaptan treatment by fitting TKV measurements to multiple timepoints, showed that more than 40% of ADPKD subjects did not significantly decrease their annualized TKV growth rate, a biomarker of disease severity and prognosis, during a 4-year period of treatment with tolvaptan. The tolvaptan responder group, defined by a higher mean pretreatment, baseline TKV growth rate than non-responders, also had a slower rate of decline in estimated glomerular filtration rate during tolvaptan treatment. Thus, baseline TKV growth rate appeared to predict tolvaptan efficacy in individual patients, and the difference in TKV growth rate prior to and during tolvaptan can be used to evaluate tolvaptan efficacy in individuals. The

small sample size, retrospective study design, and lack of external validation may limit generalization of the results. Prospective studies of larger populations of ADPKD patients are required to confirm these findings. Genetic and molecular biomarkers may complement baseline TKV growth rate and baseline Uosm for predicting tolvaptan treatment response.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jcm14051449/s1, Supplementary Figure S1. Kidney volume trajectory plots for 32 subjects; Supplementary Table S1. Comparison of study cohort (n = 32) to all patients on tolvaptan; Supplementary Table S2. Multivariate logistic regression models of parameters associated with responder and non-responder status.

Author Contributions: Conceptualization, M.R.P., J.M.C., D.S. and J.D.B.; methodology, Z.H., H.D., X.H. and C.Z.; software, X.H., Z.H. and C.Z.; validation, A.S., H.D., Z.H., Y.K. and M.R.P.; formal analysis, A.W., A.R., Z.H. and H.D.; resources, M.R.P. and J.D.B.; data curation, S.D., H.D. and Z.H.; writing—original draft preparation, H.D. and M.R.P.; writing—review and editing, all authors; project administration, M.R.P.; funding acquisition, M.R.P.; H.D. and Z.H. contributed equally as co-first authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Shaw Foundation and departmental funds.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Weill Cornell Medicine (protocol code 1610017623 approved on 6 July 2019 and 0304006105).

Informed Consent Statement: Patient consent was waived due to this study being a retrospective review using existing clinical information with low risk and impracticalities.

Data Availability Statement: Raw data can be shared under a data sharing agreement negotiated with Weill Cornell Medicine.

Conflicts of Interest: Daniil Shimonov declares speaking honoraria from Otsuka Inc. Jon D. Blumenfeld declares funding by Vertex for an unrelated ADPKD project. Arindam RoyChoudhury declares receiving payment for reviewing services from American Medical Association. All other authors have nothing to declare.

References

1. Irazabal, M.V.; Rangel, L.J.; Bergstralh, E.J.; Osborn, S.L.; Harmon, A.J.; Sundsbak, J.L.; Bae, K.T.; Chapman, A.B.; Grantham, J.J.; Mrug, M.; et al. Imaging classification of autosomal dominant polycystic kidney disease: A simple model for selecting patients for clinical trials. *J. Am. Soc. Nephrol.* **2015**, *26*, 160–172. [CrossRef] [PubMed]
2. Higashihara, E.; Nutahara, K.; Itoh, M.; Okegawa, T.; Tambo, M.; Yamaguchi, T.; Nakamura, Y.; Taguchi, S.; Kaname, S.; Yokoyama, K.; et al. Long-Term Outcomes of Longitudinal Efficacy Study with Tolvaptan in ADPKD. *Kidney Int. Rep.* **2022**, *7*, 270–281. [CrossRef]
3. Torres, V.E.; Chapman, A.B.; Devuyst, O.; Gansevoort, R.T.; Grantham, J.J.; Higashihara, E.; Perrone, R.D.; Krasa, H.B.; Ouyang, J.; Czerwiec, F.S.; et al. Tolvaptan in patients with autosomal dominant polycystic kidney disease. *N. Engl. J. Med.* **2012**, *367*, 2407–2418. [CrossRef] [PubMed]
4. Torres, V.E.; Chapman, A.B.; Devuyst, O.; Gansevoort, R.T.; Perrone, R.D.; Dandurand, A.; Ouyang, J.; Czerwiec, F.S.; Blais, J.D.; Investigators, T.T. Multicenter, open-label, extension trial to evaluate the long-term efficacy and safety of early versus delayed treatment with tolvaptan in autosomal dominant polycystic kidney disease: The TEMPO 4:4 Trial. *Nephrol. Dial. Transpl.* **2018**, *33*, 477–489. [CrossRef] [PubMed]
5. Torres, V.E.; Chapman, A.B.; Devuyst, O.; Gansevoort, R.T.; Perrone, R.D.; Koch, G.; Ouyang, J.; McQuade, R.D.; Blais, J.D.; Czerwiec, F.S.; et al. Tolvaptan in Later-Stage Autosomal Dominant Polycystic Kidney Disease. *N. Engl. J. Med.* **2017**, *377*, 1930–1942. [CrossRef]
6. Torres, V.E.; Chapman, A.B.; Devuyst, O.; Gansevoort, R.T.; Perrone, R.D.; Lee, J.; Hoke, M.E.; Estilo, A.; Sergeyeva, O. Multicenter Study of Long-Term Safety of Tolvaptan in Later-Stage Autosomal Dominant Polycystic Kidney Disease. *Clin. J. Am. Soc. Nephrol.* **2020**, *16*, 48–58. [CrossRef]

7. Higashihara, E.; Torres, V.E.; Chapman, A.B.; Grantham, J.J.; Bae, K.; Watnick, T.J.; Horie, S.; Nutahara, K.; Ouyang, J.; Krasa, H.B.; et al. Tolvaptan in autosomal dominant polycystic kidney disease: Three years' experience. *Clin. J. Am. Soc. Nephrol.* **2011**, *6*, 2499–2507. [CrossRef]
8. Demoulin, N.; Nicola, V.; Michoux, N.; Gillion, V.; Ho, T.A.; Clerckx, C.; Pirson, Y.; Annet, L. Limited Performance of Estimated Total Kidney Volume for Follow-Up of ADPKD. *Kidney Int. Rep.* **2021**, *6*, 2821–2829. [CrossRef]
9. Sharma, K.; Caroli, A.; Quach, L.V.; Petzold, K.; Bozzetto, M.; Serra, A.L.; Remuzzi, G.; Remuzzi, A. Kidney volume measurement methods for clinical studies on autosomal dominant polycystic kidney disease. *PLoS ONE* **2017**, *12*, e0178488. [CrossRef]
10. Goel, A.; Shih, G.; Riyahi, S.; Jeph, S.; Dev, H.; Hu, R.; Romano, D.; Teichman, K.; Blumenfeld, J.D.; Barash, I.; et al. Deployed Deep Learning Kidney Segmentation for Polycystic Kidney Disease MRI. *Radiol. Artif. Intell.* **2022**, *4*, e210205. [CrossRef]
11. Kim, Y.; Ge, Y.; Tao, C.; Zhu, J.; Chapman, A.B.; Torres, V.E.; Yu, A.S.; Mrug, M.; Bennett, W.M.; Flessner, M.F.; et al. Automated Segmentation of Kidneys from MR Images in Patients with Autosomal Dominant Polycystic Kidney Disease. *Clin. J. Am. Soc. Nephrol.* **2016**, *11*, 576–584. [CrossRef] [PubMed]
12. Kline, T.L.; Korfiatis, P.; Edwards, M.E.; Blais, J.D.; Czerwiec, F.S.; Harris, P.C.; King, B.F.; Torres, V.E.; Erickson, B.J. Performance of an Artificial Multi-Observer Deep Neural Network for Fully Automated Segmentation of Polycystic Kidneys. *J. Digit. Imaging* **2017**, *30*, 442–448. [CrossRef] [PubMed]
13. Potretzke, T.A.; Korfiatis, P.; Blezek, D.J.; Edwards, M.E.; Klug, J.R.; Cook, C.J.; Gregory, A.V.; Harris, P.C.; Chebib, F.T.; Hogan, M.C.; et al. Clinical Implementation of an Artificial Intelligence Algorithm for Magnetic Resonance-Derived Measurement of Total Kidney Volume. *Mayo Clin. Proc.* **2023**, *98*, 689–700. [CrossRef] [PubMed]
14. Raj, A.; Tollens, F.; Hansen, L.; Golla, A.K.; Schad, L.R.; Norenberg, D.; Zollner, F.G. Deep Learning-Based Total Kidney Volume Segmentation in Autosomal Dominant Polycystic Kidney Disease Using Attention, Cosine Loss, and Sharpness Aware Minimization. *Diagnostics* **2022**, *12*, 1159. [CrossRef]
15. Shin, T.Y.; Kim, H.; Lee, J.H.; Choi, J.S.; Min, H.S.; Cho, H.; Kim, K.; Kang, G.; Kim, J.; Yoon, S.; et al. Expert-level segmentation using deep learning for volumetry of polycystic kidney and liver. *Investig. Clin. Urol.* **2020**, *61*, 555–564. [CrossRef]
16. van Gastel, M.D.A.; Edwards, M.E.; Torres, V.E.; Erickson, B.J.; Gansevoort, R.T.; Kline, T.L. Automatic Measurement of Kidney and Liver Volumes from MR Images of Patients Affected by Autosomal Dominant Polycystic Kidney Disease. *J. Am. Soc. Nephrol.* **2019**, *30*, 1514–1522. [CrossRef]
17. Dev, H.; Zhu, C.; Sharbatdaran, A.; Raza, S.I.; Wang, S.J.; Romano, D.J.; Goel, A.; Teichman, K.; Moghadam, M.C.; Shih, G.; et al. Effect of Averaging Measurements from Multiple MRI Pulse Sequences on Kidney Volume Reproducibility in Autosomal Dominant Polycystic Kidney Disease. *J. Magn. Reson. Imaging* **2023**, *58*, 1153–1160. [CrossRef]
18. Hu, Z.; Sharbatdaran, A.; He, X.; Zhu, C.; Blumenfeld, J.D.; Rennert, H.; Zhang, Z.; Ramnauth, A.; Shimonov, D.; Chevalier, J.M.; et al. Improved predictions of total kidney volume growth rate in ADPKD using two-parameter least squares fitting. *Sci. Rep.* **2024**, *14*, 13794. [CrossRef]
19. Higashihara, E.; Yamamoto, K.; Kaname, S.; Okegawa, T.; Tanbo, M.; Yamaguchi, T.; Shigemori, K.; Miyazaki, I.; Yokoyama, K.; Nutahara, K. Age- and height-adjusted total kidney volume growth rate in autosomal dominant polycystic kidney diseases. *Clin. Exp. Nephrol.* **2019**, *23*, 100–111. [CrossRef]
20. Borrego Utiel, F.J.; Espinosa Hernandez, M. How to Estimate Kidney Growth in Patients with Autosomal Dominant Polycystic Kidney Disease. *J. Am. Soc. Nephrol.* **2023**, *34*, 944–950. [CrossRef]
21. Oporto, F.J.; Espín, J.; Gómez, C.A.; Cosano, G.M.; Sánchez-Pozo, M.C.; Rocha, J.L. Cost-Minimization Analysis of Tolvaptan Treatment for ADPKD in Southern Spain. *Kidney Int. Rep.* **2025**, *10*, 265–266. [CrossRef] [PubMed]
22. Inker, L.A.; Eneanya, N.D.; Coresh, J.; Tighiouart, H.; Wang, D.; Sang, Y.; Crews, D.C.; Doria, A.; Estrella, M.M.; Froissart, M.; et al. New Creatinine- and Cystatin C-Based Equations to Estimate GFR Without Race. *N. Engl. J. Med.* **2021**, *385*, 1737–1749. [CrossRef] [PubMed]
23. Huh, H.; Kim, Y.S.; Chung, W.; Kim, Y.L.; Kim, Y.; Han, S.; Jung, Y.; Na, K.Y.; Lee, K.B.; Oh, Y.K.; et al. Evaluating the Safety and effectivenesS in adult KorEaN patients treated with Tolvaptan for management of autosomal domInAnt poLycystic kidney disease (ESSENTIAL): Short-term outcomes during the titration period. *Kidney Res. Clin. Pract.* **2023**, *42*, 216–228. [CrossRef] [PubMed]
24. Ivaturi, V.; Gobburu, J.; Leslie, B.; Wang, X.; Jadhav, P. Urine Osmolality Is a Potential Marker of Longer-Term Efficacy of Tolvaptan in ADPKD: A Post-Hoc Analysis. *Kidney360* **2024**, *5*, 996–1001. [CrossRef]
25. Arjune, S.; Oehm, S.; Todorova, P.; Gansevoort, R.T.; Bakker, S.J.L.; Erger, F.; Benzing, T.; Burst, V.; Grundmann, F.; Antczak, P.; et al. Copeptin in autosomal dominant polycystic kidney disease: Real-world experiences from a large prospective cohort study. *Clin. Kidney J.* **2023**, *16*, 2194–2204. [CrossRef]
26. Dev, H.; Zhu, C.; Barash, I.; Blumenfeld, J.D.; He, X.; RoyChoudhury, A.; Wu, A.; Prince, M.R. Feasibility of Water Therapy for Slowing Autosomal Dominant Polycystic Kidney Disease Progression. *Kidney360* **2024**, *5*, 698–706. [CrossRef]

27. Baert, L. Hereditary polycystic kidney disease (adult form): A microdissection study of two cases at an early stage of the disease. *Kidney Int.* **1978**, *13*, 519–525. [CrossRef]
28. Cornec-Le Gall, E.; Audrézet, M.P.; Rousseau, A.; Hourmant, M.; Renaudineau, E.; Charasse, C.; Morin, M.P.; Moal, M.C.; Dantal, J.; Wehbe, B.; et al. The PROPKD score: A new algorithm to predict renal survival in autosomal dominant polycystic kidney disease. *J. Am. Soc. Nephrol.* **2016**, *27*, 942–951. [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.

tomography

MDPI

Article

Deep Learning for Ultrasonographic Assessment of Temporomandibular Joint Morphology

Julia Lasek [1], Karolina Nurzynska [2,*], Adam Piórkowski [3], Michał Strzelecki [4] and Rafał Obuchowicz [5]

[1] Faculty of Geology, Geophysics and Environmental Protection, AGH University of Krakow, 30-059 Krakow, Poland; julial@agh.edu.pl
[2] Department of Algorithmics and Software, Silesian University of Technology, 44-100 Gliwice, Poland
[3] Department of Biocybernetics and Biomedical Engineering, AGH University of Krakow, 30-059 Krakow, Poland
[4] Institute of Electronics, Lodz University of Technology, 93-590 Lodz, Poland; michal.strzelecki@p.lodz.pl
[5] Department of Diagnostic Imaging, Jagiellonian University Medical College, 30-663 Krakow, Poland
* Correspondence: karolina.nurzynska@polsl.pl

Abstract: Background: Temporomandibular joint (TMJ) disorders are a significant cause of orofacial pain. Artificial intelligence (AI) has been successfully applied to other imaging modalities but remains underexplored in ultrasonographic evaluations of TMJ. Objective: This study aimed to develop and validate an AI-driven method for the automatic and reproducible measurement of TMJ space width from ultrasonographic images. Methods: A total of 142 TMJ ultrasonographic images were segmented into three anatomical components: the mandibular condyle, joint space, and glenoid fossa. State-of-the-art architectures were tested, and the best-performing 2D Residual U-Net was trained and validated against expert annotations. The algorithm for joint space width measurement based on TMJ segmentation was proposed, calculating the vertical distance between the superior-most point of the mandibular condyle and its corresponding point on the glenoid fossa. Results: The segmentation model achieved high performance for the mandibular condyle (Dice: 0.91 ± 0.08) and joint space (Dice: 0.86 ± 0.09), with notably lower performance for the glenoid fossa (Dice: 0.60 ± 0.24), highlighting variability due to its complex geometry. The TMJ space width measurement algorithm demonstrated minimal bias, with a mean difference of 0.08 mm and a mean absolute error of 0.18 mm compared to reference measurements. Conclusions: The model exhibited potential as a reliable tool for clinical use, demonstrating accuracy in TMJ ultrasonographic analysis. This study underscores the ability of AI-driven segmentation and measurement algorithms to bridge existing gaps in ultrasonographic imaging and lays the foundation for broader clinical applications.

Keywords: deep learning; temporomandibular joints; ultrasonography; artificial intelligence

1. Introduction

Temporomandibular joint (TMJ) disorders are among the most significant causative factors of orofacial pain [1]. According to the latest meta-analysis, approximately 34% of the global population is affected by conditions related to TMJ dysfunction [2]. Geographically, TMJ disorders are most prevalent in South America, potentially due to socio-demographic and genetic factors, though these remain poorly understood [3].

Research consistently indicates that the etiology of TMJ disorders is multifactorial, with individual prevalence varying based on specific contributing factors [4–7]. Commonly cited causative factors include parafunctional habits, psychological conditions such as depression and anxiety (often accompanied by increased muscle tension), and joint degeneration due to

hypermobility, trauma, and inflammation. Additionally, autoimmune diseases, rheumatoid disorders, and fibromyalgia have been linked to prolonged TMJ pain [8–13].

Among the morphological and functional parameters used to assess TMJ health, the width of the joint space is particularly critical for determining the status of the orthognathic system [14]. Changes in joint space width, especially reductions, are strongly associated with morphological disorders such as arthrosis and other degenerative TMJ conditions [15,16]. Traditional imaging techniques, such as computed tomography (CT), cone-beam computed tomography (CBCT), and magnetic resonance imaging (MRI), are widely used for evaluating TMJ space, demonstrating high levels of agreement in results [17]. However, these techniques present several limitations in the context of temporomandibular joint (TMJ) assessment. While CT and CBCT are highly effective for evaluating bony structures, they offer limited contrast for soft tissue visualization, which is critical for assessing the articular disk and synovial membrane of the TMJ [18]. Additionally, these modalities rely on ionizing radiation, raising concerns about repeated exposure, particularly for patients with chronic TMJ disorders. CT and CBCT imaging are also prone to artifacts caused by metallic dental restorations, which can compromise image quality and diagnostic accuracy [19]. Furthermore, these techniques usually provide static images, making it difficult to evaluate dynamic TMJ movements that are crucial for understanding joint function. If pseudo-dynamic assessment is held, this technique is associated with multiplied radiation dose and, therefore, is considered not safe [20]. MRI, on the other hand, excels in soft tissue visualization without radiation exposure but is not without its challenges. It is costly, time-intensive, and less accessible compared to other imaging modalities. Motion artifacts during MRI scans can degrade image quality, and its static nature also limits the ability to capture real-time joint dynamics [21,22]. Moreover, MRI is contraindicated for patients with metallic implants, pacemakers, or claustrophobia [23]. These limitations underscore the need for alternative imaging approaches, such as artificial intelligence-enhanced ultrasonography, which offers real-time, non-invasive, cost-effective, and radiation-free imaging, addressing many of the challenges posed by traditional techniques. Ultrasonography has shown potential in clinical settings, achieving similar results and emerging as a candidate for routine use [24]. However, ultrasonography measurements are subject to operator fatigue and bias, leading to concerns about reproducibility. Such measurements often require significant expertise, limiting their reliability when performed by less experienced clinicians [25,26].

Over the past few years, artificial intelligence (AI) has shown significant potential in addressing various challenges associated with medical imaging. It is worth noting that, due to the high variability and noise levels in ultrasound images, the number of AI algorithms supporting the analysis of such images is limited compared to images from other modalities (CT, MRI, X-ray) [27]. Nevertheless, machine learning methods are applied in the analysis of such images. These methods utilize radiomics, including frequently used image texture features [28], to support image-based diagnosis in a diverse range of applications [29–31]. Recently, deep learning approaches have also been employed for processing and analyzing such images, e.g., to improve ultrasound image quality [32] or to assist in the diagnostic process [33].

AI has also been effectively utilized in the evaluation of the TMJ. Vinayahalingam et al. [34] demonstrated the potential of deep learning techniques for the automated segmentation of TMJ structures on MRI. In addition to CBCT, deep learning-based segmentation of the TMJ has also been widely explored in MRI [35–39]. A review of studies on TMJ segmentation across different modalities is summarized in Table 1. Choi et al. [40] utilized AI-enhanced analysis of panoramic radiography and joint noise data to improve the diagnosis of degenerative joint

diseases. Furthermore, Kazimierczak et al. [41] showcased the impact of AI in optimizing CBCT imaging quality for TMJ evaluation.

Table 1. Summary of deep learning studies on TMJ segmentation.

Author(s)	Modality	Segmented Structures
Vinayahalingam et al. (2023) [34]	CBCT	Mandibular condyle, glenoid fossa
Kumar & Baskaran (2023) [35]	X-ray	Mandibular condyle, glenoid fossa
Ito et al. (2022) [36]	MRI	Articular disk
Choi et al. (2024) [37]	CBCT	Mandibular condyle
Li et al. (2022) [38]	MRI	Mandibular condyle, articular eminence, articular disk
Yoon et al. (2024) [39]	MRI	Mandibular condyle, temporal bone, articular disk
Our Study	US	Mandibular condyle, joint space, glenoid fossa

Although AI has been successfully applied to the assessment of the temporomandibular joint using various imaging modalities, to our knowledge, there are currently no AI-based solutions for the ultrasound assessment of the TMJ (Table 1), despite its crucial importance in modern medicine [42]. In light of the above, this study aimed to develop and validate an AI-driven method for the automatic and reproducible measurement of TMJ space width using ultrasonographic imaging. Recognizing the growing role of AI in medical imaging, we sought to evaluate the effectiveness of the developed method using deep learning segmentation architectures for this relatively under-explored imaging modality. To achieve this, a dataset comprising 142 ultrasonographic images of the TMJ was curated and annotated by a radiologist, segmenting the images into three anatomically distinct components: the mandibular condyle, the joint space, and the glenoid fossa. In an effort to advance the field, we have made this dataset publicly accessible, enabling further research and validation of AI-based solutions for the ultrasonographic evaluation of TMJ.

2. Materials and Methods

Between April 2020 and June 2022, a total of 300 patients undergoing evaluation for TMJ-related pain were referred for ultrasonographic assessment of the temporomandibular joints. The diagnostic ultrasonographic procedures were conducted at the Department of Radiology and Ultragen Medical Clinic, with patients referred from various clinical departments within the Clinical Hospital.

To ensure consistency and reliability of the results, all measurements were performed by a radiologist with a minimum of six years of experience in ultrasonographic examinations, utilizing ultrasonographic technology. The radiologist was directly involved in the joint visualization and measurement process and personally oversaw the examinations. Ethical approval for the study was obtained from the Institutional Ethics Committee of Jagiellonian University (Approval No. 1072.6120.138.17). The study included patients presenting with facial pain localized around the TMJ. These patients were referred for initial ultrasonographic evaluation as part of their diagnostic workup imaging, which is a part of standard evaluation procedure. This study's dataset has been published and can be retrieved at https://doi.org/10.5281/zenodo.14760859 (accessed on 29 January 2025).

2.1. Exclusion Criteria

1. Age Restrictions: Patients under 18 years of age or over 65 years of age were excluded from the study to ensure the study focuses on the adult population with typical TMJ anatomy;
2. Previous TMJ Surgery: Patients with a history of surgical intervention on the TMJ, as such procedures could alter joint morphology and affect ultrasonographic findings;
3. Severe TMJ Deformities: Cases with congenital or severe acquired TMJ deformities, such as ankylosis, which may limit accurate ultrasonographic evaluation;
4. Systemic Conditions: Patients with systemic disorders affecting the TMJ, such as advanced rheumatoid arthritis, severe osteoarthritis, where deformities were too severe;
5. Severe Dental Conditions: Presence of extensive dental malocclusion or recent dental surgeries, which may confound TMJ-associated pain assessment;
6. Inflammatory or Infectious Conditions: Active TMJ infections or severe inflammatory conditions (e.g., abscesses) that may require immediate medical or surgical intervention;
7. Inability to Undergo Ultrasonography: Patients with conditions limiting their ability to comply with proper positioning or imaging protocols;
8. Inadequate Imaging Quality: Exclusion of scans with poor imaging quality;

2.2. Examination Protocol

Patients were positioned in a supine, relaxed state with their jaws in habitual occlusion, ensuring that any splints or oral devices were removed prior to the examination. A layer of ultrasound gel was applied to the skin to optimize acoustic coupling. The examination was performed using the GE HealthCare ultrasound system (GE Versana Premier [GE Healthcare, Chicago, IL, USA]) equipped with a shear wave device and an 8–18 MHz hockey stick probe (GE Healthcare, Chicago, IL, USA).

The TMJ was evaluated using morphological imaging sections oriented parallel to the joint line. This approach enabled detailed visualization of joint features, including the presence of exudate, irregularities in the bony surface, overall joint congruence (i.e., the alignment of the joint surfaces), and the position of the articular disk. Assessments were conducted during habitual occlusion and dynamic movements of the jaw, such as opening and closing the mouth upon command (Figure 1).

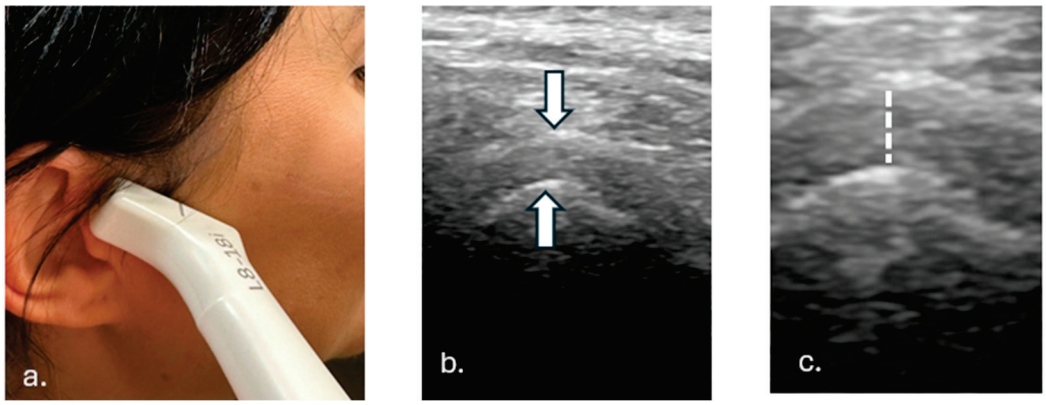


Figure 1. Figure 1 presents methodology of examination of the TM Joint. (**a**). Hockey stick probe alignment parallel to the joint gap. (**b**). Obtained US image of the joint with indicated articulation surfaces (arrows). (**c**). Joint gap width–searched parameter (marked with dotted line).

2.3. Data

Following the exclusion of unsuitable cases, the final dataset consisted of 142 exams. Each image was preprocessed by being cropped to the diagnostic field of view. The images

were saved in 8-bit grayscale format. The dataset was partitioned using the holdout method with a random split into three subsets: a training set containing 86 exams (60%), a validation set consisting of 21 exams (15%), and a test set comprising 35 exams (25%). The validation set was used for hyperparameter tuning and model optimization, while the test set was reserved for final performance evaluation.

2.4. Data Annotation

The segmentation process was designed to annotate three anatomically distinct components of the TMJ: the mandibular condyle (MC), representing the inferior portion of the TMJ; the joint space (JS); and the glenoid fossa of the temporal bone (GF), constituting the superior aspect of the TMJ. In the first phase, voxel-wise segmentation of these structures was meticulously performed by a radiologist, creating a reference dataset to serve as the initial ground truth for preliminary U-Net model training. This model was then used to generate the initial AI-generated segmentations.

In the second phase, the radiologist reviewed and refined the initial manual segmentations with guidance from the AI-generated outputs. These refinements were conducted carefully to maintain the anatomical accuracy of the segmented structures while addressing any inconsistencies between the left and right boundaries. Importantly, these adjustments did not alter critical parameters, such as the measurement of the TMJ space width, but were undertaken to enhance the uniformity of the dataset, thereby optimizing its utility as high-quality training data for further model development. This step provided the final segmentations for the final model training.

2.5. Model Evaluation

The evaluation of the model was conducted by comparing its predictions with the ground truths provided on the test set. All metrics were computed using Python version 3.11.9. Segmentation performance was quantified using the following metrics:

Dice Coefficient:

$$\text{Dice} = \frac{2 \times |P \cap G|}{|P| + |G|} \tag{1}$$

where P and G represent the predicted and ground truth segmentations, respectively. Dice measures the overlap between the two sets, with a value of 1 indicating perfect agreement.

Precision:

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

Precision evaluates the fraction of correctly predicted positive pixels among all predicted positives, where TP represents true positives and FP false positives.

Recall:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

Recall (or Sensitivity) measures the fraction of correctly predicted positive pixels among all actual positive pixels, where FN represents false negatives.

Volume Similarity (VS):

$$\text{VS} = 1 - \frac{|V_P - V_G|}{V_P + V_G} \tag{4}$$

where V_P and V_G are the volumes of the predicted and ground truth segmentations, respectively. VS assesses the relative agreement in size between the predicted and actual segmentations, with a value of 1 indicating perfect volumetric match.

Hausdorff Distance (HD):

$$\mathrm{HD}(P,G) = \max\left\{ \sup_{p\in P} \inf_{g\in G} d(p,g), \sup_{g\in G} \inf_{p\in P} d(g,p) \right\}$$

Hausdorff Distance (HD) quantifies the maximum boundary deviation between predicted (P) and ground truth (G) segmentations, computed as the greatest distance from any point in one set to the closest point in the other.

2.6. Postprocessing

Postprocessing was applied to refine the predicted segmentation masks by analyzing connected components for each class (excluding the background). For each class, contiguous regions (connected components) were identified, and their sizes were calculated. The largest component was used as a reference, and smaller components were retained only if their size was at least 60% of the largest component. This thresholding effectively removed spurious regions while preserving meaningful structures. The processed masks were reconstructed by retaining the relevant components for each class. This method improved segmentation quality by reducing noise and focusing on significant structures.

2.7. TMJ Space Width Measurement Algorithm

To measure the TMJ space width (JSW), a computational algorithm was developed to identify and calculate the vertical distance between the mandibular condyle and the glenoid fossa in segmented data. The algorithm first locates the superior-most point of the MC, defined as the pixel with the highest vertical position within its region. If multiple pixels share this vertical position, the algorithm selects the median pixel along the horizontal axis. Subsequently, the algorithm identifies the closest point on the GF directly aligned vertically with the selected MC point within the same column. The JSW is then determined by calculating the vertical distance between these two points.

3. Results

3.1. Architecture Selection

To identify the most effective segmentation architecture, we evaluated multiple state-of-the-art deep learning models, including Residual U-Net [43], U-Net++ [44], SegResNet (with and without Variational Autoencoder) [45], V-Net [46], Attention U-Net [47], and DeepLabV3 [48]. All training and inference experiments were conducted on an NVIDIA RTX 3080 GPU (NVIDIA Corporation, Santa Clara, CA, USA). Each model was trained using an automated hyperparameter tuning approach to optimize performance. Model selection was performed using Ray Tune [49], along with the Asynchronous Successive Halving Algorithm, which dynamically pruned underperforming configurations, enabling computationally efficient exploration of the hyperparameter space. The entire optimization process was systematically tracked and logged using Weights & Biases [50]. The hyperparameter search space was defined to optimize key model components, including optimizer type, learning rate, number of feature channels, activation function, normalization method, kernel size, and dropout rate. The specific parameters varied depending on the architecture to ensure compatibility with the network's design. The optimal hyperparameters for each architecture were selected based on the mean Dice score on the validation set.

Subsequently, the best-performing architectures were evaluated on the test set using clinically relevant and computational metrics: segmentation accuracy (Dice coefficient, Hausdorff Distance [mm], precision, recall, volume similarity), training efficiency (number of epochs, epoch-wise training time [s]), inference speed (per-image inference latency [s]), and model complexity (number of parameters, multiply–accumulate operations [MACs]).

A comprehensive comparison of these architectures is presented in Table 2. Given its demonstrated segmentation accuracy and efficiency, the Residual U-Net was identified as the optimal model for further refinement and final validation.

Table 2. Summary of evaluated segmentation architectures, including segmentation accuracy, training efficiency, inference speed, and model complexity.

	Attention U-Net	U-Net ++	DeepLabv3	SegResNet	SegResNetVAE	Residual U-Net	V-Net
Dice	0.64	0.72	0.72	0.73	0.73	0.75	0.66
Hausdorff Distance [mm]	5.36	2.14	2.08	2.52	2.76	1.42	3.10
Precision	0.67	0.74	0.75	0.73	0.73	0.77	0.60
Recall	0.65	0.72	0.72	0.74	0.74	0.75	0.79
Volume Similarity	0.87	0.90	0.88	0.90	0.91	0.91	0.79
Number of Epochs	72	76	35	38	73	59	22
Epoch Training Time [s]	25.24	89.72	163.22	22.85	26.25	15.39	42.38
Inference Time [ms]	0.09	0.18	0.45	0.09	0.09	0.23	0.10
Parameters ($\times 10^6$)	31.83	133.78	281.4	11.61	16.64	35.31	110.95
MACs ($\times 10^9$)	1.01	2.41	55.8	24.25	0.88	3.15	9.36

3.2. Optimal Hyperparameter Configuration

The final model was a 2D Residual U-Net implemented within the MONAI framework [51]. This model utilized a single-channel input and a four-channel output, with a network configuration consisting of feature channels [16, 32, 64, 128], with strides of 2 and incorporated 16 residual units per block, instance normalization, PReLU activation, and dropout (0.3) for regularization (Figure 2).

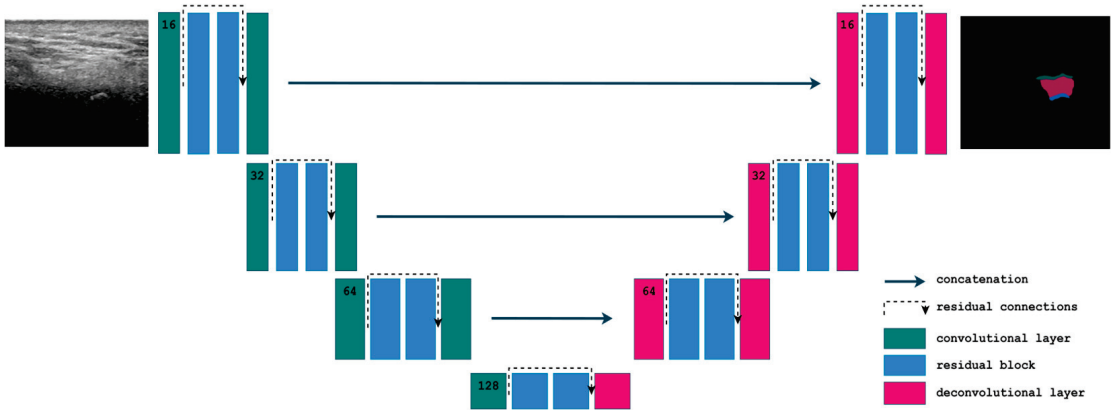


Figure 2. Schematic representation of the Residual U-Net architecture.

The model was trained for 150 epochs with a batch size of 2, using employed Dice Loss with softmax normalization and was optimized via the Adam optimizer with a starting learning rate of 1×10^{-3}, using a ReduceLROnPlateau scheduler all implemented within the PyTorch framework (version 2.5.1). To further improve segmentation performance and model stability, the final model was trained multiple times (with the same hyperparameters), and the three best runs—determined based on the validation Dice score—were selected for ensembling, enhancing robustness, and reducing variance in predictions. Figure 3 provides an overview of a selected subset of explored hyperparameter configurations and their corresponding mean Dice scores on the validation set.

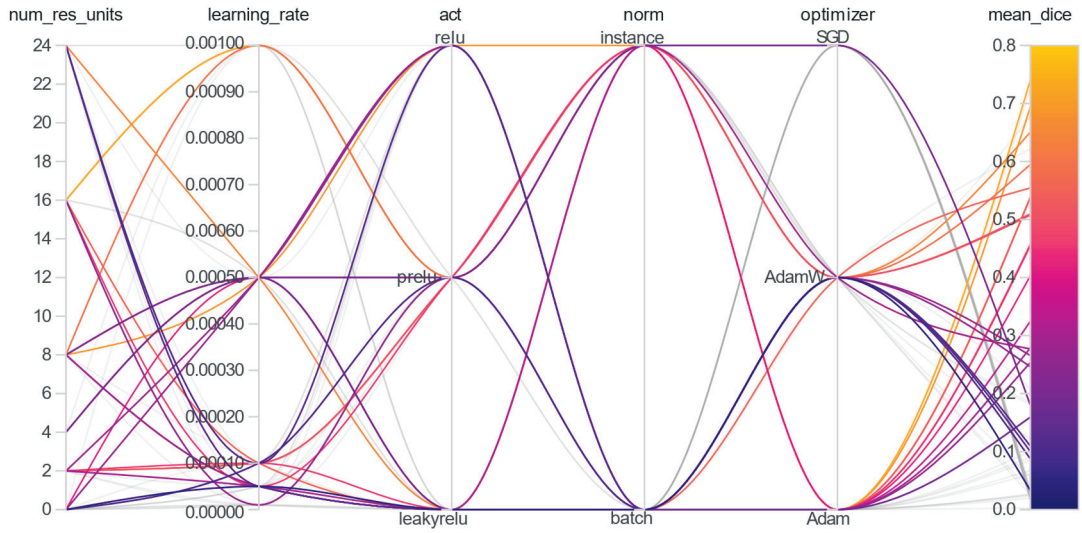


Figure 3. The plot visualizes the performance of different hyperparameter combinations for a subset of experiments.

3.3. Final Model Segmentation Performance

The segmentation performance of the model was evaluated across three anatomical structures: the mandibular condyle, the joint space, and the glenoid fossa of the temporal bone. The results, summarized in Table 3 and Figure 4, present the Dice coefficient, precision, recall, and volume similarity (VS) for each structure.

Table 3. Segmentation performance metrics for the mandibular condyle (MC), joint space (JS), and glenoid fossa (GF) on the test set. Metrics are reported as mean ± standard deviation.

Metric	GF	JS	MC
Dice	0.60 ± 0.24	0.86 ± 0.09	0.91 ± 0.08
Precision	0.60 ± 0.27	0.90 ± 0.09	0.95 ± 0.07
Recall	0.63 ± 0.25	0.84 ± 0.15	0.89 ± 0.11
VS	0.86 ± 0.10	0.90 ± 0.08	0.94 ± 0.07

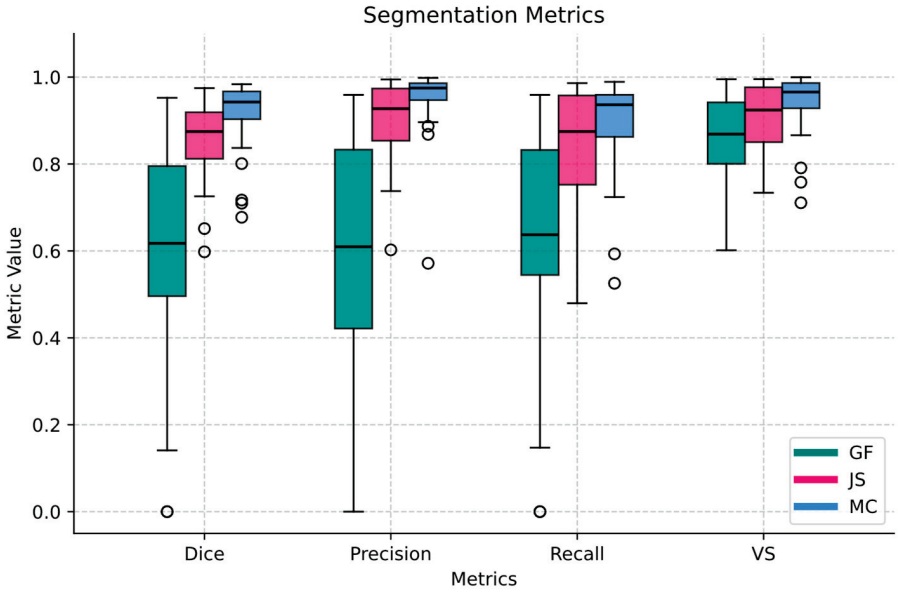


Figure 4. Segmentation performance metrics for the mandibular condyle (MC), joint space (JS), and glenoid fossa (GF) on the test set. Metrics include Dice coefficient, precision, recall, and volume similarity present as boxplots. Circles represent outliers.

The MC achieved the highest performance across all metrics, with a Dice coefficient of 0.91 ± 0.08, precision of 0.95 ± 0.07, recall of 0.89 ± 0.11, and VS of 0.94 ± 0.07. These values indicate highly accurate segmentation with minimal volume discrepancies. The joint space (JS) also demonstrated strong segmentation performance, with a Dice coefficient of 0.86 ± 0.09, precision of 0.90 ± 0.09, recall of 0.84 ± 0.15, and VS of 0.90 ± 0.08. While the high precision suggests accurate predictions, the slightly lower recall indicates occasional under-segmentation. In contrast, the GF exhibited the lowest performance, with a Dice coefficient of 0.60 ± 0.24, precision of 0.60 ± 0.27, recall of 0.63 ± 0.25, and VS of 0.86 ± 0.10. The larger variability in these metrics highlights challenges in segmenting this structure, potentially due to its intricate geometry and less-defined boundaries.

Overall, the model demonstrated robust segmentation for the mandibular condyle and joint space, albeit with greater variability in performance for the glenoid fossa.

3.4. TMJ Space Width Measurement

The algorithm for measuring TMJ space width was evaluated on the same test set used to assess the segmentation model. Measurements were derived from manual segmentations (referred to as reference measurements), which were performed by a radiologist to serve as the ground truth, and AI model predictions (referred to as AI measurements) to assess the model's performance in comparison to expert-defined standards. The mean TMJ space width obtained from the AI predictions was 1.80 mm, with a median of 1.67 mm. In comparison, the reference measurements yielded a mean of 1.88 mm, with the same median of 1.67 mm. The mean difference between the AI and reference measurements was 0.077 mm, indicating minimal systematic bias. The standard deviation of the differences was 0.33 mm, reflecting the variability between the two sets of measurements.

The mean percentage error was 8.96%, and the mean absolute error was 0.18 mm. These results suggest that the algorithm performs consistently and accurately. Such findings highlight the algorithm's potential for reliable and automated quantification of TMJ space width in clinical and research applications. A statistical summary of the results is provided in Table 4.

Table 4. Summary of the key metrics for TMJ space width measurements. The table includes the mean and median values for measurements derived from AI predictions and manual segmentations, along with statistical comparisons.

Metric of TMJ Space Width	Value
AI Predicted Mean (mm)	1.80
AI Predicted Median (mm)	1.67
Reference Mean (mm)	1.88
Reference Median (mm)	1.67
Mean Difference (mm)	0.08
Standard Deviation of Differences (mm)	0.33
Mean Percentage Error (%)	8.96
Mean Absolute Error (MAE, mm)	0.18
Root Mean Square Error (RMSE, mm)	0.34

The quality of the automatic measurements was further assessed through the graphs in Figure 5, which include the Bland–Altman plot and the concordance plot. The Bland–Altman plot evaluated the agreement between the AI-predicted TMJ space width measurements and the reference measurements obtained from manual segmentations. The x-axis represents the mean of the AI and reference measurements, while the y-axis shows the difference between the two measurements. Individual data points, displayed in blue, represent the differences across the test set. The dashed red line denotes the mean differ-

ence, highlighting any systematic bias, while the dashed green lines represent the upper and lower 95% limits of agreement. Notably, the mean difference lies close to zero, reflecting minimal systematic bias between the AI predictions and the reference standard. Furthermore, nearly all measurements (except two) fall within the limits of agreement, underscoring the consistency of the AI model. The two outliers were both underestimations, which may be attributed to the possibility that the AI model incorrectly segmented the glenoid fossa, potentially marking a different anatomical structure or delineating the region much too broadly. It is worth noting that one of the outliers corresponded to an exceptionally large reference measurement compared to the overall dataset, which could indicate that the deviation was influenced by the specific characteristics of this particular measurement. This suggests that the issue may be a result of the limited diversity within the dataset. In a more diverse and representative dataset, such extreme cases might be less frequent or entirely absent, thereby reducing the overall impact of outliers on the evaluation of the model's performance.

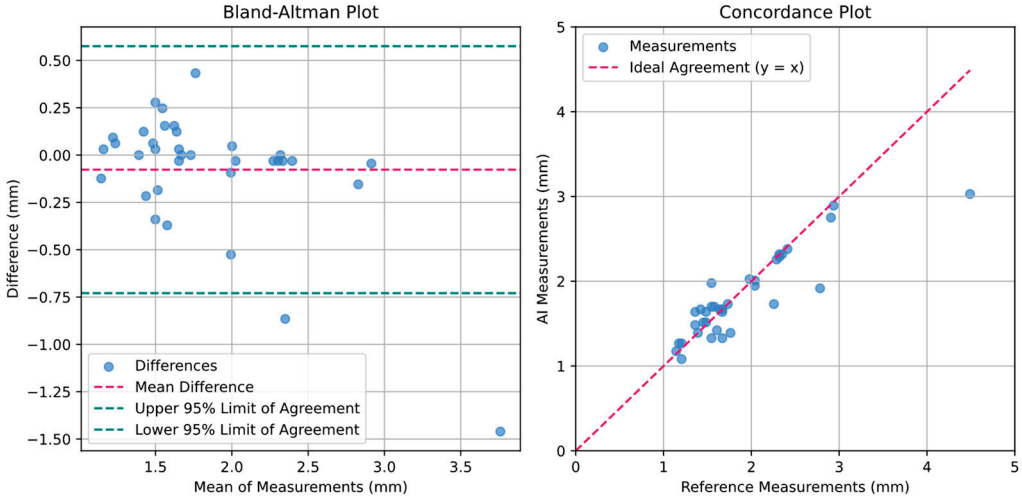


Figure 5. Bland–Altman and concordance plots for assessing the agreement between AI-predicted TMJ space width measurements and reference measurements.

The concordance plot assessed the agreement between the AI-predicted measurements and the reference measurements. The x-axis corresponds to the reference measurements, and the y-axis corresponds to the AI-predicted measurements. Individual data points are displayed in blue, while the dashed red line represents the line of perfect agreement. The alignment of data points along this line demonstrates a high degree of agreement between the two measurement sets.

Both graphs underscore the high concordance and accuracy of the AI model in predicting TMJ space width when compared to the reference standard, confirming the robustness and reliability of the proposed algorithm.

The results of both the AI model for TMJ segmentation and the TMJ space width measurement are visually presented in Figure 6, highlighting performance across varying levels of accuracy. The figure includes representative cases from the test set, categorized by best, average, and worst Dice scores. Specifically, segmentation performance is illustrated for the highest Dice scores ((A) GF: 0.89, JS: 0.96, MC: 0.97), the average Dice scores ((B) GF: 0.78, JS: 0.91, MC: 0.84), and the lowest Dice scores ((C) GF: 0.00, JS: 0.65, MC: 0.84). The mandibular condyle is segmented with high accuracy, highlighting the model's reliability in identifying this anatomical structure. Deviations in TMJ space width measurements are primarily attributable to under-segmentation of the glenoid fossa, as illustrated in the example study (C). Occasionally, the model may incorrectly segment a different structure

due to the challenges of interpreting ultrasonographic signals. One potential source of segmentation error is the influence of surrounding ligaments, which can sometimes appear as hyperechoic (bright) structures on an ultrasound, potentially leading to misclassification. Despite these occasional inaccuracies, the model demonstrates strong overall performance. With further refinements and the incorporation of additional training data, the AI model holds significant potential as a valuable tool for automating TMJ assessments.

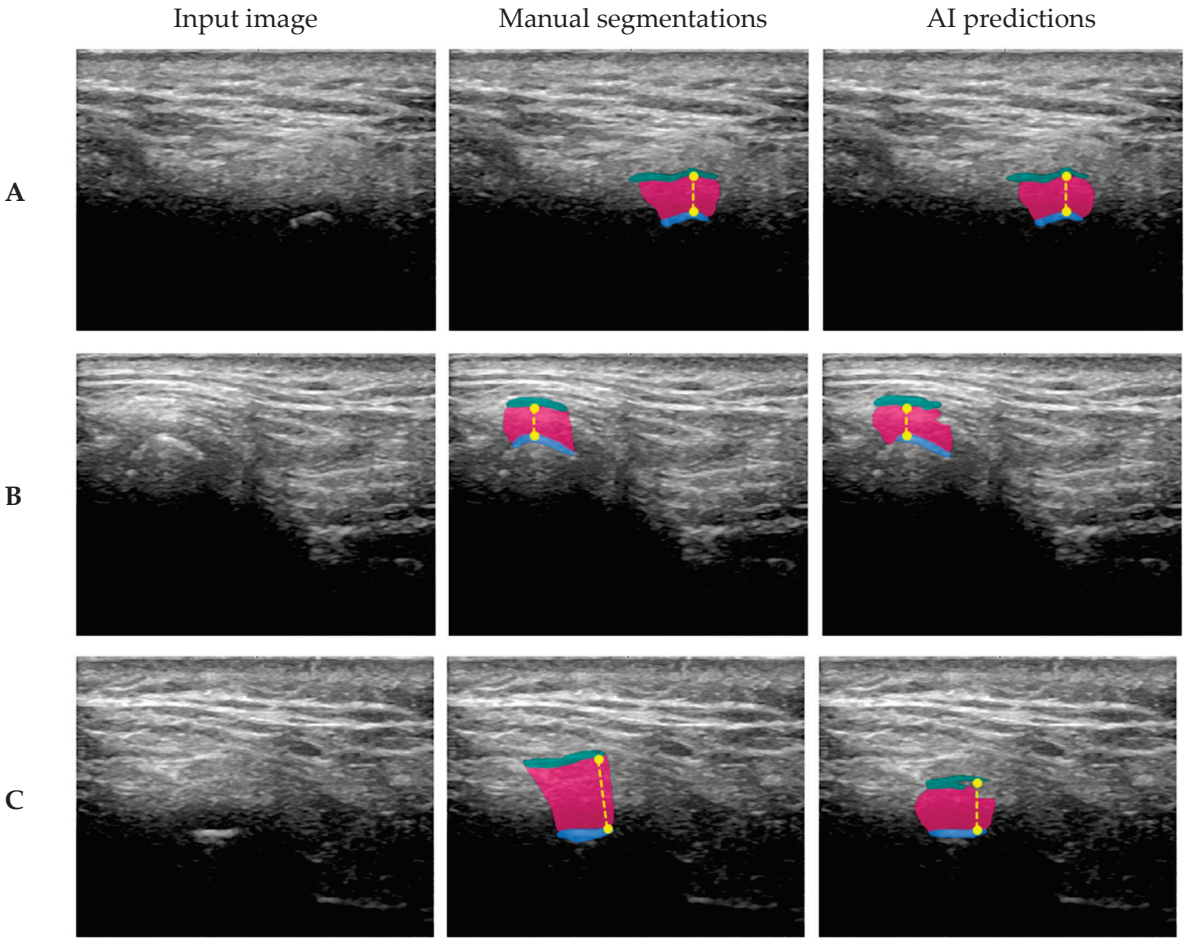


Figure 6. The visualization showcases TMJ segmentation and joint space width (yellow dashed line) measurement results, highlighting the mandibular condyle (MC), joint space (JS), and glenoid fossa (GF) with distinct colors: MC in blue, JS in pink, and GF in green. Segmentation performance is demonstrated across three scenarios: (**A**) the highest Dice (GF: 0.89, JS: 0.96, MC: 0.97), (**B**) the average Dice (GF: 0.78, JS: 0.91, MC: 0.84), (**C**) and the lowest Dice (GF: 0.00, JS: 0.65, MC: 0.84).

4. Discussion

Among the various radiological modalities used for assessing the TMJ, ultrasonography (US) has gained popularity among imaging professionals due to its affordability, ease of use, and non-invasive nature [52]. However, concerns remain regarding its accuracy, as the imaging setup can be difficult to replicate. Variations in patient anatomy and the subjective approach of the examiner necessitate standardization to improve reliability [53,54]. In this study, the mean distance between the TMJ articulating surfaces measured using ultrasound (US) was 1.9 mm. This finding does not align with the results reported by Noh et al. [26], who observed a mean distance of 3 mm in their studied population. This discrepancy may be attributed to the inclusion of patients with varying conditions in our study. Specifically, cases with disk displacement and mild arthrosis were included to develop a more robust system capable of handling diverse datasets. While artificial

intelligence (AI) has been employed in various studies for detecting and differentiating pathologies, its application in TMJ ultrasonography remains unexplored. Lee et al. [55] utilized convolutional neural networks (CNNs) to automatically detect effusion in the TMJ, diagnosed via MRI, achieving a specificity of 87.25% with their proposed model. Similarly, Xu et al. [56] reviewed CNN-based detection of osteoarthrosis in X-ray images, reporting diagnostic accuracies ranging from 78.0% to 92.4%, highlighting the potential of AI as a promising diagnostic tool. In a study by Almăşan et al. [57], artificial neural networks (ANNs) and CNNs were used to detect TMJ osteoarthritis from cone-beam computed tomography (CBCT) images, achieving sensitivity and specificity values of up to 98% and 83%, respectively. Eşer et al. [58] applied the YOLO approach to CBCT image analysis, reporting a sensitivity of 1, further demonstrating the potential of this method. Farook and Dudley [59] reviewed the overall performance of AI techniques based on radiographs and highlighted their strong diagnostic accuracy compared to MRI, particularly for edge detection in X-ray images. Zou et al. [60] used a deep learning model (ANN) to evaluate clinical features, medical status, and psychosocial conditions, comparing the results with those of physicians with varying levels of experience. Their study revealed that the ANN achieved an accuracy of over 90%, outperforming less experienced physicians. Additionally, Yıldız et al. [61] applied machine learning algorithms to evaluate non-radiographic clinical parameters, achieving an accuracy of 86% in their study. The comparative analysis of existing studies underscores the significant advancements in AI applications for CBCT, X-ray, MRI, and even clinical data. However, there remains a notable gap in the application of AI for ultrasonographic analysis of the TMJ. Various methods, predominantly CNNs, have been employed for segmentation and diagnosis in other imaging modalities [34]. Addressing this gap is crucial and could provide valuable insights into the field.

This work represents a step toward standardization in TMJ imaging by utilizing an AI training set based on the expertise of highly experienced physicians. The proposed AI solution has the potential to serve as a reliable second opinion for a broader cohort of medical professionals, offering a significant contribution to the field of TMJ diagnostics and filling a critical gap in current knowledge.

Seven deep neural network architectures designed for image segmentation were compared in this study. Residual U-Net achieved the best segmentation results, as indicated by the Dice, HD, precision, recall, and VS metrics presented in Table 2. This network is based on the classic U-Net architecture but incorporates residual blocks, a solution commonly used in residual networks. These blocks help to mitigate the vanishing gradient problem, which hinders training, especially in networks with a large number of deep layers. A residual block includes a skip connection that adds the original input to the output of the convolutional block, providing an alternative pathway for the gradient. Compared to other architectures, Residual U-Net also exhibited a relatively short inference time (0.23 ms), a moderate number of model parameters (35.3×10^6), and a computational complexity of 3.15×10^9 MACs. The beneficial impact of residual blocks on segmentation accuracy is further supported by the good results obtained with SegResNet and SegResNet VAE, where most quality metrics were only slightly worse than those of Residual U-Net (with the exception of HD, which was notably higher). Additionally, these networks exhibited the shortest inference time (0.09 ms), and SegResNet VAE had the lowest computational complexity, equal to 0.88×10^9 MACs.

While this study utilized 142 ultrasonographic examinations, the relatively small dataset poses limitations to the generalizability of the findings. The lack of diversity in demographic and clinical characteristics may hinder the applicability of the AI model to broader patient populations with varying anatomical or pathological profiles. Future studies with larger, more diverse datasets are essential to validate the robustness and adapt-

ability of the proposed model. Such datasets should include a wide range of ethnicities, age groups, and TMJ conditions to ensure broader clinical applicability. A study on a larger and more diverse dataset to validate the robustness of the AI model is planned as a future work.

The lower performance of the segmentation model for the glenoid fossa (Dice: 0.60 ± 0.24) highlights challenges related to the complex geometry and less-defined boundaries of this structure. This variability could reduce the reliability of joint space width measurements in cases with atypical anatomical features. The inclusion of patients with conditions like disk displacement and mild arthrosis aimed to create a robust system but introduced significant variability in the dataset. This might have influenced model training and performance, potentially limiting its application in studies focused on specific TMJ disorders.

Another critical limitation is the reliance on manual annotations from a single, albeit experienced, expert radiologist. While this ensured high-quality annotations, it introduced the possibility of human bias in defining ground truth, which could have influenced model training and evaluation. To address this, future studies could explore using annotations from multiple experts, complemented by consensus strategies or automated refinement tools, to minimize bias and improve annotation reliability.

5. Conclusions

The present AI method for TMJ ultrasonographic assessment has significant clinical implications, particularly in addressing the challenges of variability and reproducibility in TMJ imaging. The developed AI model demonstrates high accuracy in segmenting critical anatomical structures, such as the mandibular condyle and joint space, and provides reliable measurements of TMJ space width with minimal bias compared to manual assessments. This capability is important for standardizing ultrasonographic evaluations, which are traditionally prone to operator fatigue and variability, thereby ensuring consistent and objective results across different clinical settings.

By automating the segmentation and measurement processes, the AI system has the potential to reduce the time and expertise required for TMJ ultrasonographic analysis. This is especially valuable in resource-limited settings where access to highly trained radiologists may be restricted. Moreover, the absence of ionizing radiation and the real-time capabilities of ultrasonography make this AI-driven approach a safer and more practical option for routine TMJ evaluations, particularly for monitoring chronic conditions that require repeated imaging.

Furthermore, its ability to streamline the imaging workflow and offer standardized results could enable broader adoption of ultrasonography as a routine diagnostic tool for TMJ disorders, bridging the gap between high-cost imaging modalities like MRI and CT and the need for accessible, non-invasive diagnostics in everyday practice. This positions the AI method as a transformative tool in the field of TMJ imaging, paving the way for more efficient, and patient-centered care.

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

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by Institutional Ethics Committee of the Jagiellonian University no. 1072.6120.138.17, dated 28 September 2017.

Informed Consent Statement: The patients provided their written informed consent to participate in the examination and the study.

Data Availability Statement: The dataset is available at https://doi.org/10.5281/zenodo.14760859 (accessed on 29 January 2025).

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

Abbreviations

The following abbreviations are used in this manuscript:

TMJ	temporomandibular joint
US	ultrasonography
CBCT	cone-beam computed tomography
MC	mandibular condyle
JS	joint space
GF	glenoid fossa

References

1. List, T.; Jensen, R.H. Temporomandibular Disorders: Old Ideas and New Concepts. *Cephalalgia* **2017**, *37*, 692–704. [CrossRef] [PubMed]
2. Zieliński, G.; Pająk-Zielińska, B.; Ginszt, M. A Meta-Analysis of the Global Prevalence of Temporomandibular Disorders. *J. Clin. Med.* **2024**, *13*, 1365. [CrossRef] [PubMed]
3. Yost, O.; Liverman, C.T.; English, R.; Mackey, S.; Bond, E.C. Section: 2: Definitions and Scope: What Are TMDs? In *Temporomandibular Disorders: Priorities for Research and Care*; The National Academies Press: Washington, DC, USA, 2020. [CrossRef]
4. Murphy, G.; Haider, M.; Ghai, S.; Sreeharsha, B. The Expanding Role of MRI in Prostate Cancer. *Am. J. Roentgenol.* **2013**, *201*, 1229–1238. [CrossRef] [PubMed]
5. Chisnoiu, A.M.; Picos, A.M.; Popa, S.; Chisnoiu, P.D.; Lascu, L.; Picos, A.; Chisnoiu, R. Factors Involved in the Etiology of Temporomandibular Disorders-A Literature Review. *Med. Pharm. Rep.* **2015**, *88*, 473–478. [CrossRef] [PubMed]
6. Moreno-Fernández, A.M.; Jiménez-Castellanos, E.; Iglesias-Linares, A.; Bueso-Madrid, D.; Fernández-Rodríguez, A.; De Miguel, M. Fibromyalgia Syndrome and Temporomandibular Disorders with Muscular Pain. A Review. *Mod. Rheumatol.* **2017**, *27*, 210–216. [CrossRef]
7. Golanska, P.; Saczuk, K.; Domarecka, M.; Kuć, J.; Lukomska-Szymanska, M. Temporomandibular Myofascial Pain Syndrome—Aetiology and Biopsychosocial Modulation. A Narrative Review. *Int. J. Environ. Res. Public Health* **2021**, *18*, 7807. [CrossRef] [PubMed]
8. Obuchowicz, R.; Strzelecki, M.; Piórkowski, A. Clinical Applications of Artificial Intelligence in Medical Imaging and Image Processing—A Review. *Cancers* **2024**, *16*, 1870. [CrossRef] [PubMed]
9. Sharma, S.; Gupta, D.; Pal, U.; Jurel, S. Etiological Factors of Temporomandibular Joint Disorders. *Natl. J. Maxillofac. Surg.* **2011**, *2*, 116. [CrossRef]
10. Ibi, M. Inflammation and Temporomandibular Joint Derangement. *Biol. Pharm. Bull.* **2019**, *42*, 538–542. [CrossRef]
11. Lin, Y.-C.; Hsu, M.-L.; Yang, J.-S.; Liang, T.-H.; Chou, S.-L.; Lin, H.-Y. Temporomandibular Joint Disorders in Patients with Rheumatoid Arthritis. *J. Chin. Med. Assoc.* **2007**, *70*, 527–534. [CrossRef]
12. Boening, K.; Wieckiewicz, M.; Paradowska-Stolarz, A.; Wiland, P.; Shiau, Y.-Y. Temporomandibular Disorders and Oral Parafunctions: Mechanism, Diagnostics, and Therapy. *BioMed Res. Int.* **2015**, *2015*, 354759. [CrossRef]
13. Giannakopoulos, H.E.; Quinn, P.D.; Granquist, E.; Chou, J.C. Posttraumatic Temporomandibular Joint Disorders. *Craniomaxillofacial Trauma Reconstr.* **2009**, *2*, 91–101. [CrossRef] [PubMed]
14. Pihut, M.; Gala, A.; Obuchowicz, R.; Chmura, K. Influence of Ultrasound Examination on Diagnosis and Treatment of Temporomandibular Disorders. *J. Clin. Med.* **2022**, *11*, 1202. [CrossRef] [PubMed]
15. Alqhtani, N.R.; Alkhaldi, M.S.; Alanazi, A.F.; Alabdulsalam, A.S.; Alenazi, A.; Zaman, M.U.; Alzahrani, A.; Alshadwi, A.; Rafedah, A.A.; AlOtaibi, M. Temporomandibular Joint Space Dimensions among Saudi Patients with Temporomandibular Disorders: MRI-Based Retrospective Study. *Int. J. Clin. Pract.* **2022**, *2022*, 5846255. [CrossRef] [PubMed]

16. Zhu, J.; Gong, Y.; Zheng, F.; Yin, D.; Liu, Y. Relationships between Functional Temporomandibular Joint Space and Disc Morphology, Position, and Condylar Osseous Condition in Patients with Temporomandibular Disorder. *Clin. Oral Investig.* **2024**, *28*, 193. [CrossRef]
17. Martins, E.; Silva, J.; Pires, C.; Ponces-Ramalhao, M.; Lopes, J. Coronal Joint Spaces of the Temporomandibular Joint: Systematic Review and Meta-Analysis. *J. Clin. Exp. Dent.* **2015**, *7*, e435–e440. [CrossRef]
18. Dhabale, G.S.; Bhowate, R.R. Cone-Beam Computed Tomography for Temporomandibular Joint Imaging. *Cureus* **2022**, *14*, e31515. [CrossRef]
19. Veerappan, R.; Gopal, M. Comparison of the Diagnostic Accuracy of CBCT and Conventional CT in Detecting Degenerative Osseous Changes of the TMJ: A Systematic Review. *J. Indian Acad. Oral Med. Radiol.* **2015**, *27*, 81. [CrossRef]
20. Hilgenberg-Sydney, P.B.; Bonotto, D.V.; Stechman-Neto, J.; Zwir, L.F.; Pachêco-Pereira, C.; Canto, G.D.L.; Porporatti, A.L. Diagnostic Validity of CT to Assess Degenerative Temporomandibular Joint Disease: A Systematic Review. *Dentomaxillofacial Radiol.* **2018**, *47*, 20170389. [CrossRef] [PubMed]
21. Tomas, X.; Pomes, J.; Berenguer, J.; Quinto, L.; Nicolau, C.; Mercader, J.M.; Castro, V. MR Imaging of Temporomandibular Joint Dysfunction: A Pictorial Review. *RadioGraphics* **2006**, *26*, 765–781. [CrossRef] [PubMed]
22. Emshoff, R.; Jank, S.; Bertram, S.; Rudisch, A.; Bodner, G. Disk Displacement of the Temporomandibular Joint: Sonography Versus MR Imaging. *Am. J. Roentgenol.* **2002**, *178*, 1557–1562. [CrossRef]
23. Vogl, T.J.; Günther, D.; Weigl, P.; Scholtz, J.-E. Diagnostic Value of Dynamic Magnetic Resonance Imaging of Temporomandibular Joint Dysfunction. *Eur. J. Radiol. Open* **2021**, *8*, 100390. [CrossRef]
24. Kundu, H. Assessment of TMJ Disorders Using Ultrasonography as a Diagnostic Tool: A Review. *J. Clin. Diagn. Res.* **2013**, *7*, 3116. [CrossRef]
25. Liao, L.-J.; Lo, W.-C. High-Resolution Sonographic Measurement of Normal Temporomandibular Joint and Masseter Muscle. *J. Med. Ultrasound* **2012**, *20*, 96–100. [CrossRef]
26. Kumar, L.K.S.; Zachariah, G.P.; Chandran, S. Ultrasonography: A Step Forward in Temporomandibular Joint Imaging. A Preliminary Descriptive Study. *Clin. Pract.* **2019**, *9*, 1134. [CrossRef] [PubMed]
27. Obuchowicz, R.; Lasek, J.; Wodziński, M.; Piórkowski, A.; Strzelecki, M.; Nurzynska, K. Artificial Intelligence-Empowered Radiology—Current Status and Critical Review. *Diagnostics* **2025**, *15*, 282. [CrossRef] [PubMed]
28. Szczypiński, P.M.; Strzelecki, M.; Materka, A.; Klepaczko, A. MaZda—A Software Package for Image Texture Analysis. *Comput. Methods Programs Biomed.* **2009**, *94*, 66–76. [CrossRef] [PubMed]
29. Chrzanowski, L.; Drozdz, J.; Strzelecki, M.; Krzeminska-Pakula, M.; Jedrzejewski, K.S.; Kasprzak, J.D. Application of Neural Networks for the Analysis of Intravascular Ultrasound and Histological Aortic Wall Appearance—An In Vitro Tissue Characterization Study. *Ultrasound Med. Biol.* **2008**, *34*, 103–113. [CrossRef]
30. Kusunose, K. Radiomics in Echocardiography: Deep Learning and Echocardiographic Analysis. *Curr. Cardiol. Rep.* **2020**, *22*, 89. [CrossRef]
31. Obuchowicz, R.; Kruszyńska, J.; Strzelecki, M. Classifying Median Nerves in Carpal Tunnel Syndrome: Ultrasound Image Analysis. *Biocybern. Biomed. Eng.* **2021**, *41*, 335–351. [CrossRef]
32. Narang, A.; Bae, R.; Hong, H.; Thomas, Y.; Surette, S.; Cadieu, C.; Chaudhry, A.; Martin, R.P.; McCarthy, P.M.; Rubenson, D.S.; et al. Utility of a Deep-Learning Algorithm to Guide Novices to Acquire Echocardiograms for Limited Diagnostic Use. *JAMA Cardiol.* **2021**, *6*, 624. [CrossRef] [PubMed]
33. De Siqueira, V.S.; Borges, M.M.; Furtado, R.G.; Dourado, C.N.; Da Costa, R.M. Artificial Intelligence Applied to Support Medical Decisions for the Automatic Analysis of Echocardiogram Images: A Systematic Review. *Artif. Intell. Med.* **2021**, *120*, 102165. [CrossRef] [PubMed]
34. Vinayahalingam, S.; Berends, B.; Baan, F.; Moin, D.A.; Van Luijn, R.; Bergé, S.; Xi, T. Deep Learning for Automated Segmentation of the Temporomandibular Joint. *J. Dent.* **2023**, *132*, 104475. [CrossRef] [PubMed]
35. Kumar, K.V.; Baskaran, S. Tri-Stage Deep Learning Model for Mandibular Condyle Segmentation and Osteoarthritis Identification. In Proceedings of the 2023 International Conference on Network, Multimedia and Information Technology (NMITCON), Bengaluru, India, 1–2 September 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6.
36. Ito, S.; Mine, Y.; Yoshimi, Y.; Takeda, S.; Tanaka, A.; Onishi, A.; Peng, T.-Y.; Nakamoto, T.; Nagasaki, T.; Kakimoto, N.; et al. Automated Segmentation of Articular Disc of the Temporomandibular Joint on Magnetic Resonance Images Using Deep Learning. *Sci. Rep.* **2022**, *12*, 221. [CrossRef] [PubMed]
37. Choi, Y.-T.; Song, H.-J.; Lee, J.-S.; Im, Y.-G. Temporomandibular Joint Segmentation Using Deep Learning for Automated Three-Dimensional Reconstruction. *J. Oral Med. Pain* **2024**, *49*, 109–117. [CrossRef]
38. Li, M.; Punithakumar, K.; Major, P.W.; Le, L.H.; Nguyen, K.-C.T.; Pacheco-Pereira, C.; Kaipatur, N.R.; Nebbe, B.; Jaremko, J.L.; Almeida, F.T. Temporomandibular Joint Segmentation in MRI Images Using Deep Learning. *J. Dent.* **2022**, *127*, 104345. [CrossRef] [PubMed]

39. Yoon, K.; Kim, J.-Y.; Kim, S.-J.; Huh, J.-K.; Kim, J.-W.; Choi, J. Multi-Class Segmentation of Temporomandibular Joint Using Ensemble Deep Learning. *Sci. Rep.* **2024**, *14*, 18990. [CrossRef] [PubMed]
40. Choi, E.; Shin, S.; Lee, K.; An, T.; Lee, R.K.; Kim, S.; Son, Y.; Kim, S.T. Artificial Intelligence-Enhanced Diagnosis of Degenerative Joint Disease Using Temporomandibular Joint Panoramic Radiography and Joint Noise Data. *Sci. Rep.* **2025**, *15*, 1823. [CrossRef]
41. Kazimierczak, W.; Kędziora, K.; Janiszewska-Olszowska, J.; Kazimierczak, N.; Serafin, Z. Noise-Optimized CBCT Imaging of Temporomandibular Joints—The Impact of AI on Image Quality. *J. Clin. Med.* **2024**, *13*, 1502. [CrossRef]
42. Mourad, L.; Aboelsaad, N.; Talaat, W.M.; Fahmy, N.M.H.; Abdelrahman, H.H.; El-Mahallawy, Y. Automatic Detection of Temporomandibular Joint Osteoarthritis Radiographic Features Using Deep Learning Artificial Intelligence. A Diagnostic Accuracy Study. *J. Stomatol. Oral Maxillofac. Surg.* **2025**, *126*, 102124. [CrossRef] [PubMed]
43. Kerfoot, E.; Clough, J.; Oksuz, I.; Lee, J.; King, A.P.; Schnabel, J.A. Left-Ventricle Quantification Using Residual U-Net. In *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges*; Pop, M., Sermesant, M., Zhao, J., Li, S., McLeod, K., Young, A., Rhode, K., Mansi, T., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2019; Volume 11395, pp. 371–380, ISBN 978-3-030-12028-3.
44. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. *UNet++: A Nested U-Net Architecture for Medical Image Segmentation*; Springer International Publishing: Berlin/Heidelberg, Germany, 2018.
45. Myronenko, A. *3D MRI Brain Tumor Segmentation Using Autoencoder Regularization*; Springer International Publishing: Berlin/Heidelberg, Germany, 2018.
46. Milletari, F.; Navab, N.; Ahmadi, S.-A. *V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation*; IEEE: Piscataway, NJ, USA, 2016.
47. 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.
48. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
49. Liaw, R.; Liang, E.; Nishihara, R.; Moritz, P.; Gonzalez, J.E.; Stoica, I. Tune: A Research Platform for Distributed Model Selection and Training. *arXiv* **2018**, arXiv:1807.05118.
50. Biewald, L. Experiment Tracking with Weights and Biases 2020. [Online]. Available online: http://wandb.ai/ (accessed on 29 January 2025).
51. Cardoso, M.J.; Li, W.; Brown, R.; Ma, N.; Kerfoot, E.; Wang, Y.; Murrey, B.; Myronenko, A.; Zhao, C.; Yang, D.; et al. MONAI: An Open-Source Framework for Deep Learning in Healthcare. *arXiv* **2022**, arXiv:2211.02701.
52. Talmaceanu, D.; Lenghel, L.M.; Bolog, N.; Hedesiu, M.; Buduru, S.; Rotar, H.; Baciut, M.; Baciut, G. Imaging Modalities for Temporomandibular Joint Disorders: An Update. *Med. Pharm. Rep.* **2018**, *91*, 280–287. [CrossRef] [PubMed]
53. Manfredini, D.; Guarda-Nardini, L. Ultrasonography of the Temporomandibular Joint: A Literature Review. *Int. J. Oral Maxillofac. Surg.* **2009**, *38*, 1229–1236. [CrossRef] [PubMed]
54. Kirkhus, E.; Gunderson, R.B.; Smith, H.-J.; Flatø, B.; Hetlevik, S.O.; Larheim, T.A.; Arvidsson, L.Z. Temporomandibular Joint Involvement in Childhood Arthritis: Comparison of Ultrasonography-Assessed Capsular Width and MRI-Assessed Synovitis. *Dentomaxillofacial Radiol.* **2016**, *45*, 20160195. [CrossRef]
55. Lee, Y.-H.; Jeon, S.; Won, J.-H.; Auh, Q.-S.; Noh, Y.-K. Automatic Detection and Visualization of Temporomandibular Joint Effusion with Deep Neural Network. *Sci. Rep.* **2024**, *14*, 18865. [CrossRef]
56. Xu, L.; Chen, J.; Qiu, K.; Yang, F.; Wu, W. Artificial Intelligence for Detecting Temporomandibular Joint Osteoarthritis Using Radiographic Image Data: A Systematic Review and Meta-Analysis of Diagnostic Test Accuracy. *PLoS ONE* **2023**, *18*, e0288631. [CrossRef] [PubMed]
57. Almășan, O.; Leucuța, D.-C.; Hedeșiu, M.; Mureșanu, S.; Popa, Ș.L. Temporomandibular Joint Osteoarthritis Diagnosis Employing Artificial Intelligence: Systematic Review and Meta-Analysis. *J. Clin. Med.* **2023**, *12*, 942. [CrossRef] [PubMed]
58. Eşer, G.; Duman, Ş.B.; Bayrakdar, İ.Ş.; Çelik, Ö. Classification of Temporomandibular Joint Osteoarthritis on CONE BEAM Computed Tomography Images Using Artificial Intelligence System. *J. Oral Rehabil.* **2023**, *50*, 758–766. [CrossRef] [PubMed]
59. Farook, T.H.; Dudley, J. Automation and Deep (Machine) Learning in Temporomandibular Joint Disorder Radiomics: A Systematic Review. *J. Oral Rehabil.* **2023**, *50*, 501–521. [CrossRef] [PubMed]
60. Zou, W.; Mao, B.; Fadlullah, Z.M.; Qi, K. Assisting in Diagnosis of Temporomandibular Disorders: A Deep Learning Approach. *IEEE Access* **2022**, *10*, 124076–124082. [CrossRef]
61. Yıldız, N.T.; Kocaman, H.; Yıldırım, H.; Canlı, M. An Investigation of Machine Learning Algorithms for Prediction of Temporomandibular Disorders by Using Clinical Parameters. *Medicine* **2024**, *103*, e39912. [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.

tomography

MDPI

Article

Rosette Trajectory MRI Reconstruction with Vision Transformers

Muhammed Fikret Yalcinbas [1,*], Cengizhan Ozturk [1,2], Onur Ozyurt [3], Uzay E. Emir [4] and Ulas Bagci [5]

[1] Institute of Biomedical Engineering, Boğaziçi University, Istanbul 34684, Turkey; cozturk@bogazici.edu.tr
[2] Center for Targeted Therapy Technologies (CT3), Boğaziçi University, Istanbul 34984, Turkey
[3] Wolfson Brain Imaging Centre, Department of Clinical Neurosciences, University of Cambridge, Cambridge CB2-0QQ, UK; oo309@medschl.cam.ac.uk
[4] Department of Radiology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA; uzay_emir@med.unc.edu
[5] Machine and Hybrid Intelligence Lab, Northwestern University, Chicago, IL 60611, USA; ulas.bagci@northwestern.edu
* Correspondence: m.yalcinbas@std.bogazici.edu.tr

Abstract: Introduction: An efficient pipeline for rosette trajectory magnetic resonance imaging reconstruction is proposed, combining the inverse Fourier transform with a vision transformer (ViT) network enhanced with a convolutional layer. This method addresses the challenges of reconstructing high-quality images from non-Cartesian data by leveraging the ViT's ability to handle complex spatial dependencies without extensive preprocessing. Materials and Methods: The inverse fast Fourier transform provides a robust initial approximation, which is refined by the ViT network to produce high-fidelity images. Results and Discussion: This approach outperforms established deep learning techniques for normalized root mean squared error, peak signal-to-noise ratio, and entropy-based image quality scores; offers better runtime performance; and remains competitive with respect to other metrics.

Keywords: MRI; machine learning; medical imaging

1. Introduction

1.1. MRI, Cartesian and Non-Cartesian

Magnetic resonance imaging (MRI) is a crucial diagnostic tool in modern medicine, capable of providing detailed images of anatomical structures [1]. MRI data are collected in k-space, where signals are recorded based on their spatial frequency and phase information [2]. These k-space data are then transformed into the spatial domain through the Fourier transform [3]. The resulting images provide high-resolution, contrast-rich views of anatomical structures, aiding in accurate diagnosis and treatment planning. The Cartesian trajectory of k-space corresponds to a sampling scheme in which data are collected along regular grid lines. While such fully sampled images are ideal, some degree of undersampling is required to reduce scan times and minimize patient discomfort [4]. Non-Cartesian trajectories, such as radial or spiral paths, yield improved image quality compared to Cartesian undersampling, allowing faster MRI acquisition via efficient k-space coverage [5].

1.2. Cartesian K-Space MRI Reconstruction Methods

Traditional Cartesian reconstruction methods involve the direct application of the inverse fast Fourier transform (IFFT) to uniformly sampled k-space data. Despite its widespread use, standard FFT-based reconstruction is prone to artifacts from motion and undersampling. Compressed sensing has emerged as a key alternative, exploiting the sparsity of undersampled data to reconstruct images by solving an optimization problem that enforces

sparsity constraints, leading to high-quality reconstructions with reduced scan times [6]. Compressed sensing is often used in conjunction with parallel imaging, the combination being abbreviated as PICS. Parallel imaging techniques, such as sensitivity encoding (SENSE) and generalized autocalibrating partially parallel acquisitions (GRAPPA), leverage the spatial sensitivity profiles of multiple receiver coils to accelerate MRI acquisitions [7].

Deep learning methods have had success in reconstructing undersampled Cartesian MRI [8–11], improving upon classical techniques at high accelerations by allowing greater complexity in the employed model [12] and extracting complex mappings from training data [13]. Architectures such as convolutional neural networks (CNNs) and generative adversarial networks have been employed to restore missing or corrupted k-space data, leveraging large datasets to enhance reconstruction speed and accuracy [14,15].

1.3. Non-Cartesian K-Space MRI Reconstruction Methods

Reconstructing high-quality images from non-Cartesian data is challenging, as traditional Fourier-based methods struggle to handle the irregularity and gaps in the sampling pattern [16]. Standard non-Cartesian reconstruction methods typically involve arranging the data onto a Cartesian grid before applying the FFT, a process known as gridding [17]. The sparsity of non-Cartesian data also makes it a suitable candidate for compressed sensing, non-Cartesian GRAPPA, and SENSE [5]. Methods such as compressed sensing overcome the limitations of Fourier techniques by balancing fidelity to acquired data with the assumption of sparsity in the image domain [7].

Deep neural networks gained traction in this problem [18] due to their ability to handle non-Cartesian data without the need for regridding. Hammernik et al. [14] employed VarNet for the reconstruction of complex multi-channel MR data, leveraging its ability to seamlessly combine variational inference techniques with deep learning to achieve impressive performance in reconstruction tasks [19]. Aggarwal et al. introduced MoDL [20], an integration of model-based and deep learning approaches. The MoDL architecture applies iterative unrolling optimization algorithms within a deep learning framework that enables accurate and efficient reconstructions. Vision transformers have also demonstrated state-of-the-art performance in non-Cartesian MRI reconstruction [21]. Unlike traditional CNNs, which rely on convolutional kernels to process spatial information, ViTs decompose the input data into a sequence of patches and apply transformer models to learn complex representations [22]. Table 1 summarizes the advantages and disadvantages of the reconstruction techniques discussed. It should be noted that non-Cartesian reconstruction with deep neural networks can require specialized algorithms and additional preprocessing [23]. These limitations, combined with the need for domain-specific expertise to implement and optimize non-Cartesian deep learning techniques, present challenges to their practical application. This is particularly relevant for accelerated acquisition scenarios, which are essential for further reducing scan times while maintaining high image quality.

Table 1. Comparison of MRI reconstruction techniques for rosette trajectory imaging.

Technique	Advantages	Disadvantages
IFFT [3]	– Robust initial approximation from k-space to image domain. – Computationally efficient for uniformly sampled Cartesian data.	– Struggles with non-Cartesian trajectories. – Susceptible to artifacts due to irregular and undersampled data.
CS [4,6]	– Balances fidelity to acquired data with image sparsity. – Effective for undersampled data, reducing scan times. – Works well with Cartesian and non-Cartesian data.	– Requires complex optimization, leading to higher computational costs. – Sensitive to parameter tuning and model assumptions.

Table 1. *Cont.*

Technique	Advantages	Disadvantages
VarNet [14,19], MoDL [20]	– Learns complex mappings from data, improving reconstruction at high acceleration rates. – Handles both Cartesian and non-Cartesian data without regridding.	– Requires large datasets and computational resources for training. – Long inference times
ViT [21,24]	– Models long-range dependencies, capturing complex spatial patterns. – No need for extensive pre-processing of non-Cartesian data.	– Requires augmented training data, increasing training time and resource consumption.

1.4. The Rosette Trajectory

The rosette trajectory, a non-Cartesian trajectory that traces a petal-like path through k-space, provides particularly effective k-space coverage [25]. The efficient sampling pattern of the rosette trajectory can yield a higher signal-to-noise ratio (SNR) [26] and shorter scan times [27] compared to spiral or radial trajectories for MRI acquisition under high acceleration factors. Standard reconstruction techniques for rosette k-space involve adapting traditional gridding methods to handle the unique structure of the trajectory, including the development of dedicated density compensation functions (DCF) to account for the varying sampling density [28].

Rosette has been shown to work well alongside CS and high acceleration factors [25]. Mahmud et al. achieve good performance with an acceleration factor of 6 in 7 Tesla CS rosette spinal cord imaging [29]. Li et al. use rosette and CS to achieve 10 percent better results than radial and spiral trajectories at an acceleration factor of 10 [26]. Alcicek et al. and Bozymski et al. demonstrate the feasibility of using a novel rosette trajectory with CS to achieve patient-friendly and high-resolution magnetic resonance spectroscopy imaging [30,31].

While the rosette trajectory is a promising candidate for high-acceleration reconstruction, training on raw MRI data is computationally expensive. In order to achieve flexible high-accuracy reconstruction for accelerated rosette imaging with good runtime performance, we utilize the direct Fourier transform to approximate the final result and the ViT network to transform this approximation into the expected image. The addition of a convolutional layer improves the reconstructed image by removing artifacts that arise from the ViT network. This approach enhances the reconstruction of rosette trajectory MRI, making it not only faster but also more adaptable across different imaging scenarios. As a benchmark, we apply VarNet [14] and MoDL [20].

2. Materials and Methods

2.1. Method Overview

The method proposed in this work utilizes a strategy that combines traditional and modern deep learning techniques. The pipeline, shown in Figure 1, operates in two key phases: an initial approximation phase and a refinement phase. During the initial approximation phase, the IFFT is applied to multi-coil rosette-sampled k-space data, after which the resulting coil images are combined through root square sum to produce a rough image. This step, though computationally simple, introduces imprecision and noise due to the undersampled and non-uniform nature of the k-space trajectory. In the refinement phase, the rough image is passed to the ViT network, which has been trained with our data to predict the PICS reconstruction.

2.2. Vision Transformer

The network architecture of the employed ViT is based on the accelerated MRI reconstruction work by Lin and Heckel [21], adapted for approximating the compressed sensing algorithm for MRI reconstruction. Their model is based on the ViT, originally proposed by Dosovitskiy et al. [32] for image classification tasks. This involves processing the input image as a sequence of patches in order to apply the transformer encoder, which was designed for sequential data. A trainable linear transformation maps each patch to a d-dimensional feature vector known as a patch embedding. To compensate for the lack of positional information, learnable position embeddings are used to encode the absolute position. Finally, a classification token is added to the beginning of the sequence. The encoder consists of N encoder layers, each containing a multi-head self-attention (MHSA) block and a multilayer perceptron block that transforms each feature vector independently. The Swin ViT block, used in the Swin architecture, replaces the MHSA block with a window-based self-attention (WBSA) block that parses the region of attention in sections [24]. In both cases, layer normalization is applied before each block, and residual connections are added after each block for stable training. The output representation of the classification token is used for the final classification of the input image. The ViT architecture by Lin and Heckel adapts this system for image reconstruction by removing the classification token and replacing the classification head with a reconstruction head tailored for mapping the transformer output back to a visual image [21]. Unlike previous approaches that combine transformers with convolutions [33], this architecture uses only the standard transformer encoder.

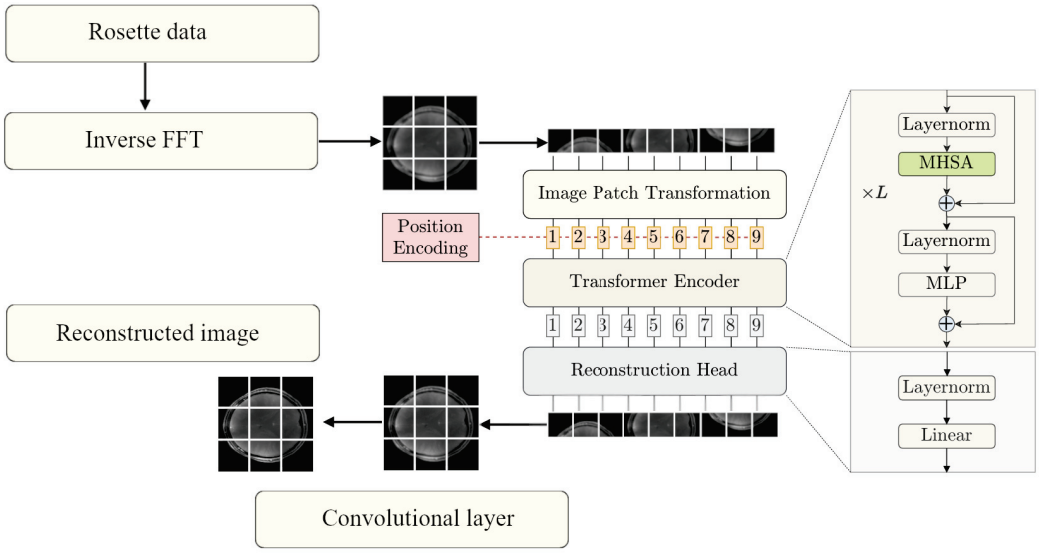
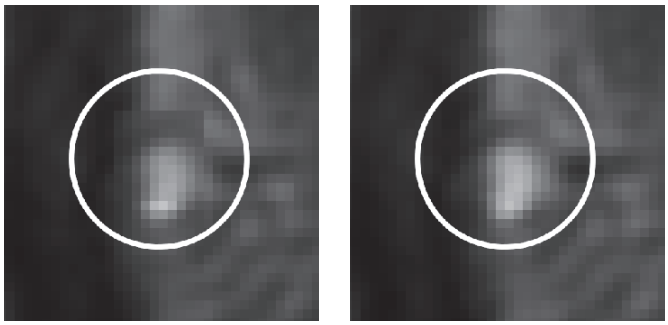


Figure 1. Pipeline of IFFT to ViT reconstruction (ViT stage adapted from Lin and Heckel [21]).

In our experiments, the MHSA ViT model produced artifacts in the form of miscolored patches with defined boundaries at regions of high contrast, irrespective of training augmentation. The inclusion of the convolutional layer corrects this issue. Figure 2 shows the improvement granted by the addition of the layer.



(**a**) ViT (**b**) ViT + convolution

Figure 2. Effect of convolution in removing patch artifacts. Subfigure (**a**) shows ViT reconstruction without convolution, (**b**) shows reconstruction with the convolutional layer added.

2.3. Dataset and Preprocessing

The experimental dataset comprises 6 sets of 500 T1-weighted time series axial brain MRI scans acquired using the rosette k-space trajectory. Imaging was performed on a Siemens MAGNETOM Terra 7 T system with a Nova Medical 8Tx/32Rx head coil. The acquisition details [34] are as follows:

- Repetition time (TR): 2.4 s;
- Echo time (TE) (dual): 1 and 9 milliseconds;
- Acceleration factor: 4;
- Total petals: 189;
- K_{max}: 1000/m;
- w_1: 400 Hz;
- w_2: 400 Hz;
- Nominal in-plane resolution: 0.468 mm;
- Slice thickness: 2 mm;
- Flip angle: 7 degrees;
- Image resolution: 512 × 512.

We employed both augmented and non-augmented datasets for ViT training and evaluation. Reference images for comparison were generated using the PICS algorithm implemented through the Berkeley Advanced Reconstruction Toolbox (BART) [35] (https://mrirecon.github.io/bart) (accessed on 20 June 2024). The k-space data were converted to an initial approximation of the final image using the IFFT. These intermediate images are then used to train the network to minimize the difference between such inputs and the CS reference.

Four training sets were used for training, and one set was used for validation, with 500 scans in each set. For singly augmented ViT, this results in 4000 inputs, and for triple augmentation, this number becomes 8000.

The augmentations used for ViT training data include:

- Random horizontal flip, probability = 0.5;
- Random vertical flip, probability = 0.5;
- Random rotation, 0 to 180 degrees;
- Color jitter, brightness/contrast/saturation, range = 0.8 to 1.2;
- Random resized crop, scale = 0.3 to 1.1.

By simulating various imaging conditions during training, the network learns to extract meaningful features while remaining flexible to variations in patient anatomies, imaging conditions, and acquisition scenarios.

2.4. Evaluation Methods

Reconstruction performance was assessed using several image quality metrics. Table 2 shows the respective formulae. The metrics are applied to six randomly generated 50×50 pixel patches (biased towards the center to avoid empty space) and averaged to generate the final metric score.

- The structural similarity index measure (SSIM) measures image similarity between a reference image and a processed image [36]. Higher scores are preferred.
- Normalized root mean square error (NRMSE) in the context of image quality is the square root of the mean squared error [37] between two images normalized by the sum of the observed values. Lower error is preferred.
- Normalized mutual information (NMI) measures shared information, where the scale between no mutual information and full correlation is given as 0 to 1 [38].
- Relative contrast is the ratio between the difference in maximum and minimum intensity and the sum of the same values.
- Peak signal-to-noise ratio (PSNR) measures the ratio between the maximum possible pixel value and the noise power [39]. Higher PSNR values indicate better image quality.
- Shannon entropy quantifies the information content of an image using a measure of uncertainty [40].
- The entropy focus criterion (EFC) provides an estimate of corruption and blurring in terms of energy—lower values are preferred [41].

SSIM, NRMSE, and NMI are used to investigate the fidelity of the reconstructed image in terms of structure, intensity, and shared information. Relative contrast and PSNR were chosen to measure the individual quality of the resulting image by quantifying signal strength and intensity range. Shannon entropy was used to assess the addition or loss of information, neither of which is desirable. The EFC criterion is useful as an indicator of clarity for an individual image.

Table 2. Formulae for image quality metrics, where $\mathbf{X}$ is the reconstructed region of interest and $\mathbf{Y}$ is the corresponding reference.

Metric	Formula
SSIM	$$\mathrm{SSIM}(\mathbf{X},\mathbf{Y}) = \frac{(2\mu_\mathbf{X}\mu_\mathbf{Y} + c_1)(2\sigma_{\mathbf{XY}} + c_2)}{(\mu_\mathbf{X}^2 + \mu_\mathbf{Y}^2 + c_1)(\sigma_\mathbf{X}^2 + \sigma_\mathbf{Y}^2 + c_2)}$$ where $\mu_\mathbf{X}$ and $\mu_\mathbf{Y}$ are the means of $\mathbf{X}$ and $\mathbf{Y}$, $\sigma_\mathbf{X}^2$ and $\sigma_\mathbf{Y}^2$ are the variances, $\sigma_{\mathbf{XY}}$ is the covariance, and c_1 and c_2 are constants to stabilize the division.
NRMSE	$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2}}{\max(\mathbf{X}) - \min(\mathbf{Y})}$$ where x_i and y_i are the pixel values of the reference and processed images, respectively, and N is the total number of pixels.
NMI	$$\mathrm{NMI} = \frac{2I(\mathbf{X},\mathbf{Y})}{H(\mathbf{X}) + H(\mathbf{Y})}$$ where $I(\mathbf{X},\mathbf{Y})$ is the mutual information between regions $\mathbf{X}$ and $\mathbf{Y}$, and $H(\mathbf{X})$ and $H(\mathbf{Y})$ are the entropies of images $\mathbf{X}$ and $\mathbf{Y}$, respectively.
Relative Contrast	$$\text{Relative Contrast} = \frac{\max(\mathbf{X}) - \min(\mathbf{X})}{\max(\mathbf{X}) + \min(\mathbf{X})}$$ where $\max(\mathbf{X})$ and $\min(\mathbf{X})$ are the maximum and minimum pixel intensities of the region.

Table 2. *Cont.*

Metric	Formula
PSNR	$$\mathrm{PSNR}(\mathbf{X},\mathbf{Y}) = 10\log_{10}\left(\frac{L^2}{\mathrm{MSE}(\mathbf{X},\mathbf{Y})}\right)$$ where L is the maximum pixel value and $\mathrm{MSE}(\mathbf{X},\mathbf{Y})$ is the mean squared error between $\mathbf{X}$ and $\mathbf{Y}$.
Shannon Entropy	$$H(\mathbf{X}) = -\sum_i p_i \log_2 p_i$$ where p_i represents the probability of intensity level i occurring in the image $\mathbf{X}$.
EFC	$$\mathrm{E} = -\sum_{j=1}^{N} \frac{x_j}{x_{\max}}, x_{\max} = \sqrt{\sum_{j=1}^{N} x_j^2}$$ where x_i represents pixel intensities in the image.

2.5. Visualization

The reconstructed images were bias field corrected with the open source Advanced Normalization Tools toolbox (https://github.com/ANTsX (accessed on 12 November 2024)).

2.6. Training Procedure

The ViT training was completed using the PyTorch (2.5.1) library (created at Facebook, now Meta AI) on an Nvidia A6000 GPU. For gradient descent, the Adam optimizer [42] was used with a maximum learning rate of 0.0003. The "1cycle" learning rate scheduler [43] was used alongside Adam to adjust the learning rate. A patch size of 10×10 pixels was used, defining the size of the square segments into which the input image is divided. The depth of the sequential layers of self-attention and feedforward networks was set as 10, while the number of attention heads were set at 16. The embedding dimension was set at 80, partly due to memory concerns. The convolutional layer is set at a kernel size of 3 and stride of 1 to preserve the output image dimension. The small kernel size preserves the quality of the image while correcting boundary artifacts generated by the ViT. The networks converged in 15 to 20 epochs, at which point the training was stopped. Ground truth is defined as the PICS reconstruction. The benchmark VarNet and MoDL networks are used with a setup informed by Blumenthal et al. [35], with batch size set at 20.

3. Results

3.1. Image Scores

Table 3 shows various image quality metric results for the ViT pipeline against VarNet and ModL. The highest SSIM score is achieved by MoDL, the best NRMSE and PSNR scores are achieved by the MHSA ViT architecture. For NMI, the MoDL score is the highest. The high Shannon entropy of all models, especially VarNet, imply the addition of redundant or distorting information. The lower entropy levels of the ViT reconstructions are preferable. Similarly, ViT achieves the best scores for EFC. Relative contrast is higher than the reference for all methods, suggesting a shift in intensity.

ANOVA shows statistically significant differences between the assessed methods for all metrics except relative contrast. Table 4 shows the Bonferroni-corrected paired *t*-test results for significance. The advantages of the different architectures are all assessed as significant, while the score gaps between the MHSA and WBSA block pipelines are

significant for NRMSE, PSNR, and NMI. Augmentation in MHSA ViT is shown to be a significant improvement for all metrics except Shannon entropy.

The quality scores demonstrate the strengths of the ViT network across multiple metrics. The triply augmented ViT achieves a respectable SSIM score while outperforming both VarNet and MoDL in terms of NRMSE and PSNR. Notably, the ViT results have better NRMSE and PSNR scores even with little or no augmentation to the training data. These result suggest that the ViT excels at preserving pixel intensity fidelity and reducing reconstruction noise. However, the lower NMI score indicates room for improvement in capturing fine inter-pixel relationships, particularly in regions of high structural complexity. Augmentation improves performance, with higher PSNR values demonstrating reduced noise and enhanced clarity. The lower Shannon entropy score for ViT suggests that a lesser amount of redundant information has been added during reconstruction. Likewise, the EFC scores indicate superior clarity for the ViT pipeline. Figure 3 illustrates sample triply augmented MHSA ViT results from the network. Figure 4 shows the same image reconstructed by VarNet, MoDL, and ViT pipeline; the final subfigure shows the amplified difference between the PICS and ViT reconstructions.

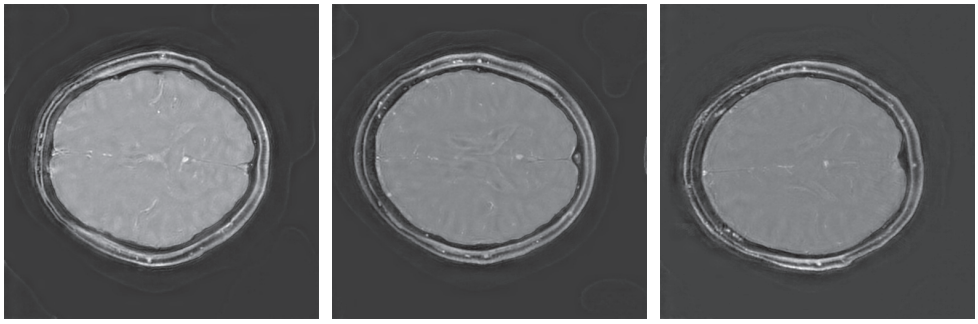
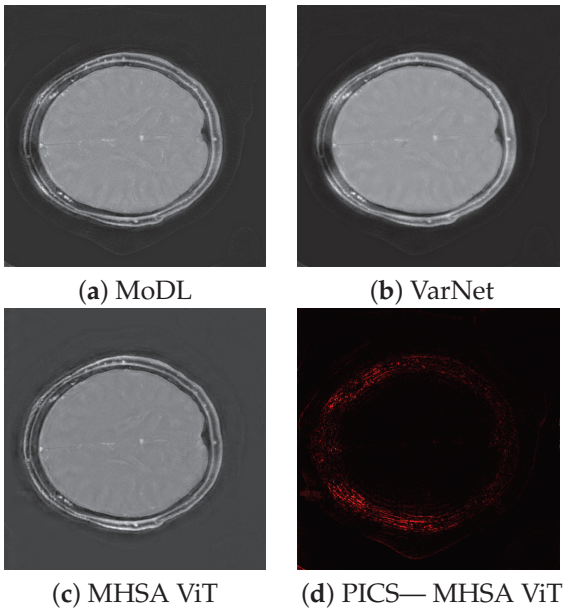


Figure 3. Triply augmented MHSA ViT reconstructions.



(**a**) MoDL (**b**) VarNet

(**c**) MHSA ViT (**d**) PICS— MHSA ViT

Figure 4. Comparison of reconstruction methods, showing results for MoDL, VarNet, MHSA ViT, and difference between PICS reference and MHSA ViT.

Table 3. Image quality scores for reconstruction methods. Arrows indicate whether higher or lower values are preferred.

Method	SSIM ↑	NRMSE ↓	PSNR ↑	NMI ↑	R. Contrast	Shannon	EFC ↓
Reference	-	-	-	-	0.332	3.840	2.960
VarNet	0.944	0.322	22.740	0.598	0.430	5.003	4.023
MoDL	**0.987**	0.060	37.248	**0.616**	0.472	4.861	3.429
Vision T.							
Non-aug. MHSA	0.974	0.048	40.134	0.501	0.441	4.697	3.244
Aug. MHSA (X1)	0.975	0.040	42.124	0.510	0.445	4.672	3.280
Aug. MHSA (X3)	0.980	**0.033**	**43.799**	0.536	0.445	**4.631**	**3.245**
Aug. WBSA (X3)	0.980	0.037	42.685	0.544	0.439	4.663	3.285

Table 4. Statistical significance (p-values) of pairwise comparisons between reconstruction methods. Significant differences are based on the Bonferroni-corrected threshold of $\alpha = 0.0083$.

Metric	MHSA(X3) vs. MoDL	MHSA(X3) vs. VarNet	WBSA(X3) vs. MoDL	WBSA(X3) vs. VarNet	MHSA(X3) vs. WBSA(X3)	MHSA X3 vs. X1
SSIM	$p < 0.0083$	$p < 0.0083$	$p < 0.0083$	$p < 0.0083$	$\mathbf{p > 0.0083}$	$p < 0.0083$
NRMSE	$p < 0.0083$	$p < 0.0083$	$p < 0.0083$	$p < 0.0083$	$p < 0.0083$	$p < 0.0083$
PSNR	$p < 0.0083$	$p < 0.0083$	$p < 0.0083$	$p < 0.0083$	$p < 0.0083$	$p < 0.0083$
NMI	$p < 0.0083$	$p < 0.0083$	$p < 0.0083$	$p < 0.0083$	$p < 0.0083$	$p < 0.0083$
Shannon	$p < 0.0083$	$p < 0.0083$	$p < 0.0083$	$p < 0.0083$	$\mathbf{p > 0.0083}$	$\mathbf{p > 0.0083}$
EFC	$p < 0.0083$	$p < 0.0083$	$p < 0.0083$	$p < 0.0083$	$\mathbf{p > 0.0083}$	$p < 0.0083$

3.2. Network Runtime Performance

Table 5 highlights runtime and resources used for the models. The ViT pipelines demonstrate a marked advantage in processing speed, requiring approximately 60 s per 10 images, compared to VarNet and MoDL, which take approximately 7 min each. This efficiency arises from the network's strategy of refining an already approximate reconstruction rather than computing directly from raw k-space data. We extrapolate from acquired data, assuming approximate linearity due to slice by slice processing, to provide estimates for processing a full 3D dataset comprised of 128 and 512 slices.

The GPU memory footprint of the ViT pipeline lies between VarNet and MoDL. While its resource requirements (4895 MB) exceed those of VarNet, it remains significantly lower than MoDL's usage. Due to programming optimizations, the maximum total memory requirements remain the same regardless of slice count.

Table 5. Reconstruction performance metrics.

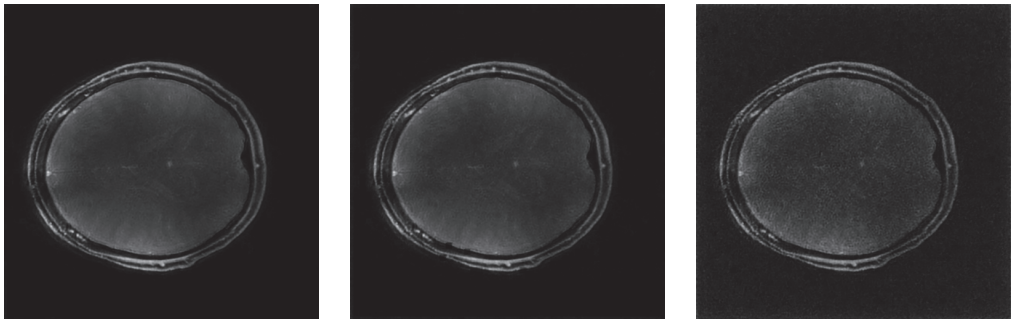
Network	Total CPU Time (Hours:Minutes:Seconds)	Max GPU Memory Used (MB)
VarNet (10 slices)	00:06:58	2785
VarNet (20 slices)	00:13:51	
VarNet (estimate for 128 slices)	01:28:00	
VarNet (estimate for 512 slices)	05:54:00	
MoDL (10 slices)	00:07:05	6369
MoDL (20 slices)	00:14:08	
MoDL (estimate for 128 slices)	01:30:00	
MoDL (estimate for 512 slices)	06:00:00	
MHSA ViT (10 slices)	00:00:45	4895
MHSA ViT (20 slices)	00:01:25	
MHSA ViT (estimate for 128 slices)	00:09:00	
MHSA ViT (estimate for 512 slices)	00:36:00	
WBSA ViT (10 slices)	00:01:09	4895
WBSA ViT (20 slices)	00:01:26	
WBSA ViT (estimate for 128 slices)	00:10:00	
WBSA ViT (estimate for 512 slices)	00:37:00	

3.3. Noise Independence

Figure 5 shows the performance of the proposed ViT networks in the presence of k-space noise added with the BART toolbox. Gaussian variance levels were selected to provide minimal, mild, and moderate levels of noise with reference to the magnitude of the k-space data. The image quality metrics, listed in Table 6, show that while a large decay is observed for high noise conditions, the models tolerate a reasonable level of noise without excessive loss. Mild patch artifacts that occur at a variance of 1×10^{-11} indicate that the model is only robust up to a moderate level of corruption.

Table 6. Image quality scores for different noise levels and architectures. Arrows indicate whether higher or lower values are preferred.

ViT	Gaussian Variance	SSIM ↑	NRMSE ↓	PSNR ↑	NMI ↑
MHSA	5×10^{-11}	0.732	0.151	29.377	0.199
MHSA	1×10^{-11}	0.917	0.090	34.078	0.308
MHSA	5×10^{-12}	0.944	0.063	37.389	0.360
MHSA	1×10^{-12}	0.970	0.041	41.582	0.460
MHSA	No noise	0.980	0.033	43.799	0.536
WBSA	5×10^{-11}	0.745	0.118	31.702	0.200
WBSA	1×10^{-11}	0.822	0.203	27.471	0.259
WBSA	5×10^{-12}	0.955	0.053	39.139	0.385
WBSA	1×10^{-12}	0.974	0.041	41.509	0.484
WBSA	No noise	0.980	0.037	42.685	0.544



(**a**) (**b**) (**c**)

Figure 5. MHSA ViT reconstructions for noisy inputs. Values indicate Gaussian variance (**a**) Minimal noise (5×10^{-12}), (**b**) Mild noise (1×10^{-11}), (**c**) Moderate noise (5×10^{-11}).

4. Discussion

The results of this study illustrate the potential of using a vision transformer pipeline for the reconstruction of rosette trajectory MRI data. The NRMSE and PSNR scores achieved against methods such as VarNet and MoDL suggest that ViT is effective in preserving image fidelity (NRMSE, PSNR) and reducing noise (PSNR). While interpreting entropy metrics is non-trivial, the Shannon entropy and EFC scores of ViT are preferable with respect to the principle of maximizing reconstruction fidelity. The competitive scores achieved with minimal or no augmentation highlight the ViT's ability to generalize well from approximations and small datasets, which is crucial in medical imaging where large, well-annotated datasets are not always available. The SSIM and NMI scores indicate that while the method excels in certain metrics, there may be room for improvement in capturing fine structural details or maintaining overall structural integrity. The slight edge of the MSHA block over WBSA in this respect suggest that a global attention scope is preferable at this level of feature detection.

In addition to VarNet and MoDL, alternative AI-based methods such as RAKI and automap have shown success in MR image reconstruction, though they often require

complex preprocessing steps to learn the inverse Fourier transformation [44]. Generative adversarial networks have likewise achieved impressive results with high quality and artifact reduction but are known to be unstable during training [45]. Vision transformers offer a balanced intermediate, with a predictable if moderately intensive training process and straightforward preprocessing, especially if combined with IFFT. This renders the model flexible in terms of development and deployment.

The noise independence experiments suggest that the model has a baseline resilience against noise, but the artifacts that arise at moderate Gaussian variance suggest that fine tuning or augmentation with noisy data could be advisable to improve performance under such conditions. The increased runtime efficiency of the framework suggests its suitability for real-time clinical applications, such as dynamic or 4D MRI, where processing speed is critical. However, structural fidelity remains a concern, especially for clinical contexts with subtle pathological features. The strong performance of the method at a resolution of 512 by 512 pixels suggests that features can be increased at lower resolutions to improve performance. Finally, although augmentation is possible, acquiring varied and high-quality primary datasets is often challenging for more complex clinical applications such as multimodal imaging. Performance under fine-tuning conditions is therefore an important avenue of investigation for future applications.

Investigating advanced augmentation techniques can improve the model's generalization capabilities. The augmentation methods used, although simple (flipping, rotating, and color jitter), appear to have bolstered the model's ability to generalize across different types of k-space data, as seen from the improved PSNR and NRMSE scores. Future work could investigate the impact of more complex augmentations, such as elastic transformations or non-uniform intensity scaling, which might further increase the effectiveness of the model across clinical scenarios.

While the results demonstrate strong performance on axial brain images, the generalizability of these findings to other MRI protocols (such as differing field strengths, acquisition times, or anatomical features) remains to be explored. Additionally, evaluating the ViT's performance across different k-space trajectories, such as radial or spiral acquisitions, would further validate its effectiveness across different MRI applications. Incorporating more sophisticated initial approximations or iterative refinement steps within the ViT framework may also yield improvements. Finally, extending the evaluation of these methods to different types of MRI data, including dynamic and functional MRI, and assessing their performance in real-time applications will be crucial for translating these advancements into practical clinical tools. This approach could be particularly advantageous in exploring 3D or 4D MRI applications, as the transformer self-attention capability can naturally extend to higher dimensions.

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

Funding: Ulas Bagci acknowledges the following NIH grants: R01-HL171376, U01-CA268808, and R01-CA240639.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of the University of Cambridge, PRE_2017_057.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The MRI data used in this study are available at https://drive.google.com/drive/folders/1lTaz1xMRGX6jEKLzU6-r6ifaOo7UVTwI?usp=sharing (accessed on 10 March 2025).

Acknowledgments: This work was supported by TÜBİTAK (Scientific and Technological Research Council of Türkiye) and the TÜBİTAK 2244-118C129 fellowship program sponsored by Onur Ozyurt and the Telemed Company. The authors would like to acknowledge valuable support from the Wellcome Trust Collaborative Award (223131/Z/21/Z) awarded to Uzay Emir and Onur Ozyurt, computational resources generously allowed by the Ulas Bagci AI Group Machine and Hybrid Intelligence Lab and the Quest High-Performance Computing Cluster at Northwestern University, the Boğaziçi University Department of Biomedical Engineering, the Boğaziçi University Life Sciences and Technologies Application and Research Center, and the Center for Targeted Therapy Technologies at Boğaziçi University for research infrastructure support.

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

References

1. Geethanath, S.; Vaughan, J.T., Jr. Accessible magnetic resonance imaging: A review. *J. Magn. Reson. Imaging* **2019**, *49*, e65–e77. [CrossRef] [PubMed]
2. Moratal, D.; Vallés-Luch, A.; Martí-Bonmatí, L.; Brummer, M.E. k-Space tutorial: An MRI educational tool for a better understanding of k-space. *Biomed. Imaging Interv. J.* **2008**, *4*, e15. [CrossRef] [PubMed]
3. Gallagher, T.A.; Nemeth, A.J.; Hacein-Bey, L. An introduction to the Fourier transform: Relationship to MRI. *Am. J. Roentgenol.* **2008**, *190*, 1396–1405. [CrossRef] [PubMed]
4. Hollingsworth, K.G. Reducing acquisition time in clinical MRI by data undersampling and compressed sensing reconstruction. *Phys. Med. Biol.* **2015**, *60*, R297. [CrossRef]
5. Wright, K.L.; Hamilton, J.I.; Griswold, M.A.; Gulani, V.; Seiberlich, N. Non-Cartesian parallel imaging reconstruction. *J. Magn. Reson. Imaging* **2014**, *40*, 1022–1040. [CrossRef]
6. Geethanath, S.; Reddy, R.; Konar, A.S.; Imam, S.; Sundaresan, R.; Ramesh Babu, D.R.; Venkatesan, R. Compressed Sensing MRI: A Review. *Crit. Rev. Biomed. Eng.* **2013**, *41*, 183–204. [CrossRef]
7. Ye, J.C. Compressed sensing MRI: A review from signal processing perspective. *BMC Biomed. Eng.* **2019**, *1*, 8. [CrossRef]
8. Pal, A.; Rathi, Y. A review and experimental evaluation of deep learning methods for MRI reconstruction. *J. Mach. Learn. Biomed. Imaging* **2022**, *1*, 1. [CrossRef]
9. Zhang, H.M.; Dong, B. A review on deep learning in medical image reconstruction. *J. Oper. Res. Soc. China* **2020**, *8*, 311–340. [CrossRef]
10. Chen, Y.; Schönlieb, C.B.; Liò, P.; Leiner, T.; Dragotti, P.L.; Wang, G.; Rueckert, D.; Firmin, D.; Yang, G. AI-based reconstruction for fast MRI—A systematic review and meta-analysis. *Proc. IEEE* **2022**, *110*, 224–245. [CrossRef]
11. Liang, D.; Cheng, J.; Ke, Z.; Ying, L. Deep magnetic resonance image reconstruction: Inverse problems meet neural networks. *IEEE Signal Process. Mag.* **2020**, *37*, 141–151. [PubMed]
12. Chandra, S.S.; Bran Lorenzana, M.; Liu, X.; Liu, S.; Bollmann, S.; Crozier, S. Deep learning in magnetic resonance image reconstruction. *J. Med. Imaging Radiat. Oncol.* **2021**, *65*, 564–577. [PubMed]
13. Montalt-Tordera, J.; Muthurangu, V.; Hauptmann, A.; Steeden, J.A. Machine learning in magnetic resonance imaging: Image reconstruction. *Phys. Medica* **2021**, *83*, 79–87. [CrossRef] [PubMed]
14. Hammernik, K.; Klatzer, T.; Kobler, E.; Recht, M.P.; Sodickson, D.K.; Pock, T.; Knoll, F. Learning a variational network for reconstruction of accelerated MRI data. *Magn. Reson. Med.* **2018**, *79*, 3055–3071. [CrossRef]
15. Lv, J.; Zhu, J.; Yang, G. Which GAN? A comparative study of generative adversarial network-based fast MRI reconstruction. *Philos. Trans. R. Soc.* **2021**, *379*, 20200203.
16. Zhou, B.; Schlemper, J.; Dey, N.; Salehi, S.S.M.; Sheth, K.; Liu, C.; Duncan, J.S.; Sofka, M. Dual-domain self-supervised learning for accelerated non-Cartesian MRI reconstruction. *Med. Image Anal.* **2022**, *81*, 102538.
17. Fessler, J.A. On NUFFT-based gridding for non-Cartesian MRI. *J. Magn. Reson.* **2007**, *188*, 191–195.
18. Wang, S.; Xiao, T.; Liu, Q.; Zheng, H. Deep learning for fast MR imaging: A review for learning reconstruction from incomplete k-space data. *Biomed. Signal Process. Control.* **2021**, *68*, 102579. [CrossRef]
19. Sriram, A.; Zbontar, J.; Murrell, T.; Defazio, A.; Zitnick, C.L.; Yakubova, N.; Knoll, F.; Johnson, P. End-to-end variational networks for accelerated MRI reconstruction. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, 4–8 October 2020; Proceedings, Part II 23; Springer: Berlin/Heidelberg, Germany, 2020, pp. 64–73.

20. Aggarwal, H.K.; Mani, M.P.; Jacob, M. MoDL: Model-based deep learning architecture for inverse problems. *IEEE Trans. Med. Imaging* **2018**, *38*, 394–405.
21. Lin, K.; Heckel, R. Vision Transformers Enable Fast and Robust Accelerated MRI. In Proceedings of the 5th International Conference on Medical Imaging with Deep Learning, Zurich, Switzerland, 6–8 July 2022; Konukoglu, E., Menze, B., Venkataraman, A., Baumgartner, C., Dou, Q., Albarqouni, S., Eds.; PMLR: New York, NY, USA, 2022, Volume 172, pp. 774–795.
22. Parvaiz, A.; Khalid, M.A.; Zafar, R.; Ameer, H.; Ali, M.; Fraz, M.M. Vision Transformers in medical computer vision—A contemplative retrospection. *Eng. Appl. Artif. Intell.* **2023**, *122*, 106126.
23. Chen, Z.; Chen, Y.; Xie, Y.; Li, D.; Christodoulou, A.G. Data-consistent non-Cartesian deep subspace learning for efficient dynamic MR image reconstruction. In Proceedings of the 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), IEEE, Kolkata, India, 28–31 March 2022; pp. 1–5.
24. 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, Paris, France, 2–6 October 2021; pp. 10012–10022.
25. Shen, X.; Özen, A.C.; Sunjar, A.; Ilbey, S.; Sawiak, S.; Shi, R.; Chiew, M.; Emir, U. Ultra-short T2 components imaging of the whole brain using 3D dual-echo UTE MRI with rosette k-space pattern. *Magn. Reson. Med.* **2023**, *89*, 508–521. [CrossRef] [PubMed]
26. Li, Y.; Yang, R.; Zhang, C.; Zhang, J.; Jia, S.; Zhou, Z. Analysis of generalized rosette trajectory for compressed sensing MRI. *Med. Phys.* **2015**, *42*, 5530–5544. [CrossRef] [PubMed]
27. Villarreal, C.X.; Shen, X.; Alhulail, A.A.; Buffo, N.M.; Zhou, X.; Pogue, E.; Özen, A.C.; Chiew, M.; Sawiak, S.; Emir, U.; et al. An accelerated PETALUTE MRI sequence for in vivo quantification of sodium content in human articular cartilage at 3T. *Skelet. Radiol.* **2024**, *54*, 601–610.
28. Bucholz, E.K.; Song, J.; Johnson, G.A.; Hancu, I. Multispectral imaging with three-dimensional rosette trajectories. *Magn. Reson. Med. Off. J. Int. Soc. Magn. Reson. Med.* **2008**, *59*, 581–589. [CrossRef] [PubMed]
29. Mahmud, S.Z.; Denney, T.S.; Bashir, A. Feasibility of spinal cord imaging at 7 T using rosette trajectory with magnetization transfer preparation and compressed sensing. *Sci. Rep.* **2023**, *13*, 8777.
30. Alcicek, S.; Craig-Craven, A.R.; Shen, X.; Chiew, M.; Ozen, A.; Sawiak, S.; Pilatus, U.; Emir, U. Multi-site ultrashort echo time 3D phosphorous MRSI repeatability using novel rosette trajectory (PETALUTE). *bioRxiv* **2024**, *2024*, 579294.
31. Bozymski, B.; Shen, X.; Özen, A.; Chiew, M.; Thomas, M.A.; Clarke, W.T.; Sawiak, S.; Dydak, U.; Emir, U. Feasibility and comparison of 3D modified rosette ultra-short echo time (PETALUTE) with conventional weighted acquisition in 31P-MRSI. *Sci. Rep.* **2025**, *15*, 6465.
32. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
33. Zhang, C.; Jiang, W.; Zhang, Y.; Wang, W.; Zhao, Q.; Wang, C. Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery. *IEEE Trans. Geosci. Remote. Sens.* **2022**, *60*, 1–20.
34. Nossa, G.; Monsivais, H.; Hong, S.; Park, T.; Erdil, F.S.; Shen, X.; Özen, A.C.; Ilbey, S.; Chiew, M.; Steinwurzel, C.; et al. Submillimeter fMRI Acquisition using a dual-echo Rosette-k-space trajectory at 3T. In Proceedings of the International Society for Magnetic Resonance in Medicine (ISMRM), Toronto, ON, Canada, 20–25 May 2023.
35. Blumenthal, M.; Luo, G.; Schilling, M.; Holme, H.C.M.; Uecker, M. Deep, deep learning with BART. *Magn. Reson. Med.* **2023**, *89*, 678–693.
36. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612.
37. Hodson, T.O. Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geosci. Model Dev. Discuss.* **2022**, *2022*, 1–10.
38. Studholme, C.; Hill, D.L.; Hawkes, D.J. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognit.* **1999**, *32*, 71–86.
39. Huynh-Thu, Q.; Ghanbari, M. Scope of validity of PSNR in image/video quality assessment. *Electron. Lett.* **2008**, *44*, 800–801.
40. Tsai, D.Y.; Lee, Y.; Matsuyama, E. Information entropy measure for evaluation of image quality. *J. Digit. Imaging* **2008**, *21*, 338–347.
41. Atkinson, D.; Hill, D.L.; Stoyle, P.N.; Summers, P.E.; Keevil, S.F. Automatic correction of motion artifacts in magnetic resonance images using an entropy focus criterion. *IEEE Trans. Med. Imaging* **1997**, *16*, 903–910.
42. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
43. Smith, L.N.; Topin, N. Super-convergence: Very fast training of neural networks using large learning rates. In Proceedings of the Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, SPIE, Baltimore, MD, USA, 7–8 May 2019; Volume 11006, pp. 369–386.

44. Lin, D.J.; Johnson, P.M.; Knoll, F.; Lui, Y.W. Artificial Intelligence for MR Image Reconstruction: An Overview for Clinicians. *J. Magn. Reson. Imaging* **2021**, *53*, 1015–1028. Available online: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.27078 (accessed on 10 November 2024).
45. Mescheder, L.; Geiger, A.; Nowozin, S. Which training methods for GANs do actually converge? In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 3481–3490.

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.

cancers

MDPI

Comment

Comment on Dalboni da Rocha et al. Artificial Intelligence for Neuroimaging in Pediatric Cancer. *Cancers* 2025, *17*, 622

Elif Keles [1,*], Mehmet Numan Colakoglu [2] and Max Bengtsson [1,3]

[1] Department of Radiology, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA; maxbengtsson2025@u.northwestern.edu
[2] Department of Radiology, Ankara Bilkent City Hospital, 06800 Ankara, Türkiye; mehmetnumancolakoglu@gmail.com
[3] McCormick School of Engineering, Northwestern University, Evanston, IL 60208, USA
* Correspondence: elif.keles@northwestern.edu

As researchers actively working in the field of pediatric neuroimaging and artificial intelligence, we were pleased to see your recent review, Artificial Intelligence for Neuroimaging in Pediatric Cancer, by Rocha et al. [1], which reflects the growing interest and rapid progress in this area. Reviews like this are critical for synthesizing knowledge and guiding future research directions.

However, we felt compelled to write this letter to clarify several points in Table 1, as some of the reported segmentation results may inadvertently misrepresent the original studies or oversimplify important distinctions—particularly regarding model evaluation metrics, dataset characteristics, and segmentation targets. These discrepancies are not limited to a single line but involve several studies where important methodological details—such as whether the reported Dice scores reflect whole-tumor or subregion segmentation, or whether metrics like Pearson's correlation were used—were either unclear or omitted. These inconsistencies, while subtle, can lead to confusion in how readers interpret and compare model performance across studies.

Our intention is to support clarity, consistency, and transparency in the evaluation of AI models in pediatric brain tumor segmentation—especially as the field moves toward clinical integration. We respectfully offer the following clarifications for improved accuracy and transparency.

Examples of clarification:

- Boyd et al. [2] reported a whole-tumor Dice score of 0.88 for low-grade gliomas, based on external datasets from CBTN (n = 60) and DFCI/BCH (n = 100), using T2-weighted MRI. The model employed a stepwise transfer learning strategy, starting from BRATS 2021 adult glioma weights (n = 1251) and fine-tuned on CBTN data (n = 124). This methodological context is essential for interpreting the reported performance.
- Mulvany et al. [3] presented Dice scores ranging from 0.657 to 0.967, reflecting different tumor subregions. The 0.931 Dice score corresponds specifically to whole-tumor segmentation on the PED BraTS 2024 validation dataset (n = 91), with a training set of n = 261 cases.
- Vossough et al. [4] reported a Dice score of 0.9 for whole-tumor segmentation on an internal test set (n = 293) and a Pearson correlation coefficient of 0.98, which reflects volume agreement rather than spatial overlap. These are distinct metrics and should not be used interchangeably.
- Our study [5] reported a Dice score of 0.642 as an average across tumor subregions (enhancing, non-enhancing, cystic, and edema). The whole-tumor Dice score was

0.681 on CBTN (n = 30 LGG cases) and 0.866 on the held-out PED BraTS 2024 dataset (n = 26). This distinction between whole-tumor and subregion metrics is important when making cross-study comparisons.

To support consistent and accurate comparisons of AI segmentation models, distinguishing between whole-tumor Dice scores and subregion Dice scores in tables and discussions would improve clarity and transparency [6,7].

We appreciate your consideration of this clarification. Providing clear segmentation performance metrics, dataset details, and model specifications is important for advancing AI applications in pediatric neuroimaging [7,8].

To facilitate accurate and meaningful comparisons across segmentation models, we suggest that future tables and summaries distinguish between

- Whole-tumor and subregion Dice scores,
- Spatial overlap (Dice) versus volume-based (e.g., Pearson r) metrics;
- Internal versus external validation cohorts, with appropriate dataset specifications.

We hope these clarifications are helpful and contribute to the broader goal of advancing transparency and reproducibility in pediatric neuroimaging AI research.

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

References

1. Dalboni da Rocha, J.L.; Lai, J.; Pandey, P.; Myat, P.S.M.; Loschinskey, Z.; Bag, A.K.; Sitaram, R. Artificial Intelligence for Neuroimaging in Pediatric Cancer. *Cancers* **2025**, *17*, 622. [CrossRef] [PubMed]
2. Boyd, A.; Ye, Z.; Prabhu, S.; Tjong, M.C.; Zha, Y.; Zapaishchykova, A.; Vajapeyam, S.; Hayat, H.; Chopra, R.; Liu, K.X.; et al. Expert-level pediatric brain tumor segmentation in a limited data scenario with stepwise transfer learning. *medRxiv* **2023**. [CrossRef] [PubMed]
3. Mulvany, T.; Griffiths-King, D.; Novak, J.; Rose, H. Segmentation of Pediatric Brain Tumors using a Radiologically informed, Deep Learning Cascade. *arXiv* **2024**, arXiv:2410.14020.
4. Vossough, A.; Khalili, N.; Familiar, A.M.; Gandhi, D.; Viswanathan, K.; Tu, W.; Haldar, D.; Bagheri, S.; Anderson, H.; Haldar, S.; et al. Training and Comparison of nnU-Net and DeepMedic Methods for Autosegmentation of Pediatric Brain Tumors. *Am. J. Neuroradiol.* **2024**, *45*, 1081–1089. [CrossRef] [PubMed]
5. Bengtsson, M.; Keles, E.; Durak, G.; Anwar, S.; Velichko, Y.S.; Linguraru, M.G.; Waanders, A.J.; Bagci, U. A new logic for pediatric brain tumor segmentation. In Proceedings of the 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI), Houston, TX, USA, 14–17 April 2025; pp. 1–5.
6. Familiar, A.M.; Kazerooni, A.F.; Vossough, A.; Ware, J.B.; Bagheri, S.; Khalili, N.; Anderson, H.; Haldar, D.; Storm, P.B.; Resnick, A.C.; et al. Towards consistency in pediatric brain tumor measurements: Challenges, solutions, and the role of artificial intelligence-based segmentation. *Neuro-Oncology* **2024**, *26*, 1557–1571. [CrossRef] [PubMed]
7. Jha, D.; Durak, G.; Sharma, V.; Keles, E.; Cicek, V.; Zhang, Z.; Srivastava, A.; Rauniyar, A.; Hagos, D.H.; Tomar, N.K.; et al. A Conceptual Framework for Applying Ethical Principles of AI to Medical Practice. *Bioengineering* **2025**, *12*, 180. [CrossRef] [PubMed]
8. Sabuncu, M.R.; Wang, A.Q.; Nguyen, M. Ethical Use of Artificial Intelligence in Medical Diagnostics Demands a Focus on Accuracy, Not Fairness. *NEJM AI* **2025**, *2*, AIp2400672. [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.

Journal of
Clinical Medicine

MDPI

Article

Automatic Segmentation of the Nasolacrimal Canal: Application of the nnU-Net v2 Model in CBCT Imaging

Emre Haylaz [1], Ismail Gumussoy [1,*], Suayip Burak Duman [2], Fahrettin Kalabalik [1], Muhammet Can Eren [1], Mustafa Sami Demirsoy [3], Ozer Celik [4] and Ibrahim Sevki Bayrakdar [5]

[1] Department of Oral and Maxillofacial Radiology, Faculty of Dentistry, Sakarya University, Sakarya 54050, Turkey
[2] Department of Oral and Maxillofacial Radiology, Faculty of Dentistry, Inonu University, Malatya 44000, Turkey
[3] Independent Researcher, Sakarya 54100, Turkey
[4] Mathematics and Computer Science, Faculty of Science, Eskişehir Osmangazi University, Eskişehir 26040, Turkey
[5] Department of Oral and Maxillofacial Radiology, Faculty of Dentistry, Eskişehir Osmangazi University, Eskişehir 26040, Turkey
* Correspondence: ismailgumussoy@sakarya.edu.tr

Abstract: Background/Objectives: There are various challenges in the segmentation of anatomical structures with artificial intelligence due to the different structural features of the relevant region/tissue. The aim of this study was to detect the nasolacrimal canal (NLC) using the nnU-Net v2 convolutional neural network (CNN) model in cone beam-computed tomography (CBCT) images and to evaluate the successful performance of the model in automatic segmentation. **Methods:** CBCT images of 100 patients were randomly selected from the data archive. The raw data were transferred to the 3D Slicer imaging software in DICOM format (Version 4.10.2; MIT, Massachusetts, USA). NLC was labeled using the polygonal type of manual method. The dataset was split into training, validation and test sets in a ratio of 8:1:1. nnU-Net v2 architecture was applied to the training and test datasets to predict and generate appropriate algorithm weight factors. The confusion matrix was used to check the accuracy and performance of the model. As a result of the test, the Dice Coefficient (DC), Intersection over Union (IoU), F1-Score and 95% Hausdorff distance (95% HD) metrics were calculated. **Results:** By testing the model, DC, IoU, F1-Scores and 95% HD metric values were found to be 0.8465, 0.7341, 0.8480 and 0.9460, respectively. According to the data obtained, the receiver-operating characteristic (ROC) curve was drawn and the AUC value under the curve was determined to be 0.96. **Conclusions:** These results showed that the proposed nnU-Net v2 model achieves NLC segmentation on CBCT images with high precision and accuracy. The automated segmentation of NLC may assist clinicians in determining the surgical technique to be used to remove lesions, especially those affecting the anterior wall of the maxillary sinus.

Keywords: artificial intelligence; cone beam-computed tomography; machine learning; neural networks; nasolacrimal duct

1. Introduction

The nasolacrimal canal (NLC), part of the lacrimal drainage system, consists of a bony and a membranous canal. The bony part is flanked laterally by the sulcus lacrimalis and medially by the processus lacrimalis. It extends from the lower end of the lacrimal sac to the meatus nasi inferior, forming an angle of approximately 15 to 30 degrees with the

coronal plane. The average width of the intraosseous canal is 4 mm, and its length ranges between 13 and 28 mm. The mucosa lining the lacrimal drainage system forms folds and sinuses, with the most crucial being the Hasner valve at the meatus nasi inferior [1–3].

Epiphora is a condition characterized by the overflow of tears due to partial or complete blockage of the NLC, which may be congenital or acquired [3–5]. Primary acquired dacryostenosis, an idiopathic fibroinflammatory obstruction, is the leading cause. Secondary acquired dacryostenosis often results from infections, trauma, surgery, tumors or radiation therapy [4,6,7].

Over the years, various imaging methods have been utilized to examine the morphological changes in the NLC, as well as to detect, treat and follow up duct and sac diseases. These methods include conventional dacryocystography (DCG) [8], digital subtraction dacryocystography (DS-DCG) [9], computed tomography (CT) [4], computerized tomography–dacryocystography (CT–DCG) [10], magnetic resonance imaging (MRI) [11], and cone beam-computed tomography (CBCT) [12]. Accurate detection of the NLC is essential for maxillofacial surgical procedures, as damage to the lower opening of the NLC or the anterior wall of the lacrimal sac may occur during surgery. Therefore, selecting an imaging method that is both accurate and effective is crucial [13,14].

The widespread use of three-dimensional (3D) imaging methods, particularly cone beam-computed tomography (CBCT) in dentistry, has addressed the limitations of two-dimensional (2D) imaging [15]. In addition, the effectiveness and success of artificial intelligence algorithms have been enhanced by the use of CBCT and other 3D imaging methods. Diagnostic images have become a primary resource for developing AI systems, which can be used to detect pathological changes, segment craniofacial anatomical structures, and classify maxillofacial tumors and cysts [16–20]. For example, Widiasri et al. [21] developed an AI model for detecting alveolar bone thickness and the mandibular canal to aid in dental implant planning. In another study, AI algorithms were utilized to assess growth and development based on cervical vertebra stages [22]. Zhang et al. [23] successfully predicted postoperative facial swelling in orthognathic surgery patients using a convolutional neural network (CNN)-based model. These capabilities are invaluable for clinicians during both intraoperative and postoperative decision-making. AI technology has the potential to minimize errors, assist in surgical treatment planning, reduce diagnosis and treatment times and improve overall operational efficiency [24–26].

However, most current deep learning algorithms require manual intervention for lesion detection and segmentation, making the process time-consuming and less practical for routine clinical use [27,28]. To address these limitations and improve usability, researchers are focusing on developing algorithms that can fully automate these steps [16,29,30].

To the best of our knowledge, no previous study has evaluated the automatic segmentation of NLC on CBCT images using the nnU-Net v2. This model simplifies and streamlines the system by employing a systematic approach, avoiding the need for additional network structures and eliminating system complexity. Furthermore, the nnU-Net v2 model can be adapted into a semi-automatic system by incorporating manual inputs, enabling external intervention to address deficiencies and enhance network performance [31].

The aim of this study is to develop a CNN algorithm utilizing the nnU-Net v2 architecture and evaluate its performance in automatically detecting the NLC in axial CBCT images. A successful model could serve as a reliable tool for clinicians, aiding in the assessment of the relationship between pathologies affecting the paranasal sinuses, nasal region and the NLC. This could improve preoperative planning and intraoperative guidance, ultimately enhancing surgical outcomes.

Additionally, the development of an automatic segmentation tool for the NLC could streamline the diagnostic process, reducing the need for manual segmentation and saving

time for radiologists and surgeons. This could lead to increased efficiency in clinical workflows. Our hypothesis is that a CNN algorithm utilizing the nnU-Net v2 architecture will achieve high accuracy and efficiency in automatically detecting the NLC in axial CBCT images.

2. Materials and Methods

2.1. Study Design

In this study, the CBCT data of patients who applied to Inonu University Faculty of Dentistry, Oral and Maxillofacial Radiology Clinic for any reason were used. Since the aim is to develop and validate an AI for automatic NLC detection, this study consists of three phases: development, training and testing. CBCT scans used full convolutional neural network algorithms (nnU-Net v2) for 3D automated detection of NLC. The AI Checklist in Medical Imaging (CLAIM) and Standards for Reporting of Diagnostic Accuracy Studies (STARD) were used in the preparation of this article. The study was retrospectively designed and all procedures were carried out in accordance with the Declaration of Helsinki and similar ethical standards. The Inonu University Non-Interventional Clinical Research Ethics Committee approved the study protocol (decision no: 2024/6575). As a routine protocol, all patients in the CBCT archive had been informed and provided written consent regarding the use of their data for scientific research.

2.2. Data

This study's sample size was determined using the G*Power software (version 3.1.9.7; Franz Faul, University of Kiel, Kiel, Germany). With an α error probability of 0.05 and a power of the study of 0.95, the actual power of the study was calculated to be 95% when at least 100 samples were included [32]. The final sample was obtained from 100 CBCT images (200 right and left NLC) of 50 males and 50 females randomly selected using archival records. The ages of the patients ranged from 18 to 72 years, and the mean age was 52 years. The inclusion criteria are as follows:

- Individuals over 18 years of age.
- Individuals without any syndrome or bone disease.
- Clearly identified images of NLC's bone boundaries.

The exclusion criteria are as follows:

- Individuals with known pre-existing infection, neoplasm and malformations associated with NLC.
- Individuals who have undergone surgical operations and trauma involving the maxillofacial region and NLC.
- Images with motion or metal artifacts that prevent NLC from being displayed and degrade diagnostic quality.

2.3. Obtaining and Evaluating CBCT Images

The scans were performed using a NewTom 5G CBCT machine (Quantitative Radiology, Verona, Italy); 110 kVp, 1–11 mA, 3.6 s, 8 × 8 cm^2, 12 × 8 cm^2 and 15 × 12 cm^2 field of view (FOV) were obtained with 0.2 mm^3–0.3 mm^3 voxel size parameters. CBCT images were taken between 2021 and 2023 for various reasons.

2.4. Ground Truth

After reconstruction of the raw data, DICOM images were transferred to the 3D Slicer imaging software (Version 4.10.2; MIT, Cambridge, MA, USA) for manual segmentation. NLC was manually labeled in the coronal, sagial and axial planes using 3D Slicer imaging software, an open source program. The labeled DICOM data were converted to NIfTI

(Neuro Imaging Information Technology Initiative) format and exported for processing. The ground truth was defined by the consensus of two maxillofacial radiologists (E.H. and I.G.), both of whom had at least 6 years of experience in maxillofacial radiology. Upon completion of the manual segmentation process, it was checked by senior individuals (I.S.B. and F.K.) who had at least 10 years of experience in maxillofacial radiology and reached full agreement on all labels.

2.5. Testing Data

After screening the best-performing model through comprehensive comparison, 100 CBCT datasets were divided into training, validation and test sets according to the ratio of 8:1:1 (Figure 1).

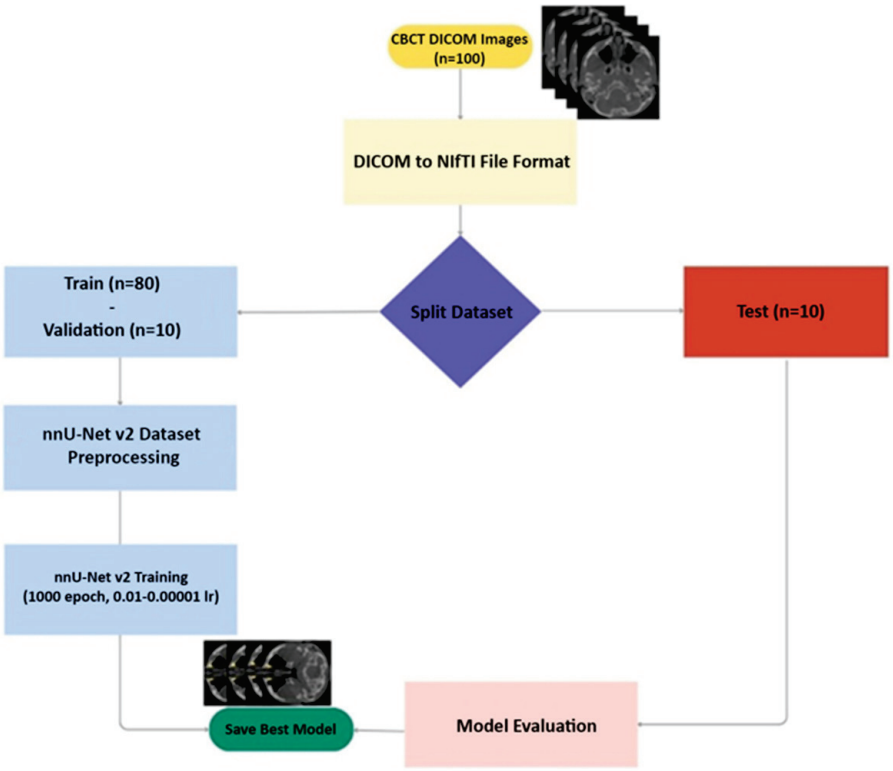


Figure 1. Workflow model of automatic segmentation of the nasolacrimal canal.

2.6. Model

The optimal model was used for 10-fold cross-validation on the training set, and the validation and test sets were evaluated. Training model and parameters: nnU-Net v2 based FCNN model, 1000 epochs, 0.00001 learning rate. The algorithm of the nnU-Net v2 model for automatic segmentation of NLC was developed in a Python environment (v3.6.1; Python Software Foundation, Wilmington, DE, USA) using the PyTorch library. The CranioCatch AI software (Version 2.1; CranioCatch, Eskisehir, Turkey) was used in the deep learning model development and training process described by Bayrakdar et al. [32].

U-Net was introduced to the world in 2015 by Olaf Ronneberber, Philip Fischer and Thomas Brox in an article titled "U-Net: Convolutional Networks for Biomedical Image Segmentation" for better segmentation, particularly in biomedical images [33]. The FCNN (Fourier Convolution Neural Network) model used in this study, nnU-Net (Neural Networks U-Net) v2, is an improved version of the U-Net architecture. Large datasets are needed for training in classical convolutional neural network models. The images in these datasets are labeled and presented to the network, and the network recognizes the images with this label information. This labeling process is particularly challenging

for biomedical images due to their pixel-based nature, requiring significant human and hardware resources. Unlike classical CNN models, U-Net offers a unique architecture for pixel-based image segmentation, addressing these challenges.

2.7. Evaluation

The model's performance in the automated 3D segmentation of the NLC on CBCT volumes was evaluated using a confusion matrix. The confusion matrix is an essential performance evaluation tool for measuring algorithm accuracy and establishing success parameters. Probabilities of segmentation models were derived from the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values in the confusion matrix. The test results yielded the confusion matrix for the nnU-Net v2 model.

Recall (sensitivity) and precision metrics of the proposed model are calculated with the expressions in the confusion matrix (Table 1). In this present study, the results are presented with Dice Coefficient (DC), F1-Score and Intersection over Union Intersection (IoU) metrics. Additionally, the area under curve (AUC) value and the 95% Hausdorff distance (95% HD) were calculated in mm. These metrics are frequently used in the literature to measure the success of segmentation, and information about the metrics is given in this section.

Table 1. Metrics used to evaluate the performance of the nnU-Net v2 model, and the results obtained.

Metrics	Metric Formula	Metric Value
True Positive		16,297.7
False Positive		4214.2
False Negative		1624.5
Precision	TP/(TP + FP)	0.7888
Recall (Sensitivity)	TP/(TP + FN)	0.9168
Dice Coefficients (DC)	(2 × T P)/(2 × T P + F P + F N)	0.8465
Intersection over Union (IoU)	(|A∩B|)/(|A∪B|)	0.7341
F1-Score	2 × (Precision × Recall)/ (Precision + Recall)	0.8480
95% Hausdorff Distance (95%HD) mm	$d_{H95}(A, B) = max(d_{95}(A, B), d_{95}(A, B))$	0.9460

Notes: TP: True Positive, FP: False Positive, FN: False Negative.

3. Results

Within the scope of this study, the nnU-Net v2 model, which is one of the deep learning networks, was used for NLC segmentation. A total of 80% of the dataset was determined as the training group, 10% as the validation group, and 10% as the test group. The training was completed in an average of 1000 epochs. In all models, "Adam" was used as the optimization algorithm. ReLU activation is used in the proposed model. The training parameters of the nnU-Net v2 model are given in Table 2. The predictive analysis of the proposed nnU-Net v2 model is shown in Figure 2.

The values of the precision and recall metrics were calculated as 0.7888 and 0.9168, respectively (Table 1). According to the results, the highest DC, IoU, F1-Score and 95% Hausdorff distance values after the segmentation performed in the nnU-Net v2 architecture were found to be 0.8465, 0.7341, 0.8480 and 0.9460, respectively (Table 1). The DC and IoU metrics of the test data are shown in Figures 3 and 4.

Table 2. Training parameters and value of the proposed nnU-Net v2 model.

Parameter	Value
Model	NnU-Net v2
Epoch	1000
Batch Size	2
Learning Rate	0.00001
Optimization	ADAM
Activation	ReLU

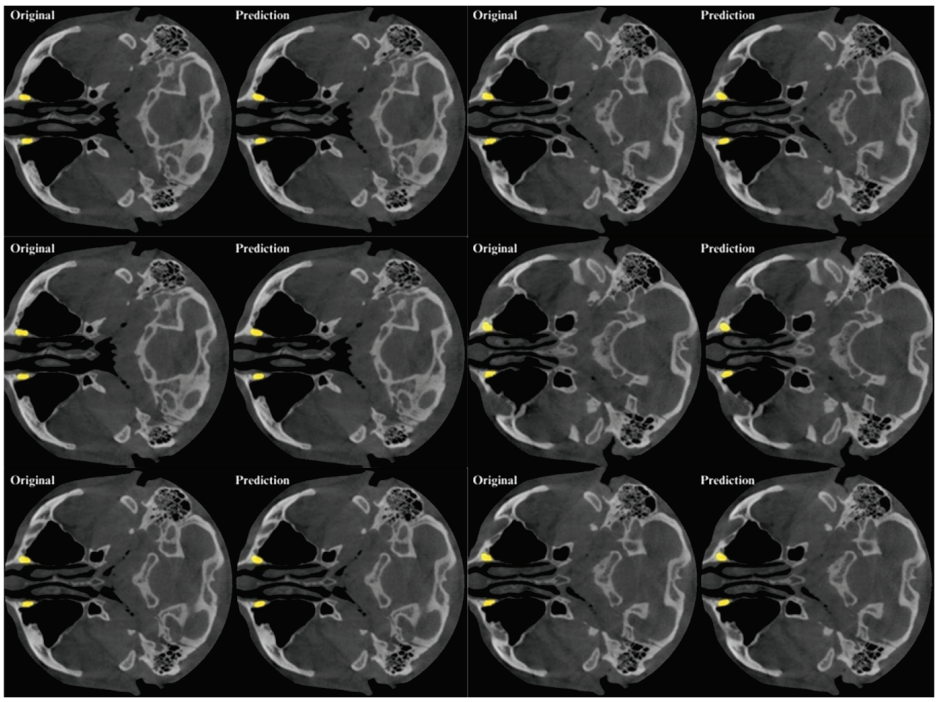


Figure 2. Automatic segmentation of nasolacrimal canal using artificial intelligence model in axial CBCT slices.

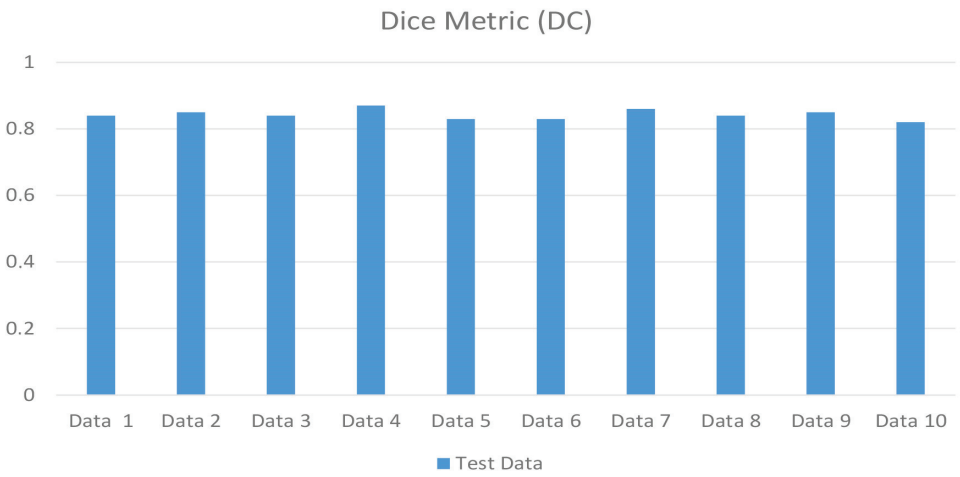


Figure 3. Dice confidence (DC) scores of the test data.

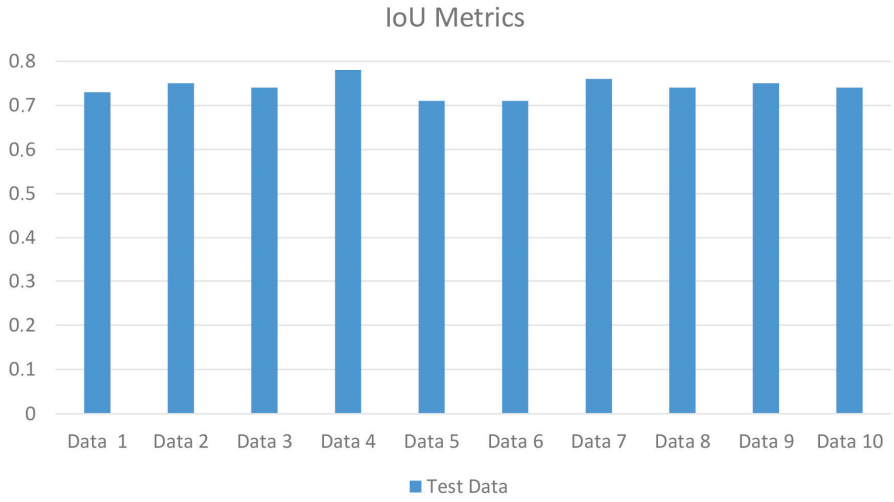


Figure 4. IoU metrics of the test data.

According to the test results, the receiver-operating characteristic (ROC) curve was drawn and the AUC value under the curve was determined to be 0.96 (Figure 5). In order to have more information about the training of the model, a graph of the dice score and loss function values at each number of cycles was created, starting from the first number of cycles of the model (Figure 6).

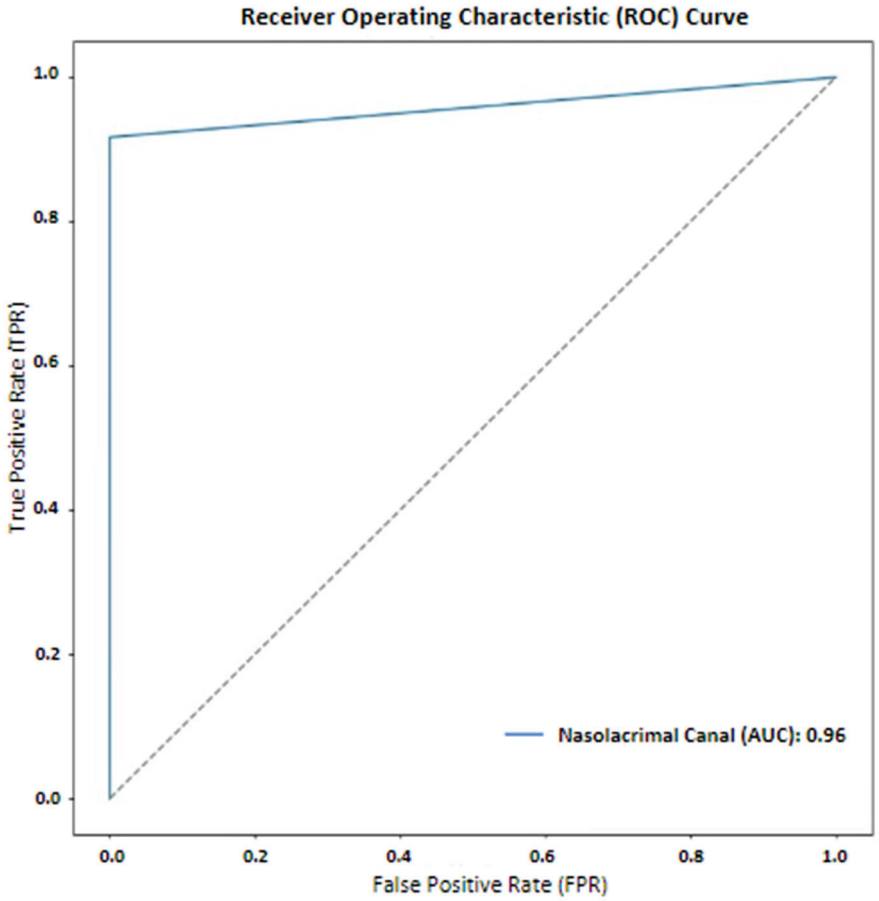


Figure 5. Receiver-operating characteristic (ROC) curve and AUC value.

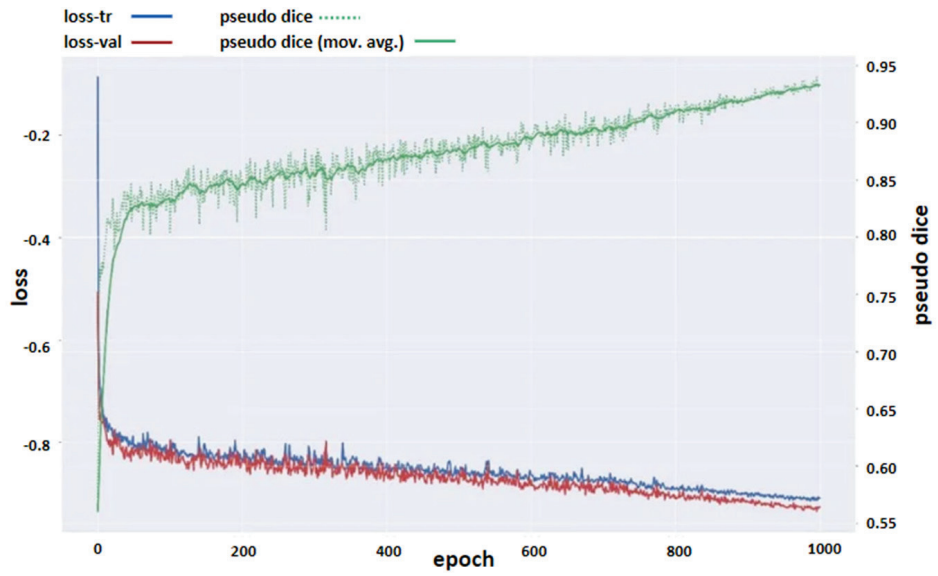


Figure 6. Dice score and loss function values at each number of cycles of the model.

4. Discussion

Paranasal sinuses and nasal anatomical structures are closely related to the NLC. Due to this anatomical proximity, paranasal sinus diseases and pathologies involving the nasal region often affect the NLC [34]. In cases where the anterior wall and floor of the maxillary sinus are involved, surgical intervention can become challenging. In such situations, the anterior wall of the maxillary sinus can be accessed using the prelacrimal approach [35]. This technique may lead to the removal of the NLC's bony component and displacement of the mucosa. To determine whether NLC dislocation or resection is required, the distance between the anterior wall of the maxillary sinus and the NLC should be considered [36]. Automatic segmentation of the NLC can help clinicians practically and easily assess the distance between the NLC and the anterior wall of the maxillary sinus, aiding in determining the appropriate surgical technique for planned interventions in the maxillary sinus and nasal region. Moreover, accurate segmentation of the NLC provides crucial guidance for maxillofacial surgical procedures and pathology evaluations, serving as an essential tool for preoperative and intraoperative planning and ultimately enhancing surgical outcomes. This study introduces a new automatic segmentation tool that uses the nnU-Net v2-based AI model to achieve high accuracy and efficiency in NLC segmentation. It is innovative in that it is one of the first studies in the literature to use AI for NLC segmentation.

To the best of our knowledge, there are no studies on NLC segmentation using AI in the literature yet. This study aimed to evaluate the performance of the nnU-Net v2 model in segmenting the NLC from CBCT scans. The DC was 0.8465, indicating successful NLC segmentation. An IoU metric of 0.7341 further confirmed accurate segmentation, as an IoU above 0.5 is typically deemed successful [37]. The F1-Score of 0.8480 demonstrated the model's high sensitivity and precision. The ROC curve indicated an AUC value of 0.96, highlighting the model's strong ability to accurately distinguish the NLC from CBCT images. These findings confirmed our hypothesis (H1). This suggests the model's potential for use in diagnosing NLC, planning treatment and evaluating NLC-related diseases before maxillofacial surgery.

Manual labeling of CBCT scans is labor-intensive, time-consuming and costly. Even with semi-automatic labeling, operator-dependent errors are unavoidable. Precise segmentation can aid forensic science by providing detailed anatomical information that is critical

for identifying individuals and understanding trauma mechanisms. In addition, precise automatic segmentation can process large datasets in a fraction of the time, significantly reducing the workload for researchers and clinicians. Moreover, automatic segmentation provides consistent and reproducible results. This consistency is crucial for comparative studies and longitudinal research where precise measurements are essential. Thus, there is a growing need for the fully automated labeling of CBCT scans [38,39]. Preda et al. [40] investigated automatic segmentation of the maxillofacial complex using CBCT data. They reported that automated segmentation takes less than a minute and is 204 times faster than manual segmentation. In their study, they reported the mean values of DSC, 95% HD and IoU metrics as 0.926, 0.621 and 0.862, respectively. The findings showed high similarity between automatically and manually segmented CBCT data. These results highlight that automated segmentation increases efficiency in the workflow and saves time [41].

AI and DL algorithms are used for different purposes in medical imaging. These include image segmentation and preprocessing, disease detection and diagnosis, personalized treatment planning, predictive analytics, quality control, monitoring and follow-up [42]. In particular, DL models can be used to follow the radiotherapy process by automating the segmentation of organs and tumors. Moreover, DL algorithms enable adaptive radiotherapy, where treatment plans can be continuously optimized over time according to changes in a patient's anatomy. This adaptive approach ensures that radiation is delivered with the highest possible accuracy, improving treatment outcomes while minimizing side effects [43]. Different deep learning models, the nnU-Net model and its variants are widely used in the literature for similar studies (Table 3). This model has been employed for segmenting various anatomical structures and detecting pathologies, such as mandibular canal abnormalities [34], maxillary sinus pathologies [41] and the classification of jaw lesions [44], have been carried out with this model.

Table 3. A summary of several studies in the literature related to automatic segmentation of dental structures.

Authors	Aim	Sample	Segmentation Model	Imaging Method	Evaluation Metrics
Ozturk [37]	The aim of this study is to develop a deep learning-based method to perform maxillary sinus segmentation using CBCT images.	100 Scans	U-Net	CBCT	F-1 Score: 0.9784 IoU: 0.9275
Preda et al. [40]	This present study investigated the accuracy, consistency and time-efficiency of a novel deep convolutional neural network (CNN)-based model for the automated maxillofacial bone segmentation from CBCT images.	144 Patients	U-Net	CBCT	DC: 0.926 %95 HD: 0.621 IoU: 0.862
Shi et al. [44]	This study proposes an automated method to measure condylar changes in patients with skeletal class II malocclusion following surgical orthodontic treatment.	48 Patients	nnU-Net	CBCT	Maxilla DC: 0.9263 Mandible DC: 0.9387 Condyle DC: 0.971
Yağmur et al. [45]	The aim of this study is to evaluate the mandibular canal with CBCT using a deep learning approach.	300 Patients	nnU-Net v2	CBCT	DC: 0.76
Ascı et al. [46]	The purpose of this study was to evaluate the effectiveness of dental caries segmentation on the panoramic radiographs taken from children in primary dentition, mixed dentition and permanent dentition with AI models developed using the deep learning method.	6075 Patients	U-Net	Panoramic Radiographs	Sensitivity: 0.8269 Precision: 0.9123 F-1 Score: 0.8675

Table 3. *Cont.*

Authors	Aim	Sample	Segmentation Model	Imaging Method	Evaluation Metrics
İçöz et al. [47]	The aim of this study was to evaluate the effectiveness of an AI system in the detection of roots with apical periodontitis on digital panoramic radiographs.	306 Scans	YOLOv3	Panoramic Radiographs	Sensitivity: 98% Specificity: 56% F-1 Score: 71%
Chang et al. [48]	The aim of this study was to develop an automated method for diagnosing periodontal bone loss for staging periodontitis on dental panoramic radiographs using the deep learning hybrid method for the first time.	340 Scans	Mask R-CNN	Panoramic Radiographs	Periodontal Bone Level IoU: 0.88 Accuracy: 0.92 DC: 0.93 Cementoenamel Junction Level IoU: 0.84 Accuracy: 0.87 DC: 0.91 Teeth and Implants IoU: 0.83 Accuracy: 0.87 DC: 0.91

Zhu et al. [49] utilized the nnU-Net model for diagnosing impacted teeth, dental caries, crowns, missing teeth and residual roots. However, there are no studies yet on the automatic segmentation of NLC. Three-dimensional imaging of the NLC not only enhances anatomical understanding but also facilitates the creation of virtual models to guide surgical planning in the maxillofacial region. In this context, the precise and comprehensive segmentation of the NLC is a crucial preliminary step.

Shi et al. [50] employed the nnU-Net model to segment CBCT data from 48 patients with class II malocclusion who underwent orthognathic surgery. They evaluated the success of automatic segmentation in identifying condylar changes. The DC values for the automatic segmentation of the maxilla, mandible and condylar changes were 0.9263, 0.9387 and 0.971, respectively. This study concluded that artificial intelligence exhibits high sensitivity in detecting condylar changes [50]. Oztürk et al. [37] performed maxillary sinus segmentation on CBCT images with the U-Net network. They used F1-Score and IoU metrics to evaluate the model's performance. The segmentation results demonstrated an F1-Score of 0.9784 and an IoU value of 0.9275. These metric values confirmed the model's success in maxillary sinus segmentation [37]. As highlighted in the aforementioned studies, artificial intelligence-based segmentation serves diverse purposes, and its performance can be evaluated using various metrics.

Morita et al. [17] assessed the segmentation of eight anatomical parameters in the facial region using a 2D U-Net model on CT data. The mean DC values for maxilla, mandible, right zygoma segmentation and left zygoma segmentation were found to be 0.909 ± 0.036, 0.984 ± 0.017, 0.936 ± 0.029 and 0.926 ± 0.049, respectively. The model showed high success. However, the average DC values for the nasal bone (0.838 ± 0.084) and frontal bone (0.858 ± 0.060) showed lower accuracy compared to the others. They attributed this to the fact that the nasal bone is thinner and smaller than other bones. They also reported that sutures on the nasal and frontal bones would affect segmentation success [17]. In this present study, the DC value for NLC segmentation was comparable to the values reported by Morita et al. [17] for the nasal and frontal bones, yet it exhibited lower success rates compared to other anatomical structures.

Yağmur et al. [45] utilized the nnU-Net v2 model, similar to the methodology employed in this study. The DC value for mandibular canal segmentation in their study was reported as 0.76. This current study demonstrated that NLC segmentation achieved higher success using the same model. However, success performance was found to be lower in this current study compared to other studies [40,41,51,52]. This discrepancy can be attributed to the anatomical connection of the NLC's bony component with adjacent spaces. Especially,

its connection to the lower meatus is wide and angulated, making manual segmentation difficult. Consequently, accurate segmentation proves challenging in cases where distinct bony boundaries are not clearly discernible.

Chang et al. [48] developed an automatic method for staging periodontitis using a deep learning hybrid framework. They proposed a novel approach that combines deep learning architecture for detection with conventional CAD (Computer-Aided Diagnosis) processing for classification. The deep learning model was employed to detect the radiographic bone level as a simple structure for the entire jaw on panoramic radiographs. They found that this hybrid framework, which integrates deep learning and traditional CAD methods, demonstrated high accuracy and excellent reliability in automatically diagnosing periodontal bone loss and staging periodontitis.

As evident from previous studies, the U-Net network architecture has been employed for various purposes in medicine and dentistry [17,37,45,46]. High achievement performance with correct training and appropriate data is one of the reasons why this model is preferred. U-Net network architecture provides success even with very limited data and has superior performance speed compared to other models. Due to these qualified features, it is frequently preferred in the biomedical field, especially for segmentation purposes [46,52].

This study had several limitations. Firstly, while the developed model allowed for the bone canal border of the NLC to be clearly determined, it did not achieve the same success rate in distinguishing the membranous canal. The use of multimodal segmentation models (CT-MR, CBCT-MR) to improve soft tissue contrast may be considered in the future, which can enhance the model's robustness and performance. Furthermore, issues related to low resolution could be addressed. Secondly, CBCT data were obtained from a single institution. Data from different centers would increase heterogeneity and allow for the results to be generalized. Finally, there is the difficulty of separating the NLC from the spaces it connects. Therefore, using more advanced multi-stream (multi-angle, multi-scale and multi-modality) models would be beneficial for improving the accuracy of NLC segmentation in 3D images.

5. Conclusions

This study evaluated the nnU-Net v2 model for segmenting the NLC in CBCT images using a dataset of 100 patients. The DC value of 0.8465 and the AUC value of 0.96 demonstrated high accuracy. The complex anatomical structure of the maxillofacial region makes the manual labeling of CBCT images laborious and time-consuming in practice. The NLC is an important anatomical structure that must be considered before surgical procedures in this region, particularly due to its proximity to the anterior wall of the maxillary sinus. The proposed nnU-Net v2 model demonstrated promising performance in segmenting the NLC on CBCT images. Our results suggest that the automated segmentation of the nasolacrimal canal can facilitate clinical decision-making, improve work efficiency and save time during diagnosis and treatment.

Future efforts will focus on addressing the current limitations by expanding the dataset with additional CBCT images and labels. Furthermore, specific adjustments to the nnU-Net v2 model, such as hyperparameter tuning, enhanced data augmentation techniques and architectural modifications (e.g., attention mechanisms), are planned to further improve the DC and other performance metrics. These refinements aim to develop a more robust and accurate model for clinical applications.

Author Contributions: Conceptualization, I.G., E.H. and M.S.D.; Methodology, I.G., S.B.D. and M.S.D.; Software, I.G., E.H., O.C. and M.C.E.; Validation, I.S.B., F.K. and S.B.D.; Formal Analysis, I.G. and O.C.; Investigation, I.G., S.B.D. and M.S.D.; Data Curation, M.C.E. and E.H.; Writing—Original Draft Preparation, I.G., E.H. and O.C.; Writing—Review and Editing, F.K., M.S.D. and I.S.B.; Visual-

ization, E.H. and M.C.E.; Supervision, I.G. and E.H.; Project Administration, I.G. All authors have read and agreed to the published version of the manuscript.

Funding: This study was not funded by any organization.

Institutional Review Board Statement: Ethical approval was obtained from the Health Sciences Ethics Committee of Inonu University, Malatya, Türkiye (reference date/no: 9 October 2024; 2024/6575). All procedures performed in studies involving human participants were in accordance with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Statement: For this type of study, formal consent is not required.

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 that they have no conflicts of interest.

References

1. Lee, S.; Lee, U.Y.; Yang, S.W.; Lee, W.J.; Kim, D.H.; Youn, K.H.; Kim, Y.S. 3D morphological classification of the nasolacrimal duct: Anatomical study for planning treatment of tear drainage obstruction. *Clin. Anat.* **2021**, *34*, 624–633. [CrossRef]
2. Jañez-Garcia, L.; Saenz-Frances, F.; Ramirez-Sebastian, J.M.; Toledano-Fernandez, N.; Urbasos-Pascual, M.; Jañez-Escalada, L. Three-dimensional reconstruction of the bony nasolacrimal canal by automated segmentation of computed tomography images. *PLoS ONE* **2016**, *11*, e0155436. [CrossRef] [PubMed]
3. Maliborski, A.; Różycki, R. Diagnostic imaging of the nasolacrimal drainage system. Part I. Radiological anatomy of lacrimal pathways. Physiology of tear secretion and tear outflow. *Med. Sci. Monit.* **2014**, *20*, 628. [PubMed]
4. Czyz, C.N.; Bacon, T.S.; Stacey, A.W.; Cahill, E.N.; Costin, B.R.; Karanfilov, B.I.; Cahill, K.V. Nasolacrimal system aeration on computed tomographic imaging: Sex and age variation. *Ophthalmic Plast. Reconstr. Surg.* **2016**, *32*, 11–16. [CrossRef] [PubMed]
5. Atum, M.; Alagöz, G. Blood cell ratios in patients with primary acquired nasolacrimal duct obstruction. *Ophthalmol. J.* **2020**, *5*, 76–80. [CrossRef]
6. Kousoubris, P.D.; Rosman, D.A. Radiologic evaluation of lacrimal and orbital disease. *Otolaryngol. Clin. N. Am.* **2006**, *39*, 865–893. [CrossRef] [PubMed]
7. Ansari, S.A.; Pak, J.; Shields, M. Pathology and imaging of the lacrimal drainage system. *Neuroimaging Clin. N. Am.* **2005**, *15*, 221–237. [CrossRef]
8. Francisco, F.; Carvalho, A.; Francisco, V.; Francisco, M.; Neto, G. Evaluation of 1000 lacrimal ducts by dacryocystography. *Br. J. Ophthalmol.* **2007**, *91*, 43–46. [CrossRef] [PubMed]
9. Saleh, G.; Gauba, V.; Tsangaris, P.; Tharmaseelan, K. Digital subtraction dacryocystography and syringing in the management of epiphora. *Orbit* **2007**, *26*, 249–253. [CrossRef]
10. Ali, M.J.; Singh, S.; Naik, M.N.; Kaliki, S.; Dave, T.V. Interactive navigation-guided ophthalmic plastic surgery: The utility of 3D CT-DCG-guided dacryolocalization in secondary acquired lacrimal duct obstructions. *Clin. Ophthalmol.* **2016**, *11*, 127–133. [CrossRef]
11. Coskun, B.; Ilgit, E.; Onal, B.; Konuk, O.; Erbas, G. MR dacryocystography in the evaluation of patients with obstructive epiphora treated by means of interventional radiologic procedures. *AJNR Am. J. Neuroradiol.* **2012**, *33*, 141–147. [CrossRef]
12. Nakamura, J.; Kamao, T.; Mitani, A.; Mizuki, N.; Shiraishi, A. Analysis of lacrimal duct morphology from cone-beam computed tomography dacryocystography in a Japanese population. *Clin. Ophthalmol.* **2022**, *16*, 2057–2067. [CrossRef]
13. Janssen, A.G.; Mansour, K.; Bos, J.J.; Castelijns, J.A. Diameter of the bony lacrimal canal: Normal values and values related to nasolacrimal duct obstruction: Assessment with CT. *AJNR Am. J. Neuroradiol.* **2001**, *22*, 845–850.
14. Moran, I.; Virdee, S.; Sharp, I.; Sulh, J. Postoperative complications following LeFort 1 maxillary advancement surgery in cleft palate patients: A 5-year retrospective study. *Cleft Palate Craniofac. J.* **2018**, *55*, 231–237. [CrossRef] [PubMed]
15. Gumussoy, I.; Duman, S.B. Morphometric analysis of occipital condyles using alternative imaging technique. *Surg. Radiol. Anat.* **2020**, *42*, 161–169. [CrossRef] [PubMed]
16. Abdolali, F.; Zoroofi, R.A.; Otake, Y.; Sato, Y. Automatic segmentation of maxillofacial cysts in cone beam CT images. *Comput. Biol. Med.* **2016**, *72*, 108–119. [CrossRef] [PubMed]
17. Morita, D.; Mazen, S.; Tsujiko, S.; Otake, Y.; Sato, Y.; Numajiri, T. Deep-learning-based automatic facial bone segmentation using a two-dimensional U-Net. *Int. J. Oral Maxillofac. Surg.* **2023**, *52*, 787–792. [CrossRef]
18. Lee, J.-H.; Kim, D.-h.; Jeong, S.-N.; Choi, S.-H. Diagnosis and prediction of periodontally compromised teeth using a deep learning-based convolutional neural network algorithm. *J. Periodontal Implant. Sci.* **2018**, *48*, 114–123. [CrossRef]

19. Lee, J.-H.; Kim, D.-H.; Jeong, S.-N.; Choi, S.-H. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. *J. Dent.* **2018**, *77*, 106–111. [CrossRef]
20. Setzer, F.C.; Shi, K.J.; Zhang, Z.; Yan, H.; Yoon, H.; Mupparapu, M.; Li, J. Artificial intelligence for the computer-aided detection of periapical lesions in cone-beam computed tomographic images. *J. Endod.* **2020**, *46*, 987–993. [CrossRef]
21. Widiasri, M.; Arifin, A.Z.; Suciati, N.; Fatichah, C.; Astuti, E.R.; Indraswari, R.; Putra, R.H.; Za'in, C. Dental-yolo: Alveolar bone and mandibular canal detection on cone beam computed tomography images for dental implant planning. *IEEE Access* **2022**, *10*, 101483–101494. [CrossRef]
22. Kök, H.; Acilar, A.M.; İzgi, M.S. Usage and comparison of artificial intelligence algorithms for determination of growth and development by cervical vertebrae stages in orthodontics. *Prog. Orthod.* **2019**, *20*, 41. [CrossRef] [PubMed]
23. Zhang, W.; Li, J.; Li, Z.-B.; Li, Z. Predicting postoperative facial swelling following impacted mandibular third molars extraction by using artificial neural networks evaluation. *Sci. Rep.* **2018**, *8*, 12281. [CrossRef] [PubMed]
24. Zhang, L.; Li, W.; Lv, J.; Xu, J.; Zhou, H.; Li, G.; Ai, K. Advancements in oral and maxillofacial surgery medical images segmentation techniques: An overview. *J. Dent.* **2023**, *138*, 104727. [CrossRef] [PubMed]
25. Koç, U.; Sezer, E.A.; Özkaya, Y.A.; Yarbay, Y.; Beşler, M.S.; Taydaş, O.; Yalçın, A.; Evrimler, Ş.; Kızıloğlu, H.A.; Kesimal, U.; et al. Elevating healthcare through artificial intelligence: Analyzing the abdominal emergencies data set (Tr_Abdomen_Rad_Emergency) At Teknofest-2022. *Eur. Radiol.* **2024**, *34*, 3588–3597. [CrossRef]
26. Mutlu, F.; Çetinel, G.; Gül, S. A fully-automated computer-aided breast lesion detection and classification system. *Biomed. Signal Process. Control* **2020**, *62*, 102157. [CrossRef]
27. Yilmaz, E.; Kayikcioglu, T.; Kayipmaz, S. Computer-aided diagnosis of periapical cyst and keratocystic odontogenic tumor on cone beam computed tomography. *Comput. Methods Programs Biomed.* **2017**, *146*, 91–100. [CrossRef] [PubMed]
28. Nicolielo, L.F.P.; Van Dessel, J.; Van Lenthe, G.H.; Lambrichts, I.; Jacobs, R. Computer-based automatic classification of trabecular bone pattern can assist radiographic bone quality assessment at dental implant site. *Br. J. Radiol.* **2018**, *91*, 20180437. [CrossRef] [PubMed]
29. Abdolali, F.; Zoroofi, R.A.; Otake, Y.; Sato, Y. Automated classification of maxillofacial cysts in cone beam CT images using contourlet transformation and Spherical Harmonics. *Comput. Methods Programs Biomed.* **2017**, *139*, 197–207. [CrossRef] [PubMed]
30. Lee, J.H.; Kim, D.H.; Jeong, S.N. Diagnosis of cystic lesions using panoramic and cone beam computed tomographic images based on deep learning neural network. *Oral Dis.* **2020**, *26*, 152–158. [CrossRef]
31. Isensee, F.; Jäger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. Automated design of deep learning methods for biomedical image segmentation. *arXiv* **2019**, arXiv:1904.08128.
32. Bayrakdar, I.S.; Elfayome, N.S.; Hussien, R.A.; Gulsen, I.T.; Kuran, A.; Gunes, I.; Al-Badr, A.; Celik, O.; Orhan, K. Artificial intelligence system for automatic maxillary sinus segmentation on cone beam computed tomography images. *DentoMaxilloFacial Radiol.* **2024**, *53*, 256–266. [CrossRef] [PubMed]
33. 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: Cham, Switzerland, 2015; pp. 234–241.
34. Razavi, M.; Shams, N.; Afshari, F.; Nanai, S. Investigating the Morphology of the Nasolacrimal Canal in Cone Beam Computed Tomography Images and Its Relationship with Age and Gender. *Maedica* **2024**, *19*, 303. [CrossRef] [PubMed]
35. Duman, S.B.; Gumussoy, İ. Assesment of Prelacrimal Recess in Patients With Maxillary Sinus Hypoplasia Using Cone Beam Computed Tomography. *Am. J. Rhinol. Allergy* **2021**, *35*, 361–367. [CrossRef] [PubMed]
36. Simmen, D.; Veerasigamani, N.; Briner, H.R.; Jones, N.; Schuknecht, B. Anterior maxillary wall and lacrimal duct relationship-CT analysis for prelacrimal access to the maxillary sinus. *Rhinology* **2017**, *55*, 170–174. [CrossRef]
37. Ozturk, B.; Taspinar, Y.S.; Koklu, M.; Tassoker, M. Automatic segmentation of the maxillary sinus on cone beam computed tomographic images with U-Net deep learning model. *Eur. Arch. Oto-Rhino-Laryngol.* **2024**, *281*, 6111–6121. [CrossRef] [PubMed]
38. Jaskari, J.; Sahlsten, J.; Järnstedt, J.; Mehtonen, H.; Karhu, K.; Sundqvist, O.; Hietanen, A.; Varjonen, V.; Mattila, V.; Kaski, K. Deep learning method for mandibular canal segmentation in dental cone beam computed tomography volumes. *Sci. Rep.* **2020**, *10*, 5842. [CrossRef]
39. Morgan, N.; Van Gerven, A.; Smolders, A.; de Faria Vasconcelos, K.; Willems, H.; Jacobs, R. Convolutional neural network for automatic maxillary sinus segmentation on cone-beam computed tomographic images. *Sci. Rep.* **2022**, *12*, 7523. [CrossRef] [PubMed]
40. Preda, F.; Morgan, N.; Van Gerven, A.; Nogueira-Reis, F.; Smolders, A.; Wang, X.; Nomidis, S.; Shaheen, E.; Willems, H.; Jacobs, R. Deep convolutional neural network-based automated segmentation of the maxillofacial complex from cone-beam computed tomography: A validation study. *J. Dent.* **2022**, *124*, 104238. [CrossRef]
41. Jung, S.-K.; Lim, H.-K.; Lee, S.; Cho, Y.; Song, I.-S. Deep active learning for automatic segmentation of maxillary sinus lesions using a convolutional neural network. *Diagnostics* **2021**, *11*, 688. [CrossRef] [PubMed]

42. Obuchowicz, R.; Strzelecki, M.; Piórkowski, A. Clinical Applications of Artificial Intelligence in Medical Imaging and Image Processing—A Review. *Cancers* **2024**, *16*, 1870. [CrossRef] [PubMed]
43. Damilakis, J.; Stratakis, J. Descriptive overview of AI applications in x-ray imaging and radiotherapy. *J. Radiol. Prot.* **2024**, *44*, 041001. [CrossRef]
44. Liu, W.; Li, X.; Liu, C.; Gao, G.; Xiong, Y.; Zhu, T.; Zeng, W.; Guo, J.; Tang, W. Automatic classification and segmentation of multiclass jaw lesions in cone-beam CT using deep learning. *DentoMaxilloFacial Radiol.* **2024**, *53*, 439–446. [CrossRef] [PubMed]
45. Yağmur, Ü.S.; Namdar, P.F. Evaluation of the mandibular canal by CBCT with a deep learning approach. *Balk. J. Dent. Med.* **2024**, *28*, 122–128. [CrossRef]
46. Asci, E.; Kilic, M.; Celik, O.; Cantekin, K.; Bircan, H.B.; Bayrakdar, İ.S.; Orhan, K. A Deep Learning Approach to Automatic Tooth Caries Segmentation in Panoramic Radiographs of Children in Primary Dentition, Mixed Dentition, and Permanent Dentition. *Children* **2024**, *11*, 690. [CrossRef] [PubMed]
47. Icoz, D.; Terzioglu, H.; Ozel, M.A.; Karakurt, R. Evaluation of an artificial intelligence system for the diagnosis of apical periodontitis on digital panoramic images. *Niger. J. Clin. Pract.* **2023**, *8*, 1085–1090. [CrossRef]
48. Chang, H.J.; Lee, S.J.; Yong, T.H.; Shin, N.Y.; Jang, B.G.; Kim, J.E.; Huh, K.H.; Lee, S.-S.; Heo, M.-S.; Choi, S.-C.; et al. Deep Learning Hybrid Method to Automatically Diagnose Periodontal Bone Loss and Stage Periodontitis. *Sci. Rep.* **2020**, *10*, 7531. [CrossRef]
49. Zhu, J.; Chen, Z.; Zhao, J.; Yu, Y.; Li, X.; Shi, K.; Zhang, F.; Yu, F.; Shi, K.; Sun, Z.; et al. Artificial intelligence in the diagnosis of dental diseases on panoramic radiographs: A preliminary study. *BMC Oral Health* **2023**, *23*, 358. [CrossRef] [PubMed]
50. Shi, J.; Lin, G.; Bao, R.; Zhang, Z.; Tang, J.; Chen, W.; Chen, H.; Zuo, X.; Feng, Q.; Liu, S. An automated method for assessing condyle head changes in patients with skeletal class II malocclusion based on Cone-beam CT images. *DentoMaxilloFacial Radiol.* **2024**, *53*, 325–335. [CrossRef] [PubMed]
51. Choi, H.; Jeon, K.J.; Kim, Y.H.; Ha, E.-G.; Lee, C.; Han, S.-S. Deep learning-based fully automatic segmentation of the maxillary sinus on cone-beam computed tomographic images. *Sci. Rep.* **2022**, *12*, 14009. [CrossRef] [PubMed]
52. Wu, X.; Zhao, L. Study on iris segmentation algorithm based on dense U-Net. *IEEE Access* **2019**, *7*, 123959–123968. [CrossRef]

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

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

MDPI Books Editorial Office
E-mail: books@mdpi.com
www.mdpi.com/books

Disclaimer/Publisher's Note: The title and front matter of this reprint are at the discretion of the Topic Editors. The publisher is not responsible for their content or any associated concerns. The statements, opinions and data contained in all individual articles are solely those of the individual Editors and contributors and not of MDPI. MDPI disclaims responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-4516-3